

# VU Research Portal

## Theory and Application of Dynamic Spatial Time Series Models

Andree, Bo Pieter Johannes

2020

### **document version**

Publisher's PDF, also known as Version of record

### **document license**

CC BY-ND

[Link to publication in VU Research Portal](#)

### **citation for published version (APA)**

Andree, B. P. J. (2020). *Theory and Application of Dynamic Spatial Time Series Models*. [PhD-Thesis - Research and graduation internal, Vrije Universiteit Amsterdam]. Rozenberg Publishers and the Tinbergen Institute.

### **General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

### **Take down policy**

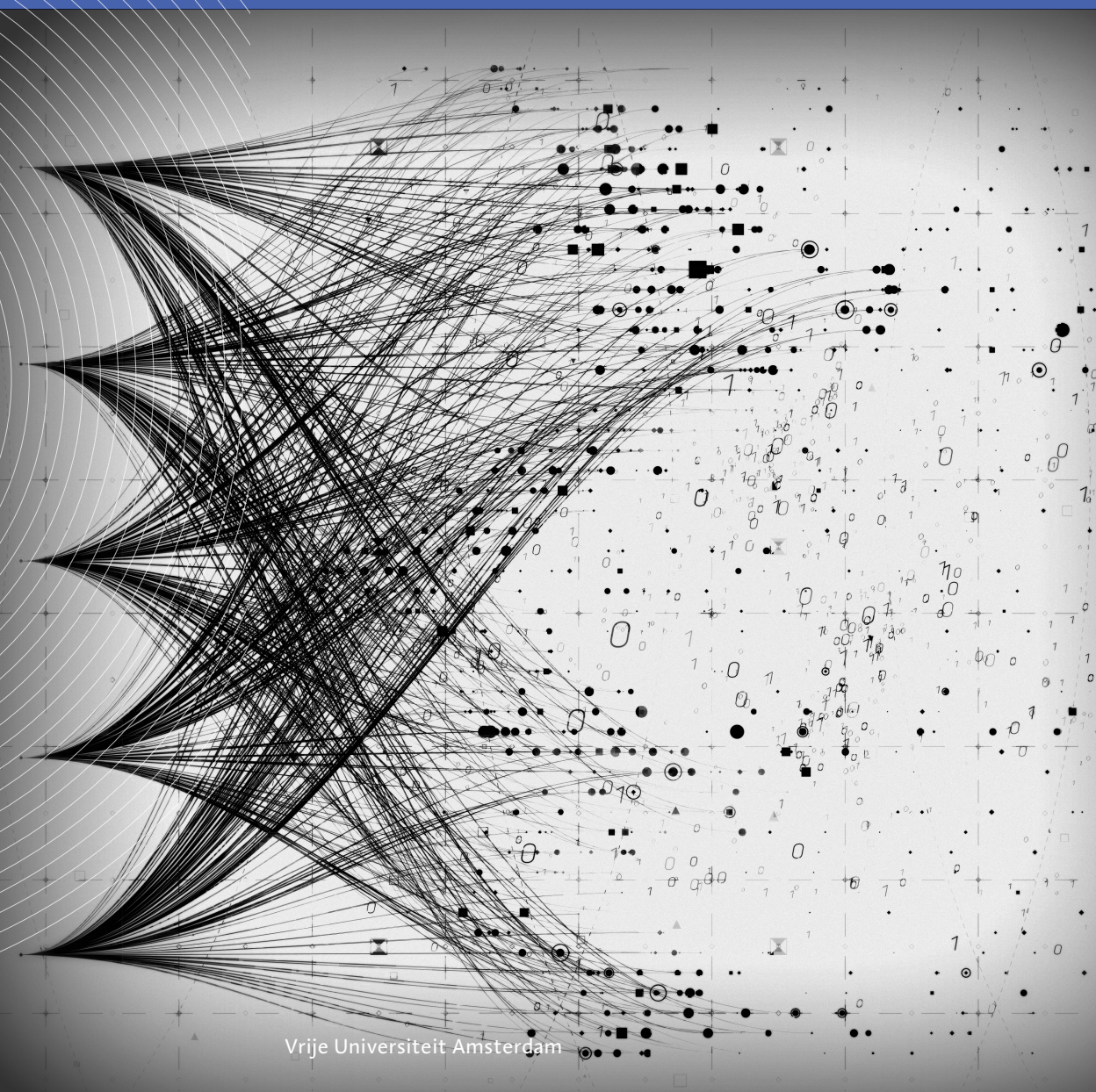
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

### **E-mail address:**

[vuresearchportal.ub@vu.nl](mailto:vuresearchportal.ub@vu.nl)

# Theory and Application of Dynamic Spatial Time Series Models

Bo Pieter Johannes Andrée







THEORY AND APPLICATION OF DYNAMIC  
SPATIAL TIME SERIES MODELS

ISBN: 978 90 361 0607

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. **762** of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.



VRIJE UNIVERSITEIT

**THEORY AND APPLICATION OF DYNAMIC  
SPATIAL TIME SERIES MODELS**

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan  
de Vrije Universiteit Amsterdam,  
op gezag van de rector magnificus  
prof.dr. V. Subramaniam,  
in het openbaar te verdedigen  
ten overstaan van de promotiecommissie  
van de School of Business and Economics  
op dinsdag 26 mei 2020 om 11.45 uur  
in de aula van de universiteit,  
De Boelelaan 1105

door

Bo Pieter Johannes Andrée

geboren te Leiden

promotor: prof.dr. H.J. Scholten

copromotor: dr. E. Koomen

## **Dedication**

This book is dedicated to the future generations that will share this planet. Many issues exist in our world, I hope my generation will pass it on in a better state than we have received it.



“The mind cannot foresee its own advance.”

— Friedrich Hayek



Image by Mathew Schwartz, a lesson from the great architects of the past to aspiring thinkers of the future; good design retains its quality, it lasts without adjustments (attributed to Nadia Piffaretti).

# Contents

<b>Preface</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Background Theory</b>	<b>7</b>
2.1 Linear estimators . . . . .	8
2.1.1 The linear Least Squares Estimator . . . . .	10
2.1.2 The linear Maximum Likelihood Estimator . . . . .	22
2.2 General Extremum Estimators . . . . .	28
2.2.1 General Consistency . . . . .	28
2.2.2 General asymptotic Normality . . . . .	33
2.3 Further complications when modeling dynamic spatial time series . . . . .	37
<b>3 Spatial Heterogeneity</b>	<b>43</b>
3.1 Introduction . . . . .	44
3.2 The importance of spatial heterogeneity in agricultural policy . . . . .	46
3.3 Methodology . . . . .	49
3.3.1 Different spatial policies . . . . .	50
3.3.2 Spatial economic model . . . . .	52

3.3.3	Modeling production quantities . . . . .	60
3.4	The case of Miscanthus in the Netherlands . . . . .	61
3.5	Results . . . . .	65
3.5.1	Economic performance of production systems . .	65
3.5.2	Assessing the impacts of different policies . . . . .	67
3.5.3	Comparing different policies . . . . .	69
3.6	Discussion and conclusions . . . . .	75
3.7	Appendix . . . . .	80
3.7.1	A. Energy data . . . . .	80
3.7.2	B. Crop rotation schemes . . . . .	81
3.7.3	C. Modeling the dairy farming production system	81
3.7.4	D. Frequency distribution of agro-economic perfor- mance . . . . .	82
3.7.5	E. Spatial distribution of minimum required subsidies	83
<b>4</b>	<b>Parametric Spatial Nonlinearities</b>	<b>85</b>
4.1	Introduction . . . . .	86
4.2	Linear and nonlinear spatial autoregressive models . . . .	90
4.2.1	Linear dynamics: the SAR Model . . . . .	90
4.2.2	The Smooth Transition Spatial Autoregressive model	92
4.3	Asymptotic theory for the ST-SAR model . . . . .	96
4.3.1	Existence and measurability of the MLE . . . . .	97
4.3.2	Consistency and of the MLE . . . . .	98
4.3.3	Set-consistency of the MLE allowing for possible parameter identification failure . . . . .	102



4.3.4	Asymptotic normality of the MLE . . . . .	103
4.3.5	Model selection under possible parameter identification failure . . . . .	105
4.4	Monte Carlo study . . . . .	112
4.5	The empirics of nonlinear spatial dependencies . . . . .	117
4.5.1	Application I: Dutch residential densities . . . . .	117
4.5.2	Application II: interest rates in the Euro region . . . . .	123
4.6	Conclusion . . . . .	134
4.7	Appendix . . . . .	135
4.7.1	Proofs to main theorems . . . . .	135
4.7.2	Additional results . . . . .	140
4.7.3	Proofs for additional results . . . . .	141
4.7.4	Additional Monte Carlo results and figures . . . . .	145
4.7.5	Time-line of events related to European Long term Interest Rates . . . . .	151
<b>5</b>	<b>Non-parametric Cross-sectional Nonlinearities</b>	<b>153</b>
5.1	Introduction . . . . .	154
5.2	Methods . . . . .	157
5.3	Data . . . . .	160
5.3.1	Forest cover . . . . .	161
5.3.2	Air pollution . . . . .	162
5.3.3	Carbon emission and economic development . . . . .	163
5.3.4	Treatment of missing data . . . . .	164
5.3.5	Other controls and final data . . . . .	167

5.3.6	Transformation to degradation intensities . . . . .	168
5.4	Empirical results . . . . .	171
5.4.1	Individual model results . . . . .	173
5.4.2	Heterogeneity in environmental output . . . . .	176
5.4.3	Average curvature . . . . .	179
5.4.4	Heterogeneity in curvature and tipping points . .	180
5.4.5	Exploring degradation dynamics under simple 2030 scenario's . . . . .	182
5.5	Discussion and conclusion . . . . .	186
5.6	Appendix . . . . .	190
5.6.1	Additional results and figures . . . . .	190
5.7	Supplementary note to the chapter . . . . .	194
5.7.1	Introduction . . . . .	194
5.7.2	The modeling framework . . . . .	195
5.7.3	The role of out-of-sample performance in the inter- pretation . . . . .	206
5.7.4	Conclusion . . . . .	216
<b>6</b>	<b>Vector Spatial Time Series</b>	<b>223</b>
6.1	Introduction . . . . .	224
6.2	Spatial Vector Autoregressive Moving Average model . .	227
6.2.1	Vector Autoregressive Moving Average model . .	229
6.2.2	Spatial Vector Autoregressive Moving Average model	231
6.3	Model properties . . . . .	232
6.3.1	Causal SVAR and it's SMA representation . . . .	233

6.3.2	Invertible SMA as a SVAR . . . . .	234
6.3.3	Stability in canonical state space . . . . .	235
6.3.4	Uniqueness . . . . .	235
6.3.5	Impulse Response Functions . . . . .	236
6.4	Estimation . . . . .	237
6.4.1	Parameterizing spatial weight matrices using Gaussian kernels . . . . .	237
6.4.2	Penalized Maximum Likelihood Estimator . . . .	240
6.4.3	Small sample distribution of the (P)MLE . . . .	244
6.5	Application to subnational pollution and household expenditure data in Indonesia . . . . .	247
6.5.1	Data . . . . .	248
6.5.2	Estimation approach . . . . .	250
6.5.3	Results . . . . .	251
6.6	Conclusion . . . . .	258
6.7	Appendix . . . . .	261
6.7.1	Restrictions . . . . .	261
6.7.2	Stability in terms of the companion matrix . . . .	263
6.7.3	Small sample distribution of the (P)MLE . . . .	265
6.7.4	Pollution data . . . . .	270
6.7.5	Additional regression results . . . . .	271
6.7.6	Additional Impulse Response analysis results . . .	275
<b>7</b>	<b>Probability and Causality in Spatial Time Series</b>	<b>277</b>
7.1	Introduction . . . . .	278



---

7.2	Causality and probability . . . . .	283
7.3	Limit divergence on the space of modeled probability mea- sures . . . . .	293
7.4	Limit Squared Hellinger distance . . . . .	300
7.5	Concluding remarks . . . . .	304
<b>8</b>	<b>Conclusion</b>	<b>309</b>
8.1	Final remarks . . . . .	314
	<b>Bibliography</b>	<b>318</b>

# Preface

This morning's headline on CNN read "*30 Days that changed the world*". It is now 10 days since the WHO has declared a global pandemic. Over the past month, the world has been ravaged by an aggressive virus, businesses have come to a sudden stop, and financial markets have shown unprecedented turmoil. The Dow Jones is down -35% in the month, Gold is down -7.5%, Crude Brent is down -55%. At least there is one silver lining, incoming data is showing us that pollution and carbon output is also down along with markets.

In continuation of the trend, central banks and governments are unleashing a new storm of interest rate cuts, tax cuts, loan guarantees and new spending, tapping emergency powers in an attempt to cushion the shock to companies and workers and reassure investors. Will "*unlimited liquidity*" preserve the foundations of a functioning economy for the future? Future generations will be to judge.

While much of the moment seems gloomy, this must all somehow also lead to new thinking. I finished high school during the downturn of the 2008 financial crisis, and now sign this book amidst a new deepening divide. I realize that my thinking around the importance of feedback, spillovers, and nonlinearity have been greatly shaped by the events following 2008, and so will the thinking of those that come after me be shaped by today's events. We have never had more brains connected and focused on shared problems. I cannot help but turn to David Hilbert for wisdom. I am rereading the preamble to his "Mathematical Problems" and find

comforting words (adapted):

“History teaches the continuity of the development of science. We know that every age has its own problems, which the following age either solves or casts aside as profitless and replaces by new ones. If we would obtain an idea of the probable development of knowledge in the immediate future, we must let the unsettled questions pass before our minds and look over the problems which the science of today sets and whose solution we expect from the future.

As long as a branch of science offers an abundance of problems, so long is it alive; a lack of problems foreshadows extinction or the cessation of independent development. Just as every human undertaking pursues certain objects, so also research requires its problems. It is by the solution of problems that the investigator tests the temper of his steel; he finds new methods and new outlooks, and gains a wider and freer horizon.”

— Hilbert, David (1902).

He goes on to warn us about the dangers of conducting research in isolation from experience, and shapes our expectations about probable development of knowledge:

“In the meantime, while the creative power of pure reason is at work, the outer world comes into play, forces upon us new questions from actual experience, opens up new branches of science, and while we seek to conquer these new fields of knowledge for the realm of pure thought, we often find the answers to old unsolved problems and thus at the same time advance most successfully the old theories. And it seems to me that the numerous and surprising analogies and that apparently pre-arranged harmony which the mathematicians so often perceives in the questions, methods and ideas of the various branches of his science, have their origin in this ever-recurring interplay between thought and experience.”

— Hilbert, David (1902).

Looking back on my own research, I realize heavily that this ever-recurring interplay between thought and experience is an infinite process, and that any one person’s individual efforts are only ever a finite undertaking. So was writing this book. This is good, because it leaves room for future

books to address the problems set by today's science. However, it implies that the work here is by no means comprehensive, which would require an entire book series to be written. Luckily, good books and papers already exist that cover related topics in detail.

First, the publication of Cliff and Ord (1969) marked a turning point in the treatment of spatial autocorrelation in quantitative geography. The issues related to spatial correlation in regression disturbances were explored further and spatial econometrics as a subfield of econometrics was rapidly developed, for a large part Europe in the early 1970s because of the need to analyze sub-country data in regional econometric models (Cliff and Ord, 1972; Hordijk, 1974; Hordijk and Paelinck, 1976; Paelinck and Klaasen, 1979). Apart from the classic work of Anselin (1988), a good introduction to spatial econometrics is provided by LeSage and Pace (2009). A bridge between spatial models for cross-sectional data and panel data is made in Elhorst (2010b). A recent book by Beenstock and Felsenstein (2019) analyzes linear spatial time series, and develops useful tests for panel co-integration. Other recent exciting developments will be discussed throughout the chapters of this book. In such a fast-developing field I will surely have missed things (or omitted them for lack of space) which a few comments below may help to fill in.

First, some reviewers have commented that the work covers surprisingly little elements from classical spatial panel econometrics, but this represents a misunderstanding of the contribution I am seeking to make; I would not expect a book on the current state of spatial econometrics to concentrate only on spatial autoregressions but rather on interesting problems that one can analyze using spatial data and econometric techniques. In a similar fashion, I do not aim to advance the field by providing an exhaustive description of existing dynamic spatio-temporal regression problems, instead my interest is in relevant emerging analysis problems that involve dynamics between multiple spatial variables over

time and on the econometric approaches to addressing those analytical problems.

Second, some books take a specific-to-general approach, and start with a simple problem gradually making it more complex across successive chapters. In this work, I instead aim to approach related problems from different angles. Naturally, the techniques introduced throughout the chapters can be combined, but I don't necessarily see the value in doing so exhaustively. It would lead to a massively complicated analysis problem and distract from the relatively simple points I am trying to make in the different chapters. Naturally, the approach of the thesis then implies that in some cases the analyses presented in the individual chapters could be extended even further. This could lead to improved results. But I believe these improvements would be locally and not globally when looking at the book as a whole.

For example, Chapter 3 highlights the importance of spatial heterogeneity. Chapter 4 then aims to capture a great deal of heterogeneity in an estimation problem using a relatively simple non-linear function. This does not imply that the data heterogeneity could not be captured by simple approaches that rely on spatial and temporal dummies. Nor does it refute that an exhaustive dummy approach may be sufficient for some analysis problems. The contribution of the chapter instead lies in the fact that the traditional dummy approach may not be optimal for some problems, such as forecasting, stochastic simulation, or analysis of the drivers behind heterogeneous dynamics and that nonlinear modeling of dependence can provide an attractive alternative in those cases.

Chapter 5 focuses on non-parametric modeling of trends in panel data, but does not focus explicitly on spatial autoregressive dependence. As one can read in the book, one important reason for appropriately modeling spatial dependence is to improve model specification. In a similar spirit, non-parametric approaches are designed for a large part to reduce

mis-specification bias. A semi-parametric model could be specified that combines both a non-parametric component for nonlinearity and a parametric spatial component for simultaneity, but this would result in a complicated model that distracts from a simple but useful point; that non-parametric techniques can be successfully applied in a panel setting to capture complex dynamics while providing interpretable results.

After paying particular focus to heterogeneity and nonlinearity, Chapter 6 analyzes data using linear parameters. While this may seem to counter some of the notions previously introduced, this chapter is not about heterogeneity and nonlinearity per se. Instead, the focus is on inter-temporal dynamics between multiple variables within a spatial system. Linear interdependencies among multiple time series are often analyzed in multivariate time series analysis, but many panel methods have traditionally been developed with inferential questions about a single dependent variable in mind. The value of the chapter thus lies in introducing methods to analyze how finite impulse responses flow through a spatial system in the presence of both spatial and temporal forms of feedback. Such an analytical framework can easily accommodate nonlinear dynamics, for example by using the tools developed in Chapter 4 in a multiple variable setting.

With regard to how this work came about, a few final words are in order. Carrying out the research and then writing this thesis was one of the most arduous task I have undertaken. However, one of the joys of having completed this is looking back at everyone who has helped me over the past years. I would first like to thank my promotor prof.dr. Henk Scholten for giving me this chance, my co-promotor dr. Eric Koomen for his instrumental role in shaping my thinking and dr. Francisco Blasques for guiding me through some of the difficult challenges on my theoretical journey. They have all become good friends. I am also thankful to the co-authors of the research papers on which the individual chapters are

based. They not only contributed writing and insights, but also made carrying out the research enjoyable. I would like to thank the members of the reading and assessment committee, prof.dr. C. Fischer, prof.dr. S.J. Koopman, prof.dr. S. Bhulai, prof.dr. L. Hordijk and prof.dr. J.P. Elhorst for their careful reading of the manuscript.

To my family, particularly my parents, sister and grandparents, thank you for your love, support, and unwavering belief in me. Without you, I would not be the person I am today and this book would not have been here. Above all I would like to thank my wife Ilona for her love and unconditional support, and for keeping me sane. Thank you for your patience and understanding. But most of all, thank you for being my best friend. I owe you everything.

Finally, despite my love for pure thought, the work reported in this thesis would not have been possible without the practical support of the Vrije Universiteit and the World Bank. Thank you for providing a space to do research. To my (ex-) World Bank colleagues, my sincere thanks and gratitude for guarding what is an incredibly valuable international intellectual space. In particular, thank you dr. Harun Dogo for your inquisitive thinking and sense of humor, dr. Nadia Piffaretti for championing quality and rigor, and prof.dr. Aart Kraay for always putting forth rigor and simplicity as the general requirements for the solution of an intellectual problem.

To all other (ex-)colleagues and friends in Amsterdam, Washington, New York and elsewhere, my sincere thanks and gratitude. Your names are too many to mention but I thank you nonetheless.

Bo Pieter Johannes Andrée  
Amsterdam,  
March 21, 2020.

# Chapter 1

## Introduction

This thesis sets out to develop econometric theory and methods to analyze dynamic interactions between observations that are interrelated across space and time. This type of modeling is becoming increasingly important as sensors and institutions continue to gather rich subnational spatial time series of remotely sensed or surveyed economic variables. Going from finance, to macro-economics or the environment, nearly all policy relevant phenomena in the socio-economic domain involve multivariate interactions across both spatial and temporal dimensions. Analyzing these problems raises a number of inquiries about the econometric methods used that are both practically and theoretically interesting. In particular, cross-sectional data is often spatially dependent. From a data generating perspective, this implies that we may be concerned with models that exhibit instantaneous forms of feedback in space. Together with possible endogenous interactions between the observations of the different variables that are collected sequentially over the time dimension, this produces complex feedback properties that may violate various assumptions made by standard econometric models. Second, as the dimensions of datasets grow, it becomes increasingly unlikely that linear relationships provide a realistic description of these phenomena. The tendency of nonlinearities and the complex feedback properties that characterize spatial time series, render many related estimation problems non-standard.



In many cases, deriving the properties of estimators for multivariate models that have complex nonlinearities over both temporal as well as spatial dimensions, can be achieved by extending the theories used to analyze the estimators of dynamic time series models. In particular, spatial feedback renders the standard Least Squares Estimator (LSE) inconsistent or inefficient depending on the situation, but estimating models that explicitly factor in the dependence and feedback between neighbors can be done within the framework of Maximum Likelihood. Other interesting problems, such as exogenous or non-contemporaneous endogenous nonlinearities, can be estimated in the Least Squares framework. In both cases, this requires modifications to the standard criterion functions used. In particular, nonlinear parametric models of spatial time series introduce new components to the likelihood function that correct for the fact that the conditional densities are derived from a nonlinear transformation of the residuals. This requires new proofs that the well-known theoretical results associated with the standard Maximum Likelihood Estimator (MLE) nonetheless apply. Non-parametric Least Squares estimation of nonlinearities over the levels of cross-sectional observations can be solved as a locally linear problem, but requires penalization techniques to ensure that convergences essentially operate within simple spaces. This may change the interpretation of the limiting result all together. We will further investigate these issues in this thesis.

Many of the ideas produced in this thesis build heavily on the theory that underlies the analysis of time series data. This is a natural angle to view many problems. Early spatial models have been developed primarily to analyze cross-sectional data. As such, the underlying theory relied on taking the number of cross-sectional observations to infinity. While this may be sufficient to establish consistency and normality theoretically, in most real world applications it occurs seldom that new cross-sectional observations are made. Often, new observations are only collected over time while the number of spatial units remains fixed. In addition, when

---

new cross-sectional observations are in fact made, it is difficult to perceive that this change does not somehow involve also an extension in the time dimension.

The analysis of spatial data over time is a concept that is gaining in popularity, but it is still relatively new. It is only since recent that a significant part of our cross-sectional datasets have grown substantially enough in the time dimension to exhibit interesting temporal dynamics. For example, with modern compute it is still not possible for everyone to analyze remotely sensed data at high temporal resolution. Many publicly available datasets are therefore summarized as annual statistics that span only a modest number of years. Economic surveys that are consistently gathered across regions are often expensive. As an effect, surveyed data usually have a similar low temporal frequency. Finance data can be available at higher frequency, but many time series only start after the digital infrastructures that support modern systems matured. When one wishes to analyze a problem that involves multiple sources of data, then the data on which the analysis rests will often be constrained in both frequency and dimension. However, we are now at a point that sufficient data can in many cases be found, resulting in interesting problems that one can analyze with basic theory. In particular, with existing time series theory it is possible to analyze the properties of complex nonlinear dynamic time series models and understand the behavior of general estimators in these settings. However, this theory was not developed with spatial dependence and possible multivariate cross-sectional nonlinearities in mind. Many of the existing spatial analysis techniques have on the other hand not been developed with non-linear, possibly observation-driven, dynamics in mind. Moreover, panel techniques often focus on a single dependent variable, and are less concerned with describing the state transitions and dynamics between multiple spatial variables over time, which is needed for multivariate spatial time series forecasting, stochastic simulation, and impulse response analysis.

Before exploring spatial relationships explicitly, we will first review several important standard theoretical results for the estimation of dependencies in cross-sectional time series. We will use this as a basis to discuss what is further needed to analyze dynamic spatial time series problems. This background theory will be confined to what is needed to read the remainder of this thesis in a relatively self-contained manner. The remainder of this thesis then touches upon five key topics:

- i Spatial heterogeneity
- ii Parametric spatial nonlinearities
- iii Non-parametric cross-sectional nonlinearities
- iv Vector spatial time series
- v Probability and causality in spatial time series

Chapter 3 analyzes spatial heterogeneity. Specifically, it uses simple linear relationships and spatial explicit data to simulate economic outcomes at high spatial resolution. The analysis highlights how economic outcomes can cluster in space due to the natural clustering of independent geophysical variables that may be of economic importance. Moreover, it reveals that simple relationships at a high spatial resolution can produce nonlinear patterns at aggregated levels.

The concepts of spatial heterogeneity, dependence, and nonlinearity form the basis of Chapter 4 that looks into parametric spatial nonlinearities. This chapter covers the econometric application of spatial autoregressive time series models and extends the theory to cover nonlinear spatial dependence. The model that is introduced allows dependence to vary smoothly across levels in the data in an idiosyncratic manner. It will be shown that this type of spatial modeling captures both spatial and temporal dynamics and performs better than the standard linear spatial autoregressive model on a number of widely used diagnostics. Moreover,

---

the chapter will show that this type of modeling can produce interesting results when both  $T$  is large and  $N$  is small, or when  $N$  is large and  $T$  is still relatively modest.

Chapter 5 drops the parametric assumption, and looks at the case of non-parametric panel relationships. In this case, the focus is on nonlinear dependence of spatial time series variables on independent data in a manner that is appropriate when a researcher wishes to impose only mild assumptions about the shape of the functional relationships. This allows for a wide range of functional relationships in the data, but, as we shall see, it is necessary to add additional structure to the criterion function to estimate these type of models. The chapter discusses how this impacts the interpretation of basic estimated quantities, and discusses how an appropriate functional form can be estimated while jointly addressing the need for possible fixed effects. It will then be shown how the resulting models can be used to produce alternative future scenarios that take into account historical nonlinear patterns.

In Chapter 6, the discussion moves away from nonlinearities, and shifts the focus toward inter-temporal dynamics between multiple variables within a spatial system. Estimation of interdependence among multiple time series is often at the center of time series analysis, but many panel methods have traditionally been developed with inferential questions about a single dependent variable in mind. The model introduced in this chapter extends the standard spatial time series model to the multiple variable setting and introduces methods to analyze how finite impulse responses flow through a spatial system in the presence of both spatial and temporal forms of feedback. This is useful to address questions about the order in which effects occur over time when variables are not only temporally, but also spatially dependent. While the chapter introduces the analytical framework in a linear way, focusing on a relatively homogeneous subset of locations, the nonlinear concepts introduced in Chapter 4 and 5

can naturally be applied in similar settings to study nonlinear impulse response behavior in heterogeneous systems.

Finally, Chapter 7 circles back to some of the fundamental concepts that are introduced in Chapter 2 that covers background theory. Only this time, the discussion stays at a more general level and focuses on the concepts of probability and causal inference in dynamical systems. The discussion highlights why using flexible models, such as the ones introduced in this thesis, are desirable in the first place when one is interested in answering basic questions about cause and effect in a multivariate setting. An argument will be provided for flexible specification of the possible time dynamics in a spatial system together with estimation strategies that minimize distance to the true probability measure that underlies the observed data. In practice, this implies a general to specific approach to exclude irrelevant dependencies. The particular case of maximizing penalized Maximum Likelihood will be discussed further, which provides additional support for the estimation strategies used throughout this thesis.

# Chapter 2

## Background Theory

Asymptotic theory is the cornerstone of inferential statistics. The limiting distribution of a basic quantity of interest delivers properties that are accurate in large samples and often reasonable when there is moderate data. In particular, limiting distributions can be used for approximate inference based on approximate confidence intervals and their associated test statistics. The benefit of the limiting distribution over exact distributional results is that it can often be derived following general rules that are valid even for complicated models that include heterogeneity, interaction and nonlinearity. The exact distributions are, however, often difficult to derive, and may not even apply in certain cases of interest. Asymptotic distribution theory is centered around the notion of an expected mean and an expected variance. The general steps to establish these quantities of interest are to establish convergence of the mean and convergence of the variance under a notion of growing data.

Because asymptotic theory is crucial for econometric analysis, it is useful to have general results with conditions that can be applied to as many estimators as possible to deliver standard and identical interpretation to a wide range of empirical results. The purpose of this chapter is to present such results in a brief and common format adapted to the setting of spatial time series. The basic exposition sets the table for the later chapters that establish and discuss properties of complex models,

including some that have not been used in existing literature. References to literature on specific results and proofs, but also to advanced textbooks that have wide coverage, will be provided in the relevant sections in later chapters.

## 2.1 Linear estimators

In introductory econometrics books, the properties of standard estimators have been extensively studied. However, basic theory only works in the simplistic setting of linear models and requires the very restrictive assumption that the model is an exact description of reality (i.e. that the model is correctly specified). Generally, as the dimensions of the data grow in time, space, and number of variables, it becomes increasingly unlikely that the same average description appropriately describes local processes across all dimensions and levels in the data. It is more likely that the derivatives that describe marginal effects between dependent and independent data vary from one local mean to another across regions or regimes. While flexibility to cope with these transitions may be a natural idea, it is not always possible to simply allow for more complex model dynamics without breaking assumptions that are made under standard theory. In particular, the linearity of the standard regression model was key to obtaining an analytical expression for a simple estimator and the assumption of correct specification of the model was used to express the estimator in terms of deviations around the *true* parameter. The linearity of the model also made it straightforward to derive stationary conditions and ensure that a Law of Large Numbers and Central Limit Theorems can be applied to obtain the consistency and asymptotic normality of the estimator. For example, the LSE of the linear autoregressive parameter  $\beta$  in the model given by  $y_t = \beta y_{t-1} + \varepsilon_t$  takes the form:

$$\hat{\beta}_T = \frac{\sum_{t=2}^T y_t y_{t-1}}{\sum_{t=2}^T y_{t-1}^2}. \quad (2.1)$$

Deriving this expression was only possible because the model is linear, which drastically reduced the complexity of the calculus involved. Due to the simplicity, the properties of this estimator can also easily be analyzed if we assume that this linear description is correct, e.g. that our parametrization corresponds exactly with the *true* model that produced the observed data. This allows us to rewrite the estimator in terms of the *true* parameter  $\beta_0$  and a remainder,  $\beta_r$ :

$$\hat{\beta}_T = \beta_0 + \beta_r, \quad \beta_r = \frac{\sum_{t=2}^T \varepsilon_t y_{t-1}}{y_{t-1}^2}. \quad (2.2)$$

Furthermore, when dependence is linear, we can straightforwardly show that if  $|\beta_0| < 1$ , then the model is stationary. Stationarity then allows us to apply the LLN and CLT to  $\beta_r$  and as a result, following these simple steps, we can conclude that:

1. The remainder,  $\beta_r$ , vanishes to 0 as the time dimension  $T$  approaches infinity:

$$\frac{\sum_{t=2}^T \varepsilon_t y_{t-1}}{\sum_{t=2}^T y_{t-1}^2} \xrightarrow{p} 0 \quad \text{as } T \rightarrow \infty,$$

hence the estimator  $\hat{\beta}_T$  is consistent toward  $\beta_0$ .

2. The remainder,  $\beta_r$ , is asymptotically normally distributed:

$$\frac{\sum_{t=2}^T \varepsilon_t y_{t-1}}{\sum_{t=2}^T y_{t-1}^2} \xrightarrow{d} N(0, \sigma^2) \quad \text{as } T \rightarrow \infty,$$

hence the estimator  $\hat{\beta}_T$  is asymptotically normally distributed around  $\beta_0$ .



These simple results have proved extremely useful over time. For good reasons, the Law of Large Numbers, which took more than a staggering 300 years to complete, has been coined the *Golden Theorem*. In many cases, these simple results are more than just interesting, and remain the work horse of standard analysis approaches that are widely used to support policies and interventions across many domains. However, they are applicable only in the limited setting of linear models and under the very restrictive assumption that this linear relationship describes reality correctly.

### 2.1.1 The linear Least Squares Estimator

Many empirical problems dealing with repeated cross-sectional data can be analyzed by the linear regression model:

$$\mathbf{y}_t = \alpha + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t \quad \forall t \in \mathbb{N}, \quad (2.3)$$

where  $\mathbf{y}_t$  is the dependent vector variable at time  $t$  containing  $i \in \{1, \dots, N\}$  values each observed at a different location,  $\mathbf{X}_t$  is a  $d$ -dimensional matrix containing the independent or explanatory variables similarly observed at locations  $i \in \{1, \dots, N\}$  and time  $t$ , and  $\boldsymbol{\varepsilon}_t$  are the unobserved residuals. The parameter  $\alpha$  is a constant, and  $\boldsymbol{\beta}$  is a vector of length  $d$  containing the marginal effects, or slope parameters, for each variable included in  $\mathbf{X}_t$ . The error term is assumed to satisfy  $\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{X}_t) = 0$ . Under this assumption, the linear regression model is a model of the conditional expectations of  $\mathbf{y}_t$  given the observed  $\mathbf{X}_t$ . In particular, one can decompose the problem as follows:

$$\mathbb{E}(\mathbf{y}_t|\mathbf{X}_t) = \mathbb{E}(\alpha + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t|\mathbf{X}_t). \quad (2.4)$$

Naturally, given that the expectation of a static parameter is simply the value of that parameter, the right hand side can be separated in

individual parts,  $\alpha$ ,  $\mathbb{E}(\mathbf{X}_t|\mathbf{X}_t)\boldsymbol{\beta}$ , and  $\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{X}_t)$ . Furthermore,

$$\mathbb{E}(\mathbf{X}_t|\mathbf{X}_t) = \mathbf{X}_t, \quad (2.5)$$

and by assumption,

$$\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{X}_t) = 0. \quad (2.6)$$

Hence, the expectation of  $\mathbf{y}_t$  conditional on observables is simply:

$$\mathbb{E}(\mathbf{y}_t|\mathbf{X}_t) = \alpha + \mathbf{X}_t\boldsymbol{\beta}. \quad (2.7)$$

This interpretation will turn out to remain incredibly useful in the nonlinear case as well, as, no matter how complex the model gets, the modeled data can often be interpreted as local conditional expectations rather than global (average) expectations, which is still an intuitively accessible concept. The key exogeneity assumption used for this, can be summarized as follows:

ASSUMPTION. 1 (Exogeneity of the Regressors).  $\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{X}_t) = 0 \ \forall \ t \in \mathbb{N}$ .

REMARK. 1. *Note that by a Law of Total Expectation, the Exogeneity of Regressors assumption also implies*

$$\mathbb{E}(\boldsymbol{\varepsilon}_t\mathbf{x}_t) = \mathbb{E}(\mathbb{E}(\boldsymbol{\varepsilon}_t\mathbf{x}_t|\mathbf{x}_t)) = \mathbb{E}(\mathbf{x}_t\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{x}_t)) = \mathbb{E}(\mathbf{x}_t0) = 0 \ \forall \ t \in \mathbb{N}.$$

Note that  $\boldsymbol{\varepsilon}_t$  is a vector of residuals at time  $t$  for locations  $i \in \{1, \dots, N\}$ . The conditional expectation condition is stated for vectors indexed by time intervals. Essentially, the parameters in the vector  $\boldsymbol{\beta}$  measure the expected changes in the cross-section  $\mathbf{y}_t$  given the changes in  $\mathbf{X}_t$ . While it may well be that  $\mathbb{E}(\varepsilon_{it}|\mathbf{x}_{it}) = 0 \ \forall \ t \in \mathbb{N}$  for certain locations (or the cross-sectional mean),  $\mathbb{E}(\boldsymbol{\varepsilon}_t|\mathbf{X}_t) = 0 \ \forall \ t \in \mathbb{N}$  may still break, if for example local errors have non-zero expectation ( $\varepsilon_{it}|\mathbf{x}_{it}) \neq 0$ , which for example occurs when there are expectations about missing components conditional on the data locally in the cross-section. One such example is clustering of residuals in regions in the cross-section, particularly if those clusters tend to remain in place over time. There are many reasons why

this assumption may be difficult to hold in practice. Advanced modeling techniques, including those discussed in later chapters, are in fact often aimed at mitigating these violations.

Let us now first consider the simple LSE that chooses the parameters that minimize the sum of squared residuals from a compact collection of potential solutions  $(\mathcal{A}, \mathcal{B})$ . Specifically:

$$(\hat{\alpha}, \hat{\beta}) = \arg \min_{(\alpha, \beta) \in (\mathcal{A}, \mathcal{B})} \sum_{t=1}^T \epsilon_t^2 = \arg \min_{(\alpha, \beta) \in (\mathcal{A}, \mathcal{B})} \sum_{t=1}^T (\mathbf{y}_t - \alpha + \mathbf{X}_t \beta)^2. \quad (2.8)$$

As always, the parameters can be found by simply taking the derivative of this Least Squares criterion with respect to its parameters, and equating 0. Supposing we omit  $\alpha$  for a moment, for example because we have demeaned the data such that the average is 0, and focus on the simple case of just one regressor, we can find  $\hat{\beta}$  using the derivative:

$$\frac{\partial \sum_{t=1}^T (\mathbf{y}_t - \beta \mathbf{x}_t)^2}{\partial \beta} = \sum_{t=1}^T (\mathbf{y}_t - \beta \mathbf{x}_t) \mathbf{x}_t, \quad (2.9)$$

which can be rearranged to obtain our estimate explicitly:

$$\hat{\beta}_T = \frac{\sum_{t=1}^T \mathbf{y}_t \mathbf{x}_t}{\sum_{t=1}^T \mathbf{x}_t^2}. \quad (2.10)$$

Deriving estimators for multiple parameters, each being a marginal effect with respect to a different variable or a simple constant, only involves longer derivations. The linear LSE can always be derived analytically. This is incredibly useful. Even in the nonlinear case we often use flexible functionals that generate parameterizations that are locally linear, in which case the same strategies can be applied for the resulting locally linear expressions only at the cost of longer equations.

The first important step now is to establish that the estimator is consistent toward the parameter of interest. That is, that it converges in probability toward the set of parameters,  $(\alpha_0, \beta_0)$ , that deliver a correct description of

the data, as  $T \rightarrow \infty$ . This requires us to assume that this set of correct parameters is in fact included in the space of considered parameters  $(\mathcal{A}, \mathcal{B})$ . We will return to this assumption in later chapters and try to get an understanding of what this truly means, and more importantly, what it means if this assumption breaks. For now, let us summarize:

ASSUMPTION. 2 (Correct Specification of the Model). *The regression  $\mathbf{y}_t = \alpha + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t \forall t \in \mathbb{N}$  is correctly specified.*

As before, this allows us to write the estimator in terms of the *true* parameter and a remainder that involves the residuals, from where we can show that this remainder term converges to 0 as  $T$  grows, leaving us with an estimator that converges to the correct result. Let us now state the exact Theorem.

THEOREM. 1 (Bernoulli's Law of Large Numbers for Independent and Identically Distributed Data). *Let  $z_1, z_2, z_T$  be an Independent and Identically Distributed random variable with finite first moment,  $\mathbb{E}|z_t| < \infty$ . Then,*

$$\frac{1}{T} \sum_{t=1}^T z_t \xrightarrow{p} \mathbb{E}(z_t) \quad \text{as } T \rightarrow \infty.$$

This Theorem tells us that disregard of the distribution of  $z$ , the sample average is a consistent estimator of the *true* mean. It is easy to see that this Theorem can also be applied to cross-sectional data, in which case we would index the observations cross-sectionally. The main issue that results is that observations are often not independent across space as by the definition of neighborhood relationships, independence is violated. This similarly applies to the endogenous time series case, in which we assume dependence of observations over time. For now, this Theorem is sufficient as we are interested in the relationship between  $\mathbf{y}_t$  and exogenous variables  $\mathbf{X}_t$  for which no process has been defined at this point. The application to the LSE follows by first noting that the criterion is a function of random variables, hence noting that it is itself is a random

variable, and then multiplying the numerator and the denominator of the remainder term by  $\frac{1}{T}$ , and applying the LLN to both components. In particular, again for the simple case,

$$\beta_r = \frac{\sum_{t=2}^T \boldsymbol{\varepsilon}_t \mathbf{x}_t}{\sum_{t=2}^T \mathbf{x}_t^2} = \frac{\frac{1}{T} \sum_{t=2}^T \boldsymbol{\varepsilon}_t \mathbf{x}_t}{\frac{1}{T} \sum_{t=2}^T \mathbf{x}_t^2}, \quad (2.11)$$

and if both  $\{\boldsymbol{\varepsilon}_t \mathbf{x}_t\}$  and  $\{\mathbf{x}_t^2\}$  are i.i.d. with finite first moment  $|\boldsymbol{\varepsilon}_t \mathbf{x}_t| < \infty$  and  $|\mathbf{x}_t^2| < \infty$ , then

$$\frac{1}{T} \sum_{t=2}^T \boldsymbol{\varepsilon}_t \mathbf{x}_t \xrightarrow{p} \mathbb{E}(\boldsymbol{\varepsilon}_t \mathbf{x}_t) \quad \text{and} \quad \frac{1}{T} \sum_{t=2}^T \mathbf{x}_t^2 \xrightarrow{p} \mathbb{E}(\mathbf{x}_t^2) \quad \text{as} \quad T \rightarrow \infty.$$

Note that by our first assumption,  $\mathbb{E}(\boldsymbol{\varepsilon}_t \mathbf{x}_t) = 0$ , and because the Least Squares criterion is continuous, and functions are limit-preserving even if their arguments are sequences of random variables, the LLN thus delivers

$$\beta_r = \frac{\frac{1}{T} \sum_{t=2}^T \boldsymbol{\varepsilon}_t \mathbf{x}_t}{\frac{1}{T} \sum_{t=2}^T \mathbf{x}_t^2} \xrightarrow{p} \frac{0}{\mathbb{E}(\mathbf{x}_t^2)} = 0 \quad \text{as} \quad T \rightarrow \infty.$$

We have now proven that the estimator is consistent because the error in our estimation converges to zero as we collect more and more data over the time dimension. Note that the above derivations shows the criticality of assuming that the regressors are exogenous  $\mathbb{E}(\boldsymbol{\varepsilon}_t | \mathbf{x}_t) = 0$ , otherwise

$$\beta_r = \frac{\frac{1}{T} \sum_{t=2}^T \boldsymbol{\varepsilon}_t \mathbf{x}_t}{\frac{1}{T} \sum_{t=2}^T \mathbf{x}_t^2} \xrightarrow{p} \frac{\eta}{\mathbb{E}(\mathbf{x}_t^2)} = \epsilon \neq 0 \quad \text{as} \quad T \rightarrow \infty.$$

With  $\eta$  and  $\epsilon$  being unknown non-zero components, hence  $\beta_r$ , and therefore  $\hat{\beta}_T$ , converge to unknown real-valued constants. In other words, we can't really tell what limit our criterion converges to, which renders the entire estimation result quite arbitrary.

Often, the finite moments of lower constituents of complex regression models are introduced as a separate assumption, and we shall see that

instead of assuming these conditions it is often possible to verify the assumptions by defining a process for the endogenous regressors and validating that certain stability conditions and moment-preserving properties hold within specified parameter ranges. For now, let us collect our simple assumption as follows:

ASSUMPTION. 3 (Finite First Moments). *Assume that*

1.  $|\varepsilon_t \mathbf{x}_t| < \infty$ ,
2. and  $|\mathbf{x}_t^2| < \infty$ .

for each  $\mathbf{x}_t$  contained in  $\mathbf{X}_t$ .

We can collect the general consistency result of the LSE.

COROLLARY. 1 (Consistency of the Correctly Specified Least Squares Estimator). *Let  $\{\mathbf{y}_t\}_{t \in \mathbb{N}}$  and  $\{\mathbf{X}_t\}_{t \in \mathbb{N}}$  be observed sequences, and the model*

$$\mathbf{y}_t = \alpha + \mathbf{X}_t \boldsymbol{\beta} + \varepsilon_t \quad \forall t \in \mathbb{N},$$

*be correctly specified. Furthermore, let  $\{\varepsilon_t \mathbf{x}_t\}_{t \in \mathbb{N}}$  and  $\{\mathbf{x}_t^2\}_{t \in \mathbb{N}}$  be i.i.d. with  $\mathbb{E}(\varepsilon_t | \mathbf{x}_t) = 0 \quad \forall t \in \mathbb{N}$  and  $|\varepsilon_t \mathbf{x}_t| < \infty$  and  $|\mathbf{x}_t^2| < \infty$  for each  $\mathbf{x}_t$  contained in  $\mathbf{X}_t$ . Then, the Least Squares estimator of  $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$  defined as*

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \arg \min_{(\alpha, \boldsymbol{\beta}) \in (\mathcal{A}, \mathcal{B})} \sum_{t=1}^T (\mathbf{y}_t - \alpha + \mathbf{X}_t \boldsymbol{\beta})^2$$

*is consistent*

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) \xrightarrow{p} (\alpha_0, \boldsymbol{\beta}_0) \quad \text{as } T \rightarrow \infty.$$

In practice, one is also interested in making statements about the probability that our estimates of individual components in  $(\alpha_0, \boldsymbol{\beta}_0)$  are different from 0. That allows us to say that estimated economic effects are *significantly* different from 0, e.g. that an intervention had effect. This requires us to know the distribution of the estimator, which in practice is unknown. Luckily, we can approximate this distribution by appealing to the Central Limit Theorem and showing that the estimator is approximately normally distributed when  $T$  is large.

THEOREM. 2 (Lindeberg-Levy's Central Limit Theorem for Independent and Identically Distributed Data). *Let  $z_1, z_2, z_T$  be an Independent and Identically Distributed random variable with  $\mathbb{E}|z_t| = \mu < \infty$  and  $\text{Var}(z_t) = \sigma^2 < \infty$ , then*

$$\sqrt{T} \left( \frac{1}{T} \sum_{t=1}^T (z_t - \mu) \right) \xrightarrow{d} N(0, \sigma^2) \quad \text{as } T \rightarrow \infty.$$

We can now use the CLT to obtain the asymptotic normality of our correct LSE of any parameter by first writing  $\sqrt{T}(\hat{\beta} - \beta_0)$  and then plugging in our estimator in terms of the *true* parameter and the remainder term:

$$\sqrt{T}(\hat{\beta} - \beta_0) = \sqrt{T}((\beta_0 + \beta_r) - \beta_0) = \sqrt{T}(\beta_r) = \sqrt{T} \left( \frac{\frac{1}{T} \sum_{t=2}^T \varepsilon_t \mathbf{x}_t - \mathbb{E}(\varepsilon_t \mathbf{x}_t)}{\frac{1}{T} \sum_{t=2}^T \mathbf{x}_t^2} \right). \quad (2.12)$$

The term  $\mathbb{E}(\varepsilon_t \mathbf{x}_t)$  can be added, as by our first assumption, exogeneity of the regressors, this term equals 0. We can now apply the CLT to the numerator:

$$\sqrt{T} \left( \frac{1}{T} \sum_{t=2}^T \varepsilon_t \mathbf{x}_t - \mathbb{E}(\varepsilon_t \mathbf{x}_t) \right) \xrightarrow{d} N(0, \sigma^2 \mathbb{E}(\mathbf{x}_t^2)) \quad \text{as } T \rightarrow \infty,$$

and the LLN to the denominator:

$$\left( \frac{1}{T} \sum_{t=2}^T \mathbf{x}_t^2 \right) \xrightarrow{p} \mathbb{E}(\mathbf{x}_t^2) \quad \text{as } T \rightarrow \infty.$$

By Slutsky's Theorem, we now have

$$\sqrt{T}(\hat{\beta} - \beta_0) \xrightarrow{d} \frac{N(0, \sigma^2 \mathbb{E}(\mathbf{x}_t^2))}{\mathbb{E}(\mathbf{x}_t^2)} N(0, \sigma^2 \mathbb{E}[\mathbf{x}_t^2]^{-1}).$$

This is the standard strategy to deliver asymptotic normality, which we can summarize in the following general result. First, note that the CLT imposes a stricter moment assumption. In particular:

ASSUMPTION. 4 (Finite Second Moments). *Assume that*

$$1. \text{Var}(\varepsilon_t \mathbf{x}_t) < \sigma^2 < \infty,$$

for each  $\mathbf{x}_t$  contained in  $\mathbf{X}_t$ .

While this assumption is stated in terms of the second moment, variance, of  $\boldsymbol{\varepsilon}_t \mathbf{x}_t$ , it is sometimes stated in terms of higher moments of the lower constituents  $\boldsymbol{\varepsilon}_t$  and  $\mathbf{x}_t$  individually. In particular, since the variance involves squared terms, it can be shown that this assumption involves the finiteness of the fourth moments of  $\boldsymbol{\varepsilon}_t$  and each  $\mathbf{x}_t$  contained in  $\mathbf{X}_t$ . Intuitively, if the fourth moments are finite, then the tails of the distributions are relatively short, so the probability that an unusually large observations occurs is small. In that regard, this is interpreted by many as an indication that Least Squares estimates are very sensitive to the presence of outliers. Similar assumptions are however made when establishing the properties of other estimators, including those that aim at outliers-robustness by assuming non-Gaussian distributions that can better accommodate tail events. It turns out that many proofs of multivariate nonlinear estimators require even higher moments to exist.

**COROLLARY. 2** (Asymptotic Normality of the Correctly Specified Least Squares Estimator). *Let  $\{\mathbf{y}_t\}_{t \in \mathbb{N}}$  and  $\{\mathbf{X}_t\}_{t \in \mathbb{N}}$  be observed sequences, and the model*

$$\mathbf{y}_t = \alpha + \mathbf{X}_t \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t \quad \forall t \in \mathbb{N},$$

*be correctly specified. Let  $\{\boldsymbol{\varepsilon}_t \mathbf{x}_t\}_{t \in \mathbb{N}}$  and  $\{\mathbf{x}_t^2\}_{t \in \mathbb{N}}$  be i.i.d. with  $\mathbb{E}(\boldsymbol{\varepsilon}_t | \mathbf{x}_t) = 0 \quad \forall t \in \mathbb{N}$  and  $|\boldsymbol{\varepsilon}_t \mathbf{x}_t| < \infty$  and  $|\mathbf{x}_t^2| < \infty$  for each  $\mathbf{x}_t$  contained in  $\mathbf{X}_t$ . Suppose furthermore that the variances  $\text{Var}(\boldsymbol{\varepsilon}_t \mathbf{x}_t) < \sigma^2 < \infty$  are finite for each  $\mathbf{x}_t$  contained in  $\mathbf{X}_t$ . Then, the Least Squares estimator of  $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$  defined as*

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \arg \min_{(\alpha, \boldsymbol{\beta}) \in (\mathcal{A}, \mathcal{B})} \sum_{t=1}^T (\mathbf{y}_t - \alpha + \mathbf{X}_t \boldsymbol{\beta})^2$$

*is asymptotically normally distributed for each parameter  $\theta \in (\alpha, \boldsymbol{\beta})$  and variable  $\mathbf{x}_t$  associated with that parameter*

$$\hat{\theta}_T \xrightarrow{\text{approx}} N \left( \theta_0, \sigma^2 \left[ \sum_{t=1}^T \mathbf{x}_t^2 \right]^{-1} \right).$$



Similar results can also be obtained when focusing on the case where  $\mathbf{x}_t$  is replaced by a lag of the endogenous variable  $\mathbf{y}_{t-1}$ . In this case, the exogenous regressors assumption is stated  $\mathbb{E}(\boldsymbol{\varepsilon}_t | \mathbf{y}_{t-1}) = 0 \ \forall t \in \mathbb{Z}$ . This implies that conditional on the past, no further information about the residuals can be available. This essentially requires that the residual process must be free from further correlations after filtering the time-dependencies conditional on lags and observable components from the dependent variable. In many cases there may still be correlations in the innovations, for example because policies impact a process not only idiosyncratically but for prolonged periods. Models therefore often include lagged residuals as explanatory variables. Apart from the need to render an observed time series free from time correlations to fulfill the assumptions needed to apply the LLN and CLT, finite moments can also not simply be assumed when the model is correct. In fact, we know that for certain parameter values the process is explosive such that  $\mathbf{y}_t$  is in fact expected to tend to infinity. To prevent this from occurring, we need an additional result that ensures that  $\mathbf{y}_t$  is Stationary. The following result specifically, is useful in standard settings.

**THEOREM. 3 (Strict Stationarity of a Linear Recursion).** *Let  $\{\mathbf{y}_t\}_{t \in \mathbb{Z}}$  be generated by:*

$$\mathbf{y}_t = \alpha + \phi \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \ \forall t \in \mathbb{Z}.$$

*If  $|\phi| < 1$  and  $\boldsymbol{\varepsilon}_t$  are innovations drawn from  $NID(0, \sigma_{\boldsymbol{\varepsilon}}^2)$ , then  $\{\mathbf{y}_t\}_{t \in \mathbb{Z}}$  is Strictly Stationary, that is the distribution of every finite sub-vector is invariant in time*

$$F_Y(\mathbf{y}_1, \dots, \mathbf{y}_\tau) = F_Y(\mathbf{y}_{t+1}, \dots, \mathbf{y}_{t+\tau}) \ \forall (t, \tau) \in \mathbb{N} \times \mathbb{N}.$$

*where  $F_Y(\mathbf{y}_{t+1}, \dots, \mathbf{y}_{t+\tau})$  represents the cumulative distribution function of the unconditional joint distribution of  $\{\mathbf{y}_t\}_{t \in \mathbb{Z}}$  at times  $t+1, \dots, t+\tau$ .*

This stationarity property is incredibly important to obtain properties of estimators because it allows us to make use of the Laws of Large Numbers

for Stationary and Ergodic data, and if the model is correctly specified the Central Limit Theorem for Stationary and Ergodic Martingale Difference Sequences, rather than appealing to the Theorems for *i.i.d.* data. This extension will be discussed in more detail in the next section. If the model remains linear, but multiple (cross-sectional) variables are included, or a single cross-sectional time series is modeled with multiple locational autoregressive parameters  $\Phi \mathbf{y}_{t-1}$  collected in the  $N \times N$  matrix  $\Phi$ , the linear Stationarity condition can be generalized as  $\|\Phi\| < 1$ , using some norm or a spectral radius. However, when the process turns nonlinear, and we can no longer condition on static parameters, proofs for Stationarity become more complex. Particularly when analyzing cross-sectional time series we not only want observations to depend possibly on unique local histories, but also on those of neighbors and possibly even on the contemporaneous values of neighbors. In these cases, models begin to exhibit more complex feedback properties for which proving stability may turn out to be a nontrivial task. At this point, one may start to make explicit distinctions between various types of stability as sometimes weaker forms of stability, that are easier to verify, may already be sufficient to obtain useful properties of estimators.

We shall return extensively to both the stability conditions and the residual dependencies in later chapters. For now, let us explore what happens to our LSE if we would want to model contemporaneous dependencies on neighbors in addition to the exogenous covariates of interest. This will highlight what can already be done with the simple theory that we have developed so far and expose some of its limitations. Suppose we extend our regression model:

$$\mathbf{y}_t = \alpha + \rho W \mathbf{y}_t + \mathbf{X}_t \beta + \boldsymbol{\varepsilon}_t \quad \forall t \in \mathbb{N}, \quad (2.13)$$

in which  $W$  is an  $N$  by  $N$  pre-defined parameter matrix with zero diagonal. We reserve discussion about this matrix, that defines contemporaneous relations with neighboring observations, for later chapters. For now it

is sufficient to see that  $\mathbf{y}_t$  occurs on both sides of the equation and the exogenous regressors assumption is thus now stated  $\mathbb{E}(\mathbf{y}_t - \alpha - \mathbf{X}_t\boldsymbol{\beta} - \rho W\mathbf{y}_t | W\mathbf{y}_t) = 0 \forall t \in \mathbb{N}$ , which obviously makes little sense to impose since  $W\mathbf{y}_t$  occurs on both sides. Only if  $\rho = 0$ , and the model is non-spatial, the expectation is zero by the fact that the residuals are *i.i.d.* In other words,  $W\mathbf{y}_t$  is an endogenous regressor. Contrary to the time series case, where the lagged term of the dependent variable can be uncorrelated with the residual term if there is no serial residual correlation, e.g. if the model is correct, in the spatially lagged case, this correlation occurs regardless of the properties of the residual term. We had already seen that the Least Squares criterion converges to an unknown limit if the exogenous regressor assumption breaks, implying that standard application of the Least Squares criterion delivers arbitrary results.

One option is to invert the equation, and ensure that  $\mathbf{y}_t$  only enters on the left side of the equation:

$$(I - \rho W)\mathbf{y}_t = \alpha + \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t \forall t \in \mathbb{N}, \quad (2.14)$$

with  $I$  being an identity matrix. At this point, our dependent variable contains unknown parameters. We can get rid of  $(I - \rho W)$  on the left side by division:

$$\mathbf{y}_t = (I - \rho W)^{-1}\alpha + (I - \rho W)^{-1}\mathbf{X}_t\boldsymbol{\beta} + (I - \rho W)^{-1}\boldsymbol{\varepsilon}_t \forall t \in \mathbb{N}. \quad (2.15)$$

This highlights that when  $\mathbf{y}_t$  is in part a function of  $W\mathbf{y}_t$ , e.g. when  $|\rho| > 0$ ,  $\mathbf{y}_t$  is a nonlinear function of the data and residuals. The model cannot be parameterized and estimated in this form because the residuals result as a product of estimation, hence their values are not available *a priori* as regressors. Chapter 4 discusses that the nonlinearity can be approximated using an infinite power series approximation, which reveals that  $\mathbf{y}_t$  not only depends on local observations and neighbors, but also

on the values of residuals and covariates of distant neighbors.

$$\mathbf{y}_t = (I + \rho W + \rho^2 W^2 + \dots) (\alpha + \mathbf{X}_t \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t) \quad \forall t \in \mathbb{N}. \quad (2.16)$$

The influence of distant neighbors will be small if  $\rho$  is not too high. This suggests that when spatial dependence is mild and residuals are small, a considerable share of the dependencies can be captured with a first order approximation of the spillover dynamics.

$$\begin{aligned} \mathbf{y}_t &\sim (I + \rho W) (\alpha + \mathbf{X}_t \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t) + \mu_t \\ &\sim (I + \rho W) (\alpha + \mathbf{X}_t \boldsymbol{\beta}) + \boldsymbol{\varepsilon}_t + \mu_t + \xi_t \quad \forall t \in \mathbb{N}, \end{aligned} \quad (2.17)$$

in which  $\mu_t$  is an approximation error that results from restricting to dependence on first order neighbors, and  $\xi_t$  is an additional approximation error that results from neglecting the residual spillovers. The magnitude of both errors increases with  $|\rho|$ , and the magnitude of  $\xi_t$  increases with the magnitude of residuals  $\boldsymbol{\varepsilon}_t$ . The aim is then to specify as many lower-level constituents of the residuals by incorporating many covariates to ensure that residuals are small, and parameterize spatial dependence on covariates directly to capture the important first order spatial dependence dynamics. The resulting simplified model can consistently be estimated using Least Squares as it is simply equal to a standard regression introduced in the previous section:

$$\mathbf{y}_t = \alpha + \mathbf{X}_t \boldsymbol{\beta} + W \mathbf{X}_t \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_t \quad \forall t \in \mathbb{N}. \quad (2.18)$$

In this equation, we made use of the fact that  $(I + \rho W)\alpha$  simply remains a linear constant and introduced a new unknown set of parameters  $\boldsymbol{\beta}_2$  to capture dependence on neighboring values of the exogenous covariates. Note that our simple estimation theorems at this point still require the correct specification assumption to be satisfied, which is unrealistic since we have already established sources of approximation error that stem from neglecting the spatial effects in residuals and dependence on distant

observations.

While the validity of the correct specification assumption can be verified by diagnosing  $\varepsilon_t$ , the approach may be seen as a dis-satisfactory as it provides no empirical strategy to dealing with residual spatial correlation or pure SAR processes in which exogenous covariates play no role. The question naturally arises if other, alternative, estimators can be thought of that are not prone to this problem and that can handle estimation of spatial disturbance terms directly. It turns out that the problem can be tackled with the framework of Maximum Likelihood.

### 2.1.2 The linear Maximum Likelihood Estimator

Given  $T$  observations  $\mathbf{y}_1, \dots, \mathbf{y}_T$  from the time series  $\{\mathbf{y}_t\}_{t \in \mathbb{Z}}$ , generated by the model

$$\mathbf{y}_t = \phi \mathbf{y}_{t-1} + \varepsilon_t \quad \forall t \in \mathbb{Z}, \quad (2.19)$$

with  $\varepsilon_t$  being drawn from a standardized normal distribution with zero mean. Suppose we have a correctly specified regression. The likelihood function  $\ell(\mathbf{y}_1, \dots, \mathbf{y}_T; \boldsymbol{\theta})$  is simply the joint density function of the sequence  $\mathbf{y}_1, \dots, \mathbf{y}_T$  under the parameter vector  $\boldsymbol{\theta} = (\phi, \sigma_\varepsilon^2)$  that defines the distribution of the data. Note that if our model would include more or other parameters, they would simply be part of this parameter vector (for example, if we would include a constant as we did earlier, it would be  $\boldsymbol{\theta} = (\alpha, \phi, \sigma_\varepsilon^2)$ ). The MLE is the parameter vector that the maximizes the likelihood function:

$$\hat{\boldsymbol{\theta}}_T = \arg \max_{\boldsymbol{\theta} \in \Theta} \ell(\mathbf{y}_1, \dots, \mathbf{y}_T; \boldsymbol{\theta}). \quad (2.20)$$

A useful property of joint density functions is that they can be factorized into the product of conditional and marginal densities:

$$\ell(\mathbf{y}_1, \mathbf{y}_2; \boldsymbol{\theta}) = \ell(\mathbf{y}_1; \boldsymbol{\theta}) \times \ell(\mathbf{y}_2; \boldsymbol{\theta}),$$

$$\ell(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3; \boldsymbol{\theta}) = \ell(\mathbf{y}_1; \boldsymbol{\theta}) \times \ell(\mathbf{y}_2 | \mathbf{y}_1; \boldsymbol{\theta}) \times \ell(\mathbf{y}_3 | \mathbf{y}_2, \mathbf{y}_1; \boldsymbol{\theta}),$$

...

$$\ell(\mathbf{y}_1, \dots, \mathbf{y}_T; \boldsymbol{\theta}) = \ell(\mathbf{y}_1; \boldsymbol{\theta}) \times \prod_{t=2}^T \ell(\mathbf{y}_t | \mathbf{y}_{t-1}, \dots, \mathbf{y}_1; \boldsymbol{\theta}). \quad (2.21)$$

Writing the likelihood as a product of conditional densities is useful because we impose the distribution of  $\mathbf{y}_t$  conditional on  $\mathbf{y}_{t-1}$  through our parameterized model. For example, in the linear autoregressive case that we have assumed, with  $\phi$  being the linear autoregressive parameter, it is

$$\mathbf{y}_t | \mathbf{y}_{t-1} \sim N(\phi \mathbf{y}_{t-1}, \sigma_\varepsilon^2). \quad (2.22)$$

It may also be possible to work with different distributions, for example distributions that can accommodate fatter tails. Different distributional assumptions or models will merely imply that the densities are of another form, which can be accounted for. Under the Gaussian assumption, it is given by the well-known formula:

$$\ell(\mathbf{y}_t | \mathbf{y}_{t-1}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp \left[ -\frac{(\mathbf{y}_t - \phi \mathbf{y}_{t-1})^2}{2\sigma_\varepsilon^2} \right]. \quad (2.23)$$

Taking logs allows us to express the products as sums, hence we have that the MLE can be written as

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \Theta} \sum_{t=2}^T -\log \sqrt{2\pi\sigma_\varepsilon^2} - \frac{(\mathbf{y}_t - \phi \mathbf{y}_{t-1})^2}{2\sigma_\varepsilon^2}. \quad (2.24)$$

Just as in the Least Squares case, we can find the estimator by calculating the derivative and setting it to zero. Since in this simple example we have assumed  $\sigma = 1$ , we will set it to unit. In practice, the variance is often estimated, in which case the derivations have to take into account that  $\sigma$  is itself a free parameter. For now, the estimator for  $\phi$  is simply:

$$\frac{\partial \ell(\mathbf{y}_1, \dots, \mathbf{y}_T; \phi)}{\partial \phi} = \sum_{t=2}^T (\mathbf{y}_t - \phi \mathbf{y}_{t-1}) \mathbf{y}_{t-1}. \quad (2.25)$$

Equating to zero and rearranging gives us a familiar expression:

$$\hat{\phi} = \frac{\sum_{t=2}^T \mathbf{y}_t \mathbf{y}_{t-1}}{\sum_{t=2}^T \mathbf{y}_{t-1}^2}. \quad (2.26)$$

In this particular case, in which we have assumed the same model and distributional form of the residuals, the MLE is identical to the LSE that we explored earlier. Since we now have an analytical expression, and have again assumed correct specification, we might expect that the proofs for consistency and normality will follow the exact same steps from here. This is almost correct. In the early Least Squares example, we worked with a model  $\mathbf{y}_t = \mathbf{X}_t \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t$ ,  $\forall t \in \mathbb{N}$  in which our dependent variable is generated by exogenous, independent, data  $\mathbf{X}_t$ . In the current example  $\mathbf{y}_t = \phi \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t$ ,  $\forall t \in \mathbb{Z}$ , our dependent variable is generated only by innovations and temporal dependence. This implies that unlike in the exogenous regressor case, where we can assume that  $\mathbf{X}_t$  is *i.i.d.*, we can now no longer assume that  $\mathbf{y}_{t-1}$  is *i.i.d.* as we model its dependence explicitly. This implies that we can no longer make use of Bernoulli's LLN and the Lidneberg-Levy's CLT. Note that this is an issue that is not related to the MLE itself, or the LSE vice-versa, it is just because we are now formally considering a time series process. We already hinted earlier that stability of the time series was intuitively important to ensure that  $\mathbf{y}_t$  does not wander off to infinity, in which case the expectations are infinite and moment assumptions would surely break. It turns out that the Stationarity property is also key to applying an LLN and CLT. Particularly, since the derivations are identical to the Least Squares example, we want to show that by application of an LLN that

$$\phi_r = \frac{\frac{1}{T} \sum_{t=2}^T \boldsymbol{\varepsilon}_t \mathbf{y}_{t-1}}{\frac{1}{T} \sum_{t=2}^T \mathbf{y}_{t-1}^2} \xrightarrow{p} \frac{0}{\mathbb{E}(\mathbf{y}_t^2)} = 0 \quad \text{as } T \rightarrow \infty,$$

and by application of a CLT to the numerator and a LLN to the denominator, that

$$\sqrt{T}(\phi_r) = \sqrt{T} \left( \frac{\frac{1}{T} \sum_{t=2}^T \varepsilon_t \mathbf{y}_{t-1} - \mathbb{E}(\varepsilon_t \mathbf{y}_{t-1})}{\frac{1}{T} \sum_{t=2}^T \mathbf{y}_{t-1}^2} \right) \xrightarrow{d} \frac{N(0, \sigma_\varepsilon^2 \mathbb{E}(\mathbf{y}_t^2))}{\mathbb{E}(\mathbf{y}_{t-1}^2)} \sim$$

$$N(0, \sigma_\varepsilon^2 \mathbb{E}[\mathbf{y}_{t-1}^2]^{-1}),$$

as  $T \rightarrow \infty$ .

We can do so by appealing to the following LLN for Strictly Stationary and Ergodic sequences and CLT for Martingale Difference Sequences.

**THEOREM. 4** (Birkhoff-Khinchin's Law of Large Numbers for Strictly Stationary and Ergodic data). *Let the random sequence  $\{z_t\}_{t \in \mathbb{Z}}$  be Strictly Stationary and Ergodic with finite first moment  $\mathbb{E}|z_t| < \infty$ , then we have*

$$\frac{1}{T} \sum_{t=2}^T z_t \xrightarrow{p} \mathbb{E} z_t \quad \text{as } T \rightarrow \infty.$$

**THEOREM. 5** (Billingsley's Central Limit Theorem for Stationary and Ergodic Martingale Difference Sequences). *Let the sequence  $\{z_t\}_{t \in \mathbb{Z}}$  be Strictly Stationary and Ergodic with first moment  $\mathbb{E}(z_t) = \mu < \infty$  and second moment  $\text{Var}(z_t) = \sigma^2 < \infty$ . Suppose furthermore that  $\{z_t\}_{t \in \mathbb{Z}}$  is a Martingale Difference Sequence of random variables,  $\mathbb{E}(z_t | z_{t-1}, z_{t-2}, \dots) \forall t \in \mathbb{Z}$ , then we have*

$$\sqrt{T} \left( \frac{1}{T} \sum_{t=2}^T z_t - \mu \right) \xrightarrow{d} N(0, \sigma^2) \quad \text{as } T \rightarrow \infty.$$

Using these Theorems, together with the stationarity property, we come to the following results.

**COROLLARY. 3** (Consistency of the MLE for the Correctly Specified Autoregressive Model). *Let the time series  $\{\mathbf{y}_t\}_{t \in \mathbb{Z}}$  be generated by the Strictly Stationary autoregressive model  $\mathbf{y}_t = \phi_0 \mathbf{y}_{t-1} + \varepsilon_t \forall t \in \mathbb{Z}$ ,  $|\phi_0| < 1$ , with exogenous innovations  $\mathbb{E}(\varepsilon_t | \mathbf{y}_{t-1}) = 0$  that satisfy  $\{\varepsilon_t\}_{t \in \mathbb{Z}} \sim \text{NID}(0, \sigma_\varepsilon^2)$  with finite variance  $\sigma_\varepsilon^2 < \infty$ . Suppose furthermore that the regression model is correctly specified  $\phi_0 \in \Theta$ , then*

$$\hat{\phi}_T \rightarrow \phi_0 \quad \text{as } T \rightarrow \infty.$$



COROLLARY. 4 (Asymptotic Normality of the MLE for the Correctly Specified Autoregressive Model). *Let the time series  $\{\mathbf{y}_t\}_{t \in \mathbb{Z}}$  be generated by the Strictly Stationary autoregressive model  $\mathbf{y}_t = \phi_0 \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \forall t \in \mathbb{Z}$ ,  $|\phi_0| < 1$ , with exogenous innovations  $\mathbb{E}(\boldsymbol{\varepsilon}_t | \mathbf{y}_{t-1}) = 0$  that satisfy  $\{\boldsymbol{\varepsilon}_t\}_{t \in \mathbb{Z}} \sim NID(0, \sigma_\varepsilon^2)$  with finite variance  $\sigma_\varepsilon^2 < \infty$ . Suppose furthermore that the regression model is correctly specified  $\phi_0 \in \Theta$ , then*

$$\sqrt{T}(\hat{\phi}_T - \phi_0) \xrightarrow{d} N(0, \sigma_\varepsilon^2 [\mathbb{E} \mathbf{y}_{t-1}^2]^{-1}) \quad \text{as } T \rightarrow \infty.$$

Note that the consistency results from applying an LLN to  $\frac{1}{T} \sum_{t=2}^T \boldsymbol{\varepsilon}_t \mathbf{y}_{t-1}$  and  $\frac{1}{T} \sum_{t=2}^T \mathbf{y}_{t-1}^2$ , which easily follows from the fact that when  $\{\mathbf{y}_t\}_{t \in \mathbb{Z}}$  is Stationary and Ergodic, the sequences  $\{\mathbf{y}_t^2\}_{t \in \mathbb{Z}}$  and  $\{\boldsymbol{\varepsilon}_t \mathbf{y}_{t-1}\}_{t \in \mathbb{Z}}$  are trivially also Stationary and Ergodic. Furthermore, as long as  $\sigma_\varepsilon^2 < \infty$  then  $\mathbb{E}|\mathbf{y}_t^2| < \infty$  and  $\mathbb{E}|\boldsymbol{\varepsilon}_t \mathbf{y}_{t-1}| < \infty$ . Application of the CLT to  $\sqrt{T} \frac{1}{T} \sum_{t=2}^T \boldsymbol{\varepsilon}_t \mathbf{y}_{t-1} - \mathbb{E}(\boldsymbol{\varepsilon}_t \mathbf{y}_{t-1})$  requires first that  $\text{Var}(\boldsymbol{\varepsilon}_t \mathbf{y}_{t-1})$  is finite and that  $\boldsymbol{\varepsilon}_t \mathbf{y}_{t-1}$  is a Martingale Difference Sequence,  $\mathbb{E}(\boldsymbol{\varepsilon}_t \mathbf{y}_{t-1} | \boldsymbol{\varepsilon}_{t-1} \mathbf{y}_{t-2}, \boldsymbol{\varepsilon}_{t-2} \mathbf{y}_{t-3}, \dots) = 0$ . The finiteness of the variance can naturally be stated in terms of a moment conditions on the innovations. In particular, if  $|\boldsymbol{\varepsilon}_t^4| < \infty$ , then  $|\mathbf{y}_t^4| < \infty$  and  $|(\boldsymbol{\varepsilon}_t \mathbf{y}_t)^2| < \infty$  are easily verified. The martingale difference property follows trivially by the fact that the *true* innovations are exogenous  $\mathbb{E}(\boldsymbol{\varepsilon}_t | \mathbf{y}_{t-1}) = 0$  and the model is correctly specified and consistent. Hence, the residuals of the regression around the correct parameter are also exogenous and *NID*. To verify the martingale difference property, we only have to define  $\mathcal{F}_{t-1} := (\boldsymbol{\varepsilon}_{t-1} \mathbf{y}_{t-2}, \boldsymbol{\varepsilon}_{t-2} \mathbf{y}_{t-3}, \dots)$  and then need that  $\mathbb{E}(\boldsymbol{\varepsilon}_t \mathbf{y}_{t-1} | \mathcal{F}_{t-1}) = 0$ . This follows by application of a Law of Total Expectation,

$$\begin{aligned} \mathbb{E}(\boldsymbol{\varepsilon}_t \mathbf{y}_{t-1} | \mathcal{F}_{t-1}) &= \mathbb{E}(\mathbb{E}(\boldsymbol{\varepsilon}_t \mathbf{y}_{t-1} | \mathbf{y}_{t-1}, \mathcal{F}_{t-1}) | \mathcal{F}_{t-1}) \\ &= \mathbb{E}(\mathbf{y}_{t-1} \mathbb{E}(\boldsymbol{\varepsilon}_t | \mathbf{y}_{t-1}, \mathcal{F}_{t-1}) | \mathcal{F}_{t-1}) \\ &= \mathbb{E}(\mathbf{y}_{t-1} \mathbb{E}(\boldsymbol{\varepsilon}_t) | \mathcal{F}_{t-1}) = \mathbb{E}(\mathbf{y}_{t-1} 0 | \mathcal{F}_{t-1}) = 0. \end{aligned} \tag{2.27}$$

Note that again, this result relies on the fact that we can substitute  $\mathbb{E}(\boldsymbol{\varepsilon}_t) = 0$  which holds by the fact that our autoregressive parameter is

consistent with respect to the correct parameter. Hence, the result only follows due to the critical assumption that our simple model correctly reflects reality. Furthermore, Stationarity of the correctly specified model was crucial but simple to show because  $\phi_0$  is a linear parameter. As soon as we replace  $\phi_0$  with a nonlinear observation-driven function, the theory that we used to obtain stationarity no longer applies.

Maximum Likelihood is a very flexible framework, and a wide variety of models can be estimated as long as the conditional densities implied by the model can be expressed to derive the log likelihood function. In the case of the spatial autoregressive model, for which the Least Squares assumptions broke down, a likelihood function is also available. In this particular case, one can derive the joint distribution of the dependent variable from that of the residuals using the determinant of the first order derivatives of the functional relationship between the two. Doing the derivations, one will find that the log likelihood function contains new components that account for the feedback term  $(I - \rho W)^{-1}$  that multiplies with the residuals. Several additional assumptions are now needed to show that the likelihood function with these additional terms is still continuous, such that it is limit-preserving. In addition, slightly more demanding stability conditions are needed to obtain Stationarity of the model. This has to factor in that stable feedback now has to account for dependence on both past observations and current neighbor values. The last difficulty is then that the added complexities to the log likelihood function result in difficult derivatives, that once set to zero, do not have analytical solutions. This prevents us from obtaining the analytical expression of the estimator, and in particular, showing that the remainder term vanishes as  $T$  grows. Analytical intractability is a key problem to solve before we can start to tackle the MLE's of the more complicate spatial autoregressive time series processes.

## 2.2 General Extremum Estimators

In practice, it is often the case that only simple econometric models lead to analytically tractable estimators. From a practical point of view, the estimation can be easily carried out using numerical methods that approximate the optima and derivatives of interest. However, from a theoretical perspective, the absence of an expression for the estimator implies that we can no longer analyze its properties in the manner that we have done earlier. As an effect, we might numerically find the parameters that maximize likelihood of the spatial autoregressive time series model, but without establishing consistency and normality, we can't really tell how the obtained results can be interpreted. This obviously calls for the need of a more general theory to establish the desired properties. In particular, we can classify most of the estimators of interest as *extremum estimators*, and state general conditions to verify their properties.

### 2.2.1 General Consistency

Given a probability space  $(\Omega, \mathcal{F}, P)$ , a random sample  $\mathbf{y}_T$ , shorthand for the entire sequence  $(\mathbf{y}_1, \dots, \mathbf{y}_T)$ , and a parameter space  $\Theta$ , we can define an extremum estimator as the measurable map  $\hat{\boldsymbol{\theta}}_T : \Omega \rightarrow \Theta$

$$\hat{\boldsymbol{\theta}}_T \in \arg \max_{\boldsymbol{\theta} \in \Theta} Q_T(\mathbf{y}_T; \boldsymbol{\theta}). \quad (2.28)$$

The criterion function  $Q_T : \mathbb{R}^T \times \Theta \rightarrow \mathbb{R}$  is real-valued and random because it is a function of the random sample  $\mathbf{y}_T$ , which is itself a measurable map  $\mathbf{y}_T : \Omega \rightarrow \mathbb{R}^T$ , hence  $Q_T$  is a map  $Q_T(\mathbf{y}_T, \cdot) : \Omega \times \Theta \rightarrow \mathbb{R}$ . When the random sample is realized and we observe  $\mathbf{y}_T(\omega) \in \mathbb{R}^T$  for some event  $\omega \in \Omega$ , then  $Q_T(\mathbf{y}_T(\omega), \cdot)$  is a real valued function  $Q_T(\mathbf{y}_T(\omega), \cdot) : \Omega \times \Theta \rightarrow \mathbb{R}$ . Hence, for every realization we get a new function to maximize, and we obtain a new maximizer that is our parameter estimate. Hence, the estimators we consider are random.

Note that the maximizer is a set in the  $\arg \max$  as at this point we have not yet said anything about the uniqueness of a maximum. Extremum estimators that take the form of a sum are called an  $M$ -estimator, while those with criterion functions  $Q_T(\mathbf{y}_T, \cdot) : \Omega \times \Theta \rightarrow \mathbb{R}$  that are differentiable on the parameter space  $\Theta$  can also be written as  $Z$ -estimators that directly set the derivative of the criterion to zero. If  $Q_T(\mathbf{y}_T, \cdot) : \Omega \times \Theta \rightarrow \mathbb{R}$  is also strictly concave, then it ensures that the point where  $\nabla Q_T(\mathbf{y}_T, \cdot) = 0$  is really the global, rather than a local, maximum of the function  $Q_T(\mathbf{y}_T, \cdot)$ . Strict concavity is not necessary, for a twice differentiable criterion one may also use the second derivative to infer which solution corresponds to the global maximum. In any case, one can define the estimate as an element

$$\hat{\boldsymbol{\theta}}_T \in \{\boldsymbol{\theta} \in \Theta : \nabla Q_T(\mathbf{y}_T, \cdot) = 0\}. \quad (2.29)$$

The first thing that one generally wants to ensure is that  $\hat{\boldsymbol{\theta}}_T \notin \emptyset$ , which can be shown by applying a Bolzano-Weierstrass Theorem. In particular, this Theorem tells us that every function that is continuous, has a maximum on a compact set. This leads us to the following standard assumption and the implied useful result.

**ASSUMPTION. 5** (Compactness of the Parameter Space). *Let  $\Theta$  be a compact space in  $\mathbb{R}^{n \in \mathbb{N}}$ .*

The compactness assumption is standard, and apart from its critical role in establishing existence and measurability, it will again play a crucial role in the uniform convergence of the estimator.

**THEOREM. 6** (Existence and Measurability of the Estimator). *Let  $\Theta$  be a compact space in  $\mathbb{R}^{n \in \mathbb{N}}$  and  $Q_T(\cdot; \cdot)$  be continuous in its arguments, then there exists a measurable map  $\hat{\boldsymbol{\theta}}_T : \Omega \rightarrow \Theta$  satisfying*

$$\hat{\boldsymbol{\theta}}_T \in \arg \max_{\boldsymbol{\theta} \in \Theta} Q_T(\mathbf{y}_T; \boldsymbol{\theta}).$$

Apart from Existence and Measurability, one typically wants the maximizing point  $\boldsymbol{\theta}_0$  to be identifiable and unique.

ASSUMPTION. 6 (Identifiable Uniqueness of the Maximizer of the Limit Criterion). *Let  $\theta_0 \in \Theta$  be the identifiable unique maximizer of the limit criterion  $Q_\infty : \Theta \rightarrow \mathbb{R}$ .*

There are different definitions with varying mathematical detail. Typically, we mean that  $\theta_0$  not only maximizes the limit criterion  $Q_\infty$ , i.e. that  $Q_\infty(\theta_0) \geq Q_\infty(\theta) \forall \theta \in \Theta$ , but that this point is well separated from other points. If  $S(\theta_0, r)$  is the set of points contained in a ball with fixed radius  $r > 0$  and center-point  $\theta_0$  and  $S^c(\theta_0, r)$  denotes its complement set in  $\Theta$ , i.e

$$S(\theta_0, r) := \{\theta \in \Theta : \|\theta - \theta_0\| < r\} \text{ and } S^c(\theta_0, r) := \{\theta \in \Theta : \theta \notin S(\theta_0, r)\},$$

then  $\theta_0$  is not only the maximizer of  $Q_\infty$ , but also the identifiable unique maximizer if

$$\sup_{\theta \in S^c(\theta_0, r)} Q_\infty(\theta) < Q_\infty(\theta_0). \quad (2.30)$$

Essentially, this says that if we draw a sphere around the correct parameter  $\theta_0$  with any positive real valued radius, then the criterion always judges that any parameter outside of that sphere is not optimal, even as the radius of that sphere becomes arbitrarily small. Note that the identifiability is thus a property of the criterion, and characterizes its ability to differentiate between possible likely solutions.

We now have the right conditions in place to establish consistency of the extremum estimator. In particular, an estimator is (weakly) consistent, *if and only if* it convergences in probability  $\hat{\theta}_T \xrightarrow{p} \theta_0$  as  $T \rightarrow \infty$ , and strongly consistent, *if and only if* it convergences *almost surely*  $\hat{\theta}_T \xrightarrow{a.s.} \theta_0$  as  $T \rightarrow \infty$ . Weak consistency states that for a specified large  $T$ , the estimator  $\hat{\theta}_T$  is likely to be near its correct value  $\theta_0$ , leaving open the possibility that one can find some arbitrary  $\epsilon > 0$  for which  $|\hat{\theta}_T - \theta_0| > \epsilon$  still happens an infinite number of times, although at infrequent intervals. Strong consistency instead states that this will in fact *almost surely* not occur. In particular, it implies that with probability 1, we have that

for any  $\epsilon > 0$  the inequality  $|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0| < \epsilon$  holds when  $T$  has become large enough. Either result can be obtained following a similar strategy, though strong consistency requires a stricter condition that may in some cases not hold while the conditions for (weak) consistency may still be verified.

The general consistency theorem for the criterion of an extremum estimator requires uniform convergence of the criterion function to a limit deterministic function. We say that the criterion function  $Q_T$  converges point-wise in probability over  $\Theta$  to a limit function  $Q_\infty$  if it holds true that  $|Q_T(\mathbf{y}_T; \boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \xrightarrow{p} 0 \forall \boldsymbol{\theta} \in \Theta$  as  $T \rightarrow \infty$ . Moreover, we say that the criterion function  $Q_T$  converges uniformly in probability over  $\Theta$  to a limit function  $Q_\infty$  if it holds true that  $\sup_{\boldsymbol{\theta} \in \Theta} |Q_T(\mathbf{y}_T; \boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \xrightarrow{p} 0$  as  $T \rightarrow \infty$ . The difference lies in the fact that the latter is expressed for the *supremum*, which can loosely be interpreted as the “worst case”-convergence across all elements in  $\Theta$ . The point-wise convergence is thus a much weaker condition than the uniform convergence, since for point-wise convergence the rate of convergence can be different for each element in  $\Theta$ . More so, while uniform convergence implies point-wise convergence, point-wise convergence does not imply uniform convergence. Unfortunately, directly establishing uniform convergence is often not easy. However, due to a remarkable result known as the stochastic Arzelà-Ascoli Theorem, it is known that point-wise convergence of the criterion function over a compact parameter space implies uniform convergence if the estimator is stochastically equicontinuous. A family of functions is equicontinuous if all the functions are continuous and they have equal variation over a given neighborhood. By itself, stochastic equicontinuity is not an easy to use concept, but a Lipschitz condition implies stochastic uniform equicontinuity. This gives us the very easy to use condition that if  $\sup_T \mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \|\partial Q_T(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}\| < \infty$ , then the sequence of random criterion functions at sample size  $T$  generated under  $\omega \in \Omega$  is stochastically equicontinuous. Furthermore, in order to obtain strong

uniform convergence of  $Q_T$  to  $Q_\infty$  we need it to be strongly stochastically equicontinuous, which requires that the derivative be uniformly bounded rather than bounded in expectation  $\sup_T \sup_{\theta \in \Theta} \|\partial Q_T(\theta)/\partial \theta\| < \infty$  *almost surely*. As a result, it is quite straightforward to verify that the criterion function of an extremum estimator is (strongly) consistent. In particular by applying a suitable LLN to obtain point-wise convergence and using the bounded expectation of the derivative of the criterion to obtain the uniform convergence. Strong consistency of the criterion can be obtained by applying an LLN to obtain point-wise convergence and using the uniform boundedness of the derivative of the criterion to obtain strong uniform convergence. This can be summarized as follows.

**THEOREM. 7** (General Consistency for M-estimators). *Let  $(\Omega, \mathcal{F}, P)$  be a probability space, and let the criterion function  $Q_T : \Omega \times \Theta \rightarrow \mathbb{R}$  be a sequence of random continuous functions that take the form*

$$Q_T(\mathbf{y}_T; \theta) = \sum_{t=2}^T q(\mathbf{y}_t, \mathbf{y}_{t-1}; \theta),$$

*and  $q$  be differentiable on a convex compact parameter space  $\Theta$ . Assume that also that  $q(\mathbf{y}_t, \mathbf{y}_{t-1}; \theta)$  is Stationary and Ergodic and has bounded first moment  $\mathbb{E}|q(\mathbf{y}_t, \mathbf{y}_{t-1}; \theta)| < \infty$ . Then the criterion satisfies a Law of Large Numbers*

$$\frac{1}{T} \sum_{t=2}^T q(\mathbf{y}_t, \mathbf{y}_{t-1}; \theta) \xrightarrow{p} \mathbb{E}(q(\mathbf{y}_t, \mathbf{y}_{t-1}; \theta)) \quad \text{as } T \rightarrow \infty \quad \forall \theta \in \Theta,$$

*hence the sequence  $\{Q_T\}_{T \in \mathbb{N}}$  converges point-wise in probability to a limit function  $Q_\infty = \mathbb{E}(q(\mathbf{y}_t, \mathbf{y}_{t-1}; \theta)) \forall \theta \in \Theta$ . If furthermore,  $q(\mathbf{y}_t, \mathbf{y}_{t-1}; \theta)$  has a derivative with bounded expectation,*

$$\mathbb{E} \sup_{\theta \in \Theta} \|\partial q(\mathbf{y}_t, \mathbf{y}_{t-1}; \theta)/\partial \theta\| < \infty,$$

*then  $\{Q_T\}_{T \in \mathbb{N}}$  is stochastically equicontinuous. Together with the point-wise convergence this implies that  $\{Q_T\}_{T \in \mathbb{N}}$  then converges uniformly in probability to the limit function  $Q_\infty$*

$$\sup_{\theta \in \Theta} |Q_T(\mathbf{y}_T; \theta) - Q_\infty(\theta)| \xrightarrow{p} 0 \quad \text{as } T \rightarrow \infty.$$

The compactness of  $\Theta$  together with the continuity of  $q$  implies that the measurable map  $\hat{\theta}_T : \Omega \rightarrow \Theta$  exists, satisfying

$$\hat{\theta}_T \in \arg \max_{\theta \in \Theta} Q_T(\mathbf{y}_T; \theta).$$

If furthermore the parameter  $\theta_0$  is the identifiably unique maximizer of the limit criterion function  $Q_\infty$

$$\sup_{\theta \in S^c(\theta_0, r)} Q_\infty(\theta) > Q_\infty(\theta_0),$$

then the uniform convergence implies that  $\hat{\theta}_T$  is consistent for  $\theta_0$  since

$$|\hat{\theta}_T - \theta_0| \xrightarrow{p} 0 \quad \text{as } T \rightarrow \infty.$$

If finally, the derivative is also uniformly bounded,

$$\sup_{\theta \in \Theta} \|\partial q(\mathbf{y}_t, \mathbf{y}_{t-1}; \theta) / \partial \theta\| < \infty \text{ a.s.}$$

then  $\{Q_T\}_{T \in \mathbb{N}}$  is strongly stochastically equicontinuous. The strong stochastic equicontinuity together with the established point-wise convergence implies that  $\{Q_T\}_{T \in \mathbb{N}}$  converges uniformly almost surely to the limit function  $Q_\infty$

$$\sup_{\theta \in \Theta} |Q_T(\mathbf{y}_T; \theta) - Q_\infty(\theta)| \xrightarrow{\text{a.s.}} 0 \quad \text{as } T \rightarrow \infty,$$

hence that  $\hat{\theta}_T$  is also strongly consistent for the identifiably unique maximizer of the limit criterion function  $\theta_0$  since

$$|\hat{\theta}_T - \theta_0| \xrightarrow{\text{a.s.}} 0 \quad \text{as } T \rightarrow \infty.$$

## 2.2.2 General asymptotic Normality

The general consistency theorem can be applied in a wide range of settings. The asymptotic normality follows by a very similar argument. In particular, we can show uniform convergence of the second derivative in turn obtained from the point-wise convergence of the second derivative together with boundedness of the third derivative that implies the stochastic equicontinuity of the second derivative.



The reason behind the central role of the second derivative can be easily understood. First, focus on the fact that we are interested in obtaining an approximate limit distribution for  $\sqrt{\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0}$ . Remember that by construction of our estimate as the optimum of the criterion, it holds that, at the estimate, the derivative of the criterion is zero  $\nabla Q_T(\mathbf{y}_T, \hat{\boldsymbol{\theta}}) = 0$ . Suppose we introduce a new point  $\boldsymbol{\theta}_T^*$  that lies between  $\hat{\boldsymbol{\theta}}_T$  and  $\boldsymbol{\theta}_0$ , then we can use the *Mean Value Theorem* to write the derivative as

$$\nabla Q_T(\mathbf{y}_T, \hat{\boldsymbol{\theta}}_T) = \nabla Q_T(\mathbf{y}_T, \boldsymbol{\theta}_0) + \nabla^2 Q_T(\mathbf{y}_T, \boldsymbol{\theta}_T^*)(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) = 0. \quad (2.31)$$

We can now obtain an expression for  $\sqrt{(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)}$  by rewriting the second equality and multiplying both sides by the square root of  $T$ .

$$\sqrt{(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)} = (\nabla^2 Q_T(\mathbf{y}_T, \boldsymbol{\theta}_T^*))^{-1} \times \sqrt{T} \nabla Q_T(\mathbf{y}_T, \boldsymbol{\theta}_0). \quad (2.32)$$

This immediately suggests that obtaining the asymptotic normality of  $\hat{\boldsymbol{\theta}}_T$  at  $\boldsymbol{\theta}_0$  follows in three steps. First, by showing that  $\sqrt{T} \nabla Q_T(\mathbf{y}_T, \boldsymbol{\theta}_0)$  converges in distribution to  $N(0, \Sigma)$ . Second, by showing that  $(\nabla^2 Q_T(\mathbf{y}_T, \boldsymbol{\theta}_T^*))^{-1}$  converges in probability to  $(\nabla^2 Q_T(\boldsymbol{\theta}_0))^{-1}$  as  $T \rightarrow \infty$ . Since  $\boldsymbol{\theta}_T^*$  is evaluated between  $\hat{\boldsymbol{\theta}}_T$  and  $\boldsymbol{\theta}_0$ , the consistency of  $\hat{\boldsymbol{\theta}}_T$  implies that  $\boldsymbol{\theta}_T^*$  approaches  $\boldsymbol{\theta}_0$ . Hence, the convergence of  $(\nabla^2 Q_T(\mathbf{y}_T, \boldsymbol{\theta}_T^*))^{-1}$  to  $(\nabla^2 Q_T(\boldsymbol{\theta}_0))^{-1}$  in probability, follows by showing that  $\nabla^2 Q_T$  converges uniformly over  $\Theta$  to  $\nabla^2 Q_\infty$ . Finally, to obtain the convergence to  $(\nabla^2 Q_T(\boldsymbol{\theta}_0))^{-1}$ , we must establish that the limit is invertible.

The first condition, that the scaled criterion derivative  $\sqrt{T} \nabla Q_T(\mathbf{y}_T, \boldsymbol{\theta}_0)$  converges in distribution to  $N(0, \Sigma)$  can be obtained by applying a CLT to the derivative  $\nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta})$ . This is straightforward because

$$\sqrt{T} \nabla Q_T(\mathbf{y}_T, \boldsymbol{\theta}_0) = \sqrt{T} \frac{1}{T} \sum_{t=2}^T \nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)$$

$$= \sqrt{T} \left( \frac{1}{T} \sum_{t=2}^T \nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0) - \mathbb{E}(\nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)) \right) \quad (2.33)$$

where, just as before,  $\mathbb{E}(\nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0))$  can be added as it equals zero by construction. Recall that the CLT will require the derivative of the criterion  $\nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)$  to be Stationary and Ergodic Martingale Difference Sequence and be bounded in second moment  $\mathbb{E}\|\nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)\|^2 < \infty$ .

The uniform convergence of the second derivative of the criterion function  $\sup_{\boldsymbol{\theta} \in \Theta} \|\nabla^2 Q_T(\mathbf{y}_T, \boldsymbol{\theta}) - \nabla^2 Q_\infty(\boldsymbol{\theta})\| \xrightarrow{p} 0$  as  $T \rightarrow \infty$  can be obtained by the same strategy as followed in the consistency Theorem. In particular, the stochastic Arzelà-Ascoli Theorem tells us we can focus the argument on the point-wise convergence of

$$\|\nabla^2 Q_T(\mathbf{y}_T, \boldsymbol{\theta}) - \nabla^2 Q_\infty(\boldsymbol{\theta})\| \xrightarrow{p} 0 \quad \forall \quad \boldsymbol{\theta} \in \Theta \quad \text{as } T \rightarrow \infty,$$

and the stochastic equicontinuity of  $\{\nabla^2 Q_T\}$ , in turn implied by a Lipschitz condition ensured by a bounds on it's derivative

$$\sup_T \mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \|\nabla^3 Q_T(\mathbf{y}_T, \boldsymbol{\theta})\| < \infty.$$

The final invertibility requirement is strongly related to the identification of  $\boldsymbol{\theta}_0$ . In particular, when  $\boldsymbol{\theta}_0$  is well-separated, then the limit criterion  $Q_\infty$  must have strong curvature around  $\boldsymbol{\theta}_0$ . This strong curvature implies that  $Q_\infty$  accelerates moving from  $\boldsymbol{\theta}_0$  to any point around it, hence that the second derivative  $\nabla^2 Q_\infty$  is non-singular and invertible. If, on the other hand,  $Q_\infty$  is flat around  $\boldsymbol{\theta}_0$ , then  $\nabla^2 Q_\infty$  is singular and hence not invertible.

We can thus summarize the general normality theorem for extremum estimators as follows.

**THEOREM. 8** (General Asymptotic Normality for M-estimators). *Let  $\Theta$  be a compact parameter space and  $\hat{\boldsymbol{\theta}}_T$  be a consistent M-estimator for an identifiable unique point  $\boldsymbol{\theta}_0 \in \Theta$ .*

Suppose that the derivative of the criterion  $\nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)$  is a Stationary and Ergodic Martingale Difference Sequence and bounded in second moment  $\mathbb{E}\|\nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)\|^2 < \infty$ , then it is asymptotically normal at

$$\sqrt{T} \left( \frac{1}{T} \sum_{t=2}^T \nabla q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0) \right) \xrightarrow{d} N(0, \Sigma) \text{ as } T \rightarrow \infty.$$

Suppose also that the second derivative is Stationary and Ergodic and bounded

$$\sup_T \mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \|\nabla^2 q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)\| < \infty,$$

then application of a Law of Large Numbers yields that the second derivative converges point-wise

$$\left\| \frac{1}{T} \sum_{t=2}^T \nabla^2 q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}) - \mathbb{E} \nabla^2 q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}) \right\| \xrightarrow{p} 0 \quad \forall \quad \boldsymbol{\theta} \in \Theta \quad \text{as} \\ T \rightarrow \infty.$$

Suppose furthermore that the third derivative is bounded

$$\sup_T \mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \|\nabla^3 q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0)\| < \infty,$$

then the point-wise convergence and the stochastic equicontinuity of the second derivative imply that it also converges uniformly

$$\sup_{\boldsymbol{\theta} \in \Theta} \left\| \frac{1}{T} \sum_{t=2}^T \nabla^2 q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}) - \mathbb{E} \nabla^2 q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}) \right\| \xrightarrow{p} 0.$$

Finally, by the invertibility of the limit  $\mathbb{E} \nabla^2 q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta})$ , implied by a strong curvature of the criterion around  $\boldsymbol{\theta}_0$  in turn ensured by the identifiability of  $\boldsymbol{\theta}_0$ , together with the established uniform convergence and asymptotic normality at  $\boldsymbol{\theta}_0$ , implies that

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, E \Sigma E') \text{ as } T \rightarrow \infty,$$

with  $E = (\mathbb{E} \nabla^2 q(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\theta}_0))^{-1}$ .

We now have a general theory that can be applied to show the consistency and normality of possibly complex models like the spatial autoregressive

time series model. All the exact derivations of the conditions and steps will not be provided here. Instead, Chapter 4 provides a more general proof that covers the linear spatial autoregressive model but also allows for the possible failure of several simplifying assumptions that have been made here. In particular, the result also covers cases in which the spatial autoregressive parameter is nonlinear, possibly observation-driven, and under a more general distributional assumption than the Gaussian one. The theory there shall also detail what happens in the case of multiple optima.

## 2.3 Further complications when modeling dynamic spatial time series

To conclude this chapter, we will look at the steps of the General Consistency and Normality Theorems more closely and discuss them further with regard to the MLE of a general, unparameterized, nonlinear autoregressive model. This broad setting covers, among possible other dynamics, the spatial autoregressive ones that we were particularly interested in. We then discuss each of the assumptions that were made and aim to provide relevant meaning to them for the cases in which one would consider certain parameterizations. We then wish to find out what might still be easily violated or difficult to show, and set several theoretical objectives to remedy those situations.

Suppose we have possibly nonlinear model that depends on both past and current values

$$\mathbf{y}_t = \psi(\mathbf{y}_t, \mathbf{y}_{t-1}, \boldsymbol{\varepsilon}_t; \boldsymbol{\theta}) \quad \forall t \in \mathbb{Z}. \quad (2.34)$$

Note that this includes also popular linear spatial time series, for example of the form

$$\mathbf{y}_t = \alpha + \rho W \mathbf{y}_t + \phi \mathbf{y}_{t-1} + \phi_2 W \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \quad \forall t \in \mathbb{Z}. \quad (2.35)$$

In this model, the values of  $\mathbf{y}_t$  depend on past local and neighbor values, and spillover contemporaneously across regions. The extension to dependence on past residuals, and possible spatial lags thereof, will be made in Chapter 4. In 6, extensions will also be made that allow dependence on past residuals of a different spatial variable. We already noted that models currently discussed can be written in the following form

$$\mathbf{y}_t = f(\mathbf{y}_{t-1}, \boldsymbol{\theta}) + v(\boldsymbol{\varepsilon}_t, \boldsymbol{\theta}) \quad \forall t \in \mathbb{Z}. \quad (2.36)$$

This highlights that the current notation is general enough to also account for an additional spatial error process. In contrast to the earlier, more simplistic, example in which the residuals were not nonlinearly transformed such that we could obtain the density directly from the dependencies implied by the regression model, we now obtain the density  $\boldsymbol{\varepsilon}_t \sim p_{\boldsymbol{\varepsilon}}(\boldsymbol{\theta})$  by inverting  $v$ . In general, for a regression model that has instantaneous dependence

$$\mathbf{y}_t = h(\boldsymbol{\theta})\mathbf{y}_t + g(\mathbf{y}_{t-1}, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}_t \quad \forall t \in \mathbb{Z}, \quad (2.37)$$

one can define the contribution to the log likelihood at time  $t$  as

$$\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1}) = \log \det |I - h(\boldsymbol{\theta})| + \log p_{\boldsymbol{\varepsilon}}\left((I - h(\boldsymbol{\theta}))\mathbf{y}_t - f(\mathbf{y}_{t-1}, \boldsymbol{\theta}), \boldsymbol{\theta}\right). \quad (2.38)$$

The log likelihood function is again defined as  $\ell_T(\mathbf{y}_T, \boldsymbol{\theta}) = \sum_{t=2}^T \ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})$ . The component  $\log \det |I - h(\boldsymbol{\theta})|$  stems from the fact that inverting the model leads to a residual dynamic  $(I - h(\boldsymbol{\theta}))^{-1}\boldsymbol{\varepsilon}_t$ . Hence, if no autoregressive dynamics would be modeled but only a spatial error process, a similar component would enter the log likelihood function. In the standard nonlinear case, without feedback,  $I - h(\boldsymbol{\theta}) = I - 0 = I$ , hence  $\log \det |I - h(\boldsymbol{\theta})| = \log \det |I| = 0$ . We thus obtain the standard nonlinear log likelihood contribution  $\log p_{\boldsymbol{\varepsilon}}(\mathbf{y}_t - f(\mathbf{y}_{t-1}, \boldsymbol{\theta}), \boldsymbol{\theta})$  in the standard case in which  $h(\boldsymbol{\theta})$  does not transform the data. When  $h(\boldsymbol{\theta})$  does transform the data, then immediately, the established continuity prop-

erties of many density functions that are used in empirical applications are complicated by the additional component  $\log \det |I - h(\boldsymbol{\theta})|$ . We must thus ensure the non-singularity and boundedness of this additional term before we can obtain any result that relies on the continuity of  $\ell_T(\mathbf{y}_T, \boldsymbol{\theta})$ . Furthermore, calculating the derivatives of the log likelihood function can get particularly complicated. To apply an LLN, we need  $\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})$  to be differentiable, Stationary and Ergodic, and be bounded in first moment. In contrast to the standard case, the verification of these properties has to take into account the additional component  $h(\boldsymbol{\theta})$  that would normally not complicate any result. In particular, proving a suitable stationarity result under  $\boldsymbol{\theta}_0$ , that cannot be assumed as this is now a property of the model, can be significantly more complex when we need to control for both temporal dependence and instantaneous feedback. The stationarity results known in time series literature that only focus on stability of  $f(\mathbf{y}_{t-1}, \boldsymbol{\theta})$  are not sufficient, neither are the stability results from spatial literature that only focus on  $h(\boldsymbol{\theta})$ . Suppose, however, that suitable forms of stability across the space-time dimension have been verified, then we could obtain the point-wise convergence

$$\frac{1}{T} \sum_{t=2}^T (\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})) \xrightarrow{p} \mathbb{E}(\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})) \text{ as } T \rightarrow \infty \forall \boldsymbol{\theta} \in \Theta,$$

hence we would establish that the sequence  $\{\ell_T(\mathbf{y}_T, \boldsymbol{\theta})\}_{T \in \mathbb{N}}$  converges point-wise to a limit deterministic function  $\ell_\infty(\boldsymbol{\theta}) = \mathbb{E}(\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1}))$ . Next we need the derivative of the log likelihood, the score, at each step  $\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})'$  to be bounded and the parameter space, that also includes any components part of  $h(\boldsymbol{\theta})$ , to be compact, to obtain strong stochastic equicontinuity and thus the uniform convergence of the log likelihood function

$$\sup_{\boldsymbol{\theta} \in \Theta} |\ell_T(\mathbf{y}_T, \boldsymbol{\theta}) - \ell_\infty(\boldsymbol{\theta})| \xrightarrow{a.s.} 0 \text{ as } T \rightarrow \infty.$$

For example, if the score is uniformly bounded we have *almost sure* uniform convergence by the strong stochastic equicontinuity and point-

wise convergence. If  $\boldsymbol{\theta}_0$  is the identifiable unique maximizer in a compact space  $\Theta$  we then obtain strong consistency of the estimator  $\hat{\boldsymbol{\theta}}_T$  for  $\boldsymbol{\theta}_0$ . Remember that if the weaker stochastic equicontinuity is obtained instead, we can still show the point-wise convergence of the log likelihood function and obtain the weak consistency result. It is not uncommon for nonlinear models to introduce identification complications, particularly when the data is in fact linear. For example, for a function  $\delta/(\gamma(\mathbf{y}_t))$ , the parameter  $\delta$  can be any value if  $\gamma(\mathbf{y}_t) = 0$ , or, if instead  $\delta = 0$ , the quantity  $\gamma(\mathbf{y}_t)$  can take on any value without affecting the value of the criterion function. We will see in Chapter 4 that the estimator can still be set-consistent, but for normality, however, identification is critical.

In particular, to establish normality we need first that the score is a Stationary and Ergodic Martingale Difference Sequence and bounded in second moment.

$$\mathbb{E}\|\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})'\|^2 < \infty.$$

The stationarity and ergodicity can be obtained by continuous differentiability of  $\ell$  and the stationarity and ergodicity of  $\{\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})\}_{t \in \mathbb{Z}}$ . The second moment can be similarly obtained from  $\{\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})\}_{t \in \mathbb{Z}}$  by deriving moment preserving properties that again have to factor in the properties of  $h(\boldsymbol{\theta})$ . The score is a Martingale Difference Sequence if the model is correctly specified and the criterion is consistent. Naturally,  $h(\boldsymbol{\theta})$  can tremendously improve the fit and help ensure the appropriateness of the Martingale Difference Sequence assumption. Application of a CLT then delivers

$$\sqrt{T} \frac{1}{T} \sum_{t=2}^T \ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})' \xrightarrow{d} N(0, \Sigma) \quad \forall \boldsymbol{\theta} \in \Theta.$$

Twice continuous differentiability of  $\ell$  delivers stationarity and ergodicity of  $\{\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})''\}_{t \in \mathbb{Z}}$ , and together with the moment bound

$$\mathbb{E}\|\ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})''\| < \infty,$$

this allows us to apply an LLN for every  $\boldsymbol{\theta} \in \Theta$  to obtain

$$\frac{1}{T} \sum_{t=2}^T \ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})'' \xrightarrow{p} \mathbb{E} \ell_t(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{y}_{t-1})'' \text{ as } T \rightarrow \infty \forall \boldsymbol{\theta} \in \Theta.$$

Three times continuous differentiability of  $\ell$  together with a uniform bound on the third derivative ensures the stochastic equicontinuity of the second derivative and thus the uniform convergence. Together with the invertibility of the limit second derivative, this deliver the asymptotic normality of the estimator. Again, the invertibility relies on parameter identification that can be complicated by identification problems of nonlinear functions. This is especially problematic since one would normally use the approximate distribution to infer whether a parameter is significantly different from zero. In many cases, nonlinear models are only asymptotically normal under the alternative assumption that the data is nonlinear. If one needs to assume nonlinearity to test for nonlinearity, then that test statistic is deeply flawed in some sense.

The proof for normality relies heavily on continuity and bounds of derivatives of the log likelihood function. In some sense this can be understood as smoothness, or, well-behavedness properties. The additional component  $\log \det |I - h(\boldsymbol{\theta})|$  can complicate the derivations significantly which may make verifying these properties quite unpleasant. At a high level, one can easily understand that the smoothness of the log likelihood function depends on the type of nonlinearities generated by  $h(\boldsymbol{\theta})$  and  $g(\mathbf{y}_{t-1}, \boldsymbol{\theta})$ . In these cases one may note that the stochastic equicontinuity, in turn implied by Lipschitz conditions, is only used as an optional tool that allows one to exploit the often easier to obtain point-wise convergence. It may naturally be possible to show uniform convergence directly. For example, instead of stochastically bounding the derivatives, such computations can be avoided when the Ergodic Theorem for random elements with values in a separable Banach space is applied. This is the strategy that we will take in Chapter 4. In other situations, one may try to show that the Lipschitz conditions are themselves ensured by higher-level smoothness



properties such as higher moment bounds of a function. For a sufficient degree of smoothness, the Lipschitz conditions may stretch out across a sufficient number of derivatives to immediately ensure that the second derivative of the log likelihood function is stochastically equicontinuous without the need of taking the derivations.

Finally, the discussions here all circled around compact parameter spaces. It was mentioned that in Chapter 4, set-consistency would be developed for the case of multiple solutions to the criterion. Nevertheless, stochastic equicontinuity relied on compactness, hence it was a crucial ingredient for normality too. We already analyzed that intuitively, there must be strong curvature around the solution to the criterion function to obtain an approximate distribution around a parameter estimate. In Chapter 6 we will consider a simple penalty modification to the criterion function that remedies a known form of non-identifiability and whose effect becomes negligible in the limit. However, non-parametric models may need penalization that does not vanish in the limit. In Chapter 5 we will analyze this. It turns out that penalties force the criterion function to favor simple solutions over complex ones. This, similar to the strategy of obtaining boundedness of derivatives through high-level smoothness conditions, essentially limits possible solutions of the criterion function to only those that are available in lower frequency domains, which again emphasizes that understanding the properties of  $h(\boldsymbol{\theta})$  and  $g(\mathbf{y}_{t-1}, \boldsymbol{\theta})$  is critical to establish the desired theoretical results needed to apply them to analyze spatial time series data.

# Chapter 3

## Spatial Heterogeneity

### Chapter Summary

Policy schemes that aim to stimulate the cultivation of biofuel crops typically ignore the spatial heterogeneity in costs and benefits associated with their production. Because of spatial heterogeneity in biophysical, and current agricultural production factors, potential gains from stimulating biofuel crops are non-uniformly distributed across space. This paper explores implications of this type of heterogeneity for the net benefits associated with different subsidy schemes. We present a simple framework based on discounted cash flows, to assess potential gains from introducing the notion of heterogeneity into stimulation schemes. We show that agricultural subsidy spending can be reduced in a Pareto efficient way and simultaneously improve the total stimulation potential of biofuel policies, when schemes: 1) are production based instead of land based; 2) accommodate differences in opportunity costs, and 3) target sites where subsidies for conventional agricultural land-use types are high. These results are robust for a range of different bioenergy prices and the relative gains of addressing these key elements in policy compared to conventional stimulation schemes increase with lower bioenergy prices, and are largest when low prices coincide with high emission reduction ambitions.<sup>1</sup>

---

<sup>1</sup>This chapter is based on “*Efficiency of second-generation biofuel crop subsidy schemes: Spatial heterogeneity and policy design*” published in the *Journal of Renewable and Sustainable Energy Reviews*, and is reproduced with kind permission from Elsevier. The full reference is Andree et al. (2017b). The material is reproduced in this chapter with kind permission from *Elsevier*. This work was part of a research initiative funded by the Dutch National Research Programme Knowledge for Climate.

### 3.1 Introduction

The constraints of finite natural resources in combination with concerns about global warming have led researchers and policymakers to pay increased attention to the topic of sustainable energy policies and the reduction of greenhouse gas emissions. The switch to biofuels as a transportation fuel source has been put forward as a possible contribution to carbon emission reduction plans and overall sustainable energy strategies (Farrell et al., 2006; Ragauskas et al., 2006; Koçar and Civaş, 2013). Second-generation ethanol production from lignocellulosic material is generally considered to avoid (partly) social and environmental impacts linked to biofuel production (Singh et al., 2010), and could become a key contributor to emission reductions. Although lignocellulosic ethanol production from biomass may become a suitable option in the future, large-scale production is not economically feasible at present and stimulation policies have to be implemented to achieve future bioenergy usage ambitions (Wiesenthal et al., 2009). Many countries are struggling to achieve 2020 goals for fuel standards. In 2015, the average European blending share of crop based ethanol and biodiesel was estimated at respectively 3.3% and 4.3%, and at about 0.6% for non-food based biofuels (Flach et al., 2015). Though the sector has achieved considerable growth worldwide in recent years (REN21, 2015), the strong decline in crude oil prices that started in the second half of 2014 has put the competitiveness of biofuels under severe pressure, and the current policy ambitions are not expected to lead to significant higher production in the next decade (OECD and FAO, 2015). Because economic benefit is arguably the most important incentive for adoption, efficient subsidy strategies are of relevance for the future of biofuels and might not only be key in reaching 2020 fuel standards, but might determine when, or whether, we ever get a viable model for large scale production.

The focus of this paper is to explore possibilities to minimize subsidy

spending and simultaneously increase the total stimulation potential of biofuel policies, while maintaining the income levels of farmers. Such possibilities allow for Pareto improvement with respect to the current situation as society can both save money on subsidies and gain from environmental benefits related to biofuel production, while profits of farmers would be unaffected by the subsidy reform. Reducing spending and increasing the stimulation potential of schemes can contribute to the overall cost-effectiveness of policies and might strengthen the case of bioenergy production in the political arena. In past years, different studies proposed heterogeneous allocation of resources under different environmental policies, for example carbon sequestration contracts (Antle et al., 2003), air pollution emission trading programs (Fowlie and Muller, 2013), vehicle emission abatement (Mérel and Wimberger, 2012), and policies that promote investment in renewable electricity generators (Fell and Linn, 2013). Current bioenergy stimulation policies typically do not recognize that there is substantial spatial variation in costs and benefits associated with biofuel crop production. This heterogeneity relates to interaction between policies stimulating the production of bioenergy, spatially heterogeneous production factors, agricultural land-use patterns, and other agricultural policies. The central thesis of this paper is that introducing the notion of spatial heterogeneity into subsidy schemes allows for more efficient allocation of subsidies, and potentially increases net social benefits by decreasing subsidy costs and increasing positive externalities. We build our analysis on the following three elements: first we assess spatial heterogeneity in Net Present Value (NPV) of current agricultural production systems; we then estimate site specific net social costs and benefits of stimulation schemes; and finally, we compare the relative efficiency of alternative subsidy schemes in terms of associated potential net benefits. We repeat the analysis for a range of different bioenergy prices to show how the results change when the relative competitiveness of conventional land use and bioenergy

production changes. We apply our analysis to explore production of a specific second-generation bioenergy crop – Miscanthus (*Miscanthus* × *Giganteus*) – in the Netherlands, a country with an advanced agricultural sector that has a high economic value per hectare. The Netherlands is currently far behind the European average for sustainable energy usage, and as we shall see in our application of the developed theory, could benefit from more effective bioenergy policy design.

The remaining part of this paper is organized as follows. In section 3.2, we discuss inefficiencies that arise due to land heterogeneity. Section 3.3 details our methods. Section 3.4 describes our application to Miscanthus in the Netherlands. Section 3.5 presents the results, followed by a discussion and conclusion in section 3.6.

### **3.2 The importance of spatial heterogeneity in agricultural policy**

Agricultural systems are strongly determined by spatially heterogeneous agro-economic, socio-economic, and local biophysical conditions (Diogo et al., 2013). Spatial economic models that build upon this heterogeneity confirm that biomass is able to provide a substantial contribution to the overall energy supply. This future bioenergy potential has been assessed on the global scale (Hoogwijk et al., 2005; Smeets and Faaij, 2007), at the European level (van Dam et al., 2007; de Wit and Faaij, 2010; Fischer et al., 2010a,b), and at national levels (Batidzirai et al., 2006; Styles and Jones, 2007; van den Broek et al., 2001). An overview of the different assessments and their respective strengths and weaknesses is given by Dornburg et al. (2008), who point out that spatial variation in production characteristics is the most important aspect in assessing bioenergy potentials. Recent studies focusing on local opportunities for biofuel production were able to pinpoint specific areas of interest

by using micro data on production characteristics (van der Hilst et al., 2010; Diogo et al., 2012). Understanding the economic implications of spatial variability in local production factors might help researchers and policymakers in the field of environmental economics and resource management work towards more efficient forms of policy. However, existing agro-economic and bioenergy stimulation policies often do not explicitly address spatial heterogeneity and abstain from insights gained from bioenergy potential assessments.

Two examples illustrate this lack of attention to spatial aspects. The governments of Canada and the United States have proposed policies in which farmers are paid for the adoption of certain management practices to sequester carbon dioxide in agricultural soils (Agriculture and Agri-Food Canada, 2003; Young, 2003). In the European Union, farmers who grow bioenergy crops can apply for a standard land based subsidy (European Commission, 2007). Such a subsidy scheme is analogous to the proposed Canadian and United States government subsidy scheme as farmers are paid for adopting site-specific practices. Market-based incentives, however, are generally seen as more efficient than command-and-control or environmental design standard policies because there are cost-efficiency differences in abatement strategies among the entities within a sector, for example when both costs and environmental benefits differ among plots (Tietenberg, 1990; Stavins, 1998). Efficient agricultural policies that aim to increase environmental benefits by influencing the management decisions of farmers, must therefore take into account the heterogeneity of the biophysical and economic factors that determine the agricultural system (Just and Antle, 1990). Paying farmers to adopt certain management practices in a land based system, disregarding the biophysical differences among their production sites, is generally seen as inefficient (Helfand and House, 1995; Babcock et al., 1996; Fleming and Adams, 1997).

We particularize by distinguishing between two types of inefficiencies in bioenergy stimulation schemes: over-funding and mis-allocation of funds. When a farmer produces biofuel under a (government-funded) carbon contract, the contract value is part of the farmer's private profit function. In the economic environment of an emission trading market, contract values are conditional on a spatially varying factor, that is, the quantity of biomass produced in specific locations, and an exogenous factor that is equal among farmers, the price of one unit of carbon. It follows that the income generated by farmers through carbon contracting is a monotonic transformation of the spatial distribution of production quantities, which has a direct relation not to local production costs but to the local biophysical conditions that determine biomass yields. Farmers with comparative advantages thus possibly receive aids that greatly exceed the marginal costs of bioenergy crop production, resulting in allocation of excessive funds and *ex post* inequalities.

Mis-allocation of funds occurs when spatial characteristics that are not part of the private profit function appear in the social welfare function. This difference can originate from both the cost and the benefit side of the economy. One mechanism through which the social cost function differs from the private cost function, is that the production of bioenergy crops can reduce subsidy distributed elsewhere in the market. Subsidies for crops are mutually exclusive, meaning that farmers can only opt for a single crop subsidy per plot at a time. Under the assumption that avoided subsidies for conventional land uses will return efficiently to society, low social costs do not necessarily coincide with low private costs in their joint spatial distribution. The possibility exists that for any given farmer, the private bioenergy production profits are below zero (a subsidy is required, disregarding the potential profits related to other land-use types), while the net social cost of sustaining the required subsidy is negative. This happens for example when farmers receive subsidies for conventional land-use types that exceed the subsidy requirements to produce bioenergy,

while the private profits, after subsidy, of both alternatives are equal. In this case there is room for a Pareto-efficient reduction in conventional subsidies. Generally, if positive private opportunity costs for biomass production at a certain point in the spatial distribution coincide with negative social costs for sustaining local bioenergy production in the joint spatial distribution, possibilities for Pareto improvement exist and it could be said that society is allocating its funds to the wrong sites.

A similar issue arises on the benefit side of social welfare. Since the production of bioenergy crops is associated with positive external effects – perennial crops sequester carbon in their root systems and the use of biofuels reduces carbon emissions – a strictly positive spatial distribution of externalities exists conditional on the local biophysical conditions. This spatial distribution, part of the spatial social benefit distribution, is not internalized in the spatial distribution of private benefits. The optimal land-use patterns from a societal perspective will thus differ from patterns that arise from private decisions. Summarizing, spatial heterogeneity in local production characteristics under a policy that inadequately accounts for spatial differences leads to inefficient outcomes due to two principles:

- A non-uniform spatial distribution of marginal production costs leads to over-funding at production sites that have comparative advantages, when subsidies are distributed uniformly across plots.
- If the spatial distributions of social and private costs and social and private benefits differ, schemes that do not promote internalization of externalities and instead follow private optimization decisions, misallocate funds from a societal perspective.

### 3.3 Methodology

We develop a spatially explicit economic assessment strategy to evaluate potential net benefits of subsidizing bioenergy at the grid-cell level.



Both the cost and the benefit side of the model build upon an explicit representation of land heterogeneity and require micro-data on biophysical conditions, market prices, and agricultural land use. By combining this information, we estimate the NPV of the currently existing agricultural land uses and of bioenergy crop production using a multiple-year time span, thus incorporating long-term decision processes related to perennial crop management and start-up investment costs. The difference in the economic performance of bioenergy crop production and conventional agricultural production types is used to determine a minimum required contract value for each plot. The net social cost of the stimulation policy is calculated by comparing the minimum required subsidies with subsidies distributed among conventional land uses. The spatial distribution of potential social benefits of bioenergy production is based on the emission offsets provided by local biomass production quantities, and the amount of carbon sequestered in the root systems of perennial crops. By comparing the net costs of subsidizing with the benefits associated with the subsidized sites, we are able to evaluate the effectiveness of different types of policies. The next section introduces the seven stimulation schemes that we shall explore in our application. Section 3.3.2 provides further details to the structure of the model.

### 3.3.1 Different spatial policies

A government that engages in bioenergy stimulation can either subsidize farmers directly through a periodical land based payment or introduce a carbon contracting system in which farmers receive funds by providing carbon emission offsets to entities, including the government, that are willing to buy such contracts to suppress their carbon rating.<sup>2</sup> Contracts or subsidies based on emission offsets are referred to as production based

---

<sup>2</sup>We use contracts and subsidies somewhat interchangeably throughout the paper because we do not explicitly differentiate between types of entities, but view society as the final entity that pays for such contracts.

schemes as they directly relate to the quantities of produced biomass. Both types of periodical payments can be homogeneous across space, as in conventional schemes, or they can be altered to account for spatial heterogeneity. Heterogeneous subsidy schemes allocate the exact amount of funds to each farmer needed to sustain the local production of bioenergy crops.

We analyze seven alternative bioenergy stimulation schemes table 3.1 that address heterogeneity to various degrees. We first distinguish between spatially heterogeneous and conventional (homogeneously distributed) subsidies. Within the category of heterogeneous subsidies, a second distinction is made between: 1) single focus (SF) schemes that subsidize plots where the local opportunity costs for biofuel production are lowest, and minimize the total funds spent on bioenergy stimulation; and 2) integrated agricultural focus (IAF) schemes that also take into account how conventional agricultural subsidies are spatially distributed. IAF schemes aim to limit total aggregate spending on agricultural subsidies by subsidizing plots where net social costs for biofuel production stimulation are lowest. This integrated approach is more efficient as it captures reductions in the aggregate agricultural subsidy spending by decreasing the, often excessive, subsidies for other types of production. Both SF and IAF schemes offer farmers a single subsidy that does not depend on the farmer's choice of production type. Under the assumption that farmers optimize profits, and that land conversion occurs accordingly, farmers that receive less subsidy after the policy reform are reconciled by increased productivity associated with the alternative production system. So, in our simple framework, both SF and IAF schemes reduce spending in an Pareto efficient way. These stimulation strategies are particularly interesting for biofuel stimulation schemes that generate positive externalities associated with emission reductions, but the implications of the results stretch out over other conventional agricultural subsidy schemes. In a sense, heterogeneous schemes undo a policy induced market failure. By offering

farmers site-specific financial support, agricultural land-use patterns are generated by profit maximization principles that follow the productivity of farmers. Under homogeneous subsidies that vary per crop type, this equilibrium land-use pattern is distorted and farmers have an incentive to move away from optimum and produce crops for which their land is not suited, as long as they are sufficiently reconciled by the crop-specific subsidy.

Both heterogeneous and conventional subsidies are analyzed under per-hectare and per-tonne contracts. Additionally, we analyze policy that is fully directed at minimizing land-use change by allocating subsidies on a command-and-control basis to the sites associated with highest biomass production potentials.

Table 3.1: Overview of the alternative subsidy schemes distinguished in this study.

	Heterogeneous		Conventional	
	Land Based	Production Based	Land Based	Production Based
Single Focus	SFLB <sup>1</sup>	SFPB <sup>2</sup>	CLB <sup>3</sup>	CPB <sup>4</sup>
Integrated Focus	IAFLB <sup>5</sup>	IAFBP <sup>6</sup>		
Minimize LUC		MLPB <sup>7</sup>		

<sup>1</sup> Single focus land based, <sup>2</sup> single focus production based, <sup>3</sup> conventional land based, <sup>4</sup> conventional production based, <sup>5</sup> integrated agricultural focus land based, <sup>6</sup> integrated agricultural focus production based, <sup>7</sup> minimized land-use change production based.

### 3.3.2 Spatial economic model

The model that we apply consists of several elements, and we first detail the overall structure before providing the equations. The underlying assumption of our approach is that farmers operate under optimal inputs and subsequently allocate their land between alternative crops in order to maximize their profits. We therefore, similar to other land-use allocation models, start with a profit maximization problem. From this maximization problem, we derive a simple land-use allocation rule that

serves as a starting point for constructing an indifference surface conditional on site-specific subsidy levels. We then formulate restrictions for homogeneity and heterogeneity of stimulation schemes. Given these restrictions and the spatial indifference surface, we derive spatial distributions for the minimum required subsidies under heterogeneous and conventional schemes. These are compared to current subsidy spending on conventional crops to derive plot-specific social costs of sustaining the required subsidy to convert a given plot to a bioenergy production site. We separately construct a spatial distribution of external benefits associated with fossil fuel savings and carbon sequestered in the root system of the bioenergy crop using a carbon price. The site-specific net social costs of sustaining biofuel production is then compared to the benefits associated with production to map site-specific potential net benefits of biofuel production. Finally, we aggregate the local net benefits of those sites that are covered under a specific stimulation scheme to enable a comparison of the total potential benefits associated with different schemes. The remaining part of this section details all of the analysis steps.

In our simplified land-use allocation model, we view the study area as consisting of a number of production sites indexed by  $i \in I$ . We assume that in order to maximize their profits, farmers make a choice between two types of land use  $l = [c, e]$ , where  $c$  denotes conventional agricultural land use and  $e$  denotes the allocation of energy crops. The economic decision process for a multiple-year period that farmers face can thus be described as the following multi-period profit-maximization problem:<sup>3</sup>

---

<sup>3</sup>In this paper we make use of the following notation: (discounted) summations of any symbol are given by its respective capital, e.g.  $Z \equiv \sum_{i=1}^I z_i$ , similarly for sets, capitals contain their lowercase as elements, we index variables by land-usetype with a superscript such that  $z^e \equiv z^{l=e}$  is identical and reads as the variable  $z$  under bioenergy production, and write constraints between braces  $\{\}$ , e.g.  $i\{A\} \in I^A$  reads as “all elements  $i$  for which rule “ $A$ ” holds, are members of the set  $I^A$ ”.

$$\arg \max_{l \in L} \Pi_i^l = \sum_{t=0}^T \frac{\pi_{it}^l}{(1+r)^t}, \quad (3.1)$$

where profits are generated according to the spatial time series:

$$\pi_{it}^l = p_{it}^l q_{it}^l (\varphi_{it}^l) + s_{it}^l - w_{it}^l h_{it}^l - k_{it}^l - o_{it}^l (q_{it}^l) \quad \forall it \in \mathbb{N} \times \mathbb{Z}. \quad (3.2)$$

Where  $p_{it}^l$  are the plot-level prices for the vector of outputs of land-use types  $l$  at time  $t$ ,  $q_{it}^l$  is the vector of outputs, which is a function of local yield factors  $\varphi_{it}^l$ ,  $w_{it}^l$  are prices for the vector of inputs,  $s_{it}^l$  are plot-level subsidies,  $k_{it}^l$  are fixed costs containing start-up investments, equipment costs and yearly fixed costs, and finally  $o_{it}^l$  are the field operation costs, which may vary with output  $q_{it}^l$ .

Let  $\Pi_i^c$  and  $\Pi_i^e$  be real variables and  $\Pi_i'$  denote any particular set of values of these two variables. Any such set is represented by a point in a two-dimensional Cartesian space. Let  $\Pi \supseteq \Pi_i'$  be the superset of all such points and let  $\Pi^c$  and  $\Pi^e$  be the subsets of  $\Pi$  including points for which  $\Pi_i'$  produces either a conventional crop production site contained in  $I^c$ , or an energy crop production site contained in  $I^e$ . To be able to assign set membership to  $I^l$  based on  $\Pi_i'$ , we require that some further logic exists. The profit maximization by land-use choice under mutually exclusive land-use types provides a rule “ $A(\Pi_i')$ ” that defines the subset  $I^e \subseteq I$ . This rule “ $A(\Pi_i')$ ” ascribes to each production site contained in  $I$ , the property of belonging to  $I^e$  or not, based on any set of values  $\Pi_i'$  for the two land use types. The profit-maximization problem in Equation (3.1) leads to the following rule:

$$i\{\Pi_i^e > \Pi_i^c\} \in I^e. \quad (3.3)$$

Note that if profits would instead have been stochastic, this and subsequent results follow similarly under expected value theory if  $\Pi_i^c$  and  $\Pi_i^e$  are consistent estimates of the first moments of their stochastic counterparts.

Note also that, though prices are given in a competitive market, a farmer has the ability to choose the level of inputs and that a government can choose to change  $s_{it}^l$ , such that any pair  $\Pi_i'$  is possible, which in turn according to the time series system Equations (3.1) to (3.3) can produce any agricultural landscape. If we keep  $\Pi_i^c$  constant, that is, assuming optimal private inputs and fixed conventional subsidies,  $\Pi_i^e(s_{it}^e) > \Pi_i^c$  has a unique asymptotic lower bound,  $\Pi_i^e(s_{it}^e) = \Pi_i^c$ , on which rational farmers are indifferent between biofuel and conventional agricultural production. All sets  $\Pi_i'$  that produce membership in  $I \setminus I^e \setminus I^c = I^0$  together constitute this spatial indifference surface on which profit maximizing farmers are indifferent between production types. From Equation (3.1) it is clear that under these assumptions it is a straightforward approach to find the local subsidy values  $s_{it}^e$  that establish this indifference surface, e.g. to find the lower bound of required subsidies to sustain biofuel stimulation.

Depending on the policy type, the spatial distribution of the bioenergy subsidy  $s_{it}^e$  can either be spatially homogeneous  $\bar{s}_{nt}^e$ , which we indicate with an overbar and index by  $n \in N$  with  $N \subseteq [1, \dots, |I|]$  being a subset of the ordered sequence from 1 to the cardinality of the entire set of plots, or spatially heterogeneous  $\check{s}_{it}^e$ , which we indicate with a check. We index by  $n$  to distinguish from heterogeneous schemes. More specifically, heterogeneous subsidies can be uniquely indexed over the entire set of plots, whereas homogeneous subsidies are equal among plots and indexed by the cardinality of the set of bioenergy production sites,  $n := |I^e|$ , i.e. homogeneous subsidies increase as the amount of plots converted to bioenergy production sites increases. Spatially homogeneous subsidies  $\bar{s}_{nt}^e$  are identical across  $i$  and must satisfy  $\bar{s}_{i-1,nt}^e := \bar{s}_{int}^e \forall i \in I$ , hence we shall continue without subscripting  $i$  for spatially homogeneous variables. The criterion carries that when  $n$  plots are subsidized, the marginal increase in the aggregate subsidy by subsidizing the next plot  $i + 1$  that is contained in an extended set of  $n + 1$  production sites, equals  $\bar{s}_{n+1,t}^e + n(\bar{s}_{n+1,t}^e - \bar{s}_{nt}^e)$ . Thus, if there are differences in subsidy

requirements between a newly subsidized plot and the most efficient plot of all formerly subsidized plots, the marginal increase in aggregate subsidy does not only increase by the subsidy amount required at the new plot but also by the efficiency difference multiplied by the amount of plots that already received aids. Spatially heterogeneous subsidies, on the other hand, are flexible and allow for any finite difference between subsidies at any two points within the distribution and satisfy  $0 \leq |\check{s}_{it}^e - \check{s}_{i+1,t}^e| < \infty$ . The marginal increase in the aggregate subsidy of subsidizing the next plot  $i+1$  when the subsidies are heterogeneous is just  $\check{s}_{i+1,t}^e$ , the subsidy required for production at the new site. The difference in the marginal increase in the aggregate homogeneous subsidy and aggregate heterogeneous subsidy is  $\bar{s}_{n+1,t}^e + n(\bar{s}_{n+1,t}^e - \bar{s}_{nt}^e) - \check{s}_{i=n+1,t}^e$ . The requirements of the newly subsidized plot under both schemes are equal, thus the marginal increase in total subsidy costs is given by  $n(\bar{s}_{n+1,t}^e - \bar{s}_{nt}^e)$  and thus depends on the amount of heterogeneity between plots that drives the difference between  $(\bar{s}_{n+1,t}^e - \bar{s}_{nt}^e)$ , and the size of the policy area  $n$  before increasing its extent. The marginal cost function of converting an additional production site to energy crop production under homogeneous schemes, thus depends on the site-specific exogenous determinants that enter any private profit function within the entire policy area, whereas the marginal cost function under a heterogeneous scheme is just a function of local variables. In our application we shall study how the impact of heterogeneity changes as the size of the policy area grows to the size of the entire set of potential production sites  $|I^e| \rightarrow |I|$ .

To derive the minimum spatially homogeneous subsidies  $\bar{s}_{nt}^e$ , it is necessary to write down the exact relationship between the total *area* of sites dedicated to bioenergy production and the decision rule of Equation (3.3), so that we can determine  $n$ . Assuming optimal private inputs under both land-use types, the total bioenergy production area is given by summing over the individual sizes of the plots for which the decision rule of Equation (3.3) holds.

$$Y^e = \sum_{i=1}^{I^e} y_i^e, \quad (3.4)$$

where  $y_i^e$  is the individual plot size and  $Y^e$  the aggregate production size dedicated to bioenergy production. Suppose that a government aims to have a total area of  $Y^e$  devoted to the production of bioenergy crops, then Equation (3.4) inversely provides the amount of required plots to achieve that level of coverage. To find the minimum spatially homogeneous subsidy  $\bar{s}_{nt}^e$  required to stimulate  $n$  plots, we need to establish indifference in the least efficient production site  $i = j, j \in [1, \dots, n]$ , e.g. the production site that requires the highest aid. Thus, in the  $j$ -th plot, we need  $\Pi_j^e = \Pi_j^c$  to hold and then solve for  $s_{jt}^e$ . As  $\Pi_j^e$  is the subsidized profit, we can write it as a function of the unsubsidized profit and a subsidy component:

$$\Pi_i^e = \tilde{\Pi}_i^e + S_i^e, \quad (3.5)$$

where

$$S_i^e = \sum_{t=0}^T \frac{s_{it}^e}{(1+r)^t}.$$

There are multiple solutions to  $s_{it}^e$  in Equation (3.5) as cash flows may vary throughout years, for example a lump sum in the first year can be the NPV equivalent of an annuity. The discounted total subsidy  $S_i^e$  is, however, unique, so it follows that indifference  $\Pi_j^e = \Pi_j^c$  holds when  $S_j^e = \Pi_j^c - \tilde{\Pi}_j^e$  and bioenergy subsidy is equal to the unsubsidized profit gap. Using the homogeneity rule, it follows also that the discounted total homogeneous subsidy for any plot equals that of the least efficient plot  $\bar{S}_n^e = S_j^e$ . For any production area size  $Y^e$  containing  $n$  plots, we can write the spatially homogeneous subsidy that minimizes the aggregate subsidy as a function of the largest unsubsidized profit gap occurring in all the bioenergy production sites  $I^e$ .<sup>4</sup>

---

<sup>4</sup>To confirm that this indeed is a minimum, consider lowering the value of  $\bar{s}_{nt}^e$  for all plots with a



$$\bar{S}_n^e = \max_{i \in I^e} (\Pi_i^c - \tilde{\Pi}_i^e). \quad (3.6)$$

One important result that we can directly derive from this is that within the current system, farmers are paid to withhold from innovation, and the introduction of new subsidy systems is required for any innovative crop before it can be produced on a large scale.<sup>5</sup>

Using the decision rule of Equation (3.3), we can similarly construct a distribution of spatially heterogeneous minimum subsidies at which farmers are tangent to choosing  $l = e$ , by finding the exact values for  $S_i^e$  that coincide with values of  $\Pi_i'$  that produce set membership in  $I^0$ . Hence, to find the plot-specific discounted spatially heterogeneous minimum subsidy  $\check{S}_i^e$ , we need to establish indifference  $\Pi_i^e = \Pi_i^c$  at all plots  $i$ . Straightforward use of Equation (3.5), gives us:

$$\check{S}_i^e = \Pi_i^c - \tilde{\Pi}_i^e. \quad (3.7)$$

The spatial distribution of net subsidy spending is calculated as the difference between conventional subsidies and bioenergy stimulation subsidies. Net subsidy spending is what society pays to produce environmental benefits through biofuels; therefore, we will refer to it as social costs, though we do not take into account other potential costs than direct spending.

$$C_i^e = S_i^e - S_i^c. \quad (3.8)$$

Apart from the net subsidy spending involved in bioenergy stimulation,

---

minor fraction just sufficient to cause  $\pi_j^e < \pi_j^c$ . This will only be sufficient to stimulate  $n - 1$  plots. Lowering the value of  $\bar{s}_{nt}^e$  for one or several plots with a minor fraction will break the condition of spatial homogeneity.

<sup>5</sup>Note that we can similarly split up net present value profits of conventional agricultural land use  $\Pi_j^c$ . Therefore to obtain indifference in a subsidized agricultural system,  $\tilde{\Pi}_i^e = \tilde{\Pi}_i^c \implies S_i^e = S_i^c$ , which means that any positive value for  $S_i^c$  forms an innovation barrier, preventing bioenergy production at otherwise competitive sites, e.g., where both unsubsidized land-use types would be equally profitable.

we consider also the benefits associated with avoided carbon emissions. The potential social benefits at a specific production site are given by the benefits of emission savings and the difference between benefits from additionally sequestered carbon.

$$B_i^e = \sum_{t=0}^T \frac{p_\varepsilon \varepsilon^e q_{it}^e + p_\varepsilon (\sigma_{it}^e - \sigma_{it}^c)}{(1+r)^t}, \quad (3.9)$$

where  $p_\varepsilon$  is the carbon price,  $\varepsilon^e$  are the emissions saved per unit of bioenergy production, and  $\sigma_{it}^l$  are the emission saving equivalents of sequestered carbon in the root system of perennial crops. Net benefits are given by the difference between social benefits and social costs.

$$\omega_i^e = B_i^e - C_i^e \quad (3.10)$$

In aggregate, we can quantify the total net benefits by summing up all the plot-level gains for sites where private profits for growing bioenergy crops exceed profits from growing conventional crops.

$$\Omega^e = \sum_{i=1}^{I^e} \omega_i^e = B_i^e - C_i^e = \sum_{t=0}^T \frac{p_\varepsilon \varepsilon^e q_{it}^e + p_\varepsilon (\sigma_{it}^e - \sigma_{it}^c)}{(1+r)^t} - (S_i^e - S_i^c). \quad (3.11)$$

In the right side equality in Equation (3.11), we see the direct relation between the spatially heterogeneous potential benefits and the spatial distribution of subsidies allocated to production sites  $S_i^e$ , where  $S_i^e = \bar{S}_n^e$  for homogeneous subsidies or  $S_i^e = \check{S}_i^e$  under heterogeneous subsidies schemes. It follows directly from Equation (3.11) that under heterogeneous production factors, the potential gains under heterogeneous subsidies are higher than those under homogeneous subsidies.<sup>6</sup> This should come as

---

<sup>6</sup>Combining Equation (3.6) and Equation (3.7) leads to  $\check{S}_n^e \leq \bar{S}_n^e$ , with  $\check{S}_n^e < \bar{S}_n^e$  if there is variation in  $\Pi_i^e - \tilde{\Pi}_i^e$  across  $i$ . Therefore,  $\Omega^e(\check{S}_n^e) > \Omega^e(\bar{S}_n^e)$  follows trivially under heterogeneous production factors.

no surprise given the expressions of the marginal costs for converting an additional farmer derived earlier, but it is an empirically interesting matter to contrast the differences in total potential net benefits of alternative schemes using real data for a range of potential production area sizes. The straightforward equations suggest that a researcher armed with micro-data on in- and output price vectors, production costs and quantities, land-use patterns, and current subsidy schemes, is able to do just that by plugging them in Equation (3.1), and evaluating the total net benefit potential under different types of policy by substituting  $S_i^e$  with values of  $\check{S}_i^e$  or  $\bar{S}_n^e$  and calculating  $\Omega^e$  for the set of sites  $I^e$  for which the subsidy is sufficient to convert profit maximizing farmers into bioenergy crop producers. Similarly, policy-objective related benefit potentials can be calculated by summing  $i$  over the set of production sites  $I_{target}^e$  for which the total bioenergy crop production  $Q^e = \sum_{t=0}^T \sum_{i=1}^{I^e} q_{it}^e$  equals a target amount  $Q_{target}^e$  of the bioenergy product. The corresponding size of the production area  $Y_t$  can be evaluated with Equation (3.4), to compare the potential size of policy areas. Finally, the subsidy is optimal when the marginal gains to society from subsidizing the least efficient plot  $j$  equals zero  $\omega_j^e = 0$ . That is where the marginal social benefits equal the marginal net costs of subsidizing. Since  $S_j^e$  is fixed, we can calculate the value of  $S_i^e$  that corresponds to the optimum subsidy pattern.

### 3.3.3 Modeling production quantities

We propose to approximate the output vector of products with crop-specific yield values, which can be directly mapped from local biophysical features. We model the yield following van Bakel et al. (2007) by attributing crop-specific damage scores related to drought  $Rd$  and water-logging  $Rw$  according to the local combination of geological and hydrological conditions. The damage scores are designed to be used with the yield function below to calculate the crop-specific expected yields.

$$q_{it}^l = \varphi_{it}^l q_{it,max}^l, \quad (3.12)$$

$$\varphi_{it}^l = 100 - Rw_{it}^l + Rd_{it}^l \left( \frac{(100 - Rw_{it}^l)}{100} \right). \quad (3.13)$$

The production quantity vector for a specific land-use type in the choice model of Equation (3.1),  $q_{it}^l$ , is the crop-specific maximum attainable yield,  $q_{it,max}^l$ , multiplied by  $\varphi_{it}^l$ , that is, the local yield conditions factor ranging from 0-100%. This procedure to quantify expected yields has been successfully applied to model a variety of crops in studies for the Netherlands (van der Hilst et al., 2010; Kuhlman et al., 2013) and Argentina (Diogo et al., 2014). In a similar NPV framework Diogo et al. (2015) were able to replicate national agricultural land-use patterns in the Netherlands with 84% degree of correspondence on a pixel by pixel comparison.<sup>7</sup> This shows that Equations (3.1) to (3.3) in combination with Equations (3.12) to (3.13), is not just practical but also appropriate to simulate land use.

### 3.4 The case of *Miscanthus* in the Netherlands

We illustrate our approach to accounting for spatial heterogeneity in bioenergy stimulation policies with an application to a second generation perennial biofuel crop – *Miscanthus* – in the Netherlands. The Netherlands is selected as a study area for several reasons. First, it has an advanced agricultural sector with high economic value per hectare and a high population density. Consequently, there is high pressure on land for both urban land uses and intensive agricultural activities, resulting in strong competition between different types of agricultural land use (Koomen et al., 2005). Because of this competitiveness, there is

---

<sup>7</sup>Weighted average, making use of the fact that 69.1% of agricultural land is dairy farming, and there was 90.1% degree of pixel by pixel correspondence for dairy farming and 71.7% for arable farming.

no unused marginal land in the study area, so we do not need to account for potential variability in the supply of agricultural land conditional on marginal changes in subsidy patterns.<sup>8</sup> Second, application of our model in the Netherlands allows us to investigate whether possibilities for Pareto improvements in current subsidy schemes are substantial even in a small, and highly competitive agricultural system. Moreover, the small size of the country has the advantage that we can assume biofuel prices to remain stable when production volumes increase; the additional production is not likely to influence these prices that follow supply and demand conditions at far larger scales. The Dutch case is also interesting for policymakers, as it is an example of a country that is still far behind current national and European ambitions for sustainable energy, and that lacks a developed agricultural production system for second generation biofuels.<sup>9</sup> Miscanthus was chosen for our case study because different studies describe it as potentially high yielding (Elbersen et al., 2005; van der Wolf, M. de; Klooster, 2006; van der Voort et al., 2008). Van der Hilst et al. (2010) show that Miscanthus is more economically viable than sugar beets for ethanol production, validating the usefulness of Miscanthus as a non-food biofuel source. Bearing in mind the arguments put forward by different critics of food-based biofuel (Gomiero et al., 2010; Tait, 2011), Miscanthus could thus be of particular interest for energy production from an ethical point of view.

Agricultural land use in our study area mainly consist of two dominant production systems, arable farming and dairy farming, both modeled with different rotation schemes for sand and clay soils. For arable farming our model is made operational by using prices and values described by

---

<sup>8</sup>An overall decrease in land supply due to ongoing urbanization is more likely and could be incorporated in our approach but we exclude this as well as it is not likely to change the competition between different types of agricultural land use, but would only adjust the total amounts per types.

<sup>9</sup>In 2012, 3.4% of fuel sold in the Netherlands originated from first and second generation sources and only 20% of these source materials were produced in the Netherlands (Dutch Emission Authority, 2013). The main sources for second generation biofuels of Dutch origin were domestic garbage, recycled fats and tallow.

Diogo et al. (2012) for in- and output vectors  $p_{it}^c$  (prices for agricultural products),  $q_{it,max}^l$  (maximum attainable yields),  $h_{it}^l$  (the types and amounts of production inputs),  $w_{it}^l$  (the prices of the various inputs),  $k_{it}^l$  (fixed costs including start-up investments and equipment costs), and  $o_{it}^l$  (the farm operation costs). The product prices are updated using 5-year averages of the product prices reported by LEI (2012).<sup>10</sup> Local production quantities  $q_{it}^l$  are obtained by transforming the maximum attainable yield quantities  $q_{it,max}^l$  using yield factors  $\varphi_i^l$  estimated using data on local soil and hydrological with Equation (3.13), assuming that these factor remain stable over time.

Since dairy farming operations do not directly sell grass, but rely on its yield as an input in milk production, the economic assessment of this production system relies on additional intermediate steps. Production quantities, and yield related costs, for dairy farming operations are modeled based on the assumption that cows require energy, supplied by grass, to produce milk. The energy (grass) supply is linked to local grass yields  $\varphi_i^{grass}$ . Energy shortages are computed at each yield level to obtain the amount of required supplementary energy. We assume that farmers supplement their grass supply with silage maize according to local energy shortages and the digestible energy content of silage maize. Silage maize is bought at opportunity costs since maize is grown in rotation, reflecting the costs of not selling it on the market. Milk is sold as the main product at similar 5-year average prices reported by LEI (2012), and excess silage maize is sold as a secondary product. Further details regarding the calculations are contained in table 3.4 in Appendix C.

Specifying the production conditions for *Miscanthus* is more complex as less documented experience is available. Soil and groundwater related yield reduction values, for example are not available for *Miscanthus*. This void was filled by relying on the expected local yield values from van der

---

<sup>10</sup>The 10-year averages for potatoes.

Hilst et al. (2010). Also, a market price for *Miscanthus* is not available as the market is undeveloped. We account for that by using a price range based on imported lignocellulosic biomass prices, averaging €3.25/GJ for pellets from Latin America, €4.50/GJ for pellets from Eastern Europe, and €5.50–6.50/GJ for pellets from Scandinavia (Hamelinck et al., 2005) and converting biomass to lignocellulosic energy equivalents (see Appendix A). Recent projections on the development of the biofuel sector taking into account the 2014 drop in crude oil prices, indicate that in the short to medium-term, high energy prices and high investments that could possibly lead to improved conversion rates are unlikely (OECD and FAO, 2015). We use data on the conventional subsidies  $s_{it}^c$  that are distributed in the European Union. Depending on the land-use type, farmers in Europe receive income support of up to €446 per hectare per year in the Netherlands according to the CAP (European Commission, 2013). Since the 2003 CAP reform, subsidies of €45 per hectare per year are available to farmers growing energy crops for 70% of their lands deployed in energy crop farming (European Commission, 2007).<sup>11</sup>

We combine all prices and other production-related values and insert them in Equation (3.1) to compute the economic profitability of land at each individual grid-cell. To construct a spatial distribution of conventional land use profits  $\Pi_i^c$ , we link conventional land-use vector  $l = c$  to agricultural land-use data (Ministerie van Economische Zaken Landbouw en Innovatie, 2013) registered at the parcel level. Since the land-use data set reflects the situation at a fixed moment in time, crop cycles are implemented to simulate the average NPV of various crop rotation schemes throughout a period of 20 years.<sup>12,13</sup> We take a weighted average of profits according to the share of each crop type in a crop rotation.

<sup>11</sup>This subsidy system is one of the oldest European policies and is still gradually being transformed. The total expenditures on CAP have declined in the past decades. In 2011, the total CAP expenditure accounted for 44% of the total European budget, while in 1986 this was around 75%. Nevertheless, the CAP remains an important source of income to farmers.

<sup>12</sup>We use an inflation-adjusted discount rate of 3%.

<sup>13</sup>The rotation scheme that we use is contained in table 3.3 in Appendix B.

By Equation (3.6), homogeneous biomass subsidies are determined by the size of the policy area through the inverse mapping of Equation (3.4). We link  $Q_{t,target}^e$  in our model to the required growth in bioenergy production to meet the bioenergy market share targets set by the European Union for 2020 and accordingly determine the required production area that provides the basis to determine the minimum homogeneous subsidy.<sup>14</sup> The benefits of emission savings per unit of biomass product  $\varepsilon^e$  are the amount of fuel savings based on the European reference of 88.3kg CO<sub>2</sub>eq/GJ (Dutch Emission Authority, 2013) per energy unit multiplied by a carbon price  $p_\varepsilon$  of €20 per ton. The carbon sequestration benefits  $\sigma_{it}^e$  of Miscanthus are based on 8.8 tons CO<sub>2</sub>eq reported by Caslin et al. (2015). Arable crops in our rotation schemes do not consist of perennial crops and are assumed to store no significant amounts of carbon in their root systems. Though we are aware of opportunities for carbon sequestering in the dairy farming sector, we omit them from our analysis as they are too strongly dependent on site-specific practices.

## 3.5 Results

### 3.5.1 Economic performance of production systems

Figure 3.1 presents our assessment of the economic performance of various crop cycles in the Netherlands for declining soil suitability. A brief discussion of the robustness of the results, along with the distributions of estimated economic performance  $\Pi_i^l$ , is provided in Appendix D. On average, clay soils perform better than sandy soils for both arable farming and dairy farming. The economic performance of arable farming is more sensitive to yield values than that of dairy farming. This results from the ability of dairy farmers to import silage maize when the grass yields on their specific plots are modest and still make profits on the sales of their

---

<sup>14</sup>All data and assumptions regarding energy usage are contained in Appendix A.



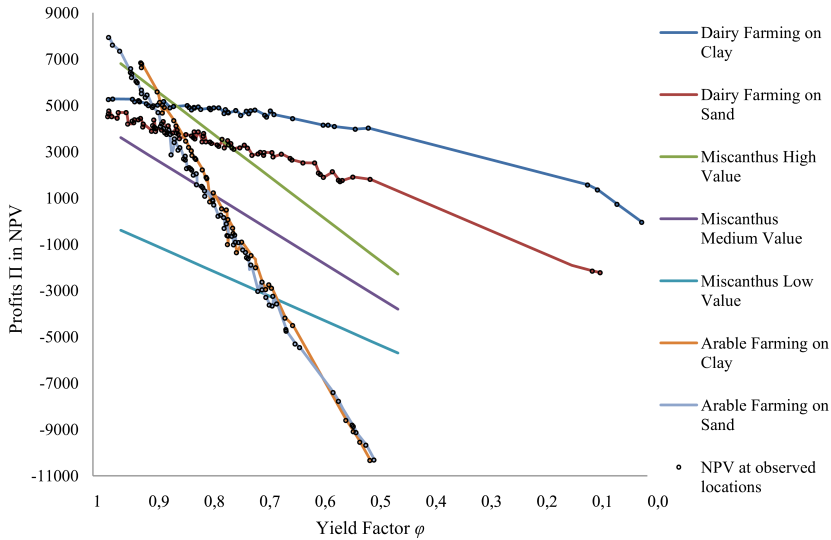


Figure 3.1: Economic performance in discounted euros per hectare of different production systems at declining yield levels. Soil and groundwater table combinations that lead to a Miscanthus yield higher than 0.95 or lower than 0.45 are non-existent in the Netherlands. Dots represent observed combinations of soil types and ground water tables, lines are linearly interpolated. The total amount of locations with a negative NPV accounts for only 2.6% of agricultural land. Possible explanations are discussed in Appendix D.

final products. Miscanthus is more sensitive to yield than dairy farming, but less than arable farming.

Figure 3.1 provides important insights into the general trend of land-use competition between Miscanthus and conventional crops.<sup>15</sup> Arable farming receives high income support through the CAP and requires high soil suitability to be profitable, as Figure 3.1 shows.<sup>16</sup> The NPV

<sup>15</sup>Plot-specific deviations from the general trend in land-use competition are possible as the yields of dairy farming production systems and arable farming production systems are not perfectly spatially correlated. The spatial analysis, on which subsequent sections build, takes this into account but is difficult to generalize here. The Histograms in Figure 3.7 in Appendix D show the factors used in Equation (3.7) to model the spatial comparison between Miscanthus and conventional land-use profit levels. The map in Figure 3.8 in Appendix E shows the resulting spatial distribution of the minimum required subsidies.

<sup>16</sup>Not deducible from Figure 3.1, the CAP support includes limited support for dairy farming production systems through subsidizing maize production, which is a small percentage of the rotation system. Arable farming production systems receive direct income support for a large part of their rotation scheme.

of dairy farming is less sensitive to change in obtainable yield, and its occurrence is centered mainly on lower suitability soils because arable farming outcompetes dairy farming on high yielding soils. Dairy farming simultaneously is less subsidized. This causes a self-selection process in which areas with productive soils receive higher subsidies while areas with low suitability intersect with land-use types that receive less financial support. The implications for bioenergy production are that the opportunity costs of producing *Miscanthus* increase on more productive soils, because: 1) high-suitability areas self-select into areas that receive higher income support, and 2) high soil suitability is relatively more in favor of the economic performance of arable farming than that of *Miscanthus*.

### 3.5.2 Assessing the impacts of different policies

Figure 3.2 shows how different subsidy schemes that reorder the sequence in which production sites are subsidized, produce differently shaped social cost and benefit curves. Single focus schemes (SFLB and SFPB) tend to mis-allocate funds as can be seen from the erratic cost curves. Integrated agricultural focus schemes (IAFLB and IAFPB) that take into account the way in which conventional subsidies  $S_i^c$  are allocated, have smoothened cost curves and a larger integral area between the social cost and benefit curves. Targeting production sites by production based opportunity costs instead of land based opportunity costs generates cost curves that are very similar, and efficiency differences are not directly apparent from the cost curves only. Policy that aims to minimize land-use change, results in a cascading succession of “separate” cost curves for regimes with similar biophysical conditions, as each biophysical regime includes production sites with low and high social costs.

Figure 3.2 finally also shows that with high market prices, marginal costs and benefits have only a few intersections clustered at a high percentage of land deployed for *Miscanthus* production. When market prices are low,

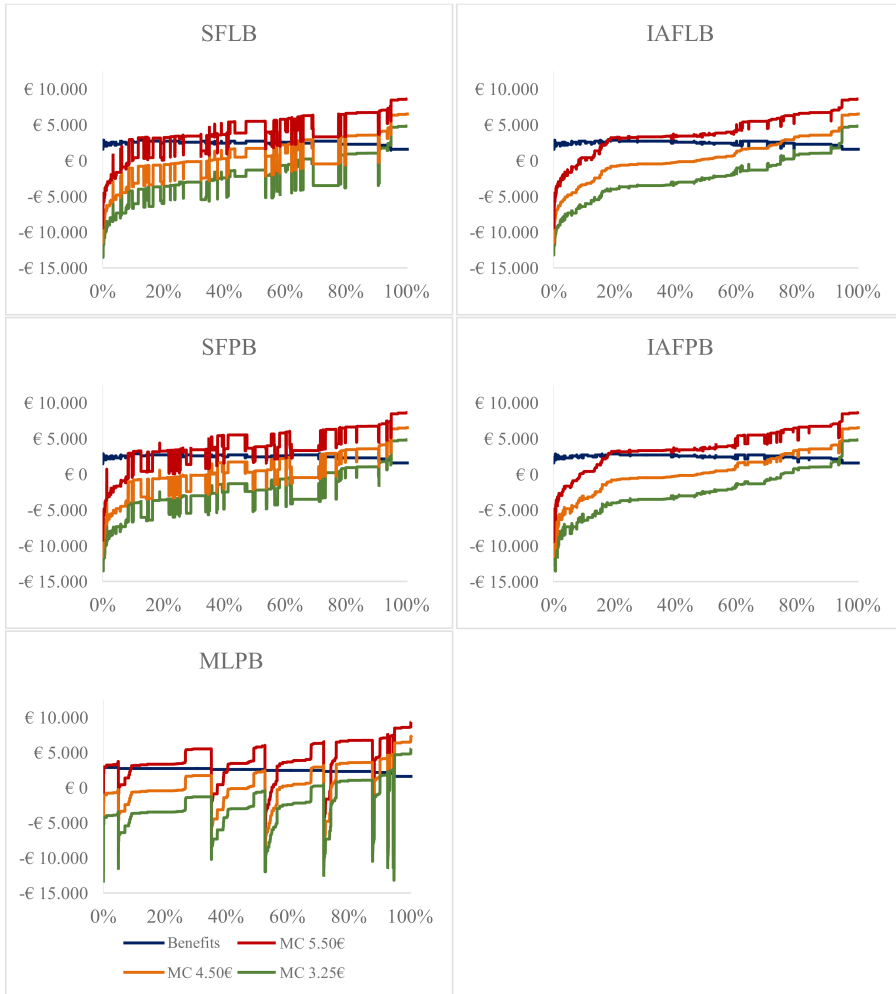


Figure 3.2: Discounted social marginal cost  $MC$  ( $C_i^e$ ) and social marginal benefit  $MB$  ( $B_i^e$ ) curves for the five heterogeneous subsidy schemes. Single Focus schemes follow the private opportunity costs for Miscanthus production, Integrate Agricultural Focus schemes follow social opportunity costs by taking into account the current allocation of conventional subsidies, ML minimizes land use change. LB and PB stand for land based and production based schemes respectively. Horizontal axis is the percentage of total agricultural land deployed for Miscanthus production.

there are however numerous intersections spread out over a large part of the graph area. This implies that the effect of spatial heterogeneity on the relative performance of SF and IAF schemes varies strongly depending on the market prices for bioenergy crop material. The main implication is that when market prices are low, and subsidy requirements are high, it pays more to subsidize the right plots first.

### 3.5.3 Comparing different policies

The relative performance of different heterogeneous policies varies with production total. Using Equation (3.10), the total net benefit potentials are calculated for the entire range of potential production sites.

Figure 3.3 depicts the course of potential net gains for an increasing total production area under different heterogeneous schemes. As the total area targeted by the policy increases, the curves diverge as spatial heterogeneity in the targeted area increases. When the total area targeted by the policy nears 70% of the entire region, the total net benefits under different heterogeneous schemes converge; when all the farmers within a region receive funds, the order of fund allocation or the selection of plots that receive funds within the region does not matter. The largest difference between IAF and SFLB schemes at €4.50/GJ occurs near a conversion rate of 68% of the region. At this point, potential net benefits of IAF schemes are 17% higher. At €4.50/GJ the differences between heterogeneous schemes are not very impressive, each policy has its own optimum and these optima produce relatively similar net benefits. But Figure 3.3 clearly shows an important aspect of heterogeneous schemes, the foregone benefits of second-best heterogeneous schemes are approximately hyperbolic with the rate of land-use conversion or aggregate subsidy spending.

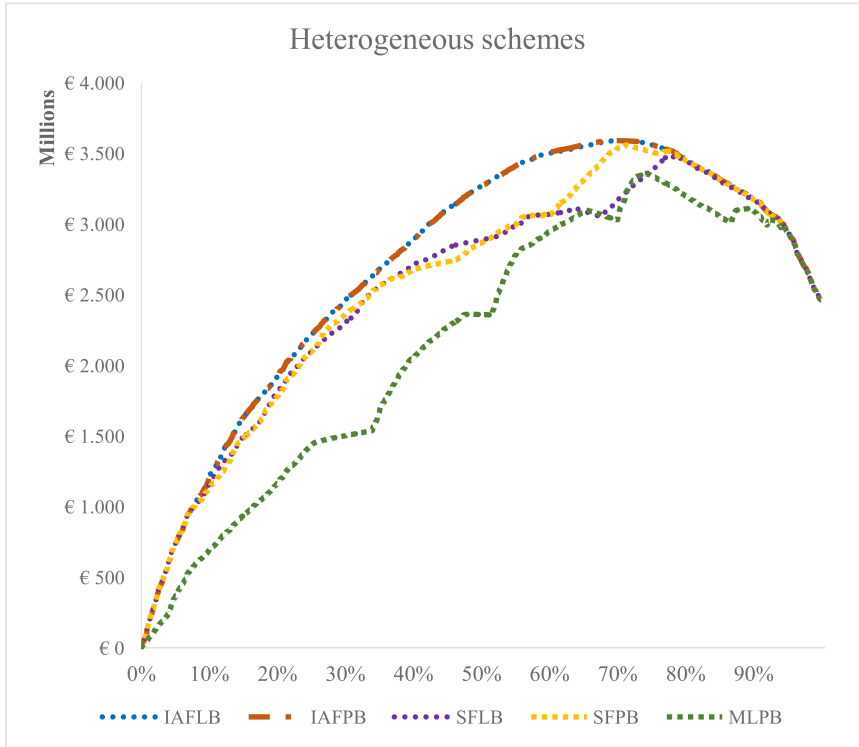


Figure 3.3: Total discounted net benefits in euros for heterogeneous subsidy schemes with bioenergy market prices of €4.50/GJ. Horizontal axis is the percentage of total agricultural land deployed for Miscanthus production.

To investigate further how different heterogeneous schemes compare, we repeated the analysis for a range of bioenergy prices with €0.10 increments. We compare the different policies to obtain information on the overall convergence or divergence in performance of the different policies when prices for bioenergy change. For robustness, we are interested in comparing the performance of policies when each policy is evaluated at its optimum and when policies are evaluated at the point where they differ the most in terms of efficiency. Therefore we computed two statistics for each price level: I) the percentage difference between maximum net benefits, evaluated for each policy at its respective optimum, and II) the

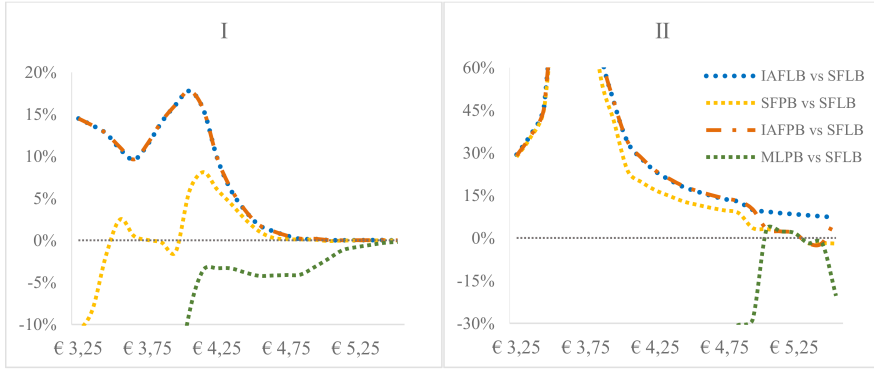


Figure 3.4: Relative net benefits of heterogeneous schemes compared to SFLB schemes for a range of bioenergy prices depicted as; I) the percentage difference between the optima of different policies, and II) the percentage difference evaluated at the widest gap between the potential benefits associated with the different schemes.

percentage difference between net benefits, evaluated at the widest gap between the benefit curves. We benchmark the policies against the SFLB scheme to see how integrated and production based schemes compare to land based heterogeneous schemes. At low market prices, the second measurement is associated with negative SFLB benefits. For these cases, relative differences are computed using an absolute valued denominator.<sup>17</sup>

Figure 3.4 shows that the relative difference between net benefits, according to both measurements, varies strongly with bioenergy prices. Ordinal performance stays, however, relatively stable over the evaluated price range. IAF schemes, measured at both optima and widest gaps, perform better than SFLB schemes, while the MLPB scheme performs less. At the lowest evaluated price, the SFPB scheme in optimum, performs less than the SFLB scheme in optimum. It performs however better at low to mid-range prices. When relative performance is measured at the widest gap between the net benefit curves, the production based version of single focus schemes performs better at any evaluated price level. The largest relative differences in optima occur at low bioenergy prices. A

<sup>17</sup>As:  $(Alternative_{scheme} - SFLB) / |(SFLB)|$

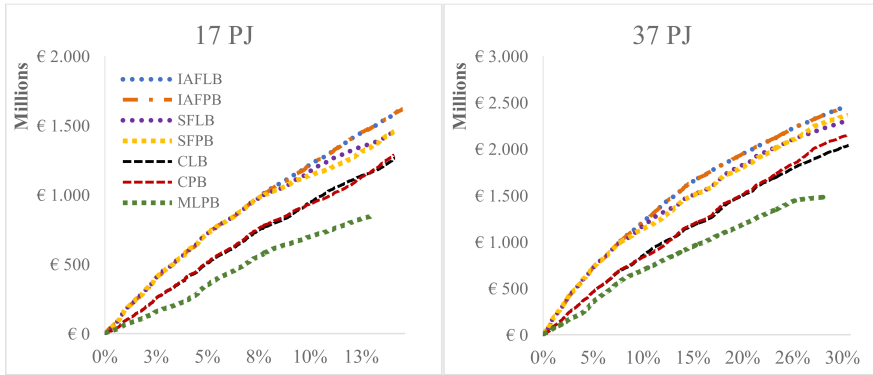


Figure 3.5: Aggregate discounted net benefits in euros for each subsidy scheme with bioenergy market prices of €4.50/GJ.

striking feature of 4I and 4II is the large spike around a bioenergy price of €3.65/GJ. At this price, evaluated at its optimum, the SFLB scheme is close to the social break-even point. This results in high relative differences. As prices increase, both figures I and II show a convergence of heterogeneous schemes. The general trend of high differences at low prices and convergence at higher prices, can be attributed to an interaction between spatial heterogeneity and bioenergy prices. When bioenergy prices are low, many plots require a subsidy. There are initially high relative rewards for subsidizing the right plots. When prices increase, fewer plots require subsidy, and subsequently there is less heterogeneity in the remaining plots that require aids.

To investigate the potential benefits of implementing spatial variation in subsidy schemes, we compare the relative performance against spatially homogeneous subsidies aimed at reaching the European 2020 fuel standards.<sup>18</sup> Figure 3.5 depicts net potential benefits associated with both heterogeneous and conventional schemes. Conventional schemes are clearly less efficient but outperform the MLPB scheme. At 17 PJ, IAF

<sup>18</sup>European fuel standards can be reached with further growth of first-generation biofuel crops, in which case a total production of 17 PJ of second-generation biofuels is required, or without further growth in first-generation biofuel crops, in which case 37 PJ is the required production growth. See Appendix A for details on the construction of these figures.

schemes produce 28% more gains than conventional land based subsidies. The net benefit curves of conventional schemes in Figure 3.5 are less steep at the 37 PJ production total than for the 17 PJ production total. This is in line with what was derived analytically from our model, the marginal increase in aggregate subsidy spending between the homogeneous schemes and heterogeneous schemes diverges as the policy area increases. At the same time, at 37 PJ, an IAF scheme increases net benefits with 20%, slightly less than at 17 PJ. This means that, while over-funding related to homogeneous subsidies increases, there is a sharper decline in the marginal benefits of IAF schemes. For the transition of 17 PJ to 37 PJ, the decrease in the marginal potential net benefits is thus stronger than the increase in the forgone benefits of conventional schemes.

The increase in foregone benefits under conventional schemes, however, might have strong implications for environmental policy. When conventional schemes are in place and international agreements become more ambitious – energy targets are replaced with more ambitious ones – subsidy schemes need to adjust to generate the increased supply required. This means that periodical aids are required to increase such that additional farmers, with higher opportunity costs, will contribute to bioenergy production as well. The implication of having conventional schemes in place is that it can form a disincentive for engaging in new and more ambitious agreements. The results show that this effect might even be slightly stronger for land based schemes than for production based schemes. Under a conventional policy, at 37 PJ, production based schemes have 5% higher potential net benefits than land based schemes. At 17 PJ, the difference is only a 2%.

The analysis also shows that minimizing land-use change comes at relatively high costs. At 17 PJ, the potential gains are 33% lower than those of IAF schemes, while the total land-use change is reduced by around 11%. At 37 PJ, the gains are 27% lower while the total land-use



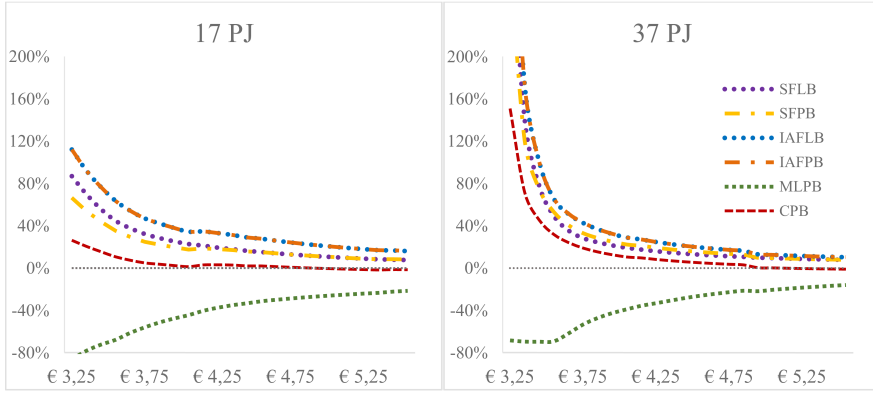


Figure 3.6: Total net benefits of different schemes at bioenergy production levels of 17PJ and 37PJ.

change is reduced by only 5%. In both cases, we do not account for possible benefits of minimizing total land-use change, but the results imply that these have to be substantial if MLPB schemes are preferred over IAF schemes. If the MLPB scheme is, however, benchmarked against conventional schemes, the additional required benefits from minimizing the land-use change need to be substantially smaller.

Final analysis shows that the results are robust to a range of different prices. At both 17PJ and 37PJ we repeated the analysis with €0.10 bioenergy price increments and noted the percentage difference between alternative subsidy schemes and CLB schemes to see how heterogeneous schemes compare to conventional schemes depending on bioenergy market prices. Figure 3.6 shows that the alternative schemes are especially more efficient when bioenergy prices are low.

We can observe from the graphs that the initial differences in relative performance of alternative schemes at low prices, are higher for increased total bioenergy production. The rate at which the curves converge is however also higher at increased total production. Whether the relative performance of heterogeneous schemes improves when bioenergy ambitions go up, thus depends on the market price of bioenergy at which

policies are evaluated. This nonlinear effect is due to the net effect of a trade-off between spatial heterogeneous interactions. When bioenergy prices are low, a large amount of plots require subsidy and subsequently the heterogeneity in subsidy requirements is high. Furthermore, when total bioenergy production increases, more plots are required to reach the target aggregate production quantity and heterogeneity within production sites increases. Together, low prices and larger aggregate production thus result in a high level of heterogeneity due to increased heterogeneity in subsidy requirements and additional heterogeneity due to the extended set of production sites required to reach total production quantities. If, however, bioenergy prices increase, heterogeneity in subsidy requirements decreases as the amount of plots that require subsidy decreases. Heterogeneity decreases disregarding aggregate production quantities, but the plots that will no longer require financial support at elevated prices, make up a larger share of the production sites at 17PJ than at an aggregate production quantity of 37PJ. As an effect, the relative gains of preventing excessive funding of plots at high prices, is larger at low production quantities, and heterogeneous schemes perform relatively better at lower total production if bioenergy prices are high. Disregard of this complexity, IAF schemes are clearly more efficient than conventional schemes at any price for both aggregate production totals, and MLPB schemes are relatively costly to society.

### 3.6 Discussion and conclusions

In this paper, we explored the role of spatial heterogeneity in biofuel stimulation schemes. Under heterogeneous subsidy allocation, we find that the potential gains from stimulating *Miscanthus* production are distributed according to the differences between potential private profits and potential net social benefits. The efficiency of a heterogeneous allocation is therefore strongly determined by the order at which sites

are targeted. We considered three types of heterogeneous schemes: 1) Single Focus schemes that allocate subsidy based on private opportunity costs, 2) Integrated Agricultural Focus schemes that additionally take into account how conventional agricultural subsidies are distributed, and 3) a scheme that Minimizes Land use change. First, our results show that conventional subsidy schemes that allocate a fixed amount of funds on a per-hectare basis, tend to over-fund a large part of the farmers who engage in biomass production, and that under heterogeneous stimulation schemes, there is scope for Pareto efficient improvements with respect to current subsidy spending. While heterogeneous schemes minimize over-funding, they may mis-allocate funds. Our results show that a scheme that targets plots based on the expected yields of bioenergy crops, is less efficient than conventional stimulation schemes in reaching the 2020 mandate. We find that schemes that follow private opportunity costs mis-allocate funds due to the heterogeneity in both external benefits of produced carbon offsets and the potential to reduce conventional agricultural subsidies. The foregone benefits of non-optimal heterogeneous allocation is hyperbolic with total land conversion. It is found that integrated agricultural focus schemes optimize benefits by reducing the conventional subsidy for other agricultural activities, minimizing social costs of sustaining biofuel stimulation and minimizing both over-funding and mis-allocation. Alternatively, one can view the differences in efficiency between single focus and integrated schemes not as properties of the schemes, but of subsidies that are currently in place for conventional agricultural activities. The differences between heterogeneous schemes are fairly small at higher bioenergy prices, but increase substantially in the lower range. At high prices, many sites do not require a subsidy, and there is less heterogeneity among potential production sites. The link between the relative importance of heterogeneity and energy prices is important considering the 2014 oil price drop. Second, we show that under both single focus schemes and conventional schemes, production

based payments generate better results than land area based payments. Again these differences are small at high bioenergy prices, but increase in the lower range of energy prices. Third, under heterogeneous schemes, the marginal costs of engaging in more ambitious environmental agreements follow only site-specific social costs, while under conventional schemes they increase rapidly due to increased over-funding of farmers that have lower minimum subsidy requirements. The most substantial increase in net benefits can be achieved when market prices for bioenergy are low and environmental targets are ambitious. This is an important finding and stresses the relevancy of spatial heterogeneity for policy since many countries struggle with meeting their ambitious objectives under current energy prices.

Our results add to the discussion around carbon contracts. In recent years, agreements such as the Kyoto Protocol and the recent Paris agreement, have encouraged the global economy to collaborate in creating emission trading markets. While the recent agreement did not specify a global carbon price, it does recognize its importance for providing incentives for emission reduction activities. Furthermore, it mentions result-based payments as an important way to provide incentives for emission reductions. It appeals to suggest that biofuel stimulation policies could be improved by capping and trading. Supporting farmers through this system can address problems related to the cost-efficiency differences arising from spatial heterogeneity and shift biofuel production to farmers that face favorable production characteristics. The cost-effectiveness of cap-and-trade has been widely discussed already in the 1970s regarding air pollution policies (Burton and Sanjour, 1967) and more recently concerning agriculture. Specifically, it has been shown that cap-and-trade programs outperform tax-based policies (Bakam et al., 2012). A large body of literature on carbon sequestration also supports integration with cap-and-trade (Parks and Hardie, 1995; Pautsch et al., 2001; Antle et al., 2003). In fact, our results corroborate that homogeneous production based

schemes are more efficient than per-hectare payments. Especially at the lower range of evaluated prices this improvement is substantial. However, a problem related to the implementation of carbon sequestration contracts is the high costs of quantifying the amount of sequestered carbon at every production site (Stavins, 1999). Obstructions of this kind seem less stringent in the case of contracts for biomass production, as the output of these activities can be more easily measured since it is primarily the final product itself that contributes to emission savings. This suggests that there is a strong case for policy targeting to reduce emissions by capping and trading. However, our study reveals that the integration of biofuel production into cap-and-trade by providing emission offsets remains prone to inefficiencies that arise from spatial heterogeneity. The results show that conventional production based schemes over- and mis-allocate funds, and schemes that explicitly address the heterogeneity in subsidy requirements, and possibly in the distribution of externalities, increase benefits substantially. Especially when total production of emission offsets increases, and market prices for bioenergy are low, the amount of foregone benefits sum up considerably.

This study contributes to the general debate on the potential contribution of the agricultural sector in reducing emissions, by offering insights in more efficient stimulation schemes. Prior to implementing such policies, more extensive cost-benefit analysis that accounts for additional factors influencing local production potential is needed. We suggest some extensions to the framework presented in this paper. One potential drawback of our pixel-by-pixel approach is that sites are treated as *identically and independently distributed*, while in reality farmers typically manage several sites under a single budget construct and can be expected to make managements decisions based on returns to the whole farm operation. On a related note, our approach does not account for economies of scale or risk aversion. As prices and yields are stochastic, profit, or expected value, maximization might deviate from the true objective function of

a risk-averse farmer. In this case, the correct objective function will instead maximize expected utility of profit, which could result in portfolio diversification. Risk is however not only related to volatility of market prices and yields, but also to irreversibility of some investments. While NPV methods are an established method for land-use valuation, there is a wide range of literature citing weaknesses that relate to this type of risk. NPV methods treat investments as a onetime only opportunity (Dixit and Pindyck, 1995), based under assumptions concerning future cash flows under a static investment strategy that a firm starts and completes as planned (Voeks, 1997). It is more realistic however to subscribe to the idea that investments become less risky into the future as the information set on which decisions are based grows, and that information can alter investment strategies along the way as it becomes available. For most investment strategies, the horizon is relatively short, as in our application, and the effect may not pose a significant problem (Pindyck, 2007). But in the case of bioenergy, for which a fully developed market does not exist, uncertainty regarding future cash flows is high. While the production of a perennial crop like *Miscanthus* allows very limited altering of the investment strategy along the way, irreversibility can be expected to play an important role in adaption and a better approach would be to explicitly balance the benefits of immediate investment to those of waiting in to reduce risk (Pindyck, 1991). Furthermore, our analysis showed that arable farming, *Miscanthus* production, and dairy farming, (here ordered by declining sensitivity to yield) all have a distinct sensitivity to yield. Risk due to stochastic yields will therefore have a distinct impact on the expected utility of profit for each land-use type.

Improving the level of detail in the assessment by incorporating the notions described above will certainly result in a more precise analysis. However, while the addition of these complexities will impact the exact amounts of subsidies required to initiate bioenergy production, we expect our general conclusions to hold these do not depend strongly on the

accuracy of point estimates, but on the ordinality of efficiency results of different schemes, which are shown to be robust for a range of different prices. Future research might consider Real Option Value methods to explore the impacts of heterogeneity under risk and irreversibility of investments (see Regan et al. (2015) for more extensive discussion), and agent based models to explore the impact of moving towards more detailed representations of farm operations.

## 3.7 Appendix

### 3.7.1 A. Energy data

Table 3.2: Input data on the energy-related variables applied in this study.

Variables	Values
Required biofuel consumption in transport	611 PJ <sup>1</sup>
Miscanthus lignocellulosic energy content	5.95 GJ <sup>2</sup>
First-generation biofuel used	14 PJ <sup>3</sup>
Second-generation biofuel used	7 PJ <sup>3</sup>
Weighing factor first-generation fuel	1 <sup>4</sup>
Weighing factor second-generation fuel	2 <sup>4</sup>

<sup>1</sup> 10% of the total energy used in the transport sector (Eurostat - Statistical Office of the European Communities, 2009), <sup>2</sup> 35% lignocellulosic energy conversion taken from van der Hilst et al. (2010), 17 GJ energy per oven dry ton taken from Brosse et al. (2012), <sup>3</sup> Dutch Emission Authority (2013), <sup>4</sup> Dutch Emission Authority (2013), only half of bioenergy production may be food-based.

### 3.7.2 B. Crop rotation schemes

Table 3.3: Crop rotations for the two major production systems in our study, specified for soil types.

	Clay	Sand
Arable farming		
Ware potatoes	17.09%	15.06%
Seed potatoes	13.03%	4.07%
Starch potatoes	0.33%	29.30%
Beets	16.10%	21.06%
Winter barley	0.77%	1.42%
Summer barley	2.74%	12.31%
Winter wheat	46.66%	10.27%
Summer wheat	2.85%	5.19%
Fallow	0.44%	1.32%
Dairy farming		
Grass	89.0%	70.0%
Silage maize	11.0%	30.0%

### 3.7.3 C. Modeling the dairy farming production system

Table 3.4: Variables and values used to model local production quantities.

Variables	Values
Average number of cows per hectare	2.1 <sup>1</sup>
Average litres of milk per cow	8147 <sup>2</sup>
Energy need per cow per day	Modelled <sup>3</sup>
Digestible energy content of feeding material	11.6 MJ per kg <sup>4</sup>
Grass supply	Modelled <sup>5</sup>
Costs of silage maize	Opportunity costs
Field operation costs	Same as for grass <sup>6</sup>
Other animal costs (healthcare and breeding)	€189 annual, per cow <sup>7</sup>
Milk processing costs	€0.21 per litre <sup>8</sup>
Herd investment costs	€ 895 per cow <sup>9</sup>
Average lifetime of cow before replacement	5 years <sup>10</sup>

<sup>1</sup> Based on figures from LEI (2012), 2.1 is slightly above the national average of 1.9 but below some locally observed values, which go up to 2.6, <sup>2</sup> based on figures from LEI (2012), <sup>3</sup> modeled following Bouwman et al. (2005), <sup>4</sup> per kg oven dry grass and pelleted whole plant corn, taken from Stanton, T.L.; LeValley (2010), <sup>5</sup> modelled per month following the method by College of Agriculture Food and Rural Enterprise (2005) and rescaled using local yield values, <sup>6</sup> taken from van der Hilst et al. (2010), <sup>7</sup> adjusted for inflation and tax, based on 3-year company survey performed by PPP Agro Advice (de Jong, 2013), <sup>8</sup> from Evers et al. (2007), <sup>9</sup> four-year average price of two-year-old calf (LEI, 2012), <sup>10</sup> from Gosselink et al. (2008).



### 3.7.4 D. Frequency distribution of agro-economic performance

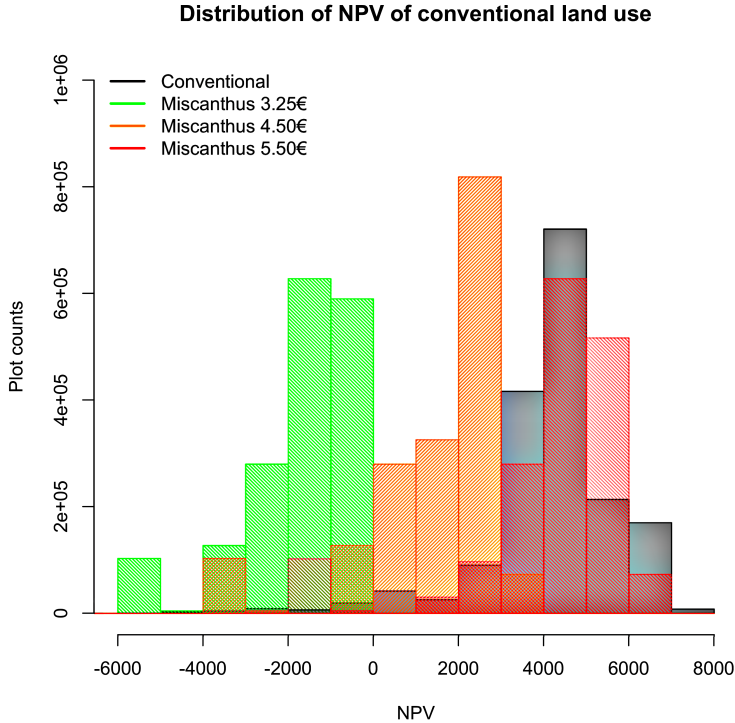


Figure 3.7: Distribution of the economic performance of the agricultural sector and Miscanthus in the Netherlands.

According to our estimations, the frequency distribution of the NPV of observed land use is left-skewed with a small number of production sites facing losses. Several factors may explain this outcome: 1) our estimation is negatively biased; 2) farmers possibly speculate on product prices and current land-use types generating long-run losses are profitable in the short run; 3) plots that face negative NPV benefit from unobserved comparative advantages such as regional specializations; 4) farmers do not fully take into account in their decision process all the costs that are included in our assessment; and 5) the agricultural sector is not fully in

equilibrium because of a high elasticity of land-use change. The aggregate amount of production sites with a negative NPV is, however, small and the overall distribution centers densely closely above zero, which is likely in a competitive market.

### 3.7.5 E. Spatial distribution of minimum required subsidies

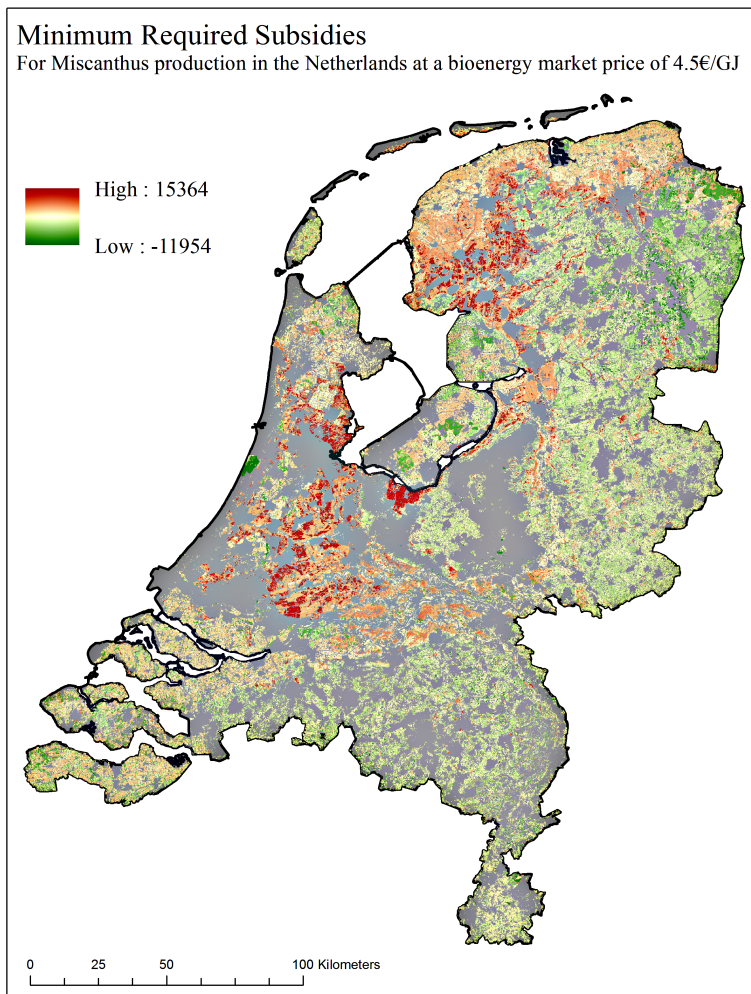


Figure 3.8: Spatial distribution of minimum required subsidies in the Netherlands.

The map shows the considerable differences in minimum required subsidies for bioenergy production. Clay soils, located mainly in the west of the Netherlands, perform economically better than the sandy soils, located in the east and north of the Netherlands. Clay soils coincide with areas where the minimum required subsidies for *Miscanthus* production are higher.

# Chapter 4

## Parametric Spatial Nonlinearities

### Chapter Summary

This paper introduces a new model for spatial time series in which cross-sectional dependence varies nonlinearly over space and time. We refer to it as the Smooth Transition Spatial Autoregressive (ST-SAR) model. We study the stochastic properties for the ST-SAR as a data generating process and obtain asymptotic theoretic properties for the maximum likelihood estimator (MLE) under correct specification and potential misspecification. The asymptotic consistency of the MLE explicitly allows for failure of parameter identification, which is a well-known issue of threshold models. To tackle the implications of the identification issue on the inference of the estimation results, we propose model selection based on in-sample and validation-sample estimates of Kullback-Leibler divergence. The methods are valid when the model is correctly specified, misspecified, over-specified and when parameters are unidentified. Simulations are presented that support the use of information criteria for model selection when the true process is linear and parameters of the model are unidentified, and when the process is nonlinear and the MLE of identified parameters is in fact well-behaved. These results are shown to be robust to additive outliers and fat-tailed errors. The model is applied to study space-time dynamics in two cases that differ in spatial and temporal extent. We study clustering in urban densities and pay particular focus to the advantages of the ST-SAR over linear spatial models as a way to appropriately filter out clustering dynamics. In our second study, we apply the ST-SAR to monthly long term interest rates, and find evidence of asymmetries and cycles in spillover dynamics. In both applications, we find strong evidence for nonlinearity. The empirical evidence highlights that the ST-SAR improves significantly over the SAR and is a powerful tool to understand and predict future values in cross-sectional time series with different dependence regimes.<sup>1</sup>

---

<sup>1</sup>This chapter is based on “*Smooth Transition Spatial Autoregressive Models*”, available as part of the *Tinbergen Institute Discussion Papers* in the Econometrics and Operations Research research group. The full reference is Andree et al. (2017a).

## 4.1 Introduction

Spatial entanglement of economic agents plays an important role in the realization of many economic processes measured over space and time. Spatial autocorrelation models are capable of describing the spatial dependence between variables measured across space and are widely adopted in different research fields; see e.g. LeSage and Fischer (2008) on regional growth, Kostov (2009) on agricultural land prices, Baltagi et al. (2014) on housing prices, Debarsy et al. (2015) on foreign direct investments and Hoshino (2016) on crime.

Standard spatial models account for spatial correlation in (un)observed variables, but commonly assume that the spatial autoregressive parameter is constant across space and over time. In particular, models that allow for spatial (auto)correlation often do not sufficiently relax linearity constraints on functional representations of spatial spillovers. Specifically, spillover-processes are represented by “global” dependence parameters (Fotheringham, 2009). The literature has stressed the importance of relying on local statistics for spatial dependence instead of global measures due to parameter heterogeneity; see Anselin (1995) and Fotheringham (2009). Local statistics allow for variation in spatial correlation across grouped cross-sectional units. Typically, spatial aggregation into groups relies on the use of econometric tools to avoid ad-hoc sample divisions. Researchers have for example relied on Geographically Weighted Regression (GWR) (Fotheringham et al., 2002; Su et al., 2012), boosted trees (Crane et al., 2012), Bayesian (Glass et al., 2016), nonparametric (Frías and Ruiz-Medina, 2016), or semiparametric (Basile et al., 2014) approaches to model heterogeneity.

All of the above approaches, however, treat the spatial dependence parameter as static. That is, the (local) parameters represent effects that are fixed (locally) across the data dimensions, rather than introducing relationships that produce varying effects as a function of data itself.

Heterogeneity in interaction is instead achieved by using trend surfaces or sample divisions that allow estimation of separate parameter vectors. For example, the Spatial Autoregressive Semiparametric Geoadditive Models discussed by Basile et al. (2014) maintain linearity assumptions with respect to the spatial autoregressive component, but have smooth locally linear dependence structures with respect to exogenous variables. The GWR isolates neighborhoods in a Cartesian coordinate system using kernels, and estimates weighted parameter vectors for those different neighborhoods. These approaches typically require many observations per neighborhood to be effective and numerous studies pointed out serious drawbacks.<sup>2</sup> We note that grouping observations not by kernels but through fixed effect approaches also has its drawbacks because convergence rates depend on the number of observations in groups tending to infinity (Bonhomme and Manresa, 2015), while in many practical situations additional observations can only be collected over time with group sizes remaining fixed.

In this paper we propose a parsimonious model in which cross-sectional dependence varies nonlinearly over both space and time. The model builds on the well-known SAR model (Anselin, 1988) and the Smooth Transition Autoregressive (STAR) framework advocated by Teräsvirta and Anderson (1992); Granger and Teräsvirta (1993); Teräsvirta (1994); Teräsvirta et al. (2010). In the resulting ST-SAR, dynamics in the cross-sectional dimension are driven by a smooth transition function around lagged cross-sectional variables that are possibly endogenous or “self-exciting”. This configuration allows for the possibility of regime-specific dynamics in spillovers with differential in intensity, and allows observations to move smoothly from one regime to another over time. Feedback loops

---

<sup>2</sup>Inadequate modeling of spatial lag and error processes (Leung et al., 2000; Fotheringham et al., 2002; Paez et al., 2002), spatial patterns revealed by GWR could be attributed to the procedure itself rather than the data generating process (Wheeler and Tiefelsdorf, 2005; Wheeler, 2007), and finally problems that relate to bandwidth selection and local violations of least-squares assumptions (Wheeler and Tiefelsdorf, 2005; Farber and Páez, 2007; Cho et al., 2010).

that amplify spillovers in the cross-section are also modeled with varying intensity in this way, both in the cross-sectional and in the temporal dimension. The entanglement structure remains exogenously determined through the specification of a spatial weights matrix following standard procedures in the spatial econometric literature.

We study the stochastic properties for the ST-SAR as a data generating process and obtain asymptotic theoretic properties for the MLE by taking the time dimension to infinity. We focus on  $t$ -distributed innovations as a generalization of the Gaussian case and as an attractive way to achieve robustness to fat tails and outliers. Our theory comes in a variety of flavors that allow for possible failure of parameter identification and potential model misspecification. In particular, we develop consistency when the model is correctly specified or possibly misspecified, and in both cases develop set-consistency when one or more parameters of the model are not identified. We establish asymptotic Gaussianity of the MLE of the correctly specified and identified parameters when the score is a martingale difference sequence and similarly when the model is mis-specified but the score it is near epoch dependent. Since the normality breaks down for unidentified parameters, estimated distributions of parameters cannot be used directly to infer the presence of nonlinear dynamics in the data. We therefore develop in-sample and out-of-sample methods that rely on unbiased estimates of log likelihood, that are still available when one or more parameters are unidentified and the model is set-consistent, to diagnose the presence and significance of nonlinearity. In particular, we investigate the usefulness of information criteria that already have been applied successfully to distinguish nonlinearity in univariate threshold time series. We highlight that information criteria consistently rank the models asymptotically according to Kullback-Leibler divergence, even if parameters are unidentified. To address possible other sources of bias, such as over-fitting, we also provide a theoretical argument for model selection based on a validation-sample estimate of Kullback-Leibler di-

vergence which is again valid when one or several parameters of the model are not identified because it relies on assumptions that are imposed directly on the differential in forecast errors and not on model parameters. Simulations support the use of information criteria for model selection when the true process is linear and parameters of the model are unidentified, and when the process is nonlinear and the MLE of identified parameters is in fact well-behaved. These results are shown to be robust to additive outliers and fat-tailed errors.

We apply our model to study two cases with different panel dimensions. In the first application, we study clustering in residential densities in a large number of districts in the Netherlands. We test two hypotheses regarding cross-sectional dependencies that cannot be captured by linear models: (i) that spatial autocorrelation decays along the urban gradient in line with the distance decay of agglomeration effects (Fotheringham, 1981; Rosenthal and Strange, 2003); and (ii) that the relation between concentrations of urban densities and household compositions of surrounding neighborhoods inverts along the urban gradient, reflecting sorting patterns that arise under single-crossing assumptions about household preferences (Epple and Sieg, 1999). We model these nonlinearities with a threshold function specified around population densities and find strong evidence for both hypotheses.

Our second application uses a long time series of long term interest rates of a sample of European sovereigns. We assess the integration of financial systems by estimating ST-SAR dynamics in co-movements. Linear dynamics in co-movement, spillovers, and cross-sectional dependence have been explored in a number of studies on financial integration Frankel et al. (2004); Caceres et al. (2016); Kharroubi et al. (2016). We pay particular focus on the time-varying properties of sovereign-specific cross-sectional dependence parameters as a way to understand convergence and dispersion in interest rates. Our spatial weights matrix



is based on pair-wise correlations and allows spillovers to flow based on non-geographic linkages. We model nonlinearities with a threshold function specified around ARMA components and find strong evidence asymmetries and cycles in the spillover dynamics.

In both applications, the ST-SAR is shown to be a powerful tool for both understanding and predicting future values in cross-sectional time series in which the dependence of observations on neighbors changes once they enter a different regime. In particular, the ST-SAR improves substantially over the SAR in terms of improving log likelihood, (corrected) AIC and forecasting power. The ST-SAR also renders the residuals free of significant correlations while the SAR residuals maintain both strong spatial clustering and temporal correlations. Our most conservative tests remain significant at the highest level.

The remainder of this paper is organized as follows. Section 4.2 considers spatial autocorrelation models and proposes our nonlinear framework. It also highlights the issues related to parameter identification. Asymptotic theory for the MLE is examined in Section 4.3. Its finite-sample behavior is studied via simulations in Section 4.4. The model is applied in Section 4.5. Section 4.6 summarizes and concludes. Additional results and proofs are located in the Appendix. Additional theoretical results are provided in the Supplementary Appendix that comes with this paper.

## **4.2 Linear and nonlinear spatial autoregressive models**

### **4.2.1 Linear dynamics: the SAR Model**

Spatial data is often highly dependent across space. In order to model this dependence, Cliff and Ord (1969) proposed the Spatial Autoregressive (SAR) model. The SAR in the context an Autoregressive Moving Average

model with Exogenous Regressors (ARMAX) model is given by:

$$\mathbf{y}_t = \rho W \mathbf{y}_t + c + \sum_{p=1}^P \mathbf{y}_{t-p} \phi_p + \sum_{k=0}^K \mathbf{X}_{t-k} \boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_t + \sum_{q=1}^Q \boldsymbol{\varepsilon}_{t-q} \mu_q \quad \forall t \in \mathbb{Z}, \quad (4.1)$$

$$\{\boldsymbol{\varepsilon}_t\}_{t \in \mathbb{Z}} \sim p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}_t, \Sigma, \boldsymbol{\lambda}),$$

where  $\mathbf{y}_t$  denotes a vector of  $N$  cross-sectional observations at time  $t$ ,  $c$  is an intercept,  $\rho$  is the spatial dependence parameter,  $W$  is the  $N \times N$  matrix of exogenous spatial weights,  $\phi_p$  is the  $p$ -th lag autoregressive parameter,  $\mathbf{X}_{t-k}$  is an  $N \times D$  matrix of  $D$  exogenous regressors at lag  $k$  with  $\boldsymbol{\beta}_k$  as the  $D \times 1$  vector of coefficients,  $\mu_q$  is a  $q$ -th lag moving average parameter and  $\boldsymbol{\varepsilon}_t$  is the disturbance vector with multivariate density  $p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}_t, \Sigma, \boldsymbol{\lambda})$  with zero mean and unknown variance-covariance matrix  $\Sigma$ . Other possible parameters are contained in the vector  $\boldsymbol{\lambda}$ . In this model structure, each entry  $y_{it}$  for  $i = 1, \dots, N$ , of the vector  $\mathbf{y}_t$  depends on the local values in the  $K$  lags of  $D$  individual-specific regressors  $\{x_{it-k,d}\}_{d=1,k=0}^{D,K}$ , as well as the neighboring entries of  $y_{jt}$  and thus indirectly on  $\{x_{jt-k,d}\}_{d=1,k=0}^{D,K}$  for  $i \neq j$ . Similarly, the (moving) error structure spills over. Spatial dependence modeling is made operational by specifying the spatial weights matrix  $W$  that defines the dependence structure between cross-sectional entries, for example as a function of geographic or economic distances. It is standard procedure to row-normalize  $W$  such that  $\sum_{j=1}^N w_{ij} = 1 \quad \forall i \in N$ , where  $w_{ij}$  is the  $i, j$ -th element from  $W$ .

The parameter  $\rho$  captures the spatially weighted effects of neighboring values  $W \mathbf{y}_t$  on  $\boldsymbol{\lambda}_t$ . In this simple framework, nonlinear feedback effects across entries can be captured, shown by rewriting the model as:

$$\mathbf{y}_t = H^{-1} \left( c + \sum_{p=1}^P \mathbf{y}_{t-p} \phi_p + \sum_{k=0}^K \mathbf{X}_{t-k} \boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_t + \sum_{q=1}^Q \boldsymbol{\varepsilon}_{t-q} \mu_q \right), \quad (4.2)$$

$$\{\boldsymbol{\varepsilon}_t\}_{t \in \mathbb{Z}} \sim \mathcal{NID}(0, \sigma_{\boldsymbol{\varepsilon}}^2),$$

where  $H := I_N - \rho W$  and  $I_N$  denotes the  $N \times N$  identity matrix. Following LeSage (2008) we obtain the following infinite power series expansion

$$\mathbf{y}_t = (I_N + \rho W + \rho^2 W^2 + \dots) \left( c + \sum_{p=1}^P \mathbf{y}_{t-p} \phi_p + \sum_{k=0}^K \mathbf{X}_{t-k} \boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_t + \sum_{q=1}^Q \boldsymbol{\varepsilon}_{t-q} \mu_q \right). \quad (4.3)$$

Equation (4.3) reveals that when  $\rho > 0$ , effects spill over to other regions  $j \neq i$  with a rate that declines as proximity to  $i$  increases, via the structure imposed by  $W$ . Feedback occurs for positive  $w_{ij}$  and  $w_{ji}$  and mutual neighbors  $i$  and  $j$ , as by construction of the matrix  $W$ , every observation is a second order neighbor of itself. A stable process therefore requires exogenous shocks to die out over space, which for the linear spatial process without time dynamics is guaranteed if  $\rho \in [-1/|\omega_{min}|, 1/\omega_{max}]$ , where  $\omega_{min}/\omega_{max}$  are the smallest/largest eigenvalues of  $W$  (Lee, 2004), or equivalently  $|\rho| < 1$ , if the rows of  $W$  sum up to one.<sup>3</sup>

The endogenous nature of this model causes inconsistencies in the least squares estimator that increase with  $N$ . However, we can consistently estimate SAR models by Quasi Maximum Likelihood methods (Q)ML or Generalized Methods of Moments (GMM), e.g. Kelejian and Prucha (2010). ML estimation of SAR models is pioneered in Ord (1975) and the asymptotics of the QML estimator are derived in Lee (2004). Finite sample distributions are investigated by Das et al. (2003); Bao and Ullah (2007).

#### 4.2.2 The Smooth Transition Spatial Autoregressive model

The linearity of the SAR model imposes the crucial simplifying assumption that the spatial dependence is fixed for any levels of both  $\mathbf{y}_{t-p}$  and  $\mathbf{X}_{t-k}$ . Anselin (1995) argues that spatial heterogeneity can complicate

<sup>3</sup>As we shall see, stability is understood in terms of bound on  $\|\rho W\|$  and depending on the configuration of  $W$ , alternative lower-level parameter restrictions can be obtained, see also Elhorst (2010a) for a discussion. In the nonlinear setting, these do not apply as  $\rho$  becomes non-scalar, several useful results on the stability of nonlinear spatial systems can be found in the appendix.

the analysis. His argument is centered on the notion that geographic phenomena often do not deviate around a constant mean, but likely move from one local average to another. For this reason, the simple SAR model may be a problematic model for describing the very phenomenon that Cliff and Ord (1969) are trying to model; see Fotheringham (2009) for a discussion. As we shall see in Section 4.5, the linearity assumption is not supported by the data.

In what follows we allow the spatial dependence parameter  $\rho$  to change as a function of a set of variables  $\mathbf{Z}_t$  that may include (spatial lags of)  $\mathbf{y}_{t-p}$  or  $\boldsymbol{\varepsilon}_{t-q}$  for any  $(p, q) \geq 1$  and or  $\mathbf{X}_{t-k}$  for any  $k \geq 0$ . In particular, we build on the popular smooth transition autoregressive (STAR) model introduced in Teräsvirta and Anderson (1992); Teräsvirta (1994).<sup>4</sup> The resulting Smooth Transition Spatial Autoregressive (ST-SAR) model with ARMAX terms takes the form<sup>5</sup>

$$\mathbf{y}_t = \rho(\boldsymbol{\theta}^\rho; \mathbf{Z}_t) \circ W \mathbf{y}_t + c + \sum_{p=1}^P \mathbf{y}_{t-p} \phi_p + \sum_{k=0}^K \mathbf{X}_{t-k} \boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_t + \sum_{q=1}^Q \boldsymbol{\varepsilon}_{t-q} \mu_q, \quad (4.4)$$

$$\{\boldsymbol{\varepsilon}_t\}_{t \in \mathbb{Z}} \sim p_\varepsilon(\boldsymbol{\varepsilon}_t, \Sigma, \lambda),$$

where the spatial dependence  $\rho(\boldsymbol{\theta}^\rho; \mathbf{Z}_t)$  is determined by

$$\rho(\boldsymbol{\theta}^\rho; \mathbf{Z}_t) = \kappa + \frac{\delta}{1 + \exp(-\gamma(\mathbf{Z}_t - \tau(\boldsymbol{\theta}^\tau; \mathbf{Z}_t)))}, \quad (4.5)$$

$$\text{and } \tau(\boldsymbol{\theta}^\tau; \mathbf{Z}_t) = \alpha + \mathbf{Z}_t \boldsymbol{\varphi}, \quad (4.6)$$

where  $\circ$  denotes element-by-row multiplication,  $\boldsymbol{\theta}^\rho$  denotes the vector of unknown parameters  $\boldsymbol{\theta}^\rho := (\kappa, \delta, \gamma, \boldsymbol{\theta}^\tau)$ , and  $\boldsymbol{\theta}^\tau := (\alpha, \boldsymbol{\varphi})$  is a parameter vector of possible additional parameters within the threshold function that may include linear coefficients w.r.t. any of the ARMAX terms

<sup>4</sup>The STAR model is well known in the time-series literature for modeling nonlinear dynamics with thresholds; see Granger and Teräsvirta (1993) for a literature review of nonlinear time-series models. For a comprehensive review of STAR models, the reader is referred to Dijk et al. (2002).

<sup>5</sup>We discuss the nonlinear model within an ARMAX framework because, as we shall see in our applications, all of terms can affect the data both directly and through the spatial dependence parameters. Allowing the terms to explicitly effect  $\mathbf{y}_t$  directly and through  $\rho(\boldsymbol{\theta}^\rho; \mathbf{Z}_t)$  simultaneously, is crucial to determine effect channels.

contained in  $\mathbf{Z}_t$ . Note that we use  $\mathbf{Z}_t$  to refer to any variable which may be an endogenous lag, moving average or exogenous variable, and we allow it to be specified differently in Equation (4.5) and Equation (4.6).

The quantity  $\mathbf{Z}_t - \tau(\boldsymbol{\theta}^\tau; \mathbf{Z}_t)$  measures deviations of  $\mathbf{Z}_t$  from a possibly time-varying quantity  $\tau(\boldsymbol{\theta}^\tau; \mathbf{Z}_t)$ . In general, we allow  $\tau(\boldsymbol{\theta}^\tau; \mathbf{Z}_t)$  to be any function of the data  $\mathbf{Z}_t$ . In this paper we consider first order polynomials around three important alternatives, the cross-sectional average  $\tau(\boldsymbol{\theta}^\tau; \mathbf{Z}_t) = \alpha + \varphi N^{-1} \sum_{i=1}^N z_{it}$ , the local average  $\tau(\boldsymbol{\theta}^\tau; \mathbf{Z}_t) = \alpha + \varphi W \mathbf{Z}_t$ , and local observations  $\tau(\boldsymbol{\theta}^\tau; \mathbf{Z}_t) = \alpha + \mathbf{Z}_t \varphi$ . Other options include modeling  $\tau(\boldsymbol{\theta}^\tau; \mathbf{Z}_t)$  as a constant  $\alpha$  only, or using wider regional averages  $W_l \mathbf{Z}_t$  with  $W_l$  as a spatial weights matrix that includes up to  $l$  higher order spatial lags.

Note that Equation (4.5)-Equation (4.6) allow the spatial dependence to change smoothly between regimes. The ST-SAR differs considerably from time-varying spatial parameter models such as the spatial score model proposed in Blasques et al. (2018) which attempts to filter the unobserved time-varying sequence of global spatial parameters  $\{\rho_t\}_{t \in \mathbb{Z}}$  by means of a score filter. The ST-SAR explores the relation between the spatial dependence parameter  $\rho$  and variables in  $\mathbf{Z}_t$ , which allow it to produce time-varying local spatial parameters. The ST-SAR parameters,  $\delta$ ,  $\tau(\boldsymbol{\theta}^\tau; \mathbf{Z}_t)$  and  $\gamma$  produce dynamics that cannot be reproduced by the time-varying spatial parameter model of Blasques et al. (2018).

It is also worth noting that the STAR dynamics nest not only the linear SAR model, but also, a threshold model (like the TAR (Tong, 2015)) with instantaneous switching between regimes. The linear SAR case is obtained when  $\gamma \rightarrow 0$ . In contrast, a TAR model is obtained when  $\gamma \rightarrow \infty$ . Depending on  $\mathbf{Z}_t$ , the transition mechanism may be endogenous or exogenous in nature. In the empirical section we shall consider both exogenous cases such as  $\mathbf{Z}_t = \mathbf{X}_{t-p}$  and endogenous examples where we allow the nonlinearities to be driven by ARMA terms. Finally, we note

that, just as in the case of the SAR model, the ST-SAR can also be re-written as

$$\mathbf{y}_t = H(\boldsymbol{\theta}^\rho; \mathbf{Z}_t)^{-1} \left( c + \sum_{p=1}^P \mathbf{y}_{t-p} \phi_p + \sum_{k=0}^K \mathbf{X}_{t-k} \beta_k + \boldsymbol{\varepsilon}_t + \sum_{q=1}^Q \boldsymbol{\varepsilon}_{t-q} \mu_q \right), \quad (4.7)$$

where  $H(\boldsymbol{\theta}^\rho; \mathbf{Z}_t) := I_N - \rho(\boldsymbol{\theta}^\rho; \mathbf{Z}_t) \circ W$ . In SAR terminology,  $(I_N - \rho W)^{-1}$  is referred to as the (global) spatial multiplier. In the ST-SAR, we highlight that  $H(\boldsymbol{\theta}^\rho; \mathbf{Z}_t)^{-1}$  varies locally and over time. As we shall see, this calls for new generalizations of stability conditions.

### Identification of the model's parameters

Just as with univariate threshold modeling, an important feature of the model is the possible failure of parameter identification (Teräsvirta et al., 2010). As pointed out, the SAR model is nested in the ST-SAR model, and letting the spatial dependence parameter  $\rho(\boldsymbol{\theta}; \mathbf{Z}_t)$  be a constant introduces well-known identification problems related to the fact that nuisance parameters are present only under the alternative assumption of nonlinearity. This is discussed for example by Davies (1977, 1987). In the current case, if  $\gamma = 0$  then the parameters inside  $\tau(\boldsymbol{\theta}^\tau; \mathbf{Z}_t)$  are not identified and  $k$  and  $\delta$  are not separately identified. Furthermore, if  $\phi = 1$  and  $\alpha = 0$  the spatial dependence may remain constant, unless different variables are used inside  $\tau(\boldsymbol{\theta}^\tau; \mathbf{Z}_t)$ . As we can see from this, the identification problem of distinguishing the SAR model from the ST-SAR model is not straightforward. For example, Likelihood Ratio testing fails because the dimensionality of the parameter space depends on the hypothesis of nonlinearity being true or false. Wald statistics can also not be applied to the individual parameters that are essentially meaningless and redundant when the process is linear. To tackle the issue, the following section develops not only consistency of the MLE, but also set-consistency which allows for the failure of parameter identification. This in turn allows one to obtain unbiased estimates of expected likelihood

using out-of-sample validation, or in-sample estimates of Kullback-Leibler divergence by using information criteria. The next section details this further and develops two tools to diagnose the presence and strength of nonlinearity that are valid when one or several parameters are not identified.

### 4.3 Asymptotic theory for the ST-SAR model

Estimation of the ST-SAR model's parameters is crucial to infer if nonlinear dynamics are present in the data. The estimated parameters will also inform us about the existence of threshold dynamics, the location of those thresholds, and the smoothness and speed of transitions. In this section we present and discuss the properties of both the log likelihood function and the ML estimator. We provide conditions for the existence, strong consistency and asymptotic normality of the MLE. We also highlight model selection procedures that can be applied to decide between linear and nonlinear descriptions of the data. Our results allow for failure of parameter identification and potential model misspecification. Proofs can be found in Appendix 4.7. Our asymptotic results all refer to increasing the time dimension rather than the spatial dimension since the applications we consider are such that  $N$  cannot grow. Additional observations are collected over time only.

For simplicity, we focus our attention on the ST-SAR model with autoregressive dependence of order one ( $p = 1$ ) and deliver a simpler exposition of the theory by focusing on a contemporaneous exogenous variable ( $k = 0$ ) and excluding MA terms ( $q = 0$ ) from this section. In any case, the same asymptotic results for both the correctly specified and the misspecified case are easily generalized to further lags for the exogenous variables and MA terms, at the cost of heavier notation, additional assumptions, and longer proofs. It is well known that the stationarity re-

sults can be generalized to models with moving average components, and can be easily extended to accommodate for (lagged) exogenous variables as long as some data generating process is defined. The endogenous case, on the other hand, is naturally the most interesting case for a study on stationarity. In general, besides extending the stationarity and moments conditions, extra parameter restrictions would need to be put in place to ensure the invertibility of the ST-SAR model and the recovery of the error term sequence.

#### 4.3.1 Existence and measurability of the MLE

Let  $\boldsymbol{\theta}$  denote the vector of parameters of our ST-SAR model,  $\boldsymbol{\theta} := (\boldsymbol{\theta}^y, \boldsymbol{\theta}^\rho)$ ,  $\boldsymbol{\theta}^\tau \in \boldsymbol{\theta}^\rho$ ,  $\boldsymbol{\theta} := (c, \boldsymbol{\beta}, \phi, \kappa, \delta, \gamma, \alpha, \varphi)'$ . Furthermore, let  $\boldsymbol{\theta}_0$  denote the parameter of interest. Naturally, the ML estimator  $\hat{\boldsymbol{\theta}}_T$  is defined as

$$\hat{\boldsymbol{\theta}}_T \in \arg \max_{\boldsymbol{\theta} \in \Theta} \sum_{t=1}^T \ell_t(\boldsymbol{\theta}), \quad (4.8)$$

where

$$\ell_t(\boldsymbol{\theta}) = \ln \det H(\boldsymbol{\theta}^\rho; \mathbf{Z}_t) + \ln p_\varepsilon \left( H(\boldsymbol{\theta}^\rho; \mathbf{Z}_t) \mathbf{y}_t - c - \mathbf{y}_{t-1} \phi - \mathbf{X}_t \boldsymbol{\beta}, \Sigma; \lambda \right). \quad (4.9)$$

The dependence of  $\ell_t(\boldsymbol{\theta})$  on the data is omitted in the notation for convenience. Equation (4.9) differs from a standard cross-section likelihood function by the log determinant  $\ln \det H(\boldsymbol{\theta}^\rho; \mathbf{Z}_t)$ , which accounts for the nonlinear spatial feedback (Anselin, 1988). In this paper we shall focus on innovations with density  $p_\varepsilon$  given by the multivariate Student's  $t$ -distribution. The  $t$ -distribution naturally generalizes the multivariate normal distribution to allow for fat tails, rendering the dynamics more robust to incidental outliers. Using the standard expression for the multivariate  $t$ -distribution with  $\lambda$  degrees of freedom we obtain

$$\ell_t(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}^\rho; \mathbf{Z}_t) + A(\boldsymbol{\theta}) - \frac{1}{2}(\lambda + N)F(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{X}_t, \mathbf{Z}_t),$$



where  $Q(\boldsymbol{\theta}^\rho; \mathbf{Z}_t)$  is the log determinant

$$Q(\boldsymbol{\theta}^\rho; \mathbf{Z}_t) := \ln \det H(\boldsymbol{\theta}^\rho; \mathbf{Z}_t),$$

$A(\boldsymbol{\theta})$  is a constant given by

$$A(\boldsymbol{\theta}) := \ln \Gamma((\lambda + N)/2) \left[ \det \Sigma^{\frac{1}{2}} (\lambda \pi)^{\frac{N}{2}} \Gamma(\lambda/2) \right]^{-1},$$

and the random element  $F(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{X}_t, \mathbf{Z}_t)$  is naturally defined as

$$F(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{X}_t, \mathbf{Z}_t) := \ln \left( 1 + \lambda^{-1} \boldsymbol{\varepsilon}_t' \Sigma^{-1} \boldsymbol{\varepsilon}_t \right),$$

$$\boldsymbol{\varepsilon}_t = H(\boldsymbol{\theta}^\rho; \mathbf{Z}_t) \mathbf{y}_t - c - \mathbf{y}_{t-1} \phi - \mathbf{X}_t \boldsymbol{\beta}.$$

We first establish the existence and measurability of the MLE  $\hat{\boldsymbol{\theta}}_T$ . This ensures that the  $\arg \max$  set in Equation (4.8) is not empty and that  $\hat{\boldsymbol{\theta}}_T$  is a random variable.

**ASSUMPTION. 7** (*Compactness of  $\Theta$* ).  $(\Theta, \mathfrak{B}(\Theta))$  is a measurable space and  $\Theta$  is a compact subset of  $\mathbb{R}^{d_\theta}$ .

**THEOREM. 9** (*Existence and Measurability*). Let **ASSUMPTION. 7** hold. Then there exists a.s. an  $\mathfrak{F}/\mathfrak{B}(\Theta)$ -measurable map  $\hat{\boldsymbol{\theta}}_T : \Omega \rightarrow \Theta$  satisfying Equation (4.4) for all  $T \in \mathbb{N}$ .

### 4.3.2 Consistency and of the MLE

The consistency of the MLE  $\hat{\boldsymbol{\theta}}_T$  w.r.t. the parameter of interest  $\boldsymbol{\theta}_0 \in \Theta$  can be obtained under standard regularity conditions. Assumptions 8-9 impose the SE (Stationary and Ergodic) nature of the data and a bounded moment for  $Q(\boldsymbol{\theta}^\rho; \mathbf{Z}_t)$  and  $F(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{X}_t, \mathbf{Z}_t)$ . **ASSUMPTION. 10** ensures that  $\boldsymbol{\theta}_0$  is identified.

**ASSUMPTION. 8.** The random sequence  $\{\mathbf{y}_t, \mathbf{X}_t\}_{t \in \mathbb{Z}}$  is SE.

**ASSUMPTION. 9.** The following moment conditions are satisfied:

$$i \ \mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |Q(\boldsymbol{\theta}^\rho; \mathbf{Z}_t)| < \infty;$$

$$ii \ \mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |F(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{X}_t, \mathbf{Z}_t)| < \infty.$$

ASSUMPTION. 10.  $\boldsymbol{\theta}_0 \in \Theta$  is the unique maximizer of the limit likelihood;

$$\mathbb{E}\ell_t(\boldsymbol{\theta}_0) > \mathbb{E}\ell_t(\boldsymbol{\theta}) \quad \forall (\boldsymbol{\theta}, \boldsymbol{\theta}_0) \in \Theta \times \Theta : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

More primitive moment conditions can also be given. For example, we will show for  $\{\mathbf{y}_t\}_{t \in \mathbb{Z}}$  that, when the model is correctly specified, Assumption 8 holds within defined parameter regions. As a counterpart to Assumption 9,  $\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |Q(\boldsymbol{\theta}^\rho; \mathbf{Z}_t)| < \infty$  can be obtained by bounding  $\rho(\boldsymbol{\theta}^\rho, \mathbf{Z}_t)W$  away from 1 in norm (see our Supplementary Appendix), which necessarily holds within stable parameter regions.<sup>6</sup>  $\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |F(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{X}_t, \mathbf{Z}_t)| < \infty$  is implied by logarithmic moment conditions on  $\mathbf{y}_t$  and  $\mathbf{X}_t$ . Again, when the model is correctly specified, then within that same stable parameter region logarithmic moments of  $\mathbf{y}_t$  and  $\mathbf{X}_t$  follow trivially because  $H(\boldsymbol{\theta}^\rho; \mathbf{Z}_t)^{-1}$  is uniformly bounded, and hence, Theorem 6.10 in Pötscher and Prucha (1997) applies, as the nonlinear ST-SAR model is bounded by a linear contracting recursion. For example, given that the innovations are Student's  $-t$  distributed,  $\lambda > 0$  is needed to ensure the existence of logarithmic moments. Finally, Assumption 10 requires  $\delta > 0$  and  $\gamma > 0$ , which holds trivially if the model is correct and not overspecified. As we shall see, even when Assumption 10 does not hold but  $\Theta$  is still compact, set-consistency can be obtained and model selection can be used to drop the unidentified parameters. When the model is misspecified, moments of the data have to be assumed.

THEOREM. 10 below establishes the strong consistency of the MLE  $\hat{\boldsymbol{\theta}}_T$  with respect to  $\boldsymbol{\theta}_0 \in \Theta$ . When the model is well specified,  $\boldsymbol{\theta}_0$  corresponds naturally to the so-called *true parameter* that indexes the distribution of the data. If the model is misspecified, then  $\boldsymbol{\theta}_0$  is often called a *pseudo-true parameter* that, by construction, is the minimizer of the expected Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between

<sup>6</sup>The moment condition  $\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |Q(\boldsymbol{\theta}^\rho; \mathbf{Z}_t)| < \infty$  is implied by positive definiteness of  $\det H(\boldsymbol{\theta}^\rho; \mathbf{Z}_t)^{-1}$  which for the SAR with a row-normalized  $W$  follows from  $|\rho| < 1$ . We note that the nonlinear case  $|\rho(\boldsymbol{\theta}^\rho; \mathbf{Z}_t)| < 1$  is not a necessary condition for  $\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |Q(\boldsymbol{\theta}^\rho; \mathbf{Z}_t)| < \infty$ ; see LEMMA. 5 in the Supplementary Appendix.

the true conditional density of the data  $p^0(\mathbf{y}_t|\mathbf{y}^{t-1})$  and the parametric conditional density implied by the ST-SAR model  $p(\mathbf{y}_t|\mathbf{y}^{t-1}, \boldsymbol{\theta})$ ; see e.g. White (1994) for details. In this sense, under model misspecification, the MLE converges at least to the parameter that delivers the best approximation to the true distribution of the data. The economic interpretation of empirical evidence is then as follows. When the model is correctly specified, the estimated parameters can directly be used as evidence for nonlinearity in an economic process. When the model is misspecified, then the parameters converge to the values for which the model best describes the data features, and as such we may conclude that the evidence for the existence of nonlinear regime-dependence in the observed data is stronger than the evidence for linear dependence which instead describes the data poorly.

**THEOREM. 10** (*Strong consistency under possible misspecification*). *Let Assumptions 7-10 hold. Furthermore, let  $\Theta$  be such that  $\Sigma$  is positive definite for every  $\boldsymbol{\theta} \in \Theta$ . Then the MLE satisfies  $\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$  as  $T \rightarrow \infty$  where*

$$\boldsymbol{\theta}_0 = \arg \min_{\boldsymbol{\theta} \in \Theta} \mathbb{E} \text{KL} \left( p^0(\mathbf{y}_t|\mathbf{y}^{t-1}) , p(\mathbf{y}_t|\mathbf{y}^{t-1}, \boldsymbol{\theta}) \right).$$

Propositions 1 and 2 give sufficient conditions for the geometric ergodicity of data generated by the ST-SAR. This allows us to impose conditions on the ST-SAR data generating process that ensure Assumptions 8-9 for the endogenous parts of the model. Propositions 1 and 2 can be easily extended to accommodate for exogenous variables  $\mathbf{X}_t$  as long as some data generating process is also defined for  $\mathbf{X}_t$ .

**PROPOSITION. 1.** *Let  $\{\mathbf{y}_t\}_{t \in \mathbb{N}}$  be generated by the ST-SAR model in (4.4) with  $\beta = 0$ ,  $\boldsymbol{\varepsilon}_t$  iid with full support, and  $\sup_{\mathbf{y} \in \mathbb{R}^N} \|\rho(\boldsymbol{\theta}_0^\rho; \mathbf{y}) \circ W\| < 1$ . Then  $\{\mathbf{y}_t\}_{t \in \mathbb{N}}$  is an aperiodic,  $\psi$ -irreducible,  $T$ -Chain.*

**PROPOSITION. 2.** *Let the conditions of Proposition 1 hold. Assume further that  $\|\boldsymbol{\varepsilon}_t\|^r < \infty$  for some  $r > 0$ , and  $\lim_{\|\mathbf{y}\| \rightarrow \infty} H(\mathbf{y})^{-1} = H_\infty \in \mathbb{R}^{p \times p}$  with  $\|H_\infty\| |\phi| < 1$ . Then  $\{\mathbf{y}_t\}_{t \in \mathbb{N}}$  is geometrically ergodic.*

$\sup_{\mathbf{y} \in \mathbb{R}^N} \|\rho(\boldsymbol{\theta}_0^p; \mathbf{y}) \circ W\| < 1$  in Proposition 1 imposes a stability condition that ensures invertibility and a uniform bound of the spatial multiplier process  $H(\mathbf{y})^{-1}$ . In the Supplementary Appendix, we show that this condition follows when the spectral radius of  $\rho(\boldsymbol{\theta}_0^p; \mathbf{y}) \circ W$  stays strictly below 1 at  $\sup_{\mathbf{y} \in \mathbb{R}^N}$ .  $\|H_\infty\| \|\phi\| < 1$  in Proposition 2 imposes a stricter contraction condition in the time dimension. COROLLARY. 5 makes use of PROPOSITION. 1 and PROPOSITION. 2 to obtain the consistency of the MLE  $\hat{\boldsymbol{\theta}}_T$  with respect to  $\boldsymbol{\theta}_0$ . Note that, this time, the parameter  $\boldsymbol{\theta}_0$  does indeed correspond to the *true parameter* that defines the *true distribution* of the data.

COROLLARY. 5 (Consistency under correct specification). *Let  $\{\mathbf{y}_t\}_{t \in \mathbb{Z}}$  be generated by the ST-SAR model Equation (4.4) under some  $\boldsymbol{\theta}_0 \in \Theta$ . Suppose that Assumptions 7 and 10 hold, and let the conditions of Propositions 1-2 be satisfied. Finally, let  $\Sigma$  be positive definite for every  $\boldsymbol{\theta} \in \Theta$ . Then the MLE satisfies  $\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$  as  $T \rightarrow \infty$ .*

Theorem 10 and Corollary 5 rely on the uniqueness of the maximizer  $\boldsymbol{\theta}_0$ . This assumption may however fail to hold. For example, if the model is misspecified, then several parameter values might provide an equally good approximation to the unknown data generating process in Kullback-Leibler divergence. In particular, we might have a non singleton set

$$\Theta_0^* = \arg \min_{\boldsymbol{\theta} \in \Theta} \mathbb{E} KL \left( p^0(\mathbf{y}_t | \mathbf{y}^{t-1}) , p(\mathbf{y}_t | \mathbf{y}^{t-1}, \boldsymbol{\theta}) \right).$$

where  $\Theta_0^*$  is now the *argmin set* composed of more than one element of  $\Theta$ . Alternatively, if the model is correctly specified, then the uniqueness assumption may fail if the true unknown data generating process is given exactly by a linear SAR since some parameters (e.g.  $\gamma$ ,  $\alpha$  and  $\phi$ ) are unidentified when  $\delta = 0$ . In this case, there exists a set

$$\Theta_0 = \left\{ \boldsymbol{\theta} \in \Theta : KL \left( p^0(\mathbf{y}_t | \mathbf{y}^{t-1}) , p(\mathbf{y}_t | \mathbf{y}^{t-1}, \boldsymbol{\theta}) \right) = 0 \right\}$$

of points that deliver a correct description of the distribution of the data.<sup>7</sup>

<sup>7</sup>Examples of the failure of the uniqueness assumption in other econometric settings can be found

### 4.3.3 Set-consistency of the MLE allowing for possible parameter identification failure

Below, we highlight that if the restrictive uniqueness condition fails, we can still show that the MLE  $\hat{\boldsymbol{\theta}}_T$  converges to the set of maximizers of the limit log likelihood function. A simple regularity condition is required which states that the *level sets* of the limit log likelihood function  $\ell_\infty$  are *regular* (see Definition 4.1 Pötscher and Prucha (1997)). The following theorem is obtained directly by application of Lemma 4.2 in Pötscher and Prucha (1997) to a time-invariant continuous limit criterion  $\mathbb{E}\ell_t : \Theta \rightarrow \mathbb{R}$  defined on a compact parameter space  $\Theta$ . This theorem holds for possibly misspecified models and ensures set consistency of the MLE  $\hat{\boldsymbol{\theta}}_T$  to the set of pseudo-true parameters  $\Theta_0^*$  of our ST-SAR model. Below, we let  $d(\cdot, \cdot)$  denote the usual metric distance from a point to a set, whereby  $d(\boldsymbol{\theta}, \Theta^*) = \inf\{\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|, \boldsymbol{\theta}^* \in \Theta^*\}$  for any  $\boldsymbol{\theta} \in \Theta$  and  $\Theta^* \subseteq \Theta$ .

**THEOREM. 11.** (Set consistency of MLE under possible misspecification and parameter identification failure) *Let Assumptions 7-9 hold and let  $\Theta$  be such that  $\Sigma$  is positive definite for every  $\boldsymbol{\theta} \in \Theta$ . Then the MLE  $\hat{\boldsymbol{\theta}}_T$  is set consistent as  $T \rightarrow \infty$ ,*

$$d(\hat{\boldsymbol{\theta}}_T, \Theta_0^*) \xrightarrow{a.s.} 0 \quad \text{as } T \rightarrow \infty$$

where  $\Theta_0^*$  is the argmin set

$$\Theta_0^* = \arg \min_{\boldsymbol{\theta} \in \Theta} \mathbb{E} \text{KL} \left( p^0(\mathbf{y}_t | \mathbf{y}^{t-1}), p(\mathbf{y}_t | \mathbf{y}^{t-1}, \boldsymbol{\theta}) \right).$$

Theorem 12 obtains the same type of set consistency of the MLE  $\hat{\boldsymbol{\theta}}_T$  applied to the setting of Corollary 5, but this time, it is stated for the case of an overspecified ST-SAR model. This is particularly relevant when the true process in fact linear (SAR). In this case, the MLE is shown to be consistent to the set of true parameters  $\Theta_0 \subseteq \Theta$  that deliver an *equivalent*, *correct* and *exact* description of the distributional properties

---

e.g. in (Freedman and Diaconis, 1982) which addresses a simple location problem with *iid* data and (Kabaila, 1983) in the context of time-series models.

of the data.

**THEOREM. 12.** (Set consistency of MLE under correct specification and parameter identification failure) *Let  $\{\mathbf{y}_t\}_{t \in \mathbb{Z}}$  be generated by the ST-SAR model Equation (4.4). Suppose that Assumption 7 holds, and let the conditions of Propositions 1-2 be satisfied. Finally, let  $\Sigma$  be positive definite for every  $\boldsymbol{\theta} \in \Theta$ . Then the MLE satisfies  $\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \Theta_0$  as  $T \rightarrow \infty$  where  $\Theta_0$  is the set of points deliver an equivalent and correct distribution of the data*

$$\Theta_0 = \left\{ \boldsymbol{\theta} \in \Theta : \text{KL} \left( p^0(\mathbf{y}_t | \mathbf{y}^{t-1}) , p(\mathbf{y}_t | \mathbf{y}^{t-1}, \boldsymbol{\theta}) \right) = 0 \right\}.$$

### 4.3.4 Asymptotic normality of the MLE

**THEOREM. 13** below obtains the asymptotic normality of the MLE. Once again we allow the ST-SAR model to be well specified or misspecified. **ASSUMPTION. 11** assumes that the score is either a martingale difference sequence (mds) or, alternatively, that it is near epoch dependent (NED) of size  $-1$  on an underlying  $\alpha$ -mixing sequence of appropriate size. It is well known that, if the model is well specified, then the score is a martingale difference sequence (mds). As such, we obtain the desired asymptotic normality application of Billingsley's central limit theorem (CLT) for an SE martingale difference sequence (mds); see Billingsley (1961). The mds assumption is also appropriate for mild forms of misspecification; see White (1994). Under strong model misspecification, the asymptotic Gaussianity of the score may still be obtained by application of a central limit for processes that are NED on an  $\alpha$ -mixing process; see e.g. Theorem 10.2 in Pötscher and Prucha (1997). The verification of the NED property can be easily achieved by appealing to preservation theorems such as Theorem 6.6 in (Pötscher and Prucha, 1997), for example in Corollary 6.8 therein it is obtained if the score is Lipschitz on some transformation of the data which is itself NED of the desired size.

In Assumption 11 the  $\alpha$ -mixing sequence is of size  $2r/(r-2)$ , for some

$r > 2$ . As we shall see, we will also require  $r$  bounded moments from the score to obtain a CLT. Finally, we note that the CLT could also be obtained for a  $\phi$ -mixing sequence of size  $r/(r-1)$ .

ASSUMPTION. 11. *The score  $\{\nabla \ell_t(\boldsymbol{\theta}_0)\}_{t \in \mathbb{Z}}$  is either a martingale difference sequence or it is near epoch dependent of size  $-1$  on an underlying  $\alpha$ -mixing sequence of size  $2r/(r-2)$ , for some  $r > 2$ .*

ASSUMPTION. 12 imposes additional moment conditions that ensure the application of a CLT to the score and a uniform law of large numbers to the second derivative of the log likelihood function. Below we let  $\nabla^i Q(\boldsymbol{\theta}_0^\rho; \mathbf{Z}_t)$  and  $\nabla^i F(\boldsymbol{\theta}_0, \mathbf{y}_t, \mathbf{X}_t, \mathbf{Z}_t)$  denote the  $i$ th derivative of  $Q(\boldsymbol{\theta}_0^\rho; \mathbf{Z}_t)$  and  $F(\boldsymbol{\theta}_0, \mathbf{y}_t, \mathbf{X}_t, \mathbf{Z}_t)$  with respect to the vector  $\boldsymbol{\theta}$ . The moment conditions are imposed on each element of the resulting vectors and matrices.

ASSUMPTION. 12. *The following moment conditions are satisfied:*

- i  $\mathbb{E}|\nabla Q(\boldsymbol{\theta}_0^\rho; \mathbf{Z}_t)|^r < \infty$ ;
- ii  $\mathbb{E}|\nabla F(\boldsymbol{\theta}_0, \mathbf{y}_t, \mathbf{X}_t, \mathbf{Z}_t)|^r < \infty$ ;
- iii  $\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |\nabla^2 Q(\boldsymbol{\theta}_0^\rho; \mathbf{Z}_t)| < \infty$ ;
- iv  $\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |\nabla^2 F(\boldsymbol{\theta}_0, \mathbf{y}_t, \mathbf{X}_t, \mathbf{Z}_t)| < \infty$ .

*If the score is an mds, then conditions (i) and (ii) hold with  $r = 2$ . If the score is NED, then  $r$  is the same as in Assumption 11.*

The moment bounds stated in ASSUMPTION. 12, will be satisfied when the data  $\mathbf{y}$  and  $\mathbf{X}$  have bounded moments of appropriate order. Again, just as for the proof of consistency, when the model is correctly specified, bounded moments for  $\mathbf{y}$  and  $\mathbf{X}$  can be obtained by applying the theorem 6.10 in Pötscher and Prucha (1997) to the dynamic model stated in Equation (4.7). In particular, when the contraction condition holds  $H(\boldsymbol{\theta}^\rho; \mathbf{y}_{t-1})^{-1}$  is bounded see LEMMA. 3 in the Appendix, and the ST-SAR is bounded by a linear recursion, and hence,  $m$  moments for  $\mathbf{y}$  can be obtained when  $\mathbf{X}$  and innovations have  $m$  moments.

THEOREM. 13 now delivers the asymptotic Gaussianity of the standardized MLE by imposing the further regularity condition that  $\boldsymbol{\theta}_0$  lies in the interior of the parameter space  $\text{int}(\Theta)$ . This theorem also assumes that  $\boldsymbol{\theta}_0$  is well identified. This is reflected in the invertibility of the limit Hessian  $\mathbb{E}\ell_t''(\boldsymbol{\theta}_0)$ .

THEOREM. 13 (Asymptotic normality of the identified parameters). *Let assumptions 1-6 hold with  $\Sigma$  positive definite for every  $\boldsymbol{\theta} \in \Theta$  and invertible Hessian  $\mathbb{E}\ell_t''(\boldsymbol{\theta}_0)$ . If  $\boldsymbol{\theta}_0 \in \text{int}(\Theta)$ , then the MLE satisfies*

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}(\boldsymbol{\theta}_0) \mathcal{J}(\boldsymbol{\theta}_0) \mathcal{I}^{-1}(\boldsymbol{\theta}_0)) \text{ as } T \rightarrow \infty,$$

where  $\mathcal{J}(\boldsymbol{\theta}_0) := \mathbb{E}\ell_t'(\boldsymbol{\theta}_0)\ell_t'(\boldsymbol{\theta}_0)^\top$  is the expectation of the outer product of the score, and  $\mathcal{I}(\boldsymbol{\theta}_0) := -\mathbb{E}\ell_t''(\boldsymbol{\theta}_0)$  denotes the Fisher information matrix.

As we shall see, the Monte Carlo simulation developed in Section 4.4 provides evidence of both the consistency and normality claims made in THEOREM. 10 and THEOREM. 13 in the correct and misspecified case.

#### 4.3.5 Model selection under possible parameter identification failure

It is well known that threshold parameters are not identified under the null (Teräsvirta et al., 2010). In univariate literature, nonlinearity tests are often based on auxiliary regressions (Dijk et al., 1999). In the ST-SAR, the expansion approach results in many components as nonlinear feedback extends both in space and time. Auxiliary statistics therefore lead to inefficient results. As an alternative, we explore model selection based on information criteria following Granger et al. (1995); Sin and White (1996). We highlight that information criteria consistently rank the models asymptotically according to Kullback-Leibler divergence, even if parameters are unidentified. To address possible other sources of bias, we also provide a theoretical argument for model selection based on a validation-sample estimate of Kullback-Leibler divergence which



is again valid when one or several parameters of the model are not identified. Simulations show support the use of information criteria for model selection when the true process is linear and parameters of the model are unidentified, and when the process is nonlinear and the MLE of identified parameters is in fact well-behaved. These results are shown to be robust to additive outliers and fat-tailed errors.

We conclude this section with details on the model selection adopted in the empirical section of this paper. We will consider both in-sample and out-of-sample model selection criteria. Furthermore, we pay special attention to selection criteria that provide an asymptotically consistent ranking of competing models even in the presence of identification issues.

Let  $L_T(\boldsymbol{\theta}) = \sum_{t=2}^T \ell_t(\boldsymbol{\theta})$  denote the sample log likelihood at  $\boldsymbol{\theta} \in \Theta$ . It is well known that model selection based on the KL divergence can be achieved by selecting the model with highest expected log likelihood  $\mathbb{E}L_T(\boldsymbol{\theta}_0^*)$  evaluated at the best (pseudo-true or true) parameter  $\boldsymbol{\theta}_0^* \in \Theta$ . Unfortunately, the sample log likelihood  $L_T(\hat{\boldsymbol{\theta}}_T)$  that is available in practice is an asymptotically biased estimator of the expected log likelihood  $\mathbb{E}L_T(\boldsymbol{\theta}_0^*)$ . This is easily shown by using a simple quadratic expansion

$$\lim_{T \rightarrow \infty} \mathbb{E} \left( L_T(\hat{\boldsymbol{\theta}}_T) - \mathbb{E}L_T(\boldsymbol{\theta}_0^*) \right) = \lim_{T \rightarrow \infty} \mathbb{E} \sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0^*)' \frac{1}{T} L_T''(\boldsymbol{\theta}_0^*) \sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0^*) \neq 0.$$

Under considerably restrictive conditions, Akaike (1973, 1974) showed originally that for a model with  $k$  parameters,

$$\lim_{T \rightarrow \infty} \mathbb{E} \sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0^*)' \frac{1}{T} L_T''(\boldsymbol{\theta}_0^*) \sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0^*) \approx k,$$

and hence, an asymptotically unbiased estimator of  $\mathbb{E}\ell_t(\boldsymbol{\theta}_0^*)$  is given by  $\frac{1}{T} \sum_{t=2}^T \ell_t(\hat{\boldsymbol{\theta}}_T) - k$ ,

$$\lim_{T \rightarrow \infty} \mathbb{E} \left( \frac{1}{T} \sum_{t=2}^T \ell_t(\hat{\boldsymbol{\theta}}_T) - k \right) = \mathbb{E}\ell_t(\boldsymbol{\theta}_0^*).$$

This follows easily for an asymptotically normal MLE of a correctly specified model since then the information equality holds  $\mathcal{J}(\boldsymbol{\theta}_0) = \mathcal{I}(\boldsymbol{\theta}_0)$ , and hence

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}(\boldsymbol{\theta}_0) \mathcal{J}(\boldsymbol{\theta}_0) \mathcal{I}^{-1}(\boldsymbol{\theta}_0)) = \mathcal{N}(0, \mathcal{I}^{-1}(\boldsymbol{\theta}_0)),$$

which implies that  $\lim_{T \rightarrow \infty} \mathbb{E} \sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0^*)' \frac{1}{T} L_T''(\boldsymbol{\theta}_T^*) \sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0^*) = \text{tr}(I_k) = k$ . Akaike also proposed the well known AIC information criteria based on the unbiased estimator  $\frac{1}{T} \sum_{t=2}^T \ell_t(\hat{\boldsymbol{\theta}}_T) - k$  given by  $\text{AIC} = 2T(k - \frac{1}{T} \sum_{t=2}^T \ell_t(\hat{\boldsymbol{\theta}}_T))$ . Since then, several authors have shown that the AIC can also be used to consistently rank models according to the KL divergence in considerably more general settings (Konishi and Kitagawa, 2008)<sup>8</sup>. The AIC and its variations can be used for consistent in-sample model selection under wider forms of misspecification, for nested or non-nested models, and, most importantly, when test statistics fail, for example because of parameter identification problems; see e.g. Granger et al. (1995); Sin and White (1996); Konishi and Kitagawa (2008).

Importantly, as  $k$  are the parameters of the model and independent from the data generating process being linear or nonlinear, it is easy to see that when the model includes unidentified parameters, they penalize the likelihood and increase the AIC while their contribution to the likelihood can be expected to remain low. The small contribution to the likelihood of unidentified parameters is eventually implied for growing data by the result of THEOREM. 12. The AIC therefore favors dropping unidentified parameters in the same way that it favors, for example, dropping autoregressive lags that do not meaningfully contribute to the implied density of a model. Simply put, in the special case that the data is linear, the linear SAR model and the larger nesting ST-SAR model that includes non-meaningful parameters attain very similar log likelihoods.

---

<sup>8</sup>See pages 61-64 for the Takeuchi Information Criterion. The original reference of Takeuchi 1976 is in Japanese and difficult to find.

At this point, the linear model can be selected on the basis that relative parsimony is favored by the AIC.

For this reason, the use of the AIC for model selection in the context of threshold models has been suggested already by Tong (1983); Li (1988); Tong (1990). Furthermore, Wong and Li (1998) showed that the AICc is an asymptotically unbiased estimator of the expected Kullback–Leibler information for SETAR models and analyzed the finite sample properties of AIC, AICc and BIC by simulation. Theoretical and simulated results on the consistency of information criteria in selecting the lag order of linear autoregressive models have been extended to the case of threshold models by Kapetanios (2001). Finally, Psaradakis et al. (2009) perform an extensive simulation study on the usefulness of the information criteria in selecting between alternative nonlinear time series models and concludes that they are effective even in small samples given that nonlinearity is substantial but that the criteria, particularly the ones with higher penalties, often favor linear models when the data do not have prominent nonlinear characteristics.

REMARK. 2. *We focus on the AIC, the corrected AIC (AICc), and a modified AIC (mAIC). The AICc, introduced by Hurvich and Tsai (1989), improves on the finite sample properties; see Brockwell and Davis (1991); McQuarrie and Tsai (1998); Burnham and Anderson (2004). The mAIC is based on the general setting put forward by Sin and White (1996).*

Unfortunately, specification issues can still influence the in-sample performance of information criteria, for example because the nonlinear model overfits linear data. For this reason, we also consider criteria based on a *validation sample*. In particular, we obtain the sample log likelihood  $\tilde{L}_{\tilde{T}}(\hat{\theta}_T) = \sum_{t=2}^{\tilde{T}} \tilde{\ell}_t(\hat{\theta}_T)$  based on a validation sample of size  $\tilde{T}$ , where  $\hat{\theta}_T$  is obtained using the estimation sample of size  $T$ . The tilde is used in  $\tilde{L}$  to emphasize that this log likelihood is calculated using the validation sample.

Lemma 1 states that, when using an (approximately) independent val-

idation sample, the sample log likelihood  $\tilde{L}_{\tilde{T}}(\hat{\boldsymbol{\theta}}_T)$  is immediately an asymptotically unbiased estimator of  $\mathbb{E}\tilde{L}_{\tilde{T}}(\boldsymbol{\theta}_0^*)$ . This can be shown using the same quadratic expansion argument as used to derive the AIC, and then letting both  $T$  and  $\tilde{T}$  diverge to infinity sequentially. In practice, for time-series data with some form of fading memory (e.g. mixing, near epoch dependence,  $L_p$ -approximability, etc), a burn-in period of  $T^*$  observations between the estimation sample  $\mathbf{y}_1, \dots, \mathbf{y}_T$  and the validation sample  $\mathbf{y}_{T+T^*+1}, \dots, \mathbf{y}_{T+T^*+\tilde{T}}$  is needed to ensure the assumption of *approximate independence* of the validation sample.

REMARK. 3. *Unbiased estimates of the out-of-sample likelihood differential can in principle be cross-validated rather than calculated over a single holdout. However, while the approximate independence of the holdout was trivially satisfied by the use of a burn-in that separates it from the estimation sample, leave-one-out or other repeated validation strategies require a correct-specification assumption on both competing SAR and ST-SAR models in order to maintain the required approximate independence of the residuals or need to implement sophisticated strategies that ensure the independence is satisfied in other ways, see Gao et al. (2016); Bergmeir et al. (2018).*

LEMMA. 1. *Let  $\ell$  be twice continuously differentiable, suppose that  $\hat{\boldsymbol{\theta}}_T \xrightarrow{as} \boldsymbol{\theta}_0^*$  as  $T \rightarrow \infty$  and assume that  $\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |\tilde{L}'_T(\boldsymbol{\theta})| < \infty$  hold. Then  $\tilde{L}_{\tilde{T}}(\hat{\boldsymbol{\theta}}_T)$  is an asymptotically unbiased estimator of  $\mathbb{E}\tilde{L}_{\tilde{T}}(\boldsymbol{\theta}_0^*)$ ,*

$$\lim_{T, \tilde{T} \rightarrow \infty} \mathbb{E} \left( \tilde{L}_{\tilde{T}}(\hat{\boldsymbol{\theta}}_T) - \mathbb{E}\tilde{L}_{\tilde{T}}(\boldsymbol{\theta}_0^*) \right) = 0$$

Lemma 1 tells us that we can rank models consistently according to the KL divergence without the need to impose penalties whose magnitude rely on intricate assumptions. Lemma 2 below highlights that the ranking is consistent regardless of potential identification issues. In particular, it shows that the models are asymptotically well ranked according to the KL divergence even in the case of a set consistent MLE for the parameters of a well-specified or misspecified ST-SAR model.

LEMMA. 2. *Let  $\ell$  be twice continuously differentiable, suppose that*

$d(\hat{\boldsymbol{\theta}}_T, \boldsymbol{\theta}_0^*) \xrightarrow{as} 0$  as  $T \rightarrow \infty$  and assume that  $\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |\tilde{L}'_T(\boldsymbol{\theta})| < \infty$  hold. Then  $\tilde{L}_{\tilde{T}}(\hat{\boldsymbol{\theta}}_T)$  is an asymptotically unbiased estimator of  $\mathbb{E} \tilde{L}_{\tilde{T}}(\boldsymbol{\theta}_0^*)$  for all  $\boldsymbol{\theta}_0^* \in \Theta_0^*$ ,

$$\lim_{T, \tilde{T} \rightarrow \infty} \mathbb{E} \left( \tilde{L}_{\tilde{T}}(\hat{\boldsymbol{\theta}}_T) - \mathbb{E} \tilde{L}_{\tilde{T}}(\boldsymbol{\theta}_0^*) \right) = 0 \quad \forall \boldsymbol{\theta}_0^* \in \Theta_0^*.$$

REMARK. 4. Let the data be generated by a linear SAR model under some  $\boldsymbol{\theta}_0 \in \Theta_0 \subset \Theta$ . Let  $\ell$  be twice continuously differentiable, suppose that  $d(\hat{\boldsymbol{\theta}}_T, \boldsymbol{\theta}_0) \xrightarrow{as} 0$  as  $T \rightarrow \infty$  and assume that  $\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |\tilde{L}'_T(\boldsymbol{\theta})| < \infty$ . Then  $\tilde{L}_{\tilde{T}}(\hat{\boldsymbol{\theta}}_T)$  is an asymptotically unbiased estimator of the expected log likelihood  $\mathbb{E} \tilde{L}_{\tilde{T}}(\boldsymbol{\theta}_0)$  at the true parameter, i.e.  $\lim_{T, \tilde{T} \rightarrow \infty} \mathbb{E}(\tilde{L}_{\tilde{T}}(\hat{\boldsymbol{\theta}}_T) - \mathbb{E} \tilde{L}_{\tilde{T}}(\boldsymbol{\theta}_0)) = 0$ .

In the special case that the linear SAR model is correctly specified, Remark 4 tells us that the linear SAR model will attain the same zero KL divergence as any larger nesting ST-SAR model. The same holds true for any other model that nests the SAR, as the larger model is also correctly specified. At this point, the linear model can be selected on the basis of being the most parsimonious model that is correctly specified.

In practice, a situation of this type will lead to very similar log likelihoods for the competing models over the validation sample. In Proposition 3 we highlight that the differences in these log likelihood values can be tested for statistical significance using the Diebold-Mariano test statistic (Diebold and Mariano, 1995). Specifically, we can test the rank position of any two models by testing if the difference in log likelihoods in the validation sample is statistically significant or not. This test is also known as a logarithmic scoring rule, see e.g. Diks et al. (2011); Amisano and Giacomini (2007); Bao et al. (2007). Below, we consider two competing models, A and B, and let  $\tilde{\ell}_t^A(\hat{\boldsymbol{\theta}}_T^A)$  and  $\tilde{\ell}_t^B(\hat{\boldsymbol{\theta}}_T^B)$  denote their respective log likelihood contributions at a certain time  $T + T^* + 1 < t \leq T + T^* + \tilde{T}$  in the validation sample. Furthermore, we let  $\tilde{\Delta}_t(\hat{\boldsymbol{\theta}}_T^A, \hat{\boldsymbol{\theta}}_T^B)$  denote the log likelihood differences  $\tilde{\Delta}_t(\hat{\boldsymbol{\theta}}_T^A, \hat{\boldsymbol{\theta}}_T^B) := \tilde{\ell}_t^A(\hat{\boldsymbol{\theta}}_T^A) - \tilde{\ell}_t^B(\hat{\boldsymbol{\theta}}_T^B)$  evaluated at the point estimates  $\hat{\boldsymbol{\theta}}_T^A$  and  $\hat{\boldsymbol{\theta}}_T^B$  respectively, and  $\tilde{\Delta}_t(\boldsymbol{\theta}_0^{*A}, \boldsymbol{\theta}_0^{*B})$  denote the log

likelihood differences evaluated at each model's pseudo-true parameter. Finally, we let  $\tilde{\sigma}_{\tilde{T}}(\hat{\boldsymbol{\theta}}_T^A, \hat{\boldsymbol{\theta}}_T^B)$  be a consistent estimator of the standard deviation of  $\tilde{\Delta}_t(\hat{\boldsymbol{\theta}}_T^A, \hat{\boldsymbol{\theta}}_T^B)$ .

PROPOSITION. 3. (Diebold-Mariano test statistic: logarithmic scoring rule) *Let  $\hat{\boldsymbol{\theta}}_T^A \xrightarrow{as} \boldsymbol{\theta}_0^{*A}$  and  $\hat{\boldsymbol{\theta}}_T^B \xrightarrow{as} \boldsymbol{\theta}_0^{*B}$  as  $T \rightarrow \infty$ . Suppose that the data is strictly stationary and ergodic. Then under the null hypothesis that model A and B fit the data equally well  $H_0 : \mathbb{E}\tilde{\Delta}_t(\boldsymbol{\theta}_0^{*A}, \boldsymbol{\theta}_0^{*B}) = 0$ , it follows that*

$$DM_{\tilde{T},T} := \tilde{T}^{\frac{1}{2}} \left( \tilde{T}^{-1} \sum_{t=T+\tilde{T}^*+1}^{\tilde{T}} \frac{\tilde{\Delta}_t(\hat{\boldsymbol{\theta}}_T^A, \hat{\boldsymbol{\theta}}_T^B)}{\tilde{\sigma}_{\tilde{T}}(\hat{\boldsymbol{\theta}}_T^A, \hat{\boldsymbol{\theta}}_T^B)} \right) \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } T, \tilde{T} \rightarrow \infty.$$

*If instead we have  $\mathbb{E}\tilde{\Delta}_t(\boldsymbol{\theta}_0^{*A}, \boldsymbol{\theta}_0^{*B}) > 0$  (model A is best) then  $DM_{\tilde{T},T} \rightarrow \infty$  as  $T, \tilde{T} \rightarrow \infty$ . Finally, if  $\mathbb{E}\tilde{\Delta}_t(\boldsymbol{\theta}_0^{*A}, \boldsymbol{\theta}_0^{*B}) < 0$  (model B is best) then  $DM_{\tilde{T},T} \rightarrow -\infty$ .*

REMARK. 5. *In Section 4.5, a more conservative finite sample correction of the statistic following a Student's-t distribution is also used, see Harvey et al. (1997).*

Just as the AIC, out-of-sample model performance evaluation has been applied in the context of threshold models in earlier literature. See for example Clements et al. (2003) who investigates out-of-sample comparison of the Mean Squared Error and concludes that, in line with to the conclusions around the use of the AIC detailed by Psaradakis et al. (2009), data need to exhibit a substantial degree of non-linearity before the SETAR model is favored over a linear model. For these reasons, we can expect both approaches to favor the ST-SAR only when the true nonlinearity is strong in the data. This is a useful feature because we would only want to accept the alternative assumption of nonlinearity over the null assumption of linearity in an empirical application if the evidence is substantial. Finally, it is important to stress that the DM-type test developed here imposes assumptions directly on the forecast errors, in particular that the likelihood differential is covariance stationary, and can therefore work in the case of unidentified parameters or even in a

model-free environments, see Diebold (2015) for reflections on this.

## 4.4 Monte Carlo study

To evaluate the empirical relevance of our estimation theory, we conduct a Monte Carlo study. Importantly, we investigate size and power of model selection based on standard information criteria. Foremost, we explore how well popular information criteria are able to distinguish between linearity and nonlinearity when the data is generated by a linear model and the ST-SAR contains unidentified nuisance parameters. We also explore how well the criteria recognize the nonlinear features of data when the true process is nonlinear.

In the following numerical investigation we focus on selection frequencies based on standard information criteria. Recall that evidence exists that information criteria perform well in small samples in the context of univariate threshold models when nonlinearity is strong but favor linear models when nonlinearity is weak (Psaradakis et al., 2009). For this reason, we simulate from a linear model to explore how well the information criteria perform when the data is linear, and only simulate from a relatively flat nonlinear dependence signal when we explore the suitability of information criteria to detect nonlinearity when the data indeed is nonlinear. The data generating process is of the general form:

$$\mathbf{y}_t = H(\boldsymbol{\theta}^\rho; \mathbf{y}_{t-1})^{-1}(\boldsymbol{\varepsilon}_t), \quad \boldsymbol{\varepsilon}_t \sim TID(1, I_N; 5), \quad (4.10)$$

We keep the ratio of distant and close-by neighbors comparable across experiments by allowing the network density of the weights matrix to increase with  $N$ . In each draw we generate a random zero diagonal row-normalized weights matrix with  $N/10$  neighbors for each observation. The process is initialized with  $H_1 = I_N$ , and the first 50 steps of the sequence are discarded to avoid dependence on the initialization. We simulate

1000 datasets and estimate parameters with Student's- $t$  likelihood.

We focus on an ST-SAR process driven by local averages as the local average should be more sensitive to additive outliers than, say, the cross-sectional mean. We simulate the linear datasets according to a linear SAR process with:

$$\rho = 0.5$$

and simulate the nonlinear data sets according to the nonlinear ST-SAR process:

$$\delta = .4, \gamma = 1.05, \alpha = -.2, \varphi = 1.4, \kappa = -.4,$$

$$\mathbf{Z}_t = \mathbf{y}_{t-1}, \tau(\boldsymbol{\theta}^\tau; \mathbf{Z}_t) = \alpha + \varphi W \mathbf{y}_{t-1}.$$

We also consider the effect of additive outliers, similar to Dijk et al. (1999), by simulating contaminated sequences (+ AO) according to the following replacement process:

$$\mathbf{y}_t^* = \mathbf{y}_t + 1. [\zeta_t > 0.5] \psi \boldsymbol{\epsilon}_t, \quad (4.11)$$

$$\{\zeta_t\} \sim UID(0, I_N), \quad \{\boldsymbol{\epsilon}_t\} \sim BID(-I_N, I_N; \pi),$$

with  $\pi = 0.05$  and  $\psi$  set to the sample equivalents of  $\sqrt{\mathbb{E}\mathbf{y}_t^2 - (\mathbb{E}\mathbf{y}_t)^2}$ , and estimating on  $\mathbf{y}_t^*$ .

In table 4.1, we pit the results of the ST-SAR with all its parameters (ST-SAR 2) against SAR estimates and focus on selection between the SAR and the ST-SAR when the process is linear (Size). The selection frequencies are also provided for contaminated data generated from the SAR (right). Both SAR and ST-SAR model are correctly specified with regard to the non-contaminated process, but the SAR is more parsimonious while the ST-SAR has additional parameters to over fit the data and possibly the outliers. The results indicate information criteria can be used to distinguish between linearity and nonlinearity with performance improving as the dimensions of the data grow.



Table 4.1: Size: selection frequencies for (contaminated) data generated from the SAR (right). The results indicate information criteria can be used to distinguish between linearity and nonlinearity with performance improving as the dimensions of the data grow. The ST-SAR is robust to over fitting outliers.

SAR DGP (AO right columns)		ST-SAR 2 vs. SAR			ST-SAR 2 vs. SAR		
		AIC	AICc	mAIC	AIC	AICc	mAIC
N=30	T=10	46	44	45	44	42	44
	T=25	33	32	32	30	29	30
	T=50	27	27	27	26	25	26
	T=100	22	22	22	23	23	23
	T=250	22	22	22	20	20	20
N=40	T=10	52	51	51	54	52	53
	T=25	32	31	31	33	33	33
	T=50	24	24	24	25	25	25
	T=100	22	22	22	23	23	23
	T=250	23	23	23	18	18	18
N=50	T=10	41	39	40	43	42	42
	T=25	25	25	25	27	26	26
	T=50	24	24	24	23	23	23
	T=100	20	19	19	19	19	19
	T=250	18	18	18	18	18	18
N=60	T=10	32	31	32	33	32	32
	T=25	24	23	24	27	27	27
	T=50	24	23	24	26	26	26
	T=100	17	17	17	16	16	16
	T=250	17	17	17	17	17	17

The results of table 4.1 show that the AIC has empirically relevant size. We have discussed that the ST-SAR converges to the set of points that deliver an equivalent and correct distribution of the data, see THEOREM. 12 and that the SAR should thus be selected on the basis of parsimony. The simulation evidence is in support of this notion. For data simulated from the linear SAR, we see that the SAR is indeed selected over the ST-SAR 2 with increasing frequency as the sample size increases. However, as data grows, the larger ST-SAR 2 is still incorrectly selected over the nested SAR with nonzero frequency. At  $T = 250$ ,  $N = 60$  we select the ST-SAR 2 in 17% of the cases. This suggests that in practice, one may want the improvement in AICc to be relatively large or prefer to keep

Table 4.2: Power: selection frequencies for data generated from the ST-SAR. The results indicate information criteria can be used to distinguish between linearity and nonlinearity with performance improving as the dimensions of the data grow.

ST-SAR DGP		ST-SAR 1 vs. SAR			ST-SAR 2 vs. SAR			ST-SAR 2 vs. ST-SAR 1		
		AIC	AICc	mAIC	AIC	AICc	mAIC	AIC	AICc	mAIC
N=30	T=10	38	38	38	45	41	43	46	45	46
	T=25	62	61	62	63	62	63	50	48	49
	T=50	80	79	80	83	82	82	57	57	57
	T=100	85	85	85	97	97	97	80	80	80
	T=250	100	100	100	100	100	100	96	96	96
N=40	T=10	51	49	50	52	50	51	44	43	44
	T=25	72	72	72	73	72	72	51	50	50
	T=50	93	93	93	91	91	91	59	59	59
	T=100	92	92	92	100	100	100	84	84	84
	T=250	100	100	100	100	100	100	99	99	99
N=50	T=10	53	52	53	54	52	53	45	43	45
	T=25	84	84	84	85	84	85	55	55	55
	T=50	98	98	98	98	98	98	66	66	66
	T=100	99	99	99	100	100	100	89	89	89
	T=250	100	100	100	100	100	100	100	100	100
N=60	T=10	63	62	63	59	58	59	45	43	44
	T=25	88	88	88	87	87	87	57	56	57
	T=50	99	99	99	99	99	99	71	71	71
	T=100	99	99	99	100	100	100	92	92	92
	T=250	100	100	100	100	100	100	100	100	100

the SAR when the improvement is modest and the data is small. In our empirical applications we find, however, very substantial improvements in the AICc while working with considerable numbers of observations. In our first empirical application we shall focus on  $T$  close to 10 but use a cross-section that is roughly 12 times that of the largest experiment covered by our simulations, while in our second application,  $T$  increases beyond what is considered here. The robustness to contamination of the process can again be seen, this time by the fact that selection rates of the ST-SAR do not inflate when additive outliers enter the process.

In table 4.2 we estimate two versions of the ST-SAR; a restricted model that is underspecified –  $\varphi$  and  $\kappa$  are fixed at 0 – (ST-SAR 1) and the correctly specified ST-SAR with all its parameters (ST-SAR 2). As before, we also estimate the SAR. Again, we find evidence that selection

frequencies, now for data generated from the ST-SAR, support the use of information criteria to distinguish between linearity and nonlinearity. In particular, while table 4.1 highlighted the ability of information criteria to correctly favor the SAR when the data is linear, table 4.2 highlights that the criteria favor the ST-SAR when the data is nonlinear. As with size, power improves as the dimensions of the data grow.

The results of table 4.2 show that the AIC has good power if the process is nonlinear. Both the misspecified ST-SAR 1 and correctly specified 2 are selected over the underspecified SAR with increasing frequency as the sample size increases. Furthermore, as data grows the larger and correct ST-SAR 2 is selected over the nested ST-SAR 1 with probability 1. We again see improvements both as  $T$  and  $N$  increase. Table 4.7 in the Appendix provides additional power results for a contaminated process. Overall, the presence of additive outliers has a small effect on power. For very small samples  $T = 10, N \leq 60$ , we observe some increase in power indicating slightly increased over fitting. However, for  $T > 10, N \leq 60$ , the outliers negatively impact power. While we reach a frequency of 92% for  $(N, T) = (60, 100)$  without contamination, we obtain only a rate of 80% for distorted data. The reduction in power contrasts the univariate STAR framework in which additive outliers can trick the threshold into fitting the contamination as a nonlinear process (Dijk et al., 1999). We find that in the cross-sectional case, the results mirrors the conclusions of the errors in variables literature. Finally, note that, as in the distribution case, the results are dependent on the strength of the nonlinear signal. In our empirical application we find strong nonlinearities.

The simulations presented here confirm the appropriateness of standard information criteria to decide between different descriptions of spatial spillover processes. Importantly, the evidence indicates that that, not only do information criteria distinguish well between linearity and nonlinearity, they also distinguish between alternative nonlinearities. The AICc comes

forward as the most conservative measure, and therefore we apply it as our primary choice criterion in the empirical section. Additional simulation results in the Appendix, fig. 4.7 in particular, further highlight that the MLE of the identified parameters is well-behaved in empirically relevant sample sizes.

## 4.5 The empirics of nonlinear spatial dependencies

This section presents two empirical cases. In our first study we use a panel of short  $T$  and large  $N$ . Our second study focuses on the opposite case of large  $T$  and small  $N$ . This allows us to explore nonlinearities both from a cross-sectional perspective, as well as from a time-varying perspective.

### 4.5.1 Application I: Dutch residential densities

The first application evaluates nonlinear spatial dynamics in the clustering of Dutch residential densities at the district level over a period of ten years. The primary focus is on the advantages of the ST-SAR compared to its linear counterpart. We investigate spatially varying features of the dependence structure, particularly in relation to a number of spatially explicit socio-economic variables. Steering urban development and preserving open, green spaces is a major policy concern in the Netherlands Koomen et al. (2008). Understanding the drivers influencing the balance between agglomeration and dispersion is essential to help define policies. These policies have a strong spatial dimension, which can be difficult to disentangle. Panel and cross-sectional methods are essential analysis tools, and we shall focus on the role of cross-sectional nonlinearities in obtaining accurate estimates.

**Economic rationale for ST-SAR dynamics in residential densities**

The dependent variable is urban density measured as addresses per hectare. We investigate two types of nonlinearities. First, we model nonlinear spatial autocorrelation to allow for differential strength in clustering. In line with the decay in agglomeration forces along the urban gradient (Fotheringham, 1981; Rosenthal and Strange, 2003), we expect autoregressive spatial dependence to fluctuate along clusters of population densities. The linearity of the SAR on the other hand assumes away any variation in autocorrelation along the urban gradient. The second nonlinearity is in the relationship between local densities and the surrounding household composition. This choice is particularly interesting because dense urban centers accommodate different households than spacious low density neighborhoods. Literature on sorting has made empirically tested predictions about the equilibrium distribution of household types across different neighborhoods (Epple and Sieg, 1999). The demand patterns for housing rooted in preference heterogeneity produces a heterogeneous relationship between concentrations in density and household composition. We focus particularly on the share of population under 14 years in surrounding areas, which proxies a mixture of social and demographic characteristics. As households with children locate in low density neighborhoods outside the city center, we can expect that dense urban cores have a positive correlation with the presence of children in surrounding areas. On the other hand, the low density areas outside main urban cores follow the inverse. A linear spatial lag forces the two opposite relationships to average out, which falsely leads to the conclusion that surrounding households are not related to urban densities, contradicting the sorting theory (Epple and Sieg, 1999). The ST-SAR specification allows us to capture the theorized positive and negative relationships simultaneously.

### Data for Dutch residential densities

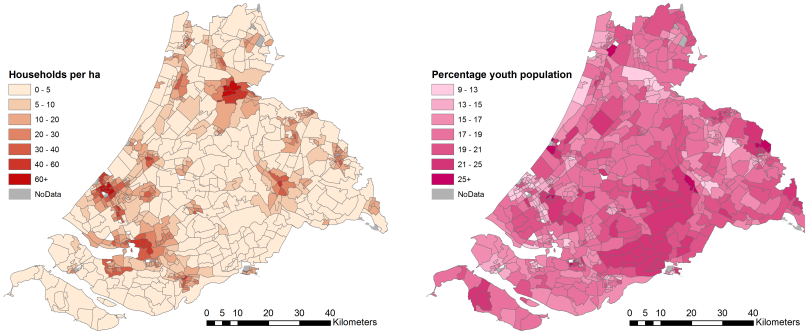


Figure 4.1: Time average spatial distribution of (left) household density and (right) population under 14.

Table 4.3: Overview of explanatory variables and parameter symbols.

Parameter	Interacting variable	Units	Range	Mean
$\beta_{cdens}$	Log company density	Continuous	-3.93 to 3.42	-0.14
$\beta_{wcdens}$	Spatial log company densities	Continuous	-2.62 to 2.30	-0.21
$\beta_{\%shh}$	Percentage of single households	Continuous	5 to 75.22	32.47
$\beta_{w\%shh}$	Spatial percentage single households	Continuous	0 to 59.75	32.15
$\beta_{w\%hhkids}$	Spatial percentage households with children	Continuous	0 to 24.11	8.75
$\beta_{w\%wim}$	Spatial percentage western immigrants	Continuous	0 to 59.03	36.78
$\beta_{\%nwim}$	Percentage non-western immigrants	Continuous	0 to 67.92	9.90
$\beta_{\%>65}$	Percentage elderly over 65	Continuous	1 to 43.23	15.09
$\beta_{w\%<14}$	Spatial percentage children	Continuous	0 to 25.38	17.81
$\rho$	Second order queen contiguity matrix	Standardized	0 to 46*	18.91**

Transition function parameters are indexed by the variables they interact with. \*The range of the spatial weights matrix is the minimum to maximum number of connections. \*\*Average number of connections.

The time series covers observations of 717 districts from 2005 to 2014 obtained from the Dutch Central Bureaus of Statistics.<sup>9</sup> Figure 4.1 shows the concentrations of urban densities and young population outside urban areas. The other regressors, that control for a variety of local demographic and economic characteristics, are taken from the same dataset. Local

<sup>9</sup>The data is available for download from the Dutch Central Bureau of Statistics: <https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data>.

values and spatial averages have been selected based on the AICc. The regressors are lagged by one time period.

### Results for Dutch residential densities

Table 4.4 presents the results. The static estimates provide strong evidence for clustering in household densities indicated by the high estimate of  $\rho$  and the high  $t$ -value. As theorized, we find weak evidence for a relationship with the surrounding household compositions indicated by the small estimate of  $\beta_{w\%14}$  and its low  $t$ -value. Household densities are strongly linked to company densities, other controls have dubious signs. The negative effect of single person households is not as expected as small households should consume little space.

The second model allows for smooth transition nonlinearities in the dependence on surrounding households. The negative value of the constant exogenous spatial lag ( $\beta_{w\%14-t-1}$ ) combined with the positive value of the upper threshold parameter ( $\delta_{w\%14-t-1}$ ) indicates that the dependencies run from negative to positive as densities increase, in line with the theory. The parameters of the transition function strongly improve the AICc (by -11939 points). The nonlinear model also improves the estimates of the control variables, both local and surrounding single person households now correlate positively with densities. The effect of company densities is substantially smaller in magnitude, indicating that the impact may easily be overestimated by the SAR. The spatial autocorrelation parameter is significant but reduced drastically in magnitude. This suggests that the nonlinearities in the relationship with spatial averages may also partially capture nonlinear spatial autocorrelations.

Model (3) controls for additional nonlinear spatial autocorrelation, further improving the AICc (-1799 points). The maps in fig. 4.2 show that spatial autocorrelation is high in the urban clusters and decays outwards, in line with theory.

Table 4.4: Estimation results for Dutch residential densities from 2005-2014. Significance at 90, 95 and 99% level are, respectively, indicated as \*, \*\* and \*\*\*.  $t$ -values in parenthesis.

	(1) SAR + WX	(2) SAR + ST-WX	(3) ST-SAR + ST-WX
$\beta_{const}$	2.309*** (26.479)	0.970*** (36.626)	0.974*** (41.905)
$\beta_{cdens_{t-1}}$	1.043*** (201.116)	0.108*** (21.832)	0.109*** (28.378)
$\beta_{wcdens_{t-1}}$	-0.825*** (-58.556)	-0.101*** (-13.821)	-0.147*** (-31.429)
$\beta_{\%shh_{t-1}}$	-0.006*** (-9.319)	0.009*** (34.352)	0.009*** (38.678)
$\beta_{w\%shh_{t-1}}$	-0.025*** (-17.731)	0.013*** (23.705)	0.009*** (19.950)
$\beta_{w\%hhkids_{t-1}}$	-0.032*** (-12.350)	0.018*** (17.606)	0.015*** (17.875)
$\beta_{\%nwim_{t-1}}$	0.019*** (35.607)	0.003*** (12.336)	0.001*** (3.162)
$\beta_{w\%14-t-1}$	0.009* (1.987)	-0.432*** (-27.409)	-0.385*** (29.141)
$\delta_{w\%14-t-1}$		1.008	0.540
$\gamma_{w\%14-t-1}$		0.209	0.362
$\phi_{w\%14-t-1}$		1.742	0.276
$\delta_\rho$	0.779***(69.605)	0.048***(7.106)	0.368
$\gamma_\rho$			1.235
$\phi_\rho$			1.358
$\lambda$	3.013	2.508	2.559
$LL$	-2078.771	3893.733	4795.443
$AICc$	4179.58	-7759.405	-9558.810

The test proposed in Proposition 3, is valid only for large  $\tilde{T}$ . However, the AICc provides ample evidence supporting the nonlinearities. In particular, the AICc improves by 13738.39 points when the nonlinearities are allowed in both the spatial lags of the exogenous and endogenous regressors. To understand how the ST-SAR improves this much, we re-fitted the models excluding the last year and compared to 1-step ahead forecast errors of the SAR+WX and the ST-SAR+ST-WX. Figure 4.3 shows that the Squared Forecast Errors (SFE) from the linear model contain a consistent mismatch in major urban areas. The SFE of the nonlinear model, however, balance evenly. This shows that the nonlinear model is better at fitting both rural and urban density regimes within one framework. Apart from the clustering of prediction errors, the predictive



power across all regions is tremendously improved by the nonlinear model as seen by the magnitude of the SFE.

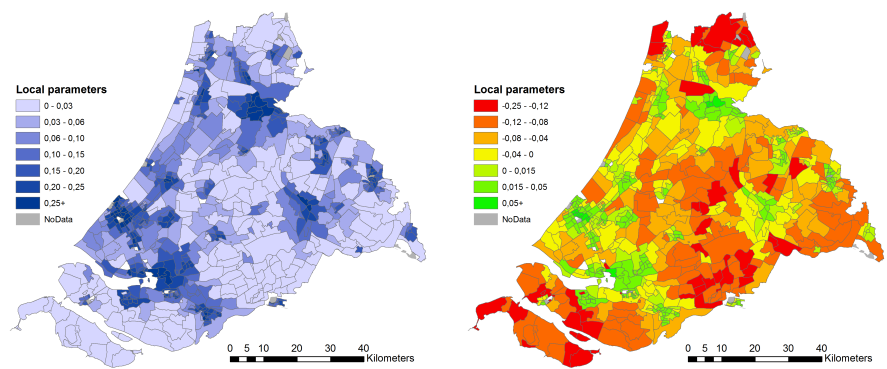


Figure 4.2: Left): time average of estimated autocorrelation parameters. Right): time average of estimated dependence on the share on population under 14 years in surrounding neighborhoods. The estimation results provide convincing evidence for weak/negative and strong/positive dependence regimes with smooth transitions in between.

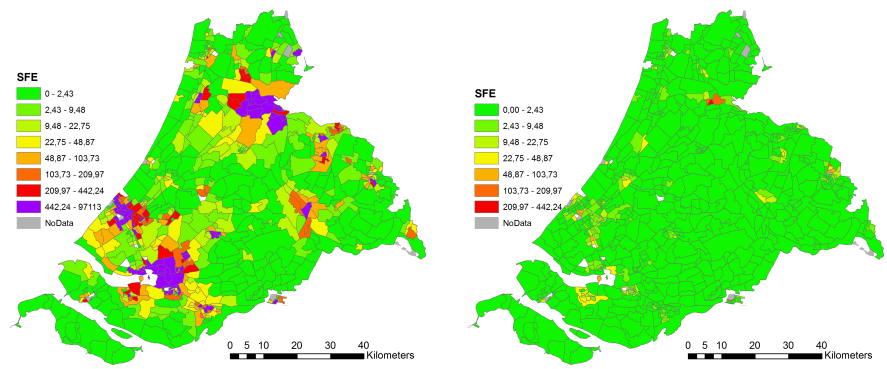


Figure 4.3: Left): SFE of the SAR (2014 as holdout data). Right): SFE of the ST-SAR. Legends are based on natural breaks of the errors of the ST-SAR. The residuals provide convincing evidence for the ability of the ST-SAR to neutralize residual clustering while the SAR does not perform well in this regard. The reduction in SFE also suggests that the ST-SAR provides better 1-step ahead forecasts.

### 4.5.2 Application II: interest rates in the Euro region

In this second empirical study we evaluate the evolution of monthly interest rates on government bonds maturing in ten years for 15 European sovereigns. The study tracks the European sovereigns over a period of 26 years, that spans the time before the European Union, the expansion of the EU, the Great Recession, and the Greek sovereign debt crisis. The primary focus is on detailing time-varying dynamics in convergence and dispersion in rates that cannot be fitted by a linear model. this application differs from the previous one in the sense that the temporal dimension is much larger. Again, we find strong evidence that favors the ST-SAR over the SAR.

#### **Economic rationale for ST-SAR dynamics in long term interest rates**

The Economic and Monetary Union (EMU) comprises a set of policies that aims at converging the economies of the member states of the European Union. The EMU prescribes euro convergence criteria, the prerequisites for a nation to join the Eurozone. Co-movement in the long term interest rates is essential to the monetary stability of the Euro region. Before the European Union, the European Economic Community relied heavily on the European Exchange Rate Mechanism (ERM) to regulate variability in exchange rates of different sovereigns as a way to achieve monetary stability. The ERM played a central role in the preparations for the Economic and Monetary Union and the subsequent introduction of the euro in 1999. The primary goal of the ERM has been to prevent large fluctuations in currency values relative to those of other European sovereigns. Empirical evidence suggests that only few, large industrial countries have some ability to choose their interest rates (Frankel et al., 2004). Interest rates are strongly affected by those of other countries (Frankel et al., 2004; Caceres et al., 2016; Kharroubi et al., 2016), but there are policy opportunities to adjust national rates. For

example, target levels and transfers of reserves (Pina, 2017), programs that increase foreign bond buying (Carvalho and Fidora, 2015), or lending rate policies (von Borstel et al., 2016). Adjustments in national interest rates have been at the center of monetary policy used as part of the European Monetary System (EMS) to lower or increase currency value such that the different currencies remained within a narrow range of one another.

Replacement of the actual currencies of all participating member states by a common currency mandates that the economies of all member states are in par with one another. After introduction of the euro, national interest rates thus still play an essential role in ensuring that fluctuations in the economies of member states remain within a narrow range. A strong adjustment in long term interest rate of a particular sovereign with respect to the common European average, signals that the underlying economy has difficulty in following the common trend. On the other hand, if all interest rates closely follow a common stochastic trend, it signals that economies are in par with one another. This can also be understood in the conventional framework where fixed or pegged interest rates are seen as a way to establish a credible nominal anchor for monetary policy, while flexible exchange rates are seen as a way to allow countries to pursue independent monetary policy (Frankel et al., 2004). Integration of financial systems and co-movements are further discussed by Caceres et al. (2016).

The cross-sectional dependencies in the de-trended changes signal the strength of commonalities in the fluctuations in the economies of member states such as in Caceres et al. (2016). Estimating spatial dependence parameters using ST-SAR has the obvious advantage that it does not only provide information on the average strength in co-movement, but it allows to study also the time-varying features in strength as well as heterogeneity across member states. The average cross-sectional averages

of the dependence parameters signal overall strength of convergence, while the standard deviations indicate an overall dispersion measure that is independent of the scale of change. In stable times, the average contraction should be high and the variance in spatial parameters should be low. Under financial instability we may expect the opposite. The parameters of the ST-SAR therefore not only provide means to filter dynamic dependencies, but also provide information on the functioning of the EMS in this specific application. Specifically, the ST-SAR provides a way to analyze whether the economies of member states are relatively in par with one another, as prescribed by the EMU's common currency mandates.

To do so, we view the interest rates as generated by the model:

$$\mathbf{v}_t = c_t + \mathbf{y}_t,$$

where  $\mathbf{v}_t$  is the observed data vector,  $c_t$  is the common stochastic trend, and  $\mathbf{y}_t$  is a vector of dynamics around the common stochastic trend. We are interested in analyzing  $\mathbf{y}_t$ , which contains the contraction and dispersion dynamics around the common stochastics. By de-trending using a common stochastic trend, synchronization due to common business cycles or seasonality is controlled for. We assume  $c_t$  to follow a random walk with  $c_t = c_{t-1} + v_t$ , and  $\{v_t\}_{t \in \mathbb{Z}} \sim p_v(v_t, \Sigma, \lambda)$ . Therefore our best expectation of  $c_t$  is  $c_t \sim \mathbb{E}^N(\mathbf{v}_t | \mathbf{v}_{t-1})$ , and the dynamics of particular interest are:

$$\mathbf{y}_t = \mathbf{v}_t - \mathbb{E}^N(\mathbf{v}_t | \mathbf{v}_{t-1}) = \mathbf{v}_t - \mathbb{E}^N(\mathbf{v}_{t-1}) \sim \mathbf{v}_t - N^{-1} \sum_1^N (\mathbf{v}_{t-1}),$$

hence we use  $\mathbf{y}_t = \mathbf{v}_t - N^{-1} \sum_1^N (\mathbf{v}_{t-1})$  as our dependent variable. We refer to  $\mathbf{y}_t$  as the de-trended data. We are interested in a description of the convergence and dispersion dynamics contained in  $\mathbf{y}_t$  as a nonlinear cross-sectional dependence process, possibly driven by the past states of  $\mathbf{y}_t$  and moving average affects.

## Data for long term interest rates across the Euro region

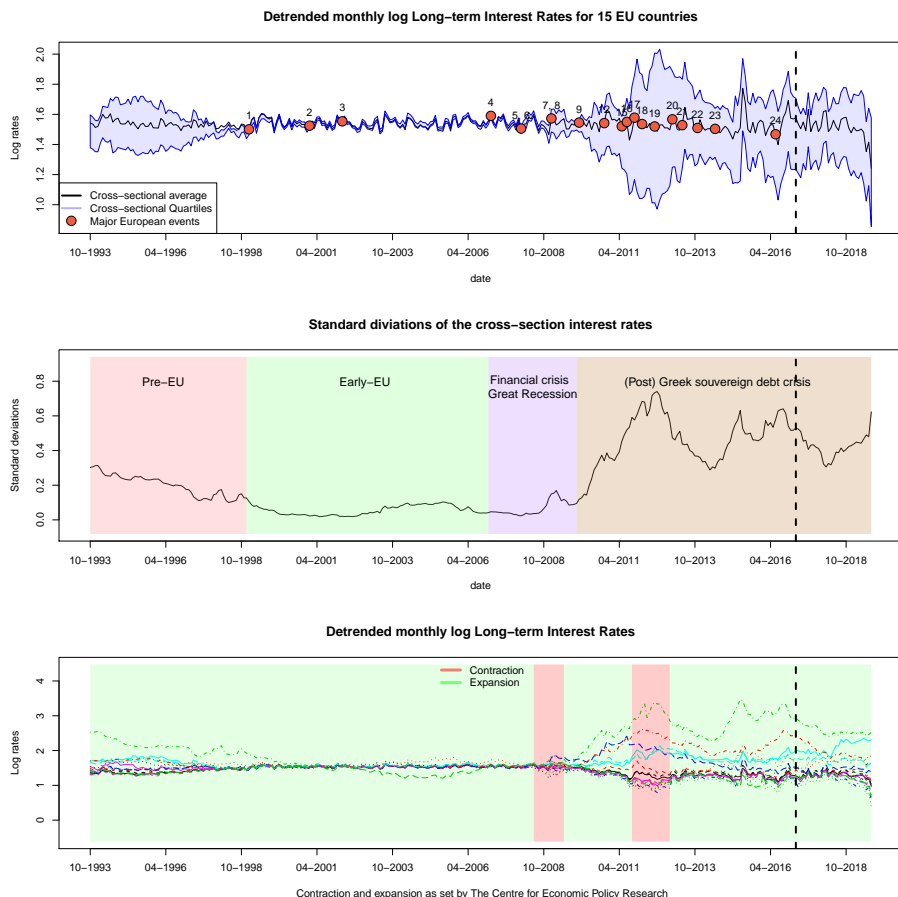


Figure 4.4: Data on monthly long term interest rates for bonds of 10-year maturity. An overview of te labeled events is contained in the Supplementary Appendix. The vertical dashed line indicates the split between training data and validation data used for our DM tests. Colors correspond to individual countries, see section 4.7.4.

The data was obtained from the European Central Bank for 311 months starting October 1993 and running through August 2019.<sup>10</sup> This period includes the formation of the European Union, its expansion, the Great Recession and the eventual Greek sovereign debt crisis. We model log

<sup>10</sup><http://sdw.ecb.europa.eu/browseChart.do?node=bbn4864>

de-trended rates. The de-trended data is visualized in fig. 4.4, the raw data including a list of labeled events and color codes is provided in the Supplementary Appendix. The time series reveal clear common patterns, especially between 1998 and 2008. Before 1998 and after 2008 there are commonalities but specifically the stressed Eurozone sovereigns (Greece, Portugal, Ireland and to some extent Spain and Italy), seem to follow a separate pattern. Our network structure is based on the correlation matrix of the de-trended data. We assign each sovereign three neighbors based on the strongest correlation. This number was determined by the AICc. The approach allows for differences in the centrality of the sovereigns within the network, and for entanglement between sovereigns that are distant from each other in a purely geographic sense. The resulting network is fully connected. We explore time lags up to order 4, and apply further restrictions guided by the AICc. As we shall see in our final model, allowing 4 lags in the ST-SAR is sufficient to render the residuals approximately free from correlations.

### Results for European long term interest rates

As a first exploration we regress models of the type:<sup>11</sup>

$$\mathbf{v}_t - N^{-1} \sum_1^N (\mathbf{v}_{t-1}) = \mathbf{y}_t = H(\boldsymbol{\theta}^\rho; (\mathbf{y}_t, \boldsymbol{\varepsilon}_t))^{-1}(\boldsymbol{\varepsilon}_t).$$

on the entire dataset. We calculate  $DM$  and  $mDM$ , respectively one-sided  $\Pr(>|z|)$  and  $\Pr(>|t|)$  against the null hypothesis that the SAR attains higher log likelihood, based on model fits on training data log likelihood evaluated on the validation sample depicted in fig. 4.4. We reserved the final 36 observations for this validation purpose, of which the first 6 observations are discarded as a burn-in.

---

<sup>11</sup>The exact threshold  $\rho(\boldsymbol{\theta}^\rho; \mathbf{y}_t, \boldsymbol{\varepsilon}_t) = \frac{\delta}{1 + \exp(-\gamma(W\mathbf{y}_{t-1} - (\alpha + \sum_{p=1}^P \mathbf{y}_{t-p}\varphi_{\phi,p} + \sum_{q=1}^Q \boldsymbol{\varepsilon}_{t-q}\varphi_{\mu,q})))} + \kappa$ .

Table 4.5: Estimation results for the different spatial models on the full dataset.  $\lambda$  fixed at 2.5.

	SAR	ST-SAR (AR)	ST-SAR (MA)	ST-SAR (ARMA)
$\beta_{const}$	0.608*** (56.670)	-0.234*** (-30.703)	0.310*** (37.294)	0.273*** (34.646)
$\kappa_\rho$	0.590*** (82.600)	0.762*** (121.424)	0.414*** (46.118)	0.366*** (36.576)
$\delta_\rho$		1.057	1.175	1.198
$\gamma_\rho$		-2.991	-0.240	-1.662
$\alpha_\rho$		0.043	-1.641	-0.665
$\varphi_{\phi,t-1}$		0.911		1.069
$\varphi_{\phi,t-2}$				0.022
$\varphi_{\phi,t-3}$		-0.055		0.043
$\varphi_{\phi,t-4}$				0.108
$\varphi_{\mu,t-1}$			14.337	0.719
$\varphi_{\mu,t-2}$			13.776	0.496
$\varphi_{\mu,t-3}$			9.801	0.440
$\varphi_{\mu,t-4}$			3.756	0.213
$LL$	2314.54	8385.81	7823.34	9520.93
$AICc$	-4623.08	-16755.59	-15626.64	-19013.76
$DM$		0.00	0.00	0.00
$mDM$		0.00	0.00	0.00

Table 4.5 presents the estimation results from both the static and non-linear spatial models for different specifications of the threshold. In the static model, we find strong evidence for spatial dependence indicated by the high estimate for  $\rho$  together with a high  $t$ -statistic. The three non-linear specifications, respectively the ST-SAR driven by past observations, moving averages, and both ARMA dynamics, all improve the AICc values by several thousand points compared to the SAR. The most elaborate ST-SAR improves the AICc by an overwhelming 14390.68 points against the SAR. The significant evidence for nonlinearity is confirmed by the finding that the DM-type tests overwhelmingly reject the null of linearity, even for the most parsimonious ST-SAR. The residuals are in strong support of the choice to allow for fat tails. As an example, the kurtosis of residuals from the linear SAR is over 14 and a Jarque-Bera tests reject Gaussianity in favor of fatter tails with a p-value of  $\sim 0$  for all four

models. Evidence for nonlinearities in the convergence and dispersion process persists across the different ST-SAR specifications. However, the SAR contains no time dynamics. We therefore extend our analysis to control for additional ARMA dynamics.

### Extensions

In our extended results, we allow for additional flexibility and explore

$$\mathbf{v}_t - N^{-1} \sum_1^N (\mathbf{v}_{t-1}) = \mathbf{y}_t = H(\boldsymbol{\theta}^\rho; (\mathbf{y}_t, \boldsymbol{\varepsilon}_t))^{-1} ARMA(\boldsymbol{\theta}^{\phi, \mu}; \mathbf{y}_t, \boldsymbol{\varepsilon}_t).$$

Table 4.6: Estimation results for the extended spatial models on the full dataset.  $\lambda$  fixed at 2.5, constant omitted from table.

	SAR + ARMA	ST-SAR (ARMA) + ARMA
$\phi_{t-1}$	-0.202*** (-15.200)	0.110*** (2.904)
$\phi_{t-2}$	0.063*** (8.050)	
$\phi_{t-3}$	0.263*** (19.080)	0.423*** (6.049)
$\phi_{t-4}$	0.268*** (24.140)	0.258*** (5.455)
$\mu_{t-1}$	1.610*** (73.140)	0.819*** (11.793)
$\mu_{t-2}$	1.517*** (55.100)	0.612*** (8.873)
$\mu_{t-3}$	0.908*** (37.270)	0.314*** (5.033)
$\mu_{t-4}$	0.223*** (15.870)	-0.140*** (-10.410)
$\kappa_\rho$	0.625*** (59.770)	0.404
$\delta_\rho$		5.405
$\gamma_\rho$		-1.524
$\alpha_\rho$		-0.602
$\varphi_{\phi, t-1}$		0.969
$\varphi_{\phi, t-3}$		-0.459
$\varphi_{\phi, t-4}$		-0.260
$\varphi_{\mu, t-1}$		-0.671
$\varphi_{\mu, t-2}$		-0.569
$\varphi_{\mu, t-3}$		-0.331
$LL$	8375.408	10146.85
$AICc$	-16728.76	-20255.54
$DM$		0.002
$mDM$		0.004



Table 4.6 presents the results. The ST-SAR dynamics remain significant as judged by the various diagnostics even when additional ARMA dynamics are added to the conditional mean equation. Importantly, the AICc of the ST-SAR improves over that of the SAR by a very significant amount, a 3526.78 point improvement. The out-of-sample validation test further confirms the evidence for nonlinearity. The estimated probability that the linear spatial model attains lower KL is below 0.002, and 0.004 for the more conservative modified test.

We also find that the residuals of the nonlinear model, similarly to our first application, are smaller and better centered at zero. This can be seen in fig. 4.9. The residuals of the SAR remain respectively below and above zero for prolonged periods and contain significant remaining correlation patterns while the ST-SAR approximately neutralizes the dynamics as revealed by the residual ACF in fig. 4.10. Jarque-Bera tests again reject Gaussianity in favor of fatter tails, supporting again the choice for the Student's- $t$  specification, with a p-value of  $\sim 0$  for both models, with the residuals of the ST-SAR (ARMA) + ARMA reaching a kurtosis of 22.

Figure 4.5 displays the evolution of the fitted spatial dependence parameters. A first striking feature is the convergence of the parameters in anticipation of the Union, continuing till around 2000. In the pre-EU period we observe separate regimes. Ireland, Portugal, Italy and Spain form a low-dependence group. Greece forms an exception and follows an individual trajectory. After 2000, the parameters corresponding to the different sovereigns linearize, indicating strong financial stability and near perfect co-movement. The onset of the Great Recession around 2008 marks an abrupt turn after which separation in a high and low regime recurs. Interestingly, the pattern after the recession reverts to the pre-EU behavior, with Greece returning to an individual trajectory and Ireland, Portugal, Italy and Spain forming a less integrated group. This breakaway is in sharp contrast to the increasing interdependence across

other member states. Divergence between the low and high dependence regimes has continued after the crisis, and the sustained strong variation in contraction parameters indicates that the Eurozone remains to struggle in attaining EU-wide financial stability. These results suggest that the EMS has still not fully succeeded in aligning all economies across the Eurozone. Figure 4.6 further visualizes the time-varying nature of the dependence regimes over cross-sections and time.

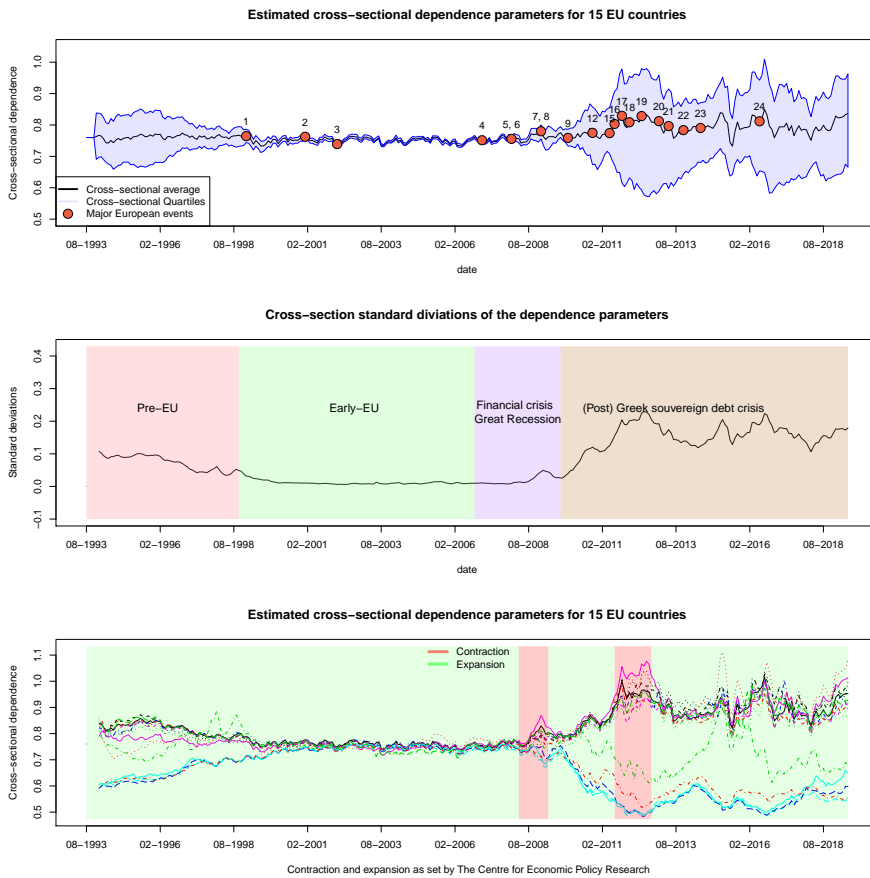
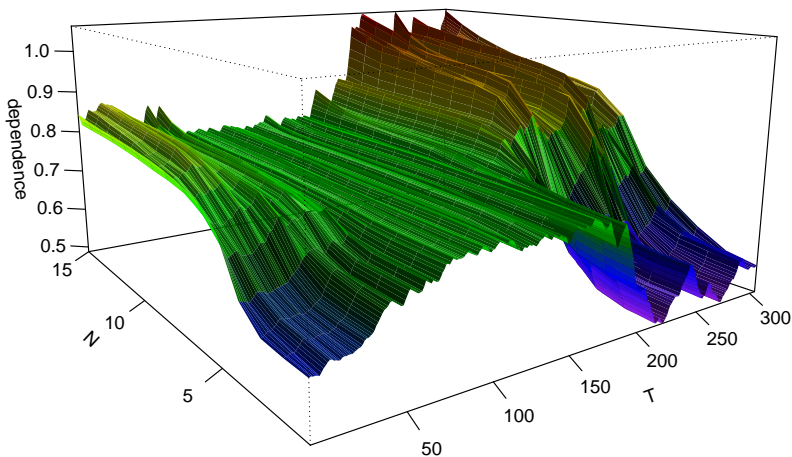


Figure 4.5: Evolution of spatial parameters estimated with the ARMA + ST-SAR (ARMA). Colors correspond to individual countries, see section 4.7.4. The estimation results highlight the nonlinear nature of dependence between sovereigns during Pre-EU times that has clearly broken into a two-regime system after the Financial Crisis.

### Local spatial dependencies throughout time



Local spatial regimes from the endogenous ST-SAR

Figure 4.6: Evolution of spatial parameters estimated with the ST-SAR.

Several interesting aspects about the convergence and dispersion dynamics can be learned from our final estimates. Local dependencies are partly driven by feedback, but also impacted by moving averages that may relate to directed financial policy or shocks. Since  $\hat{\gamma} = -1.524$  is negative, and  $\hat{\delta} = 5.405$  is positive, the spatial parameters increase with  $W\mathbf{y}_{t-1} - \hat{\tau}_{t-1}$ , thus the signs of the estimated  $\varphi$  parameters indicate the direction of individual contributions.<sup>12</sup> The complex threshold equation hints at

<sup>12</sup>The estimated threshold is  $\hat{\tau}_{t-1} = -.969\mathbf{y}_{t-1} + .459\mathbf{y}_{t-3} + .260\mathbf{y}_{t-4} + .671\epsilon_{t-1} + .569\epsilon_{t-2} + .331\epsilon_{t-3} - .602$ . Note that the signs in the table 4.6 are opposite as they enters as  $-\hat{\tau}_{t-1}$  in the likelihood function.

several subtleties. The negative signs of the moving averages suggest that positive shocks are followed by reduced dependence, while sustained exogenous policies that reduce rates result in increased contraction. If all effects are considered jointly, the following regime-dependent behavior can be distinguished:

1.  $\hat{\tau}_{t-1} < W\mathbf{y}_{t-1}$  local threshold value is below average neighbor rates, followed by intensified dependence (dispersion),
2.  $\hat{\tau}_{t-1} > W\mathbf{y}_{t-1}$  local threshold value is above average neighbor rates, followed by reduced dependence (convergence).

These regimes suggest cyclic behavior. First, high rates relative to neighboring sovereigns due to exogenous impacts (high  $\varepsilon_t - q$  for  $q = 1, \dots, 4$ ) is followed by reduced dependence to the group average, making isolated rate increases due to shock possible. Once assimilated, high relative rates (high  $\mathbf{y}_t - p$  for  $p = 1, \dots, 4$  relative to  $W\mathbf{y}_{t-1}$ ) is followed by intensified spatial dependence. Together this implies initial systemic vulnerability to exogenous shocks, but subsequent resistance to the spread of assimilated shocks. That resistance breaks when a large neighborhood is affected ( $W_{t-1}$  increases), accelerating the spread through increased feedback. Finally, the negative signs of deeper lags of  $\mathbf{y}_t - p$  indicate that initial increases in contraction are followed by a return to reduced dependence, slowing feedback.

The regimes also suggest asymmetries in spillovers. If at location  $i$  rates increase, dependence to neighbor  $j$  reduces. From the perspective of location  $j$  the opposite occurs, resulting in the opposite dynamics. This means that while a local positive impulse lowers spatial dependence locally, it increases the dependence parameters of neighbors, implying that outward spillovers accelerate while inward feedback slows down. On the other hand, lowered rates are followed by intensified inward spillovers but slower outward spillovers.

## 4.6 Conclusion

In this paper we introduced a new model for nonlinear spatial time series in which cross-sectional dependence varies smoothly over space by means of smooth-transitions between dependence regimes. In this framework, nonlinearities in cross-sectional dynamics are modeled as a function of the data. This is an advance over existing methods. Allowing for time-variation is particularly useful when modeling spatial data for large  $T$ , nonlinearities over the cross-section are particularly useful if  $N$  is large.

We have shown that the parameters of the model can be consistently estimated by maximum likelihood under appropriate regularity conditions. In particular, we provide conditions that deliver existence, strong consistency and asymptotic normality of the MLE of all static parameters that constitute the dynamic dependence structure. The theory holds for both correctly specified and misspecified models and allows for possible identification issues of the threshold parameters. Our simulation evidence suggests that the limit theory is relevant in finite samples. Furthermore, we find that information criteria are able to distinguish between the SAR specification and ST-SAR type nonlinearities. The simulations results showed that model selection is robust to overfitting of additive outliers. We have also provided a theoretical argument for model selection based on a validation-sample estimate of the Kullback-Leibler divergence. In our empirical application, both the validation test and the information criteria support nonlinearities.

The model has been applied to study space-time dynamics in two cases. We studied clustering in urban densities in a large number of districts, and convergence and dispersion in monthly long term interest rates. We found that the ST-SAR resulted in better filtering behavior over the cross-section and time dimension, improved estimates for exogenous variables,

and improved forecasts. We also found that the nonlinearities in the spatial parameters can lead to economically relevant insights, while the SAR is often criticized for its empirical interpretation. We conclude that the ST-SAR is a powerful tool for both understanding and predicting future values in cross-sectional time series.

## 4.7 Appendix

### 4.7.1 Proofs to main theorems

#### Proof of Theorem 9

*Proof.* Note first that  $L_T(\boldsymbol{\theta}) := (1/T) \sum_{t=1}^T \ell_t(\boldsymbol{\theta})$  is a.s. continuous (a.s.c.) in  $\boldsymbol{\theta} \in \Theta$  through continuity (c.) of each term  $\ell_t(\boldsymbol{\theta}) = \ln \det H(\boldsymbol{\theta}^\rho; \mathbf{Z}_t) + \ln p_\varepsilon \left( H(\boldsymbol{\theta}^\rho; \mathbf{Z}_t) \mathbf{y}_t - c - \mathbf{y}_{t-1} \phi - \mathbf{X}_t \boldsymbol{\beta}, \Sigma; \lambda \right)$ . Together with the compactness of  $\Theta$  (Assumption 1) this implies by Weierstrass' theorem that the arg max set is non-empty a.s. and hence that  $\hat{\boldsymbol{\theta}}_T$  exists a.s.  $\forall T \in \mathbb{N}$ . Note by a similar argument that  $L_T(\boldsymbol{\theta})$  is continuous in  $(\mathbf{y}_t, \mathbf{X}_t) \forall \boldsymbol{\theta} \in \Theta$  and hence measurable w.r.t. the product Borel  $\sigma$ -algebra  $\mathfrak{B}(\mathcal{Y}) \otimes \mathfrak{B}(\mathcal{X})$  that are, in turn, measurable maps w.r.t.  $\mathcal{F}$  by Proposition 4.1.7 in Dudley (2002).<sup>13</sup> Finally, the measurability of  $\hat{\boldsymbol{\theta}}_T$  follows from (Foland, 2009, p.24) and (White, 1994, Theorem 2.11) or (Gallant and White, 1988, Lemma 2.1, Theorem 2.2).<sup>14</sup>  $\square$

#### Proof of Theorem 10

*Proof.* Recall that  $L_T(\boldsymbol{\theta}) := (1/T) \sum_{t=1}^T \ell_t(\boldsymbol{\theta})$  and  $L_\infty(\boldsymbol{\theta}) = \mathbb{E} \ell_t(\boldsymbol{\theta})$  with

$$\ell_t(\boldsymbol{\theta}) = \ln \det H(\boldsymbol{\theta}^\rho; \mathbf{Z}_t) + \ln p_\varepsilon \left( H(\boldsymbol{\theta}^\rho; \mathbf{Z}_t) \mathbf{y}_t - c - \mathbf{y}_{t-1} \phi - \mathbf{X}_t \boldsymbol{\beta}, \Sigma; \lambda \right).$$

Following the usual consistency argument (found e.g. in (White, 1994, Theorem 3.4) or Theorem 3.3 in Gallant and White (1988)) we obtain

<sup>13</sup>Dudley's proposition states that the Borel  $\sigma$ -algebra  $\mathfrak{B}(\mathbb{A} \times \mathbb{B})$  generated by the Tychonoff's product topology  $\mathcal{T}_{\mathbb{A} \times \mathbb{B}}$  on the space  $\mathbb{A} \times \mathbb{B}$  includes the product  $\sigma$ -algebra  $\mathfrak{B}(\mathbb{A}) \otimes \mathfrak{B}(\mathbb{B})$ .

<sup>14</sup>The reference of Foland (2009) is used here to establish that a map into a product space is measurable if and only if its projections are measurable.

$\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0$  from the uniform convergence of the criterion function

$$\sup_{\boldsymbol{\theta} \in \Theta} |L_T(\boldsymbol{\theta}) - L_\infty(\boldsymbol{\theta})| \xrightarrow{a.s.} 0 \quad \forall f_1 \in \mathcal{F} \quad \text{as } T \rightarrow \infty \quad (4.12)$$

and the identifiable uniqueness of the maximizer  $\boldsymbol{\theta}_0 \in \Theta$  introduced in White (1994),

$$\sup_{\boldsymbol{\theta}: \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \varepsilon} L_\infty(\boldsymbol{\theta}) < L_\infty(\boldsymbol{\theta}_0) \quad \forall \varepsilon > 0. \quad (4.13)$$

The uniform convergence is obtained by application of the ergodic theorem for separable Banach spaces in Rao (1962), as in (Straumann and Mikosch, 2006, Theorem 2.7), to the sequence  $\{L_T(\cdot)\}$  with elements taking values in  $\mathbb{C}(\Theta, \mathbb{R})$ . This uniform law of large numbers  $\sup_{\boldsymbol{\theta} \in \Theta} |L_T(\boldsymbol{\theta}) - \mathbb{E}\ell_t(\boldsymbol{\theta})| \xrightarrow{a.s.} 0$  as  $T \rightarrow \infty$  follows, under a uniform moment bound  $\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |\ell_t(\boldsymbol{\theta})| < \infty$ , by the SE nature of  $\{L_T\}_{t \in \mathbb{Z}}$  which is implied by continuity of  $\ell$  on the SE sequence  $\{(\mathbf{y}_t, \mathbf{X}_t)\}_{t \in \mathbb{Z}}$  (Assumption 2) and Proposition 4.3 in Krengel (1985). The uniform moment bound  $\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |\ell_t(\boldsymbol{\theta})| < \infty$  follows immediately from Assumption 9 since

$$\begin{aligned} \mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |\ell_t(\boldsymbol{\theta})| &\leq \mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |Q(\boldsymbol{\theta}^\rho; \mathbf{Z}_t)| + |A(\boldsymbol{\theta})| \\ &+ \frac{1}{2}(\lambda + N) \mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |F(\boldsymbol{\theta}, \mathbf{y}_t, \mathbf{X}_t, \mathbf{Z}_t)| < \infty. \end{aligned}$$

Finally, the identifiable uniqueness (see e.g. White (1994)) of  $\boldsymbol{\theta}_0 \in \Theta$  in (4.13) follows from the assumed uniqueness (Assumption 10), the compactness of  $\Theta$ , and the continuity of the limit  $\mathbb{E}\ell_t(\boldsymbol{\theta})$  in  $\boldsymbol{\theta} \in \Theta$  which is implied by the continuity of  $L_T$  in  $\boldsymbol{\theta} \in \Theta \quad \forall T \in \mathbb{N}$  and the uniform convergence in (4.12).  $\square$

### Proof of Theorem 11

*Proof.* The proof follows by the same argument as laid down in the proof of Theorem 2. Only now the assumption, that  $\boldsymbol{\theta}_0$  is the unique maximizer, is missing (Assumption 10). Without uniqueness, we obtain the desired set consistency result by application of Lemma 4.3 in (Pötscher and Prucha, 1997), after noting that the continuity of the limit criterion  $L_\infty(\boldsymbol{\theta}) = \mathbb{E}\ell_t(\boldsymbol{\theta})$  in  $\boldsymbol{\theta} \in \Theta$  and the compactness of  $\Theta$  ensure that the

levels sets of  $L_\infty$  are regular (see Definition 4.1 in Pötscher and Prucha (1997)). The continuity of  $L_\infty$  is obtained directly from the continuity of  $\ell_t(\boldsymbol{\theta})$  in  $\boldsymbol{\theta} \in \Theta$  for every  $t$ , and the uniform convergence of the sample criterion  $\frac{1}{T} \sum_{t=1}^T \ell_t$  to the limit  $L_\infty$ .  $\square$

### Proof of Theorem 12

*Proof.* The desired result follows immediately by application of Theorem 11 after noting that the data generated by the ST-SAR model converges to a unique SE solution by Propositions 1 and 2.  $\square$

### Proof of Theorem 13

*Proof.* We obtain the asymptotic Gaussianity of the MLE immediately from (i) the strong consistency of  $\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0 \in \text{int}(\Theta)$ ; (ii) the a.s. twice continuous differentiability of  $\ell_T(\boldsymbol{\theta})$  in  $\boldsymbol{\theta} \in \Theta$ ; (iii) the asymptotic normality of the score

$$\sqrt{T}L'_T(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{J}(\boldsymbol{\theta}_0)), \quad \mathcal{J}(\boldsymbol{\theta}_0) = \mathbb{E}(\ell'_t(\boldsymbol{\theta}_0)\ell'_t(\boldsymbol{\theta}_0)^\top); \quad (4.14)$$

(iv) the uniform convergence of the likelihood's second derivative,

$$\sup_{\boldsymbol{\theta} \in \Theta} \|L''_T(\boldsymbol{\theta}) - L''_\infty(\boldsymbol{\theta})\| \xrightarrow{a.s.} 0; \quad (4.15)$$

and finally, (v) the non-singularity of the limit  $L''_\infty(\boldsymbol{\theta}) = \mathbb{E}\ell''_t(\boldsymbol{\theta}) = \mathcal{I}(\boldsymbol{\theta})$ . See e.g. in (White, 1994, Theorem 6.2)) for further details.

The consistency condition  $\hat{\boldsymbol{\theta}}_T \xrightarrow{a.s.} \boldsymbol{\theta}_0 \in \text{int}(\Theta)$  in (i) follows by Theorem 2 and the additional assumption that  $\boldsymbol{\theta}_0 \in \text{int}(\Theta)$ .

The smoothness condition in (ii) is trivially satisfied for the student's- $t$  density.

The asymptotic normality of the score in (4.16) follows by Theorem 18.10[iv] in van der Vaart (2000) by an application of the CLT for SE martingales in Billingsley (1961) or NED processes in Pötscher and Prucha (1997) Theorem 10.2, to obtain

$$\sqrt{T}L'_T(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{J}(\boldsymbol{\theta}_0)) \quad \text{as } T \rightarrow \infty, \quad (4.16)$$

where  $\mathcal{J}(\boldsymbol{\theta}_0) = \mathbb{E}(\ell'_t(\boldsymbol{\theta}_0)\ell'_t(\boldsymbol{\theta}_0)^\top) < \infty$ . The SE nature of  $\{L'_T(\boldsymbol{\theta}_0)\}_{t \in \mathbb{Z}}$



follows by continuity of  $L'_T$  on the SE sequence  $\{(\mathbf{y}_t, \mathbf{X}_t)\}_{t \in \mathbb{Z}}$ ; see Proposition 4.3 in Krengel (1985). Assumption 5 imposes the mds or NED nature of the score sequence  $\{\ell'_t(\boldsymbol{\theta}_0)\}_{t \in \mathbb{Z}}$ . The finite (co)variances follow from the first two moments bounds of Assumption 6.

The uniform convergence in (iv) is obtained under the moment bound

$$\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} \|\ell''_t(\boldsymbol{\theta})\| < \infty$$

and by the SE nature of  $\{\ell''_T\}_{t \in \mathbb{Z}}$ . The moment bound is ensured by Assumption 6. The SE nature is implied by continuity of  $\ell''$  on the SE sequence  $\{\mathbf{y}_t, \mathbf{X}_t\}_{t \in \mathbb{Z}}$ .

Finally, the non-singularity of the limit  $L''_\infty(\boldsymbol{\theta}) = \mathbb{E}\ell''_t(\boldsymbol{\theta}) = \mathcal{I}(\boldsymbol{\theta})$  in (v) is implied by the uniqueness of  $\boldsymbol{\theta}_0$  as a maximum of  $L''_\infty(\boldsymbol{\theta})$  in  $\Theta$ .  $\square$

### Proof of Lemma 1

*Proof.* Expand  $\tilde{L}_{\hat{T}}(\hat{\boldsymbol{\theta}}_T)$  at  $\boldsymbol{\theta}_0^*$  to obtain

$$\lim_{T, \hat{T} \rightarrow \infty} \mathbb{E} \left( \tilde{L}_{\hat{T}}(\hat{\boldsymbol{\theta}}_T) - \mathbb{E} \tilde{L}_{\hat{T}}(\boldsymbol{\theta}_0^*) \right) = \lim_{T, \hat{T} \rightarrow \infty} \mathbb{E} \tilde{L}'_{\hat{T}}(\boldsymbol{\theta}_0^*)(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0^*)$$

Next, use the uniform moment  $\mathbb{E} \sup_{\boldsymbol{\theta} \in \Theta} |\tilde{L}'_T(\boldsymbol{\theta})| < \infty$  to interchange the limit and expectation by appealing to a dominated convergence theorem, and use Slutsky's theorem to obtain,

$$\lim_{T, \hat{T} \rightarrow \infty} \mathbb{E} \tilde{L}'_{\hat{T}}(\boldsymbol{\theta}_0^*)(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0^*) = \lim_{\hat{T} \rightarrow \infty} \mathbb{E} \lim_{T \rightarrow \infty} \tilde{L}'_{\hat{T}}(\boldsymbol{\theta}_0^*) \lim_{T \rightarrow \infty} (\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0^*)$$

Finally, use the continuity of  $\ell$ , a continuous mapping theorem, and the consistency of the MLE to obtain the desired result

$$\lim_{T, \hat{T} \rightarrow \infty} \mathbb{E} \left( \tilde{L}_{\hat{T}}(\hat{\boldsymbol{\theta}}_T) - \mathbb{E} \tilde{L}_{\hat{T}}(\boldsymbol{\theta}_0^*) \right) = \lim_{\hat{T} \rightarrow \infty} \mathbb{E} \tilde{L}'_{\hat{T}}(\boldsymbol{\theta}_0^*) \times 0 = 0 \quad \text{a.s.}$$

$\square$

### Proof of Lemma 2

*Proof.* Obtained immediately by the same argument as that of Lemma 1.  $\square$

**Proof of Proposition 3**

*Proof.* Follows immediately by noting that under the null  $H_0$  :  $\mathbb{E}\tilde{\Delta}_t(\boldsymbol{\theta}_0^{*A}, \boldsymbol{\theta}_0^{*B}) = 0$ , we have

$$\begin{aligned} \lim_{T, \tilde{T} \rightarrow \infty} \tilde{T}^{\frac{1}{2}} \left( \tilde{T}^{-1} \sum_t^{\tilde{T}} \frac{\tilde{\Delta}_t(\hat{\boldsymbol{\theta}}_T^A, \hat{\boldsymbol{\theta}}_T^B)}{\tilde{\sigma}_{\tilde{T}}(\hat{\boldsymbol{\theta}}_T^A, \hat{\boldsymbol{\theta}}_T^B)} \right) &= \lim_{\tilde{T} \rightarrow \infty} \tilde{T}^{\frac{1}{2}} \left( \tilde{T}^{-1} \sum_t^{\tilde{T}} \frac{\lim_{T \rightarrow \infty} \tilde{\Delta}_t(\hat{\boldsymbol{\theta}}_T^A, \hat{\boldsymbol{\theta}}_T^B)}{\lim_{T \rightarrow \infty} \tilde{\sigma}_{\tilde{T}}(\hat{\boldsymbol{\theta}}_T^A, \hat{\boldsymbol{\theta}}_T^B)} \right) \\ &= \lim_{\tilde{T} \rightarrow \infty} \tilde{T}^{\frac{1}{2}} \left( \tilde{T}^{-1} \sum_t^{\tilde{T}} \frac{\tilde{\Delta}_t(\boldsymbol{\theta}_0^{*A}, \boldsymbol{\theta}_0^{*B})}{\tilde{\sigma}_{\tilde{T}}(\boldsymbol{\theta}_0^{*A}, \boldsymbol{\theta}_0^{*B})} - \tilde{\Delta}_t(\boldsymbol{\theta}_0^{*A}, \boldsymbol{\theta}_0^{*B}) \right) \stackrel{d}{=} \mathcal{N}(0, 1). \end{aligned}$$

The first equality is obtained by Slutsky's Theorem. The second equality by consistency  $\hat{\boldsymbol{\theta}}_T^A \xrightarrow{as} \boldsymbol{\theta}_0^{*A}$  and  $\hat{\boldsymbol{\theta}}_T^B \xrightarrow{as} \boldsymbol{\theta}_0^{*B}$  and the a.s. continuity of  $\tilde{\Delta}_t$  and  $\tilde{\sigma}_{\tilde{T}}$  on  $\Theta \times \Theta$ . The last equality, in distribution, is obtained by application of the CLT for strictly stationary and ergodic martingale difference sequences in Billingsley (1961).  $\square$

### 4.7.2 Additional results

LEMMA. 3. *Let  $A$  be an arbitrary finite-dimensional matrix. For an induced matrix norm  $\|A\| < 1$  the following inequality is implied:*

$$(1 + \|A\|)^{-1} \leq \|(I_N - A)^{-1}\| \leq (1 - \|A\|)^{-1},$$

*with  $0 < (1 + \|A\|)^{-1}$  and  $(1 - \|A\|)^{-1} < \infty$ . By the finite-dimensionality we can also write*

$$0 < c \leq \|(I_N - A)^{-1}\|_\infty \leq C < \infty,$$

*for some positive constants  $c$  and  $C$ .*

For a matrix  $H$  defined by  $H = (I_N - A)$ , LEMMA. 3 provides existence, non-negativity, and boundedness of the inverse  $H^{-1}$  for finite dimensional  $H$ . This is useful since throughout our theory as we always work with a fixed  $N$  and let only  $T$  tend to infinity.

LEMMA. 4. *Let  $A$  be an arbitrary matrix with eigenvalues  $\omega_1, \dots, \omega_n \in \mathbb{C}^{n \times n}$ , real or complex, and  $r(A) = \max\{|\omega_1|, \dots, |\omega_n|\}$  be its spectral radius. If  $r(A) < 1$  there exists  $\|A\| < 1$  for some induced matrix norm.*

LEMMA. 4 allows the condition  $\|A\| < 1$  in LEMMA. 3 to be replaced by  $r(A) < 1$  if no suitable norm can be found. In what follows we will continue stating  $\|\cdot\|$ , but remind the reader that in practice one may focus on sample estimates of  $r(\cdot)$  as a rule of thumb.

LEMMA. 5. *For any  $H^{-1} \in \mathbb{R}^{n \times n}$  defined as  $H^{-1} = (I_N - A)^{-1}$  with  $N < \infty$  and  $r(A) < 1$ , we have that the following is implied*

$$i \det(H^{-1}) > 0,$$

$$ii \log \tau(H^{-1})^N \leq \log \det(H^{-1}) \leq \log r(H^{-1})^N < \infty,$$

$$iii |\log \det(H^{-1})| < \infty,$$

Claim *iii* in LEMMA. 5 is particularly useful in establishing that ASSUMPTION. 9 holds under correct specification.

### 4.7.3 Proofs for additional results

#### Proof of Lemma 3

*Proof.* The result follows by taking norms in  $I_N = (I_N - A)(I_N - A)^{-1}$ , which gives<sup>15</sup>

$$1 \leq \|I_N - A\| \|(I_N - A)^{-1}\|.$$

This can be rearranged to obtain

$$1 \leq (1 + \|A\|) \|(I_N - A)^{-1}\|.$$

Multiplying by  $\|(I_N - A)^{-1}\|^{-1}$  gives

$$\|(I_N - A)^{-1}\|^{-1} \leq \|I_N - A\| \leq (1 + \|A\|),$$

thus

$$(\|(I_N - A)^{-1}\|)^{-1} \leq (1 + \|A\|),$$

$$(1 + \|A\|)^{-1} \leq \|(I_N - A)^{-1}\|,$$

providing the first inequality.

The second inequality follows immediately by the fact that the operator norm is sub-multiplicative. In particular,  $\|I\| = \|B \cdot B^{-1}\| \leq \|B\| \cdot \|B^{-1}\|$  implies that  $\|B\|^{-1} \leq \|B^{-1}\|$ . Hence  $(1 + \|A\|)^{-1} < \|(I - A)^{-1}\|$ .

Finiteness of  $\|(I_n - A)^{-1}\|$  follows trivially from

$$(1 - \|A\|)^{-1} < \infty. \text{ since } \|A\| < 1.$$

Non-negativity of  $(I_n - A)^{-1}$  follows by noting that all its eigenvalues are non-zero. The minimum eigenvalue of a non-singular matrix is equal to the inverse of the spectral radius of the inverse matrix, thus in this case  $\tau((I_n - A)^{-1}) = r(I - A)^{-1}$ . Having just established that  $(1 + \|A\|)^{-1} \leq \|(I_n - A)^{-1}\| \leq (1 - \|A\|)^{-1}$  it follows trivially that like-wise

$$(1 + \|A\|) \geq \|(I_n - A)\| \geq (1 - \|A\|),$$

---

<sup>15</sup>The result is similar to Proposition 6.4.1. in Lange (1999), but reworked here because both the proof and the final result are partial.

which delivers the upper bounds of  $r(I_n - A)$  by noting that  $r(I_n - A) \leq \|I_n - A\|$ , hence  $r(I - A)^{-1} > 0$ , and equally so  $\tau((I_n - A)^{-1}) > 0$ .

Finally, by noting that any two norms in finite dimension  $n < \infty$  are always within a constant factor of one another, such that we can write for some real numbers  $0 < c_1 \leq c_1 \leq c_2$  the inequality

$$c_1 \|(I_n - A)^{-1}\|_\infty \leq \|(I_n - A)^{-1}\| \leq c_2 \|(I_n - A)^{-1}\|_\infty,$$

proves the second claim by setting  $c = c_1 \|(I_n - A)^{-1}\|_\infty$  and  $C = c_2 \|(I_n - A)^{-1}\|_\infty$ .  $\square$

#### Proof of Lemma 4

*Proof.* This follows from by noting that for any matrix  $A$  and any positive number  $e > 0$ , there exists an induced matrix norm  $\|A\|$  such that

$$r(A) \leq \|A\| < r(A) + e.$$

See Proposition 6.3.2. Lange (1999). Trivially,

$$r(A) < 1 \implies 1 - r(A) > 0.$$

Choose  $e = 1 - r(A)$ , the proof is completed by noting that we can now write

$$\begin{aligned} r(A) &\leq \|A\| < r(A) + 1 - r(A), \\ r(A) &\leq \|A\| < 1. \end{aligned}$$

$\square$

#### Proof of Lemma 5

*Proof.* The proof of all three claims starts by noting that by definition (slight abuse of notation: reintroducing  $p$  and  $k$ )

$$\det(H^{-1}) = (\Pi_{i=1}^k \omega_i) (\Pi_{i=k+1}^p \omega_i \bar{\omega}_i) = (\Pi_{i=1}^k \omega_i) (\Pi_{i=k+1}^p |\omega_i|^2),$$

with  $\omega_1, \dots, \omega_N \in \mathbb{C}^{N \times N}$ , real or complex, as the eigenvalues of  $H^{-1}$ . Hence the first claim follows by showing that

$$(\Pi_{i=1}^k \omega_i) (\Pi_{i=k+1}^p |\omega_i|^2) > 0.$$

Thus we need to show that  $\omega_i > 0 \forall i \in 1, \dots, N$ , since then  $(|\omega_i|^2)^{p-k} > 0$ , and  $\omega_i^{N-p} > 0$ , hence the left side of the second equality is strictly positive. Note that LEMMA. 3 and LEMMA. 4 deliver the following inequality under assumptions of LEMMA. 5,

$$(1 + \|A\|)^{-1} \leq \|(I_N - A)^{-1}\| \leq (1 - \|A\|)^{-1},$$

which we can also write as

$$(1 - \|A\|) \leq \|(I_N - A)\| \leq (1 + \|A\|).$$

The desirable result follows by proving that  $\tau(H^{-1}) > 0$ , where  $\tau(H^{-1}) = \tau((I_N - A)^{-1}) = \min\{|\omega_1|, \dots, |\omega_N|\}$ . Applying the useful identity  $\tau(A) = (r(A^{-1}))^{-1}$ , we have

$$\tau(H^{-1}) = (r(H))^{-1},$$

hence showing that  $\tau(H^{-1}) > 0$  equals showing that  $(r(H))^{-1} > 0$ , which follows from  $r(H) < \infty$ . Using the general inequality  $r(H) \leq \|H\|$  we can write  $r(H) \leq \|(I_N - A)\| \leq (1 + \|A\|)$  thus proving  $\tau(H^{-1}) > 0$ .

Using the definition of  $\det(H^{-1})$ , and the bounds of  $H^{-1}$  we obtain the range of the determinant by allowing the finite number of  $N$  eigenvalues to be either strictly minima or maxima

$$0 < (\tau(H^{-1}))^N \leq \det(H^{-1}) \leq r(H^{-1})^N < \infty.$$

The second claim follows easily now by taking logs and applying Jensen's inequality.

Finally, the third claim follows by noting that  $0 < \det(H^{-1})$  implies that the log is defined, hence its absolute value is finite.

□

### Proof of Proposition 1

*Proof.* The result follows by Theorem 2.2 and Example 2.1 in Cline and Pu (1998). In particular, we note first that LEMMA. 3 provides the uniform bound of  $H(\mathbf{y})^{-1}$  by noting that if  $\sup_{\mathbf{y} \in \mathbb{R}^N} \|\rho(\boldsymbol{\theta}_0^e; \mathbf{y}) \circ W\| < 1$

we have

$$0 < \bar{h} \leq \sup_{\mathbf{y} \in \mathbb{R}^N} \|H(\mathbf{y})^{-1}\| \leq \bar{H} < \infty$$

with

$$\bar{h} = \left( 1 + \sup_{\mathbf{y} \in \mathbb{R}^N} \|\rho(\boldsymbol{\theta}_0^\rho; \mathbf{y}) \circ W\| \right)^{-1}, \quad \bar{H} = \left( 1 - \sup_{\mathbf{y} \in \mathbb{R}^N} \|\rho(\boldsymbol{\theta}_0^\rho; \mathbf{y}) \circ W\| \right)^{-1}.$$

Having just established that  $\sup_{\mathbf{y} \in \mathbb{R}^N} \|H(\mathbf{y})^{-1}\|$  is bounded away from zero by some constant  $\bar{h}$  and from infinity by some constant  $\bar{H} < \infty$ , and that  $H^{-1}(\mathbf{y})$  is invertible, we can now verify that the assumptions in Theorem 2.2 and Example 2.1 in Cline and Pu (1998) hold. First we note that  $H(\mathbf{y})$  and  $H(\mathbf{y})^{-1}$  are both trivially locally bounded, and that  $\boldsymbol{\varepsilon}_t$  has full support. Finally, we note that  $H(\mathbf{y})^{-1}\mathbf{y}\phi$  is also locally bounded since

$$\sup_{\|\mathbf{y}\| \leq M} \|H(\mathbf{y})^{-1}\mathbf{y}\phi\| \leq \sup_{\mathbf{y}} \|H(\mathbf{y})^{-1}\| \|\phi\| \sup_{\|\mathbf{y}\| \leq M} \|\mathbf{y}\| \leq BM\phi < \infty \quad \forall M > 0.$$

□

### Proof of Proposition 2

*Proof.* We recall that  $\sup_{\mathbf{y} \in \mathbb{R}^N} \|H(\mathbf{y})^{-1}\| \leq B < \infty$  under the assumptions of Proposition 1. Next we obtain the desired result from Theorem 3.1 of Cline and Pu (1999). First, we note that  $H(\mathbf{y})^{-1}\mathbf{y}$  is trivially unbounded in  $\mathbb{R}^N$ . Second, we have that  $H(\mathbf{y})^{-1}\mathbf{y}/(1 + \|\mathbf{y}\|)$  is bounded in  $\mathbb{R}^N$  since

$$\begin{aligned} & \|H(\mathbf{y})^{-1}\mathbf{y}\phi/(1 + \|\mathbf{y}\|)\| \\ & \leq \|H(\mathbf{y})^{-1}\| \|\mathbf{y}\| \|\phi\| / (1 + \|\mathbf{y}\|) \leq B \|\mathbf{y}\| \|\phi\| / (1 + \|\mathbf{y}\|) \leq B \|\phi\|. \end{aligned}$$

Next, we note that

$$\sup_{\|\mathbf{y}\| \leq M} \mathbb{E} \|H(\mathbf{y})^{-1}\boldsymbol{\varepsilon}_t\|^r \leq \sup_{\|\mathbf{y}\| \in \mathbb{R}^N} \|H(\mathbf{y})^{-1}\| \mathbb{E} \|\boldsymbol{\varepsilon}_t\|^r \leq B \mathbb{E} \|\boldsymbol{\varepsilon}_t\|^r < \infty$$

for every  $M > 0$ .

Additionally, it holds trivially true that

$$\lim_{\|\mathbf{y}\| \rightarrow \infty} \mathbb{E} \frac{\|H(\mathbf{y})^{-1}\boldsymbol{\varepsilon}_t\|^r}{\|\mathbf{y}\|^r} \leq \lim_{\|\mathbf{y}\| \rightarrow \infty} \frac{B \|\boldsymbol{\varepsilon}_t\|^r}{\|\mathbf{y}\|^r} \rightarrow 0.$$

Furthermore, it holds true that

$$\begin{aligned} & \lim_{\substack{\|H(\mathbf{y})^{-1}\mathbf{y}\phi\| \rightarrow \infty \\ \|\mathbf{y}-\mathbf{y}'\|/\|\mathbf{y}\| \rightarrow 0}} \left\| \frac{H(\mathbf{y})^{-1}\mathbf{y}\phi}{1 + \|\mathbf{y}\|} - \frac{H(\mathbf{y}')^{-1}\mathbf{y}'\phi}{1 + \|\mathbf{y}'\|} \right\| \\ &= \lim_{\substack{\|\mathbf{y}\| \rightarrow \infty, \|\mathbf{y}'\| \rightarrow \infty \\ \|\mathbf{y}-\mathbf{y}'\|/\|\mathbf{y}\| \rightarrow 0}} \left\| \frac{H(\mathbf{y})^{-1}\mathbf{y}\phi}{1 + \|\mathbf{y}\|} - \frac{H(\mathbf{y}')^{-1}\mathbf{y}'\phi}{1 + \|\mathbf{y}'\|} \right\| \end{aligned}$$

since  $H(\mathbf{y})^{-1}$  is uniformly bounded in  $\mathbf{y}$ , and hence,

$$\|H(\mathbf{y})^{-1}\mathbf{y}\phi\| \rightarrow \infty \quad \Leftrightarrow \quad \|\mathbf{y}\| \rightarrow \infty ,$$

$$\text{and } \left\{ \|\mathbf{y}\| \rightarrow \infty \quad \wedge \quad \|\mathbf{y} - \mathbf{y}'\|/\|\mathbf{y}\| \rightarrow 0 \right\} \quad \Leftrightarrow \quad \|\mathbf{y}'\| \rightarrow \infty .$$

As a result

$$\lim_{\substack{\|\mathbf{y}\| \rightarrow \infty, \|\mathbf{y}'\| \rightarrow \infty \\ \|\mathbf{y}-\mathbf{y}'\|/\|\mathbf{y}\| \rightarrow 0}} \left\| \frac{H(\mathbf{y})^{-1}\mathbf{y}\phi}{1 + \|\mathbf{y}\|} - \frac{H(\mathbf{y}')^{-1}\mathbf{y}'\phi}{1 + \|\mathbf{y}'\|} \right\| = \|H_\infty\phi - H_\infty\phi\| = 0.$$

Finally, we also have

$$\limsup_{\|\mathbf{y}\| \rightarrow \infty} \frac{\|H(\mathbf{y})^{-1}\mathbf{y}\phi\|}{\|\mathbf{y}\|} \leq \limsup_{\|\mathbf{y}\| \rightarrow \infty} \frac{\|H(\mathbf{y})^{-1}\| \|\mathbf{y}\| \|\phi\|}{\|\mathbf{y}\|} = \|H_\infty\| \|\phi\| < 1.$$

□

#### 4.7.4 Additional Monte Carlo results and figures

In this additional experiment, we investigate whether the MLE is well-behaved and approximately normal for increasing sample sizes in the case of identified parameters. This itself is not the most interesting result to study, but it confirms that our theory is correct. The data generating process is of the form:

$$\mathbf{y}_t = H(\boldsymbol{\theta}^\rho; \mathbf{y}_{t-1})^{-1}(\boldsymbol{\varepsilon}_t), \quad \boldsymbol{\varepsilon}_t \sim TID(1, I_N; 5), \quad (4.17)$$

We set the parameters values to



$$\delta = 1.35, \gamma = 1.05, \alpha = -.2, \varphi = 1.4, \kappa = -.4,$$

$$\mathbf{Z}_t = \mathbf{y}_{t-1}, \tau(\boldsymbol{\theta}^T; \mathbf{Z}_t) = \alpha + \varphi/N \sum_1^N (\mathbf{y}_{t-1}),$$

which satisfies the conditions for geometric ergodicity and allows for local positive and negative clustering.

We keep the ratio of distant and close-by neighbors comparable across experiments by allowing the network density of the weights matrix to increase with  $N$ . In each draw we generate a random zero diagonal row-normalized weights matrix with  $N/10$  neighbors for each observation. The process is initialized with  $H_1 = I_N$ , and the first 50 steps of the sequence are discarded to avoid dependence on the initialization. We simulate 1000 datasets and estimate the parameters of the ST-SAR with Student's- $t$  likelihood. We consider samples of size  $T = 25, 100, 250$  for  $N = 30$ . Figure 4.7 presents kernel density estimates of the distribution of the MLE for the different sample sizes.

Figure 4.7 presents the results and shows that for small sample sizes the estimators are not perfectly normal. For larger sample sizes, we see a fast convergence towards the limiting result. A second experiment with  $N = 60$  was also performed, we noticed improvements in the distributions for small  $T$  as  $N$  grows. The results indicate that for an empirically relevant signal and sample size the MLE is well-behaved. Note that these results do not directly generalize to any empirical setting. Specifically, (near)-linear signals will cause identification problems even in larger samples that break the uniqueness assumption required for normality. However, our main simulation results show that information criteria can be used to assess the presence and significance of nonlinearity.

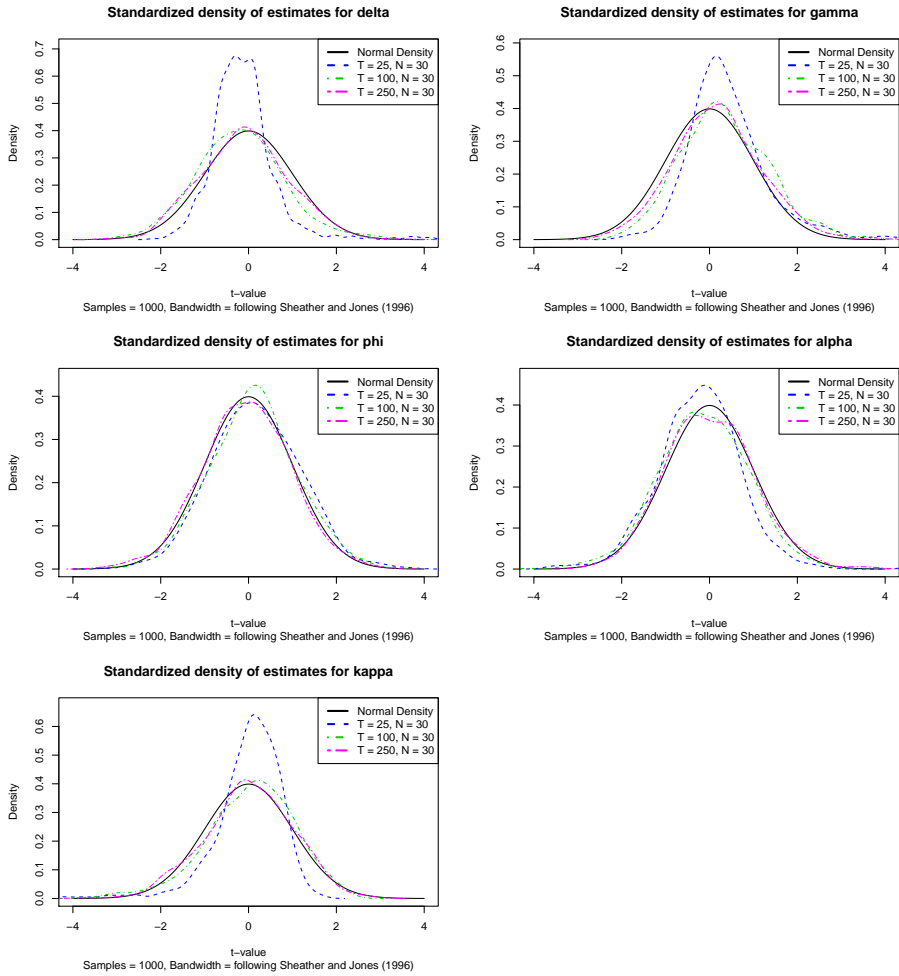


Figure 4.7: Kernel density estimates of estimated parameters from 1000 simulation replications for  $N = 30$  indicating that parameters are approximately well-behaved when identified. Note that this does not permit the use of  $t$ -statistics to test for significance, evidence for non-linearity can be obtained from AIC and DM-type tests as in our empirical analyses.



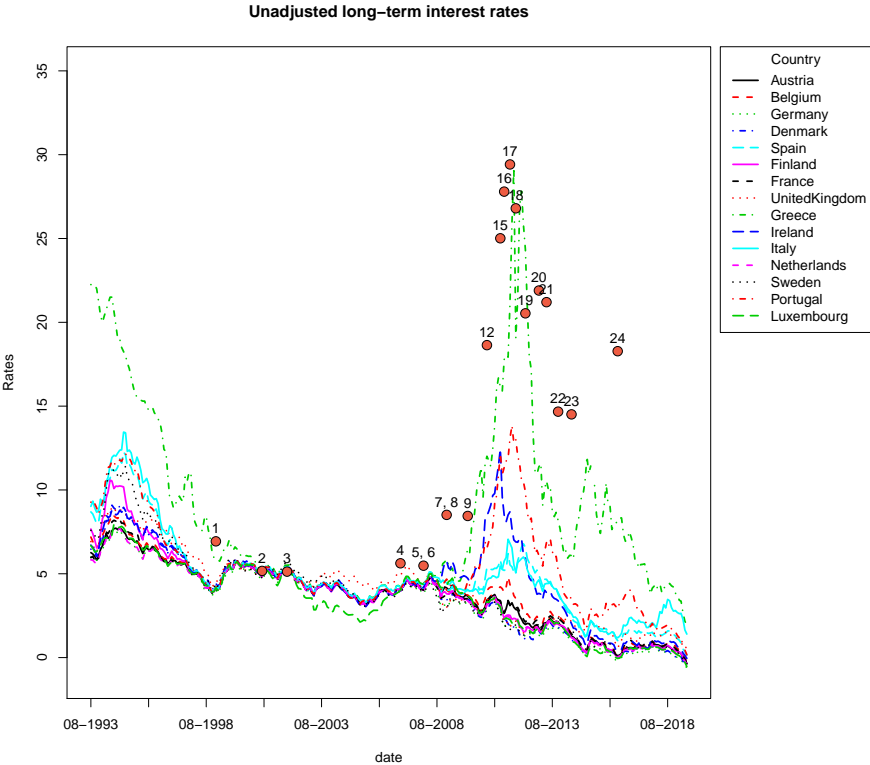


Figure 4.8: Raw data and sovereign colors used in application II.

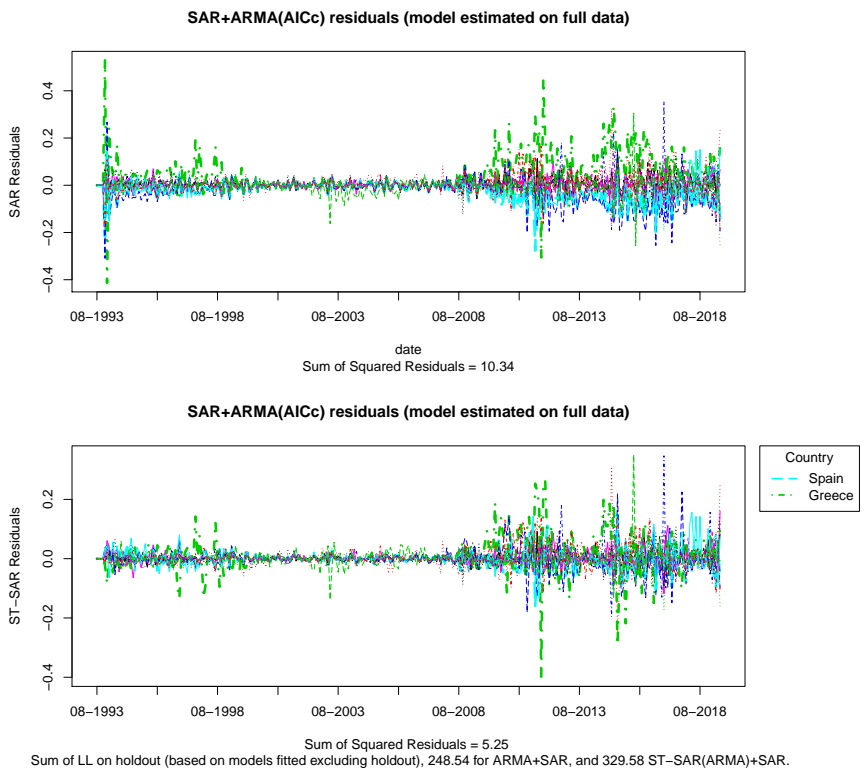


Figure 4.9: Residuals of the final SAR and ST-SAR showing that after filtering out linear spatial dynamics, the residuals of Spain and Greece are not properly centered on zero.

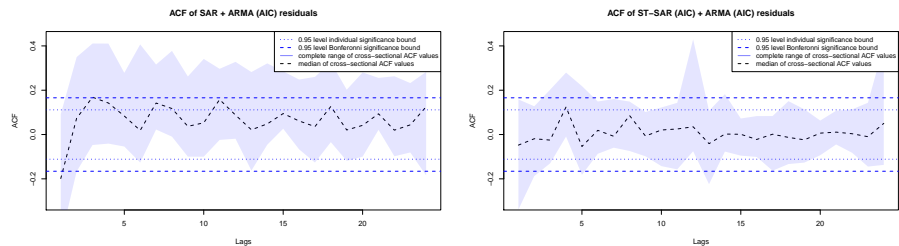


Figure 4.10: Residual correlations of the SAR and ST-SAR estimated on full data, highlighting the improved filtering of the ST-SAR.

### 4.7.5 Time-line of events related to European Long term Interest Rates

1. January 1999, start of the Euro;
2. January 2001, Greece joins the Euro;
3. January 2002, Euro coins and notes are introduced;
4. January 2007, Slovenia joins the Euro;
5. January 2008, Malta and Cyprus join the Euro;
6. November 26, 2008, 200bn European Economic Recovery Plan;
7. January 2009, Slovakia joins the Euro;
8. January 2009, Estonia, Denmark, Latvia and Lithuania join the ERM;
9. December 17, 2009, Greece hits deficit record;
10. April 19, 2010, Greece hits borrowing cost record;
11. May 2, 2010, Greece accepts 110bn bailout package;
12. November 28, 2010, Ireland accepts 85bn bailout package;
13. January 2011, Estonia joins the Euro;
14. February 14, 2011, agreement of 500bn ESM bailout fund;
15. May 3, 2011, agreement over 78bn bailout package for Portugal;
16. July 21, 2011, agreement over additional 109bn bailout package for Greece;
17. October 6, 2011, Bank of England injects additional 75bn pounds into the economy;
18. January 2012, major downgrade wave including nine Eurozone nations by S&P;
19. June 2012, Spain and Cyprus request assistance from the ESM;
20. January 23 2013, England threatens to leave the European Union;
21. May 2, 2013, ECB cuts the rate on its benchmark refinancing facility to 0.50%;
22. November 7, 2013, ECB cuts the rate on its benchmark refinancing facility to 0.25%;
23. June 2014, first negative interest rates by the ECB;
24. June 23, 2016, Brexit.



# Chapter 5

## Non-parametric Cross-sectional Nonlinearities

### Chapter Summary

The UN's Sustainable Development Goals for 2030 aim on one hand at inclusive growth and eradicating poverty, and on the other at preserving environments. The relation between development and the environment has been studied extensively since the 1990s, documenting inverted *U*-shaped relations between per capita income and indicators of environmental degradation. This paper revisits the issue with machine learning techniques and novel disaggregate data to model these relationships heterogeneously across economic indicators. Results suggest that development gradually improves the efficiency of consuming the earth's nonrenewable resources, but increased efficiency alone is not sufficient to offset growth in scale. Development shifts reliance on one nonrenewable source to another, and on average we find successive inverted *U*-shapes in deforestation, air pollution and carbon intensities, followed by a *J*-shape in per capita carbon output. Local economic circumstances further determine the shape, amplitude, and location of tipping points in environmental output. The general implications of the estimated dynamics are explored by extrapolating environmental output to 2030 under simplistic scenario's. The results are a reminder that immediate, and sustained global efforts are required to preserve our environment.<sup>1</sup>

---

<sup>1</sup>This chapter is based on a compilation of work. It draws from “*Environment and Development*” published by the *World Bank*, the full reference is Andree et al. (2019). An adapted version “*Revisiting the relation between economic growth and the environment; a global assessment of deforestation, pollution and carbon emission*” of the same authors together with Dr. Eric Koomen is published in the *Journal of Renewable and Sustainable Energy Reviews*. The reference is Andrée et al. (2019). The supplementary appendix is based on the technical background note associated with this publication, available here <https://doi.org/10.1016/j.rser.2019.06.028>. The material is reproduced here with kind permission from *Elsevier* and the *World Bank*.



## 5.1 Introduction

Will continuation of economic development increase pressure on the earth's finite resources, or does the increase in income provide the basis for environmental improvement? This question was central to several empirical studies in the early 1990s (Grossman and Krueger, 1991; Shafik and Bandyopadhyay, 1992; Panayotou, 1993). Their initial work suggested the existence of an inverted  $U$ -shaped relation between per capita income and environmental degradation indicators related to pollution (e.g.  $\text{SO}_2$ ,  $\text{NO}_x$ ), deforestation, and carbon emission. Their hypothesized environmental Kuznets curves gained massive following in research and policy as they held the attractive promise that economic development could actually benefit the environment. Early empirical examples of these curves, their possible explanations and policy implications are reviewed by Soumyananda (2004); Bo (2011). From 2000 onwards, however, substantial criticism was formulated in relation to the poor statistical foundation of these curves (Stern, 2004) and the obsession with replicating the exact inverted  $U$ -shape (Levinson, 2001). Stern (2004) points out that increases in wealth and income occur simultaneously with a structural transformation process in which the composition of inputs and methods of production gradually shift in favor of less destructive production. So it is not necessarily the increase in income that makes lower emission levels possible, but the gradual adoption of cleaner technology that can occur irrespective of development status (as documented by, for example, Stern and Common (2001); Dasgupta et al. (2002)). Environmental impact is thus determined both by efficiency of production, which may improve nonlinearly across GDP, and by total production size, which varies across panels of countries (Stern et al., 1996). If the scale of the economy is large, minute changes in the efficiency of production can result in large differences in output levels. Therefore, if a panel is constructed that includes economies of widely different scales, the variance in environmen-

tal output levels can be expected to vary with the GDP levels of the countries. To cope with this, one should acknowledge in the model design that environmental output is de facto a result of both a *scale* component, and a *technology* component.

Many empirical approaches have tried to model degradation levels directly, not distinguishing between the role of *scale* and a *technology* separately within the model, and therefore assume a degree of homogeneity in the emission-income relationship that is unrealistic for a panel of widely differing countries. A possible theoretical foundation for the environmental Kuznets curve, found in technological progress and diminishing returns to capital, is discussed by Brock and Taylor (2010). They highlight that modeling a panel relationship between emission levels and per capita income directly is not supported by their theory. Instead, they focus on combined panel data on emission intensities and abatement costs. Others have highlighted that, even when the regression deals with per capita emissions instead of levels, restricting cross-sections to undergo identical experiences over time biases results (List and Gallet, 1999). This suggests flexible approaches that allow for heterogeneous relationships may be more suitable as they allow for locally varying patterns to exist. Vollebergh et al. (2005) pay specific attention to homogeneity assumptions in their environmental impact regressions, and conclude that correctly modeling heterogeneity is essential to prevent spurious correlation in reduced-form panel estimations. Other econometric issues with environmental Kuznets curves relate to inappropriately dealing with the serial dependence and omitted variables bias Stern and Common (2001); Stern (2004). This has partly been addressed by adding control variables as in studies reviewed by Stern (1998), or by deploying fixed-effect approaches Stern (2004). Time series approaches that claim that the error correction approach provides appropriate diagnostic statistics and specification tests for the environment-economic relationship are also widespread (see, for example, Perman and Stern (2003); Stern (2004)). However, even over

time, linearity and constant variance assumptions break for moderate time dimensions because nonlinearities result from the income dependence of the derivatives of the function mapping changes in income to changes in degradation. From that perspective, the non-parametric error correction approach (Shahbaz et al., 2017) improves on previous work. More discussion on nonlinear cointegration in the context of the environmental Kuznets curve can be found in (Wagner, 2015). His general conclusion is that the diagnostics available in the standard framework are not appropriate in the nonlinear case because powers of integrated processes are themselves not integrated. In the non-parametric context on the other hand, causality and other correct-specification arguments are tightly related to the penalization technique, or bandwidth setting, that may take the limit criterion away from the *true* parameter.

In this paper, we revisit the empirical relation between economic growth and the environment using a panel data set on environmental indicators and economic development for a large set of countries applying a flexible kernel model that allows dependencies to vary smoothly throughout the data. We focus on a cross-comparable *technology* component represented by degradation intensities of average per capita wealth production to cope with the heteroskedasticity related to economic scales. The empirical strategy taken, pays tribute to the earlier literature that argued in favor of modeling outcome variables that are cross-comparable, and for using flexible models that allow relationships to vary throughout the data. To allow for a wide variety of potential nonlinearities with minimal parametric assumptions, we deploy a machine learning method that learns from similarities in the data using kernels. The method is known as Kernel Regularized Least Squares (Hainmueller and Hazlett, 2014). The key reason behind this choice is that apart from flexibility and taking full advantage of the kernel learning framework, it is still straightforward enough to back out marginal effects.

The framework is used together with out-of-sample selection of fixed effects to model remotely sensed and reported environmental data for 95 countries that include 85% of the world's population, 83% of global carbon output and 72% of all forest cover. The large sample of countries and simultaneous assessment of three environmental pressures along identical economic data using the same modeling strategy, differentiates our work from recent studies that use various methods to approach the relationship between economic development and specific environmental pressures in individual countries (e.g. (Managi and Jena, 2008; Keene and Deller, 2015; Apergis and Ozturk, 2015)), or the recent wave of research on carbon emissions in more limited samples of countries (e.g. (Apergis, 2016; Özokcu and Özdemir, 2017; Awaworyi Churchill et al., 2018)).

The remainder of this paper is as follows. We discuss estimation methods in Section 5.2. We provide only a non-technical discussion here, more technical discussion is provided in the supplementary background notes made available together with this paper. Section 5.3 details the data used for our empirical analysis in section 5.4. Finally in section 5.4.5, we use our empirical descriptions to explore the implications of continuation of growth on environmental output. Section 5.5 concludes.

## 5.2 Methods

In most explanatory analysis, the estimated model is assumed to consist of a finite set of parameters. While this makes the interpretation straightforward, it imposes strong assumptions about the behavior of the process being modeled. Specifically, linear models assume that the relationship between two variables  $Y$  and  $X$  described by a parameter  $\beta$  is constant across levels of  $Y$  and  $X$ . Such strong assumptions about the data generation process (DGP) are rarely - if ever - justified by economic

theory and can lead to seriously erroneous conclusions. The single parameter elasticities in linear models can at most approximate the average of the nonlinear elasticities locally on a function. Puu (1991) provides an excellent discussion on linear versus nonlinear dynamics in economics, and the key issue that linear approximations can be reasonable within bounds but should not be used to infer change so large that the bounds of the approximation interval are violated. Naturally, these bounds depend on the strength of the nonlinearity (Lorenz, 1993). Linear approaches can thus yield useful evidence, but only within a relatively narrow range of the overall state space, particularly if large parts of the population are likely to pass through that state. However, in general, local approximations fall short when building global arguments. Costanza et al. (1993) provide an excellent discussion on the severe limits of taking simple relationships from a local level and aggregating them up to describe the large-scale behavior of a complex system. The other way around, inferential errors induced by fixing the relationships in a complex system at the average, increase with the divergence between the average observation and the values of the observations of interest. This is undoubtedly the case in the analysis of economic development and environmental output, in which the structural behavior of the outliers, such as the poorest or most polluted, are often of foremost concern to policy makers.

Finite dimensional nonlinear parametric models may address several of these issues, but require strong predictions from underlying economic theory on the implied form of the structural relationship for the parameters to be economically meaningful. Such functions may be difficult to parameterize. Finite series approximators may also provide flexibility, but the resulting conclusions are often different from models in which the order of approximation is allowed to vary along the sample size, see Horowitz (2011) for further discussion.

Non-parametric models make fewer assumptions about the DGP, and

can produce approximations with varying flexibility (Härdle et al., 2004). The key question in this case is how flexible the empirical function should be given the data that has been observed. A regularized non-parametric model exposed to growing data is in a sense an approximator that adjusts its belief of what an appropriate description of the DGP is according to the number of observations that is seen. We apply a particular type of this model in this paper that is known as the Kernel Regularized Least Squares estimator developed by Hainmueller and Hazlett (2014). Key in this approach is that the model adjusts its understanding of the DGP as the data grows. A non-parametric model in which the size of the model is appropriately regulated results in a small size when samples are few, but may increase in dimensionality as the data grows. As a result, the approximation error declines with growing data. Regulation of the order of approximation in non-parametric models occurs through tuning parameters. Correct inference is therefore strongly dependent on values that are not estimated by the criterion, but instead set by the researcher. While the relationship between standard loss minimization and correct parameter inference, as in the linear Least Squares literature, is a basic concept well-known to many researchers, inference based on the estimators of a non-parametric model has to consider the effect of the external parameters for which results do not follow under the same consistency and normality theorems. For example, while Hainmueller and Hazlett (2014) provide consistency and normality results for their model, they state explicitly that these results are different for every level of penalization. We refer the reader to Andree et al. (2019) and the supplementary notes made available together with this paper for an in-depth discussion on this topic and a technical exposition of the model. We also provide more discussion there regarding the assumptions we make about the type of nonlinearities in the data generating process. For

the general reader, it suffices to say that our regression is of the form:

$$\mathbf{y}_t = h(X_t) + g(\epsilon_t) + \boldsymbol{\varepsilon}_t, \quad (5.1)$$

where  $\mathbf{y}_t$  is a vector of environmental degradation variables at time  $t$  which will be introduced in our next section,  $X_t$  is a matrix of economic variables at time  $t$  that are similarly introduced in the next section,  $h$  is a flexible function that is approximated using Gaussian kernels,  $\epsilon_t$  are time specific constants with  $g$  being some function that determines whether those time-specific error effects should be included, and  $\boldsymbol{\varepsilon}_t$  are vectors of residuals at time  $t$ .<sup>2</sup> The regression is estimated by minimizing Least Squares using a penalty that discourages overly complex results. The penalty is chosen using cross-validation. This also ensures that, as data grows, the estimated marginal effects can be interpreted as usual, as is explained further in the supplementary background notes.

### 5.3 Data

We combined measures of tree loss, air pollution concentrations and carbon emissions on one side, and GDP indicators of economic structure on the other. Our data is from a variety of sources and includes 95 countries measured over 1999 – 2014 containing approximately 85% of the world’s population, 83% of the world’s carbon output, and 72% of the world’s forest cover. We have removed areas below 1500 square kilometers – essentially all small island states – from the analysis. A summary of the data as it enters our regressions is given below.

---

<sup>2</sup>Note that  $\epsilon_t$  could be part of  $X_t$ , which is how we treat it in the appendix. We have written it here separately as an error component, which may be more recognizable to those that are familiar with the panel regression setting.

### 5.3.1 Forest cover

We use data from Hansen et al. (2013), which contains estimates of global tree cover extent (2000) and annual tree cover loss (2001-2014) at a spatial resolution of 30 meters.<sup>3</sup> They analyzed satellite images from Landsat 5, 7, and 8 to identify tree cover extent, defined as vegetation taller than 5 meters in height, and loss, defined as complete removal of tree cover canopy. The authors reported the tree cover loss data to have a false positive rate of 13%, a false negative rate of 12% and a ratio of total forest gain to loss over 2001-2012 of 0.34. The derived data differs from statistics reported by the UN-FAO's Forest Resource Assessment, but due to the consistent methodology and definition of forests across countries, we believe this data is better suited for a global analysis. We define forests as pixels with a minimum canopy closure density of 30%. Finally, we convert the data to area measures and sum the data by country to calculate tree cover loss as a percentage of tree cover extent in 2000. Our intention is to examine "natural dense forests", but note that the data also captures forest plantations. Our loss measure is thus only a proxy for deforestation, as there may be many other natural and anthropogenic processes (storm damage, fires, mechanical harvesting) that are reflected by the data. We refrain from including the tree cover gain data as there are significant differences in methodology that limit additivity or comparison with loss.

Forest cover loss included two outliers of respectively 10.8% and 5.4% loss in Namibia (2001, and 2005). For comparison, the median observation across time in this country was 1.7%. We have capped these numbers at 3% which, seemed an appropriate maximum for the range of forest loss after inspecting a kernel density. We applied a three-period simple moving average to further smoothen outliers. Our final data set includes

---

<sup>3</sup>The data can be found at [https://earthenginepartners.appspot.com/science-2013-global-forest/download\\_v1.2.html](https://earthenginepartners.appspot.com/science-2013-global-forest/download_v1.2.html).



the losses for countries that held 72% of forest cover in 2000. The largest missing forest patch is that of the Russian Federation.

### 5.3.2 Air pollution

We use concentrations of fine particulate matter ( $\text{PM}_{2.5}$ ), coarse dust particles of 2.5 micrometers in diameter, as a proxy for broader air pollution. The ( $0.01^\circ \times 0.01^\circ$  resolution) data is developed by van Donkelaar et al. (2016) and includes global annual ground-level  $\text{PM}_{2.5}$  (1999-2014) derived from a combination of satellite-, simulation- and monitor-based sources. The data set has been developed from satellite-derived Aerosol Optical Depth reflectance values calibrated to ground-based  $\text{PM}_{2.5}$  observations using a Geographically Weighted Regression.

Remote sensing methods aim to observe particulate matter but are prone to capturing fine dust released from barren lands that have similar reflectance properties in the high frequency spectral wavelengths. This poses a difficulty in our analysis, as countries in desert regions have high country wide average pollution levels, while large countries, or those with substantial forest cover where ambient pollution is low, have lower average concentration to what the larger population is exposed to on a regular basis. We used gridded population data that is produced using a combination of light at night data and census data, to identify patches of urban areas.<sup>4</sup> We averaged gridded pollution data that falls within urban boundaries to the country-level, defining urban areas as places where population density was higher than 300 people per square kilometer. The results in fig. 5.1 show that this procedure results in higher pollution levels in large countries with known pollution problems in cities (notably China, Nepal, Pakistan) or those with forests (notably Lao PDR, Indonesia, Senegal), and in lower concentrations in areas with

---

<sup>4</sup>The population grids are from <http://sedac.ciesin.columbia.edu/data/collection/gpw-v4>, we use the 2000 grids.



Over- or under-stating wealth does not pose a problem to the current analysis if the bias affects countries with similar GDP in a similar way, as we are purely interested in trends across wealth and not necessarily performing an unbiased wealth assessment itself. In the remainder of this paper, when we mention GDP we refer to its adjusted version and with an international dollar we refer to a single unit of GDP.

The CO<sub>2</sub> emissions estimates retrieved from WDI were produced by the U.S. Department of Energy's Carbon Dioxide Information Analysis Center (CDIAC) and include anthropogenic emissions from fossil fuel consumption and world cement manufacturing. The data set includes approximately 83% of global carbon emissions and should be quite representative of missing countries as it is close to the 85% of world population included in our samples.

### **5.3.4 Treatment of missing data**

Almost every ambitious analysis that aims to pull together various sources of data to produce insights supported by a wide range of observations, is eventually plagued by missing data of some form. The WDI data set contains a wealth of information, but some important observations are missing. Immediately, this poses a trade-off between using less, but complete, data, or using more data but having to cope with missingness by deploying an imputation strategy.

The predictive modeling community has generally found that using more information tends to result in better predictions (Kuhn and Johnson, 2013). The view is that the usefulness of the imputation can be inferred from looking at out-of-sample performance of the final model. Hence, this has led to a more relaxed opinion about using sparsely observed variables to build predictive models, and various approaches are widely available (Kuhn, 2008; Kowarik and Templ, 2016).

When interested in inference, however, it is important to understand the reason behind missingness. A common presumption is that imputations introduce additional uncertainty and possibly bias. However, complete-case studies can in fact lead to much more severe problems if the observations are missing for a reason (Westreich, 2012). A trivial example in the current context is the following one. If countries with extreme carbon emissions simply choose not to report them, then surely, we would underestimate carbon output if we drop those cases. While many remain cautious to use imputed data, complete-case analysis is in fact only unbiased under restrictive assumptions (Rubin, 1976), and our default view for better inference is to favor imputations.

Strategies to deal with (extreme) missingness are treated for example in (Little and Rubin, 2012; Graham, 2012; Salgado et al., 2016). A standard approach is the use of fully-conditional regression specifications to fill in missing data, e.g. using regression specifications based on all other available covariates (Audigier et al., 2018), and the most favorable method (but often computationally challenging in the nonlinear case ) is possibly that of using multiple imputations and pooling regression results (Rubin, 1996). This is of interest when one is also concerned about correcting the conditional variance function rather than only the conditional mean function. In either case, flexible modeling strategies tend to produce both improved prediction performance as well as better inference than linear imputation approaches (Murray, 2018). However, tractability and simplicity are not factored in when merely pitting different imputation approaches against one another based on simple diagnostics. For that reason, we adopt different approaches depending on the imputation case.

GDP has only .12% missing, manufacturing and services GDP shares have 5.57% missing. We interpolate these values linearly over the time dimension. The WDI does not report income shares in each country but sometimes reports a Gini index. We used this to back out income

shares. In particular, we make use of the fact that income shares held by a certain share of the population can be read off the Lorenz-curve and that the Gini coefficient is a measure of dispersion of the Lorenz-curve calculated from the summed surface under the Lorenz-curve and the surface under the  $45^\circ$  line. In total we are able to collect 937 observations of both Gini coefficients and income shares held by the first two quintiles. We estimate the nonlinear inverse map with high precision ( $R^2$  of .99) using the penalized non-parametric estimator. We then used the Gini observations to predict the income shares. After this first imputation step, 61.25% of the observations remain missing, but only over the time dimension. In the countries that have more observations in the time dimension, we observe that the income shares are relatively stable over time. We therefore simply interpolate the remaining missing values linearly over time.

Poverty rates has 69.54% missing values. A large part of the missing values are due to statistics that are not produced in high income countries. We fill all missing poverty and undernourishment rates above 23,000 GDP per capita with zero. The choice of this threshold is because undernourishment rates were not reported above this income value, and only Malaysia, with 24,500 GDP ppp per capita, had a positive reported poverty rate (1.3%). All other countries already attained 0% poverty rates in the published data. After this first imputation step, 49.54% of poverty remains missing because most countries are not complete in the time dimension. first interpolate these variables by taking a weighted average over time. This works well for most variables, but may yield poor results for poverty, as we have seen a tremendous improvement in most countries in past years. We improve the time dynamics in the interpolated poverty data by using information about time dynamics contained in our other variables. We vectorize the interpolated values, and fit the kernel model using the full set of undernourishment, logarithmic GDP per capita, the share of manufacturing, services, urban population shares,

and bottom 40 income shares. The model reaches an  $R^2$ 's of .91. We use this model to smoothen the interpolated poverty values by taking an average of the interpolated values and the values predicted by this nonlinear model.

A final caveat is in order. Ultimately, the data on poverty and income shares remains patchy at best. We are well aware that the sparsity and the heavily imputed nature of the variable used in the analysis may remain a controversy to some. To put the missingness of 61.25% and 49.54% of the bottom income shares and poverty rates in perspective, the World Bank and UN-FAO methodology for the calculation of the official undernourishment statistics is based on three-year linear moving averages of model results produced using Household Consumption and Expenditure Surveys that are taken every 3-5 years (Molledo et al., 2014). At this stage of knowledge, the objective of the research here is to further open up the empirical debate on the poverty-environment relationship. Recent initiatives such as the United Nation's Poverty-Environment Initiative highlight the importance of the relationship for policy, and recent research highlights the importance that poverty plays in the quality of the environment (Dogo et al., 2019). Since the hypothesis behind the environmental Kuznets curve is that poor countries may be more polluting, simply dropping incomplete cases would lead to a severe bias of the result.

### 5.3.5 Other controls and final data

We use the NDVI from the Moderate resolution Imaging Spectroradiometer (MODIS) derived from NASA's Terra satellite imagery to control for effects that relate to a variety of physical characteristics and natural assets of a country that may, for example, have an impact on ambient pollution levels or forest growth and loss dynamics.<sup>5</sup> This data set pro-

---

<sup>5</sup>Available at <https://modis.gsfc.nasa.gov/data/dataproduct/mod13.php>.

vides spatial and temporal comparisons of global vegetation conditions. The original data has a monthly frequency at a resolution of 1km. We calculated the mean NDVI value for each year in our analysis, using 2000-2015 data, and summarized the data to the country-level using the mean, minimum, and maximum value to get a broad description of the vegetation in a country.

Table 5.1 summarizes the data, all predictors are mapped into the  $[0, 1]$  interval to ensure the penalization effect is not driven by differences in variance of the different variables. We scale back the estimation results for easier interpretation.

Table 5.1: Summary of the data used in our empirical application. Statistics are not weighted and not necessarily representative of the world averages.

Statistic	Mean	St. Dev.	Min	Max
Annual % Tree loss	0.437	0.406	0.009	2.924
Urban PM <sub>2.5</sub> $\mu\text{g}/\text{m}^3$	18.924	12.609	0.311	63.498
CO <sub>2</sub> kg/\$	0.239	0.199	0.014	1.990
CO <sub>2</sub> ton p.c.	3.402	4.162	0.015	20.208
GDP ppp p.c. 2011 international \$	13,468.880	15,056.710	555.560	64,979.840
Population density, people / $\text{km}^2$	101.592	138.270	1.524	1,148.514
Undernourishment rate	15.514	13.546	0.000	64.500
Poverty 1.90\$ at 2011 international \$	21.177	22.040	0.000	84.740
Manufacturing GDP share	14.371	6.721	0.237	38.733
Services GDP share	70.065	12.734	29.279	93.881
Urban population share	53.685	22.754	12.082	97.818
Bottom 40% income share	15.970	4.130	7.510	28.024
NDVI annual mean	0.502	0.163	0.111	0.762
NDVI annual min	0.327	0.164	-0.027	0.657
NDVI annual max	0.655	0.161	0.170	0.862
Forest cover 2000 extent million ha	3.628	2.815	0.0005	9.883
Country area $\text{km}^2$	978,152	1,999,320	15,007	9,904,700

### 5.3.6 Transformation to degradation intensities

To address homogeneity concerns related to the scales of economies, we model standardized units of deforestation, pollution, and emissions, standardized per unit of GDP per capita in 2011 international dollars.

The choice to standardize degradation units by GDP per capita, and not by GDP, is for cross-comparability between countries of different economic size. In particular, using  $E$  to denote environmental pressure, the ratio  $E/GDP_{pc}$  is the environmental intensity of the average person's economic wealth, rather than the intensity of an international dollar. We favor  $E/GDP_{pc}$  over  $E/GDP$  because countries with larger populations will produce more international dollars and thus have lower intensities per dollar if everything else remains constant. The difference in efficiency of dollar production should not be used to suggest that average wealth production is environmentally more efficient. This is important because the hypothesis of the environmental Kuznets curve is that environmental pressure changes with increases of wealth, conventionally modeled as GDP per capita.

The cross-comparability is also statistically favorable because it reduces heteroskedasticity of the dependent variable across economic dimensions which allows to more robustly interpret variances, without the need of additionally modeling the conditional variance across covariates, such as discussed in Brown and Levine (2007). The general problem is that when the residuals vary strongly across the covariates, then apart from approximating a conditional mean function, one would need to approximate also the conditional variance function. The stabilization allows us to view the average standard errors as reasonable proxies, particularly given the additional approximation errors that a new non-parametric model of conditional variance would introduce.

Figure 5.2 at right shows that the variance in the log of the environmental intensities of GDP per capita is stable across GDP per capita, while the left plots with standardized intensities of international dollars contain widely differing variance. Note that the left-side is not log-transformed. While this would stabilize the data better, it does not change the relationship between the variance and GDP per capita.



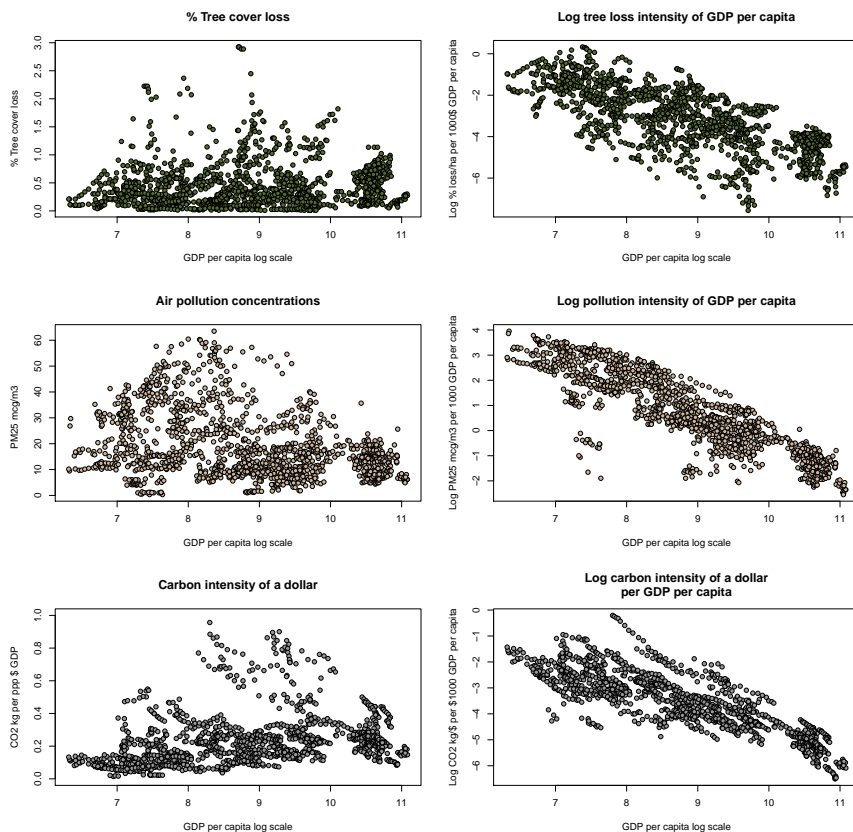


Figure 5.2: Observed degradation intensities and degradation levels across income.

Figure 5.2 immediately reveals a clear relationship between average income and the average environmental pressure for that level of average income. In particular, and unsurprisingly, the average log-linear trend is downward for all environmental pressures indicating that new per capita wealth is generated at lower environmental cost. For environmental pressure to go down on a net basis, we need that the emission intensity of average wealth declines faster than the increase in average wealth. Depending on the acceleration of one versus the other, various environmental output curves can result. Hence, the average trends in fig. 5.2 may

translate into inverted  $U$ -shapes in the environmental variables. However, since the left side is in logs, it means that minor deviations from the average may lead to large differences in degradation output explaining why inverted  $U$ -shapes are not directly visible in the left side.

## 5.4 Empirical results

Since, clearly, the environmental pressure of an average person's wealth decreases with GDP per capita, the empirical question is not whether the elasticity is non-zero, which is the question most regression analyses aim to tackle, but whether the elasticity implies that the intensity improves sufficiently fast across increasing income. Hence the empirical analysis of the Kuznets curve can be understood as an analysis of relative speed of change. In the following, we present the results in line with general models of the form

$$\log\left(\frac{E_{it}}{GDP_{pc_{it}}}\right) = \beta_{it}x_{it} + \alpha_i\epsilon_t + \varepsilon_{it}, \quad (5.2)$$

where the  $\beta_{it}$ 's are approximated by our non-parametric regularized kernel estimator. Nevertheless, the interpretation follows as one would usually interpret the above. In particular, table 5.1 lists the variables that enter the regression, including the control variables, and the tables below list how they entered the regression. For example, we use a log transformation of GDP per capita, hence the interpretation is that of a standard elasticity. For a 1 per cent increase in GDP per capita evaluated at  $it$ , the quantity  $\frac{E_{it}}{GDP_{pc_{it}}}$  is estimated to change by  $\beta_{it}$  per cent. This means that when  $\beta_{it} = -1$ , a percentage increase in average income is associated with an efficiency improvement of a percentage. For marginal changes, this means that output levels stay approximately unchanged since scale, measured in average income, also increases by a percentage. All the other effects follow log linear interpretation, which is straightforward since the variables already represent rates. For a 1 point

per cent increase, a  $\beta_{it} * 100$  per cent change in efficiency is expected, hence all parameters have an interpretation as elasticity.

The results show that the environmental output intensities are well explained by the data and that evidence for nonlinear dependencies is pervasive throughout all three models. Tables 5.2 to 5.4 show the marginal effects for individual models, summarized using the mean, quantiles and medians together with  $t$ -statistics.<sup>6</sup> For brevity, we have omitted the control variables in the tables.<sup>7</sup> All models have been checked for time fixed effects, but in all three cases the out-of-sample performance was optimal in the models without fixed components. The appendix contains conditional expectations together with confidence bands for each of the economic variables, holding the effects of all other variables constant at their mean values, which provide guidance throughout our discussion of the results. In particular, while the tables present the range of parameter estimates, it is not immediately possible to understand how effects of variables change along the levels of those variables. The conditional expectation plots plot the expected values for the outcome variables along the levels of individual covariates, and can therefore provide a sense of the ordering of the local elasticities. Variables for which the marginal effects within the inner 50% of the percentiles range have an identical and significant sign are highlighted in the tables. This reveals that many of the variables contribute both positively as well as negatively to the output intensities depending on the data levels at which effects are evaluated. This shows that nonlinearities are important. We find that income has an unambiguous effect, all three environmental intensities improve with income but not sufficiently to offset scale growth.<sup>8</sup> While increases in

---

<sup>6</sup>We obtained our results using the *R* implementation of KRLS. Our out-of-sample shrinkage strategy is not implemented by default, and requires many model fits. We found that an optimized BLAS/LAPACK implementation provided better speed than the *C++* implementation of bigKRLS.

<sup>7</sup>Annual mean, minimum and maximum NDVI values, forest cover, and country size

<sup>8</sup>The log-log specification allows for a simple interpretation. To offset the scaling effects, the marginal effect of log GDP per capita needs to be smaller than -1, which we do not observe within the 25%-75% range of effects.

GDP provide a basis for the improvement of production efficiency, it appears not to lower the net environmental output. However, as GDP increases, a structural change occurs in which poverty goes down and the share of manufacturing services, and urban population gradually increase. We will highlight several of these structural effects that are best visualized in fig. 5.7 and fig. 5.8. Figure 5.8 shows how poverty, the production composition, urban population shares, and the income distribution trend across GDP.

#### 5.4.1 Individual model results

The results for deforestation show that early increases in population density correlate with a decrease in deforestation intensity while high population densities correlate with an increase. The trend across urban populations is a weak inverted-*U*. The effects of manufacturing and services are less ambiguous, the move out of an agricultural society and specifically an increasing share in services that occurs with increasing GDP, is a strong correlate of declining deforestation rates. There is some evidence that economies with an unequal income distribution retain a higher deforestation intensity of production. The effect of the poverty variables is, however, mixed as the semi-elasticities contain both negative and positive values. The conditional expectation plot in the appendix also visualizes that there is a very mild *U*-shape along poverty rates, with deforestation efficiency slightly going down again as countries move below 20 per cent. Reducing the undernourishment rate, in an opposite manner, initially seems to increase deforestation, while the transition out of extreme poverty correlates with a decrease in deforestation intensity. In contrast with the deforestation results, we see that an increase in population density unambiguously drives pollution up. The pollution intensity trend across the urbanization rate is initially flat, but after 50% of the population has urbanized, the trend becomes negative. This

Table 5.2: Deforestation intensity results using the penalized kernel regression.

	(means)	(25%)	(50%)	(75%)
<i>Dependent: Log deforestation intensity of 1000 GDP p.c.</i>				
Log 1000 GDP per capita**	-0.453*** (-16.962)	-0.610*** (-22.856)	-0.476*** (-17.83)	-0.289*** (-10.822)
Population density	-0.001*** (-5.351)	-0.003*** (-18.303)	-0.002*** (-9.578)	0.001*** (8.449)
Undernourishment rate	0.001 (0.371)	-0.008*** (-4.48)	0.002 (0.919)	0.011*** (6.171)
Poverty 1.90\$ rate	-0.005*** (-4.118)	-0.013*** (-10.353)	-0.004*** (-2.878)	0.005*** (4.297)
Manufacturing GDP share	-0.015*** (-5.459)	-0.055*** (-19.402)	-0.016*** (-5.551)	0.023*** (8.04)
Services GDP share**	-0.032*** (-16.57)	-0.051*** (-26.109)	-0.036*** (-18.436)	-0.016*** (-7.961)
Urban population share	0.006*** (5.597)	-0.008*** (-6.929)	0.009*** (7.767)	0.019*** (17.111)
Bottom 40% income share*	-0.027*** (-5.416)	-0.06*** (-12.304)	-0.033*** (-6.671)	0.003 (0.55)
N = 1520      R <sup>2</sup> = 0.922      λ = 0.691.      *p < .1; **p < .05; ***p < .01 <i>Constant omitted, t-statistics in parenthesis. Optimal model contained no fixed effects.</i> <i>Model controls for mean, min and max NDVI, forest cover, and country size.</i> * Approximately 50% of inner marginal effects same sign, but range includes zero. ** Inner 50% of marginal significantly excludes zero.				

Table 5.3: Pollution intensity results using the penalized kernel regression.

	(means)	(25%)	(50%)	(75%)
<i>Dependent: Log pollution intensity of 1000 GDP p.c.</i>				
Log 1000 GDP per capita**	-0.691*** (-47.899)	-0.842*** (-58.388)	-0.692*** (-48.026)	-0.567*** (-39.307)
Population density**	0.002*** (18.523)	0.001*** (6.658)	0.002*** (17.063)	0.003*** (30.846)
Undernourishment rate	-0.002** (-2.172)	-0.008*** (-8.675)	-0.003*** (-3.256)	0.003*** (3.267)
Poverty 1.90\$ rate*	0.003*** (4.493)	0.000 (0.371)	0.003 (4.842)	0.008*** (11.603)
Manufacturing GDP share	-0.009*** (-6.436)	-0.023*** (-15.78)	-0.01*** (-6.756)	0.004*** (2.492)
Services GDP share*	-0.007*** (-7.00)	-0.012*** (-11.68)	-0.007*** (-6.753)	-0.001 (-1.286)
Urban population share	-0.006*** (-10.746)	-0.015*** (-24.929)	-0.008*** (-14.101)	0.001** (2.547)
Bottom 40% income share	-0.019*** (-7.776)	-0.045*** (-18.221)	-0.015*** (-6.086)	0.012*** (4.682)
N = 1520      R <sup>2</sup> = 0.978      λ = 0.691.      *p < .1; **p < .05; ***p < .01 <i>Constant omitted, t-statistics in parenthesis. Optimal model contained no fixed effects.</i> <i>Model controls for mean, min and max NDVI, forest cover, and country size.</i> * Inner 50% of marginal effects same sign, but range includes zero. ** Inner 50% of marginal significantly excludes zero.				

Table 5.4: Carbon intensity results using the penalized kernel regression.

	(means)	(25%)	(50%)	(75%)
<i>Dependent: Log carbon intensity of 1000 GDP p.c.</i>				
Log 1000 GDP per capita**	-0.630*** (-42.341)	-0.755*** (-50.71)	-0.635*** (-42.699)	-0.519*** (-34.903)
Population density	-0.000*** (-4.812)	-0.001*** (-11.919)	-0.001*** (-5.656)	0.000*** (3.356)
Undernourishment rate	0.006*** (5.927)	-0.001 (-0.523)	0.008*** (8.248)	0.014*** (14.741)
Poverty 1.90\$ rate	0.002** (2.475)	-0.002*** (-3.295)	0.001** (2.092)	0.008*** (11.496)
Manufacturing GDP share*	0.017*** (10.815)	0.001 (0.366)	0.019*** (12.344)	0.036*** (23.015)
Services GDP share	0.001 (1.198)	-0.007*** (-6.708)	0.002 (1.667)	0.01*** (9.01)
Urban population share	0.002*** (2.826)	-0.005*** (-7.41)	0.002*** (2.995)	0.008*** (12.768)
Bottom 40% income share	0.006** (2.269)	-0.018*** (-6.492)	0.002 (0.825)	0.026*** (9.297)

N = 1520

R<sup>2</sup> = 0.956 $\lambda = 0.635$ .

\*p &lt; .1; \*\*p &lt; .05; \*\*\*p &lt; .01

*Constant omitted, t-statistics in parenthesis. Optimal model contained no fixed effects.**Model controls for mean, min and max NDVI, forest cover, and country size.**\*Inner 50% of marginal effects same sign, but range includes zero.**\*\*Inner 50% of marginal significantly excludes zero.*

indicates that early urbanization is polluting, but that after reaching a tipping point, the city environment becomes cleaner. The trends across manufacturing and services are also primarily downwards. Agricultural societies have a higher pollution intensity of income, while a shift into manufacturing and services reduces the environmental output per unit of production. It remains difficult to say whether the effects reduce pollution on a net basis as this structural transformation occurs jointly with an increase in total productivity. However, for an identical amount of total GDP produced, the data seems to suggest that an agricultural economy produces the highest amount of air pollution. An economy with a high manufacturing share produces less pollution, while an entirely service orientated economy outputs the lowest amount of pollution. This may also relate to a differential in value produced by these sectors which may imply different quality of production processes and differential in the total amount of economic activity for a fixed level of GDP. Across poverty

and undernourishment, we see hyperbolic effects that suggest that the eradication of extreme hunger occurs jointly with an increase in pollution intensity while later poverty eradication eventually occurs jointly with a reduction in pollution intensity. Poverty rates are unambiguously correlated with higher pollution intensities. Again, similar to the deforestation results, it seems that societies with high income inequality are also more polluting.

Carbon intensities also trend with urbanization. We find that the carbon intensities initially increase together with the urbanization process, however after the 50% urban population tipping point, the environment becomes more efficient in carbon consumption. The shift in production composition trends oppositely with those of deforestation and manufacturing. High manufacturing and high services share in the production composition both correlate with higher carbon emission intensities. The initial decline in undernourishment rates occur together with improvements in the carbon emission intensities, poverty reduction however trends with an increase. Finally, we see that equality – a stronger bottom 40% - increases carbon output when everything else is held constant, which is again an opposite trend of what we observed for deforestation and pollution.

#### **5.4.2 Heterogeneity in environmental output**

Combined, the results show that income and poverty reduction provide a basis for improvements in the efficiency of economies in their use of finite resources. The economic composition is not unambiguous in its effects. To understand how structural transformation, together with urbanization, poverty reduction and increases in total production, interplay to produce a commonality in environmental output trends, we track the model predictions keeping the control variables at their means. We also keep the income distribution fixed at a mean value as it does not trend clearly with

GDP as seen in fig. 5.8, and keep population densities fixed at means.

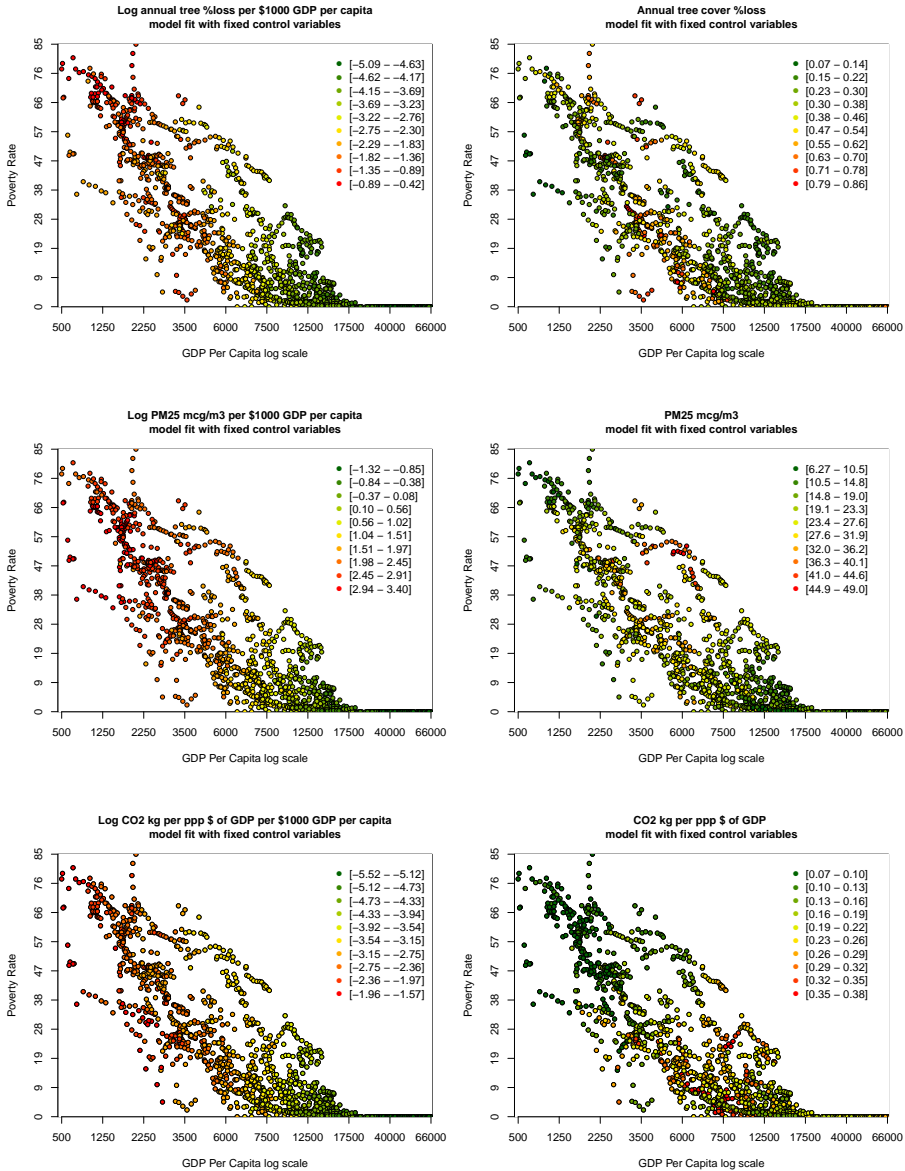


Figure 5.3: Model fits of degradation intensities of log GDP (left) and the rescaled environmental output levels (right) across poverty and income. Population densities and income equality as well as the control variables are held constant at the mean.



Figure 5.3 shows the prediction surfaces using poverty and income as Cartesian coordinates. The model predictions fit the output levels well after scaling the log intensities, see fig. 5.9. While this shows that all countries gradually grow out of poverty and improve their efficiencies following a common pattern, it also reveals that there is significant heterogeneity in the environmental output intensities that relates to differences in poverty and hunger rates, urban population shares and GDP composition. This highlights that the shape of the environmental Kuznets curve strongly depends on the development path of a country across all its dimensions. Furthermore, while the progression in output intensities follows a similar path, slight deviations from the local average may result in large differences in total environmental output. This reveals that while different development paths may relate to relatively small differences in the environmental output intensities, it may produce rather large differences in actual forest loss, air quality and carbon emissions depending on the scale of the economy.

An important takeaway is that heterogeneity in the actual output levels (right), is primarily large around the income levels where output is also highest (around \$4,000 for deforestation, \$6,000 for pollution, and \$8,000 for the carbon weight of a single dollar production value). This indicates that the theorized environmental Kuznets tipping points are also the points at which an averaged result, such as obtained from a linear regression, provides the poorest indication of relationships at the individual country level. While a few general rules could be extracted from the marginal effects, such as inequality, income and population density effects, the larger part of the environmental data seems to relate heterogeneously to economic variables.

### 5.4.3 Average curvature

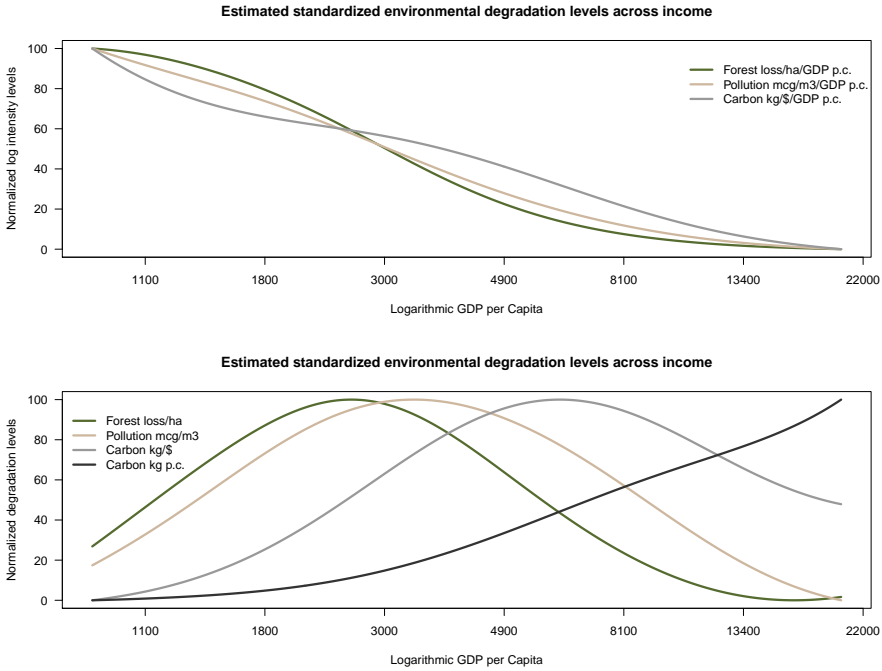


Figure 5.4: Normalized predicted environmental output levels across income. Predictors are held at expectations conditional on GDP. The  $R^2$ 's from logarithmic GDP per capita to poverty, undernourishment, manufacturing, services and urban population shares are respectively 0.801, 0.633, 0.142, 0.573 and 0.739. The conditional expectations are plotted in fig. 5.8. Population density, income equality, and controls are held at their means.

The heterogeneity in amplitude, and location of tipping points, conditional on the economic variables, implies that a single Kuznets Curve, such as it has often been treated in the literature, is a description that applies only poorly to individual country cases. However, to do some justice to the classical concept we can still construct an average development path and explore how the models fit environmental outputs to that. To do so, we derive conditional expectations for poverty, undernourishment, GDP composition, and urban population shares, using only the logarithmic GDP per capita as an explanatory variable. We then use these conditional

values to build a data set that includes all variables as local averages along with GDP itself. Again, we keep the control variables and the income distribution as well as population densities fixed. We normalize the results to compare the slopes and location tipping points across income.

Figure 5.4 shows the curvatures associated with these development paths. We have dropped the lower 2.5% of GDP observations, and the upper 20%. We focus on this range because of its particular relevance for development policy. We note that the maximum total output associated with the average tipping point is an interesting statistic, but due to heterogeneity this may be a poor approximate to predict whether a country is close to its potential tipping point after observing only environmental output. The deforestation rate associated with the average development path attains a maximum of .66% annually, while that highest pollution concentration maxes at  $28.7 \mu\text{g}/\text{m}^3$  and the carbon weight of a dollar reaches 0.271 kg.

#### 5.4.4 Heterogeneity in curvature and tipping points

The average pathways accurately describe the transition out of poverty, but they provide less insight into the effects if the economic composition changes. To better understand the importance of deviations in transitional variables, we plot the degradation levels associated with the average development path with additional differences in manufacturing shares, urban population shares and poverty rates.

Figure 5.5 shows that changing these variables, while keeping everything else at the local averages, has important impacts on the location, shape, and height of tipping points. For example, increasing the share of manufacturing by 10 points, shifts the tipping point of deforestation to the left, while economies that retain high agricultural shares reach a tipping point at higher income. This implies that an earlier transition out of agriculture may prevent high deforestation rates at higher income

and lower pollution levels at its peak. This is a slightly counter intuitive result as manufacturing has traditionally been portrayed as the main source of pollution. However, since our data only indicates the share of manufacturing in total GDP and not the quality or quantity of goods produced, higher rates may also correspond to differences in the number of manufacturing sites and the methods of production used.

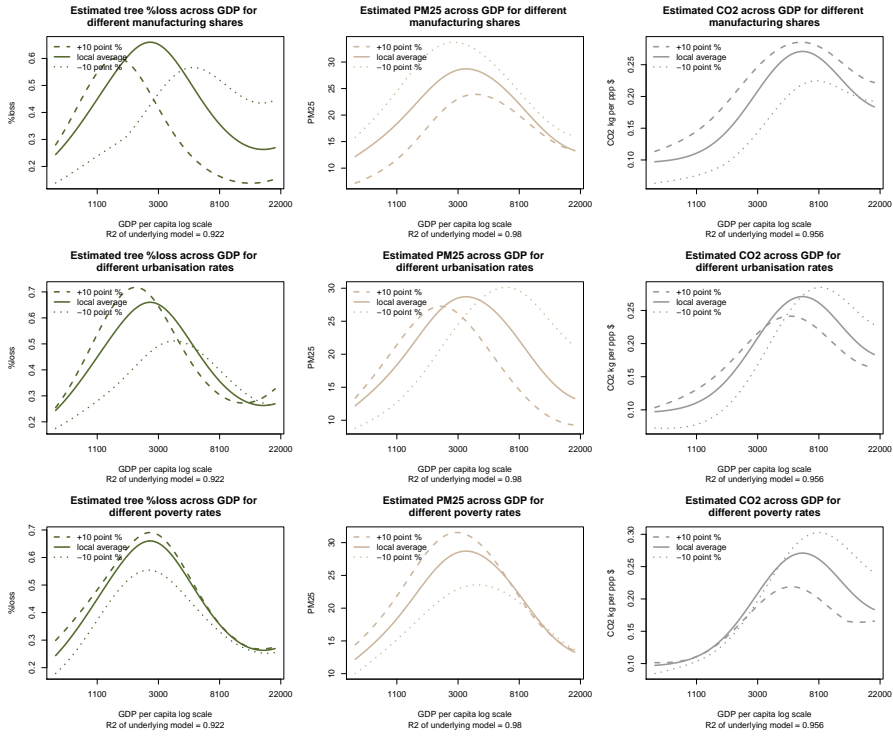


Figure 5.5: predicted environmental output levels across income. Predictors are held at expectations conditional on GDP and one variable has been incremented  $\pm 10$  points in each plot. The local average trend is identical to those in fig. 5.4.

The carbon emissions associated with this structural change are higher, suggesting a trade-off between pollution-heavy and carbon-intense production. In a similar fashion, poor countries that have a high urbanization rate have higher deforestation rates and reach a pollution peak faster. Poor countries that have lower urbanization, on the other hand, even-

tually maintain higher pollution and carbon emissions levels at higher income. This suggests that the draw-down in pollution output is not just a matter of income and productivity, it may relate to attaining critical urban population mass combined with increased income. The effects of poverty, finally, do not impact the location and shape of the environmental output levels. Countries with high poverty rates unambiguously deforest and pollute more, but emit less carbon.

#### **5.4.5 Exploring degradation dynamics under simple 2030 scenario's**

To explore whether continuation of current growth can be expected to lower environmental outputs without intervention, we extrapolate GDP into the future and calculate associated model responses under three simplistic scenarios of growth. These explorations are intended to further assess the potential impacts implied by the relationships that are captured by the models. They are by no means an attempt at accurate forecasts of future developments as these will be driven by a wide range of factors and events that are not part of historical data (e.g. unforeseen technological developments, changes in societal preferences, policy agendas). The estimated models can still be applied, however, to sketch how future environmental pressure may advance under current economic and population growth trends in the absence of policy interventions, or new technological successes, based purely on historical relationships. This still provides relevant indications of the magnitude of efforts required to meet environmental objectives.

In a base scenario, this analysis lets each sovereign grow at individual median 1999-2014 compound rates, with the highest growth rate capped at the 90% percentile (5.27% annually). In a pessimistic future, each country continues at one asymmetric deviation unit (-3.67%) below the base rate, and in an opportunistic growth scenario, countries continue

one asymmetric deviation unit (+0.89%) above the base rates.<sup>9</sup> In the opportunistic scenario, rates are capped at the 95% percentile rate (5.87% annually). Finally, we construct Business As Usual (BAU) as the average of the three results to balance between possible asymmetry. Table 5.5 summarizes the growth rates assumed in these simple scenario's.

To limit complexity, we keep population growth slightly below individual country median compound rates, resulting in 8.5 billion people by 2030 (in line with United Nations projections).<sup>10</sup> We use extrapolated GDP levels to derive fits for poverty and undernourishment levels, and the GDP composition using univariate model fits of the penalized kernel model. We let the urban population depend additionally on log population densities.<sup>11</sup> At each point in extrapolated time, we compare the conditional expectations to the predictions of our base year (2014), and compute the percentage change that we then multiply with observed 2014 values. We keep all data points within observed intervals, including a cap on the sum of agriculture and services shares. This means that effectively after reaching a level of \$64,980 per capita, our projection halts both the income effect on efficiency improvements and the effect of scale increase on the environmental output for a country (again, highlighting that this analysis does not consider future technological successes beyond what has been achieved by societies so far). In the pessimistic scenario, this does not affect any individual country result, in the base scenario this fixes Norway's output at current levels and caps those of the U.S. and Switzerland in respectively the last 4 and 5 years of the projection (final

---

<sup>9</sup>Asymmetric deviation units have been calculated as the difference between the median and respectively the 25% and 75% quantiles of growth rates. In the calculations we have dropped the two largest outliers (in absolute value) for each country.

<sup>10</sup>We reduced the population growth rates by 0.05 times the absolute point percentages globally to reduce growth everywhere, then reduced population growth rates by an additional .15 times the percentage rates in the top 40% income countries and additional .25 times in top 20% income countries. This simple scenario is designed to represent relatively higher growth in lower income countries and a slow down in developed countries, in line with UN projections.

<sup>11</sup>The  $R^2$ 's of the models are, 0.632 for undernourishment, 0.785 for poverty, 0.142 for manufacturing, 0.573 for services, and 0.814 for urban population shares. The uncertainty of the impact of changes in manufacturing remains high in our results.

9 and 10 years in the high growth scenario).

Figure 5.6 presents our results at the global level made by aggregating all country-level results and assuming that the average in-sample trend scales appropriately with missing areas. Results are also available for income segments in table 5.6. In the average scenario, global extreme poverty falls below 7.4% of the global population. The poorest 20% countries in our sample have stronger successes and go from 45% poverty to just over 33%.

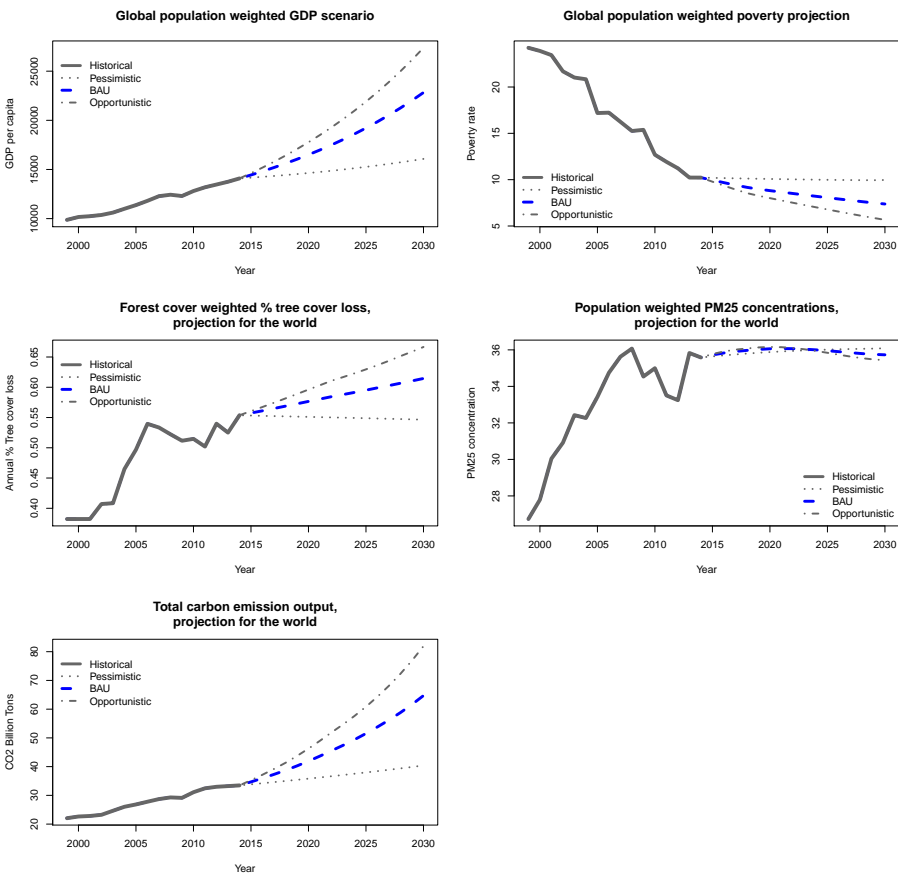


Figure 5.6: Projection for global environmental outcomes. Table 5.6 contains aggregate statistics of the BAU line and highlights the distributional changes across income.

While poverty reductions and GDP increases may improve livelihoods through economic gain, air pollution remains a serious threat to wellbeing, as the average global citizen remains exposed to  $36 \mu\text{g}/\text{m}^3$ , nearly twice the WHO prescribed guidelines. Additionally, development comes at a cost of an annual carbon output that reaches 63GT which is nearly a doubling from the 2014 levels, and a total loss of 242 million hectares of forested land. About 58% of forest loss is in countries with poverty rates above 3% or income in the bottom 60%. Many of those countries have tropical rain forests with slow regrowth rates estimated at 27% (Hansen et al., 2013). Using those statistics, loss in these countries totals 136 million ha between 2014 and 2030, netting over 3.4% of the global 2000 dense forest cover. Other insights include that success in eradicating poverty likely slows as China and India near 0% poverty and populations in poor countries grow faster than those in the developed world. Our modeled data does not signal that development alone will result in successful slowdown in natural capital depletion. At the global level, results suggest ongoing increases in global deforestation rates and carbon emissions. Global pollution exposure stabilizes regardless of the growth scenario, the results instead suggest a distributional shift toward lower income countries with improving and worsening conditions balancing out at the global scale. Air pollution concentrations rise by 28% in the bottom 20% income countries. Table 5.6 shows that the entire bottom 30% income countries of our sample in fact continues to face increasing pollution exposure. Projections of forest cover and carbon emissions on the other hand, are heavily dependent on the economic outlook. Growing wealth in the developing world together with rapid population growth may accelerate future global carbon output. A more extensive discussion on differences in trends in relation to development, and a breakdown of carbon output along income quintiles is provided by Andree et al. (2019).



## 5.5 Discussion and conclusion

In this paper, we estimated a penalized non-parametric model of environmental output across economic development. This type of modeling works well for nonlinear processes that do not result in overly complex dynamics. We deployed the framework to study environmental data in a panel of 95 countries. We modeled satellite-derived deforestation and air pollution levels and reported carbon emissions. To deal with heteroskedastic variance, we transformed our data to logarithmic degradation intensity of per capita GDP. We used a cross-validation approach to decide which fixed effects should be part of the model and this did not support the inclusion of time fixed effects.

Our results suggest that production gradually favors conserving the earth's finite resources as GDP increases, but that this alone is not sufficient to offset the scale effect of growth. Instead, structural change in the economy shapes environmental output curves. This process shares similarities between sovereigns, but remains largely heterogeneous. These results do not support a single environmental Kuznets rule. Instead, the results emphasize the importance of local economic conditions on environmental results. Across all data levels, some effects hold unambiguously. Poverty and income inequality correlate with higher pollution, higher deforestation, and lower carbon emissions; agricultural GDP shares correlate with deforestation; population densities correlate with pollution; and higher manufacturing shares correlate with increased carbon emissions. We find various tipping points in other variables, notably across urbanization rates. While local conditions may be unique, average development is associated with an inverted *U*-shape in deforestation, pollution and carbon intensities of production units. Per capita carbon emissions, however, follow a *J*-curve as the increase in per person productivity is not sufficiently offset by efficiency improvements. Disregarding the level of per capita GDP, we observe that at least one form of natural

capital degradation is high, conflicting with the belief that countries tend to "clean up" as they develop. One could argue that the scope of the impacts of externalities to production increases with development, with the burden falling to increasingly distant households both in time and space. Although local air pollution may be more intrusive on daily life, the consequences of climate change will remain globally impactful for generations to come.

We extrapolated our descriptions forward in time to highlight the daunting implications of development under continuation of current practices without improving policies. Our results are generally in line with emission paths associated to the high radiative forcing scenarios considered in IPCC's 4.9°C world (RCP 8.5). Our projections did not indicate successes on the fronts of reducing deforestation. Air quality improves in some currently severely polluted places, but worsens in poor regions.

In our results, deforestation follows an inverted *U*-shape across average development in the developing countries. This confirms and extends recent results from Crespo Cuaresma et al. (2017) that provide evidence for a partial environmental Kuznets curve for forest cover at low income. However, we find that economic growth alone is not sufficient to halt forest loss, and we find evidence that within the bottom 60% income countries, deforestation shifts to the bottom 30%, and that countries within the top 40% income do not fully stop deforesting. Others have similarly detailed forest loss in high-income countries, for example in the United States (Sleeter et al., 2012). Future efforts should also aim to understand forest regrowth dynamics across economic development, as we have only used average forest growth rates over the entire study time period as a control variable in our models, rather than investigating how regrowth possibly changes conditional on economic indicators. Generally, the temperate zones have much better regrowth rates. Taking this and projected increases in the bottom 20% into account, the African

forests seem to be at increased risk as economic successes in these areas accelerate, while the Amazon faces only marginal improvements in the immediate future in our modeled projections. Generally, deforestation is related to the economic value of land. Urban land, for instance, can be valued hundred times higher than forest land in the same area (Alig and Plantinga, 2004). Agricultural land usually also yields higher returns and policies focused on protecting forest could address this value gap (e.g. (Hyde et al., 1996)). The payment for ecosystems services schemes may provide an opportunity to conserve essential natural resources while providing an income source to landowners. However, the governance and targeting of these programs must be carefully addressed in order protect both the resources and livelihoods of those dependent upon them (Landell-Mills and Porras, 2002; Grieg-Gran et al., 2005). Extending agricultural subsidies to include renewable perennial crops has the potential to make cleaner alternatives competitive without negatively impacting farmer income or the need to increase aggregate subsidy spending, and could be a way to ensure that environmental damages are at least in part reconciled by positive externalities (Andree et al., 2017b). Other policy interventions that address forest-cover loss can focus on conservation and land-use protection, sustainable forestry, and urban growth boundaries (Alig et al., 2010). The efficacy of these interventions will likely rely upon the local circumstances surrounding forests and nearby populations, yet the potential benefits may be felt globally.

On the pollution side, our model projects rising  $PM_{2.5}$  levels in the lowest 30% income countries, with a general decrease in  $PM_{2.5}$  in middle income countries.  $PM_{2.5}$  remains far above WHO air quality guidelines in many countries, particularly in lower and middle income groups. Given population growth, these levels will expose more people to pollution-related health risks. Currently, about 90% of the global population is exposed to air quality that does not comply with the World Health Organization's Air Quality Guidelines (World Health Organization, 2016).

Tallis et al. (2018) expect that by 2050, business-as-usual development will result in over 4.8 billion people living in countries with worse air quality than in 2010. As a comparison, in our average modeled data 52% of people currently live in places where air quality has worsened by 2030. This totals to approximately 4.4 billion by 2030. Exposure to unsafe levels of particulate matter is estimated to increase the number of premature deaths related to air pollution in coming decades, killing 4.5 million people (or more) by 2030 (International Energy Agency, 2016; Lelieveld et al., 2015). Currently,  $PM_{2.5}$  levels peak in middle income countries, and while pollution levels can generally be expected to decline in these countries as their income levels grow, pollution levels will still remain dangerously high in this group. These countries include highly populated areas such as in China, India, and Bangladesh, which have already been identified as hotspots for adverse impacts of air pollution in the coming decades (Organisation for Economic Cooperation and Development, 2016; Pozzer et al., 2012). Eradicating poverty in these places may be one contribution to lowering extreme pollution, but unfortunately there is also evidence that points out that low-income households are also those that are more severely affected by pollution in economic terms Andree et al. (2019).

On the carbon end, our results suggest emission levels that could lead to the high radiative forcing scenario in IPCC's 4.9°C world (RCP 8.5) are largely in line with business-as-usual development in which developing countries follow in the footsteps of wealthier countries. Worse scenarios may in fact be considered as relevant possibilities. Specifically, this could occur if developing countries do not successfully manage to adopt cleaner technologies, or if high income countries revert (part of) their pro-climate policies. Recent studies suggest we are not alone in such a conclusion. See for example Peters et al. (2012) and comments, suggesting - in line with our findings - that reported successes in carbon reduction are short-lived and largely relate to the 2008-2009 crisis and aftermath. Emissions rapidly

increased in many places with the recovery. Furthermore, Peters et al. (2013) and comments thereon reveal that recent emissions continue to track the high end of suggested emission scenarios, making it increasingly unlikely that global warming will stay below  $2^{\circ}\text{C}$ . This is in line with our result that continuing current development puts the world on emissions associated with a  $4.9^{\circ}\text{C}$  pathway. This is further substantiated by the conclusion that developments on the fronts of negative emissions are required to reach a  $2^{\circ}\text{C}$  future Gasser et al. (2015). Combined, the evidence suggests that a worst-case scenario over  $4.9^{\circ}\text{C}$  in 2100 is both not unrealistic and overlooked in both the scientific community and the political arena.

## 5.6 Appendix

### 5.6.1 Additional results and figures

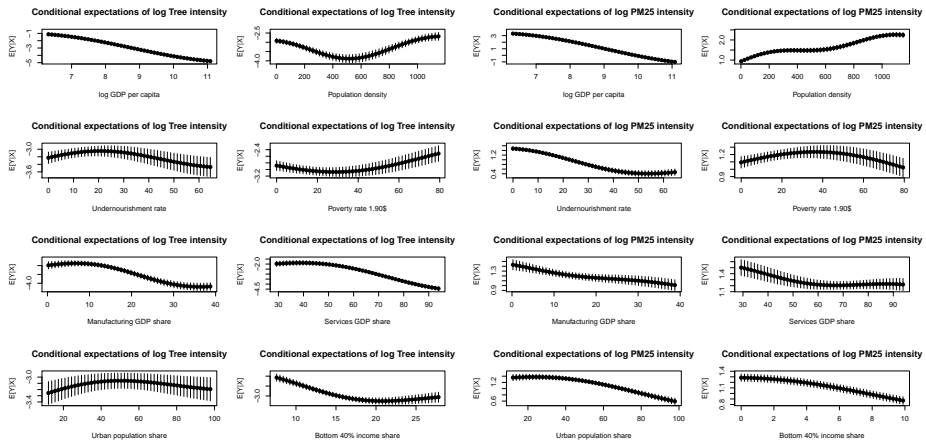


Figure 5.7: Conditional expectations of deforestation (left two columns) and pollution (right two columns) intensity of income for each variable fixing other variables constant at their mean.

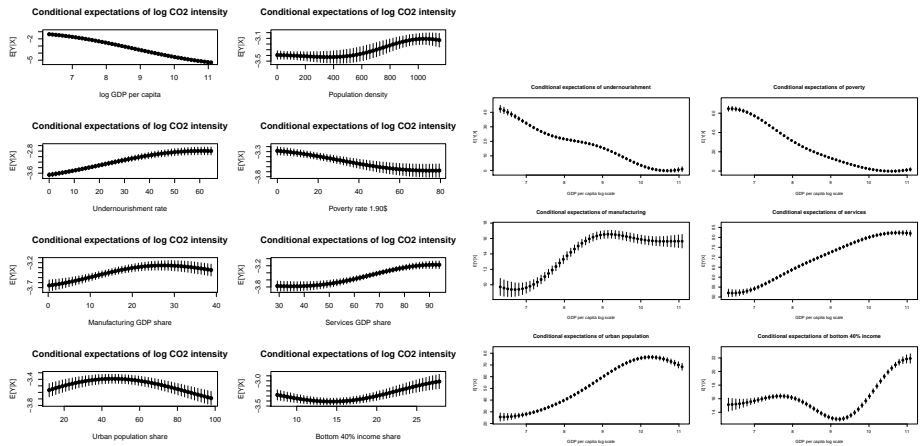


Figure 5.8: Conditional expectations of carbon intensity of income for each variable keeping other variables constant at their mean (left two columns).

Table 5.5: Summary of base rates in percentages by 5-percentiles used in the projection.

	GDP	Population
1	1.33	2.57
2	2.78	2.54
3	2.22	2.28
4	2.93	2.10
5	3.02	1.99
6	2.49	1.98
7	3.69	1.40
8	2.55	1.58
9	2.62	1.51
10	1.67	0.96
11	3.62	0.96
12	2.31	1.31
13	2.97	0.83
14	2.82	1.04
15	2.93	1.15
16	3.85	0.85
17	1.10	0.46
18	1.47	0.30
19	1.47	0.23
20	1.24	0.49

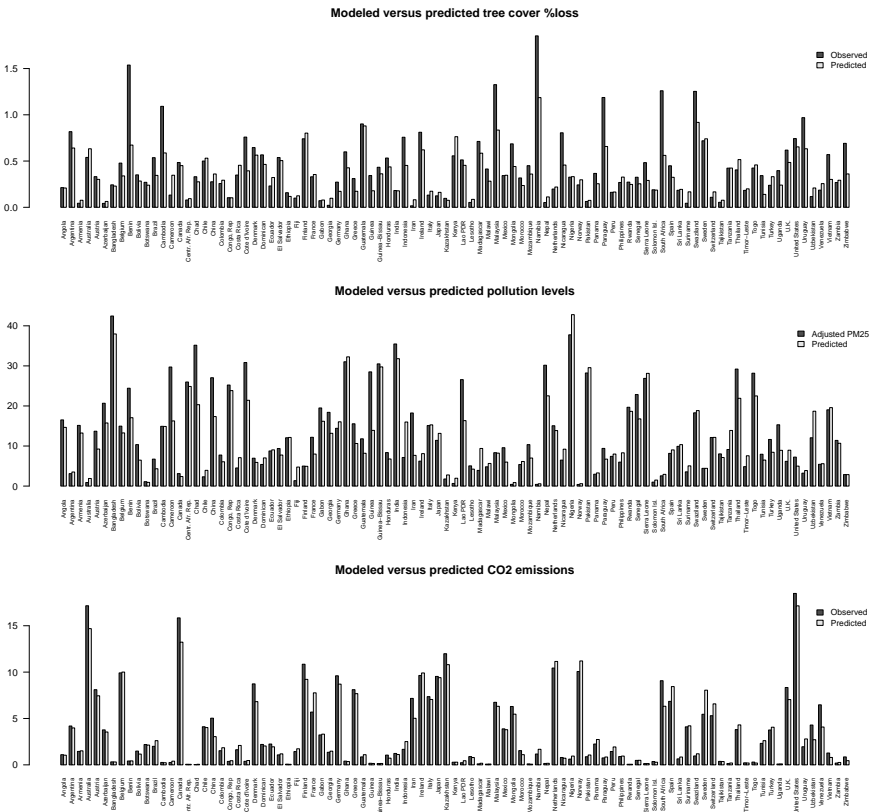


Figure 5.9: Accuracy of predicted degradation levels at high income.

Table 5.6: BAU-2030 and base year data aggregated by 5% percentiles of income. GDP per capita as population weighted averages, population in millions, number of poor people in millions, annual tree loss in square kilometers, PM<sub>2.5</sub> in population weighted average concentrations, carbon emissions in million tons. World totals are scaled to world totals using multipliers (1.16 for population and carbon based on the share of population in our data, and tree loss 1.42 based on the share of tree cover in the data).

Income Group	GDP p.c. 2014	GDP p.c. 2030	Pop. 2014	Pop. 2030	No. Poor 2014	No. Poor 2030	Treeloss 2014	Treeloss 2030	PM25 2014	PM25 2030	CO2 2014	CO2 2030
1	1,094	1,548	68	101	41	55	3,382	3,640	15	20	10	28
2	1,495	2,611	171	262	74	70	3,750	3,977	17	21	20	109
3	1,899	2,388	46	67	19	24	1,415	1,684	23	25	22	39
4	2,305	3,308	96	139	36	43	1,821	1,931	29	31	27	71
5	2,917	4,720	237	305	57	45	1,295	1,331	47	52	96	285
6	4,249	5,381	265	366	43	48	5,782	5,713	46	48	190	282
7	5,265	9,887	111	133	6	2	4,716	4,031	23	20	170	385
8	5,426	10,301	1,519	1,936	341	288	4,440	3,769	54	51	2,286	5,634
9	6,875	9,344	176	231	21	22	3,297	3,043	12	13	212	399
10	8,219	10,627	15	17	1	1	4,002	3,395	9	9	14	22
11	10,078	16,232	292	352	22	11	17,799	17,676	15	15	540	1,338
12	11,993	16,578	114	141	11	11	3,784	4,037	15	17	646	1,150
13	13,135	24,563	1,691	1,825	49	5	33,540	39,407	44	44	11,212	25,839
14	16,347	20,955	206	244	4	3	2,032	2,132	22	24	1,128	1,864
15	18,386	28,014	162	193	3	2	5,160	5,859	17	17	750	1,402
16	22,607	37,980	53	60	0	0	1,822	2,194	20	24	438	1,112
17	33,000	39,059	204	219	0	0	6,647	8,832	21	22	1,160	1,770
18	38,515	47,345	244	253	0	0	26,962	32,972	12	13	2,324	3,324
19	43,584	53,144	128	132	0	0	4,694	5,794	15	16	1,276	1,788
20	51,694	64,462	354	384	0	0	17,649	19,466	10	11	5,534	7,414
world	14,088	19,899	7,137	8,537	730	630	218,664	242,656	36	36	32,544	62,934



## 5.7 Supplementary note to the chapter

This supplementary note provides additional methodological discussion around an adaption of Kernel Regularized Least Squares to the dynamic panel context, paying specific attention to automated selection of fixed effects. The model provides an attractive approach to both nonlinearity and interpretability within one integrated framework. Importantly, it allows deriving marginal coefficients at the observational level that are similar to those of parametric models, while leveraging the flexibility provided by the similarity learning framework. Unlike in the standard regression context, the interpretation of these marginal coefficients depends on externally set tuning parameters. The paper discusses the role of the regularization parameter in the interpretation of these basic quantities of interest. The discussion highlights that rigorous hyper-tuning, and out-of-sample prediction performance of models in general, is crucial, even when one is merely interested in inference and not in prediction.

### 5.7.1 Introduction

The UN's Sustainable Development Goals for 2030 aim on one hand at inclusive growth and eradicating poverty, and on the other at preserving environments. The crucial relation between development and the environment has been studied extensively since the 1990s, and has been revisited recently in the main article associated with this background note. The paper applied a Kernel Regularized Least Squares model on disaggregate data obtained from remote sensing sources to model environmental-economic trends heterogeneously across a number of economic indicators at country level. Results suggested that local economic circumstances played an important role in determining the shape, amplitude, and location of tipping points in environmental output. This note details how the framework was adapted to the panel context,

paying particular focus to automated selection of fixed effects. The discussion here also goes more deeper into the types of assumptions that are implicitly made about environmental-economic interactions by adopting the kernel approach. The model is attractive as it provides a straightforward approach to nonlinearity and interpretability without having to rely on surrogate approaches, such as the Local Interpretable Model-agnostic Explanations method that locally interrogates a model's output surface (Ribeiro et al., 2016). The interpretation of the marginal coefficients is, however, highly dependent on an externally set hyperparameter that is not part of the estimated vector of parameters that define the model itself. Instead, consistency results for the Regularized Kernel model are toward a, possibly pseudo-true, parameter for which the limit result is separately defined for each level of penalization. This paper discusses the role of tuning the penalty, or regularization parameter, in ensuring that the marginal coefficients, and associated standard errors, admit to a standard interpretation. The model of interest, and limit theory for it, is provided in Hainmueller and Hazlett (2014). For the more general reader, most of the discussion here is posed in a general way and stretches out to many other relevant cases.

The remainder of this writing is organized as follows. Section 5.7.2 introduces the modeling framework and details the adaption to the panel setting. Section 5.7.3 provides further discussion around the tuning of the penalty and its relationship to the interpretation of the estimation result. Section 5.7.4 concludes.

### **5.7.2 The modeling framework**

Machine learning methods are often developed with different applications in mind than the classical regression models that have been developed primarily for economic inference. Because the limit results in non-parametric models depend on externally set tuning parameters that are not part of

the vector of estimated parameters, it is not immediately clear whether the estimates can be interpreted similarly as those obtained using parametric methods.

As researchers continue to tackle high dimensional problems, such as frequent in the environmental economics domain, we anticipate that machine learning methods will become more popular in the context of inference. For the sake of those less familiar with these approaches we summarize the basic assumptions and key features of the applied non-parametric kernel estimation below, relevant in the setting of the companion paper. We do not introduce new theoretical results, instead we aim to provide an overview that highlights differences with respect to parametric estimation. Particularly, we detail the role that regularization plays in correct inference. Readers that are familiar with penalized kernel models may proceed directly to section 5.7.2, in which we explain how we treat fixed effects in the estimation.

In the following, let  $\mathbb{N}$ ,  $\mathbb{Z}$ , and  $\mathbb{R}$  denote the sets of the natural, integer, and real numbers.  $\mathbb{R}_{>0}$  includes all positive, non-zero, reals. For a set  $\mathcal{A}$ , we use  $\mathfrak{B}(\mathcal{A})$  to denote the Borel  $\sigma$ -algebra over  $\mathcal{A}$ . We use  $t, \dots, T \in \mathbb{Z}$  to index time, and  $i, \dots, N \in \mathbb{N}$  to index cross-sections,  $it, \dots, NT \in \mathbb{N} \times \mathbb{Z}$  labels all locations in space-time. We use boldfaced letters, e.g.,  $\mathbf{a} \in \mathcal{A}$  to denote vectors. Furthermore,  $\times_{t=1}^{t=T} \mathcal{A} = \mathcal{A}_T$  denotes the Cartesian product of  $T$  copies of  $\mathcal{A}$ , and  $\mathcal{A}_\infty = \times_{t=-\infty}^{t=\infty} \mathcal{A}$  is the Cartesian product of infinite copies. For two maps  $f$  and  $g$ ,  $f \circ g$  is their composition resulting from a point-wise application, and  $\langle \cdot, \cdot \rangle$  denotes the inner product space.<sup>12</sup> Finally,  $\| \cdot \|_{\mathcal{A}}$  denotes a norm on  $\mathcal{A}$ .

---

<sup>12</sup>As a generalization of the dot product in the Euclidean space, to higher dimensional spaces including infinite dimensional Hilbert spaces.

### Assumptions about the data generation process

Suppose, we observe an  $n_x$ -variate  $T$ -period sequence  $\mathbf{x}_T := \{\mathbf{x}\}_{t=1}^T$  that describes the state of one economy throughout time. At each point in time, we observe  $N$  trajectories of this  $n_x$ -variate sequence, i.e., we focus on repeated cross-sectional vectors of length  $N$  describing the evolution of  $N$  economies. Each vector contains observations of for example income and the composition of the economy, all indexed over a set of locations  $i, \dots, N$ . The matrix  $\mathbf{X}_t$  consisting of  $n_x$  columns describing different variables  $\mathbf{x}$  and  $N$  rows describing the different locations, is indexed by time. We consider a second, repeated cross-sectional sequence  $\mathbf{y}$ , of degradation levels generated by:

$$\mathbf{y} := \{\mathbf{y}_t = h_0(\mathbf{X}_t), t \in \mathbb{Z}\}. \quad (5.3)$$

We can observe  $\mathbf{y}_T$ , a subset of the results of this process  $\mathbf{y}_T := \{\mathbf{y}\}_{t=1}^T$ . The function  $h_0 : \mathcal{X} \rightarrow \mathcal{Y} \subseteq \mathbb{R}$  produces environmental output for every coordinate  $\mathbf{X}_t \in \mathcal{X}$ .<sup>13</sup> We assume that  $h_0$  is a unique measurable function that for each coordinate  $\mathbf{X}_t \in \mathcal{X}$  assigns a *true* value  $\mathbf{y}_t \in \mathcal{Y}$  for all  $t \in \mathbb{Z}$ . In a sense, by assuming this particular form, we assume that the environment does not endogenously degrade itself, i.e. that  $\mathbf{y}$  does not endogenously generate itself. Instead, this assumes that the evolution of degradation levels for each economy  $\mathbf{y}$  is symptomatic to external, local economic development variables  $\mathbf{X}$ . This does not exclude the possibility that  $\mathbf{y}$  may in part affect elements of  $\mathbf{X}$ , it requires however that feedback effects are invertible, in turn implied by some form of stability, and follow levels in  $\mathbf{X}$  such that  $h_0$  describes the net relationship between  $\mathbf{X}$  and  $\mathbf{y}$ .<sup>14</sup> We also assume that  $h_0$  is smooth, particularly that it maps

<sup>13</sup>Particularly a  $\mathfrak{B}(\mathcal{X})/\mathfrak{B}(\mathcal{Y})$ -measurable mapping as kernels with a universal approximating property require at least that the target function is measurable, see for example (Micchelli et al., 2006).

<sup>14</sup>If  $\mathbf{y} = f_0(\mathbf{X}) + g_0(\mathbf{y})$  with  $g_0$  describing simultaneous feedback and  $f_0$  describing the contemporaneous exogenous effects, then one can also write  $\mathbf{y} = h_0(\mathbf{X})$  if  $g_0$  is invertible, with  $h_0(\mathbf{X}) = (I - g_0)^{-1}(f_0(\mathbf{X}))$ , hence  $h_0$  arises from the composite  $(I - g_0)^{-1} \circ f_0$  and describes the combined effect of exogenous impulses and feedback.

similar coordinates in  $\mathcal{X}$  to similar values in  $\mathcal{Y}$ . This implies that for each state of the economy at each point in time  $\mathbf{X}_t$ , we observe a level of deforestation, pollution or carbon  $\mathbf{y}_t$  that is induced by the state of the economy through the function  $h_0$ , and that for small changes in the state of an economy we expect to see small changes in environmental output. Furthermore, it assumes that for two economies that are similar in terms of composition and scale, we expect similar environmental output.

### Panel Kernel Regularized Least Squares

In the current case, the environmental Kuznets theory suggests an inverted  $U$ -shape between degradation and economic development. The relationships may of course be of a completely other form or differ across environmental variables, while ideally we keep the analysis of both relationships within a similar regression framework. We therefore postulate a very flexible regression of the form

$$\hat{\mathbf{y}} := \{\hat{\mathbf{y}}_t = h(\mathbf{X}_t; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta, t \in \mathbb{Z}\}. \quad (5.4)$$

Our modeled function  $h$  is defined as a mapping  $h : \mathcal{X} \times \Theta \rightarrow \mathcal{Y}$ , where  $\Theta$  is the parameter space. In parametric regressions,  $\Theta$  is assumed to be compact and finite dimensional. This immediately imposes structure on  $h$ , thus translating into assumptions about  $h_0$  if we maintain a belief that  $\boldsymbol{\theta}_0 \in \Theta$ . By reducing the size of  $\Theta$  we simplify the possible structure of  $h$ , i.e., chances that  $\boldsymbol{\theta}_0 \in \Theta$  become increasingly slim. While we minimize assumptions about  $h_0$  by working with  $\Theta$  as an infinite dimensional space, some assumptions about  $h_0$  are unavoidable as  $\Theta$  has to be parametrized eventually. In our example, as we shall see, one still has to specify radial basis functions.

In parametric regressions,  $\Theta$  plays a key role as the Euclidean space containing all the possible coordinates of potential parameter vectors  $\boldsymbol{\theta}$ .

In the non-parametric case, there is a subtle difference. Suppose that for every  $\boldsymbol{\theta} \in \Theta$ , there is a function  $h(\cdot; \boldsymbol{\theta}) : \mathcal{X} \rightarrow \mathcal{Y}$  that is  $\mathfrak{B}(\mathcal{X})/\mathfrak{B}(\mathcal{Y})$ -measurable. We can define  $\mathcal{H}_\Theta(\mathcal{X})$  as the Hilbert space containing an infinite collection of functions  $\{h(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$  generated by  $\Theta$ . We shall use a simplified notation to reduce cluttering and instead write that  $\boldsymbol{\theta}$  indexes the functions  $h_{\boldsymbol{\theta}} \in \mathcal{H}_\Theta$ . The common notation  $\boldsymbol{\theta}_0 \in \Theta$  is thus equivalent to saying  $h_0 \in \mathcal{H}_\Theta$ , i.e.,  $\boldsymbol{\theta}_0 \in \Theta : h(\mathbf{x}_t; \boldsymbol{\theta}_0) = h_0(\mathbf{x}_t) \forall \mathbf{x}_t \in \mathcal{X}$ . This clarifies that, while in a parametric regression problem where we are fore-mostly concerned with searching a compact parameter space  $\Theta$  for the parameter vector  $\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}_0$ , in the current framework we are explicitly interested in searching across a space of functions produced under some process of generating flexible functions from simple parameter vectors given the sample space,  $\mathcal{H}_\Theta$ , for infinite  $\boldsymbol{\theta} \in \Theta$ , for the function that best resembles the target function  $h_{\boldsymbol{\theta}} \rightarrow h_0$ . Specifically, each  $\boldsymbol{\theta}$  indexes a member in  $\mathcal{H}_\Theta(\mathcal{X})$  according to the map  $h_\mathcal{X} : \Theta \rightarrow \mathcal{H}_\Theta(\mathcal{X})$  with  $h_\mathcal{X}(\boldsymbol{\theta}) := h(\cdot; \boldsymbol{\theta}) \in \mathcal{H}_\Theta(\mathcal{X}) \forall \boldsymbol{\theta} \in \Theta$ . Hence, we can write the estimator also as

$$\hat{h}_T := \arg \min_{h_{\boldsymbol{\theta}} \in \mathcal{H}_\Theta} Q_T(\mathbf{y}_T, \mathbf{X}_T; h_{\boldsymbol{\theta}}). \quad (5.5)$$

The criterion function  $Q_T$  can also be written as  $Q_T(\mathbf{X}_T, h_0(\mathbf{X}_T), h(\mathbf{X}_T; \boldsymbol{\theta}))$ , as we started under the notion that  $\mathbf{y}_T = \{h_0(\mathbf{X}_t)\}_{t=1}^T = h_0(\mathbf{X}_T)$  which reveals the direct connection between the criterion function and target function  $h_0$ .

There are many ways to generate  $\mathcal{H}_\Theta$ . In the current framework, we focus on using a kernel  $k$  together with a local parameter  $\theta_i$  that weights the surface to produce any flexible functional form.

$$h_{\boldsymbol{\theta}} := \sum_i^N \theta_i k(x, x_i) = h(x; \boldsymbol{\theta}). \quad (5.6)$$

The functions  $h_{\boldsymbol{\theta}} \in \mathcal{H}_\Theta$ , are allowed to follow any kernel that has the universal approximation property, in this paper we adopt a Gaussian

kernel  $k(\mathbf{x}_i, \mathbf{x}_j; n_x) = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{n_x}\right)$  with  $\|\mathbf{x}_i - \mathbf{x}_j\|$  being the Euclidean distance, and  $n_x$  being a fixed bandwidth equal to the dimension of  $\mathbf{x}_T$ . We count the constant as being part of  $\mathbf{x}_T$ .

The kernel  $k$  can be understood as a measure of similarity, which is seen by applying a Cauchy-Schwarz inequality

$$k(x_i, x_j)^2 \leq k(x_i, x_i)k(x_j, x_j) \quad \forall (x_i, x_j) \in \mathcal{X},$$

revealing that if  $x_i$  and  $x_j$  are similar, then  $k(x_i, x_j)$  will be close to 1, and close to 0 when  $x_i$  and  $x_j$  are dissimilar. For a given observed collection  $(y, x \in \mathbf{x})$ ,  $h_\theta$  is thus a function resulting from placing kernels over  $x_i$  and scaling the similarity surface using local coefficients  $\theta_i$  such that the summated surface approximates the true density of the data. This produces flexible functions that can describe local relationships between  $\mathbf{y}$  and an individual covariate  $\mathbf{x}$  by assigning similar observations a similar scaling factor that maps onto similar output.

Different parameterizations of the local coefficients  $\theta_i$  may produce equally well, e.g. perfect fits, such that the problem of estimating the vector  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)'$  is generally ill-posed without adding further structure to the problem. The specific estimation strategy to learn about the trends in the data is therefore of the form

$$\hat{h}_T := \arg \min_{h_\theta \in \mathcal{H}_\Theta} Q_T(\mathbf{y}_T, \mathbf{X}_T; h_\theta) - \pi(h_\theta), \quad (5.7)$$

where  $\pi(h_\theta) > 0 \quad \forall h_\theta \in \mathcal{H}_\Theta$  is a strictly positive function that monotonically increases by a measure of complexity defined on  $h_\theta$ . The penalty is critical to ensuring identifiability and consistency of the estimator within simple subset spaces of  $\mathcal{H}_\Theta$ . At the same time, it allows to fit nonlinearities of varying smoothness while working with a fixed kernel bandwidth that produces a relatively smooth similarity surface, as  $\theta_i$  is able to scale the nonlinearities locally albeit at a cost  $\pi(h_\theta)$ . Hence, it

favors less complex solutions to the criterion function by penalizing the high frequency domain in  $\mathcal{H}_\Theta$ . Specifically, let  $K$  be an  $N \times N$  symmetric kernel matrix with entries  $k(\mathbf{x}_j, \mathbf{x}_i)$  measuring pair-wise similarities. This yields a model that is a linear combination of basis functions, each measuring similarity of one observation to another observation in the data set, and mapping it to a local output.

$$\begin{aligned} \mathbf{y}_t &= h(\mathbf{X}_t; \boldsymbol{\theta}_t) = K(\mathbf{X}_t) \boldsymbol{\theta}_t \\ &= \begin{bmatrix} k(\mathbf{x}_{1t}, \mathbf{x}_{1t}) & k(\mathbf{x}_{1t}, \mathbf{x}_{2t}) & \cdots & k(\mathbf{x}_{1t}, \mathbf{x}_{Nt}) \\ k(\mathbf{x}_{2t}, \mathbf{x}_{1t}) & & \ddots & \\ \vdots & & & \\ k(\mathbf{x}_{Nt}, \mathbf{x}_{1t}) & & & k(\mathbf{x}_{Nt}, \mathbf{x}_{Nt}) \end{bmatrix} \begin{bmatrix} \theta_{1t} \\ \theta_{2t} \\ \vdots \\ \theta_{Nt} \end{bmatrix}. \end{aligned} \quad (5.8)$$

The need for a regularization technique is obvious, the parameters  $(\theta_{1t}, \theta_{2t}, \dots, \theta_{Nt})$  can always rescale the similarity surface to match  $\mathbf{y}_t$  perfectly. Instead, the penalized estimator takes into account the complexity of the rescaling by introducing a factor  $\lambda \|h_\theta\|_K^2$  and chooses the best fitting function by minimizing:

$$\arg \min_{h_\theta \in \mathcal{H}_\Theta} \sum_i^N \sum_t^T (y_{it} - h(\mathbf{x}_{it}; \boldsymbol{\theta}))^2 + \lambda \|h_\theta\|_K^2, \quad (5.9)$$

in which  $\sum_i^N \sum_t^T (y_{it} - h(\mathbf{x}_{it}; h_\theta))^2$  are the standard sum of squared residuals.  $\lambda \|h_\theta\|_K^2 = \langle h_\theta, h_\theta \rangle_{\mathcal{H}_\Theta}$  is a penalty that increases monotonically as a function of the complexity of  $h$  under  $\boldsymbol{\theta}$ . We focus on the  $L^2$  norm. Finally,  $\lambda \in \mathbb{R}_{>0}$  is the parameter that determines the strength of the penalty. Using this kernel, we can work with an  $NT \times NT$  kernel matrix by defining the dependent variable  $Y$  as the  $NT$  length vector resulting from stacking the time observations,  $X$  as the  $NT \times n_x$  matrix resulting from stacking the columns similarly and  $\boldsymbol{\theta}$  as an  $NT$  length parameter



vector.<sup>15</sup> Using the Gaussian kernel, eq. (5.9) becomes

$$\hat{h}_{NT} = \arg \min_{h_{\theta} \in \mathcal{H}_{\Theta}} (Y - K(X)\theta)'(Y - K(X)\theta) + \lambda \theta' K(X)\theta. \quad (5.11)$$

In a panel application, the functions  $h_{\theta}$  that result from weighted kernels can produce interesting time-varying dynamics across levels of  $\mathbf{X}_t$ . This is for example appropriate when a time-varying stationary processes is of interest in which the nonlinearities change throughout the data but are not depending on time itself. Alternatively, one can work with time itself as a covariate, in which case processes that are only locally stationary can be modeled. Intuitively, the kernel approach then results in similar coefficients for similar time. In the case of non-stationary data, the kernel can approximate local conditional means in the data that may vary throughout the sample space.

### The role of the penalty

The basic idea of penalizing the criterion function has been explored in many statistical applications, and is for example at the heart of the widely adopted LASSO estimator (Tibshirani, 1996; Zou, 2006). The added structure to the criterion function is a frequentist's analogue to the role that the prior plays within the Bayesian framework. We note

---

<sup>15</sup>Specifically:

$$Y = \begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{Nt} \\ y_{12} \\ y_{22} \\ \vdots \\ y_{NT} \\ y_{1T} \\ y_{2T} \\ \vdots \\ y_{NT} \end{bmatrix}, X = \begin{bmatrix} \mathbf{x}_{1t} \\ \mathbf{x}_{2t} \\ \vdots \\ \mathbf{x}_{Nt} \\ \mathbf{x}_{12} \\ \mathbf{x}_{22} \\ \vdots \\ \mathbf{x}_{NT} \\ \mathbf{x}_{1T} \\ \mathbf{x}_{2T} \\ \vdots \\ \mathbf{x}_{NT} \end{bmatrix}, \theta = \begin{bmatrix} \theta_{11} \\ \theta_{21} \\ \vdots \\ \theta_{N1} \\ \theta_{12} \\ \theta_{22} \\ \vdots \\ \theta_{N2} \\ \theta_{1T} \\ \theta_{2T} \\ \vdots \\ \theta_{NT} \end{bmatrix}. \quad (5.10)$$

that the penalty in the current setting is not primarily a way to improve small sample performance, but that it is in fact the central feature of the learning model that determines what functional forms can be fitted. This differs from kernel approaches in which the bandwidth is the key tuning parameter. In the current approach the bandwidth is fixed to produce smooth functions, but nonlinearities are subsequently locally adjusted using the vector of weights  $\boldsymbol{\theta}$  to increase flexibility. The penalization approach is able to shrink the hypothesis space and flexibly establish a subspace in which consistency holds. By balancing between fit and complexity of the locally weighted kernel, the size of the subspace can be regulated by the penalty. In the case of penalized GLM's considered in Blasques and Duplinskiy (2018), nonzero penalties take one away from  $\boldsymbol{\theta}_0$  in the limit if the penalty effect does not vanish asymptotically.<sup>16</sup> In that sense, a penalized criterion delivers a pseudo-*true* parameter with a divergence from  $\boldsymbol{\theta}_0$  that is controlled by the penalty function. Setting an appropriate penalty therefore determines what one can infer from  $\boldsymbol{\theta}_0$ . In the current context, positive penalties are a necessity to ensure uniqueness. This might lead to the thought that penalized non-parametric estimators that require positive penalization are biased by definition. The estimate of the weights  $\hat{\boldsymbol{\theta}}$  obtained through eq. (5.11) is different for every value of  $\lambda$ . The tuning parameter  $\lambda$  thus represents the researcher's predefined level of tolerance for accepting nonlinear functions. High values of  $\lambda$  force the model to linearize it's dependencies, whereas extreme values for  $\lambda$  will set all coefficients to zero and describe the data using only an average expectation. Hence for every penalty, we find a different functional form  $\hat{h}_\lambda$  induced by the estimate  $\hat{\boldsymbol{\theta}}_\lambda$  through eq. (5.6) given a specified kernel. Since  $\lambda$  itself is not an estimated parameter, it is generally difficult, if not impossible, to tell whether eq. (5.11) yields an estimate of  $\hat{h}$  close

---

<sup>16</sup>Furthermore,  $\boldsymbol{\theta}_0$  in the standard context is the *true* parameter. In the non-parametric context, that *true* parametrization arguably does not exist, however one can think of  $\boldsymbol{\theta}_0$  as the parametrization that produces  $h_0$  through the kernel, or alternatively, selects  $h_0$  the *true* (non)linear functional form  $\mathcal{H}_\Theta$  that produces the *true* density of the data.

to  $h_0$ . Without knowing the magnitude of  $\|\hat{h}_\lambda - h_0\|$ , the method may be difficult to use for economic inference.

Schölkopf and Smola (2001), suggest to set the penalty through an out-of-sample prediction minimization problem to remove the dependence of the results on the external influence of the researcher that determines the level of penalization a priori. Hainmueller and Hazlett (2014) suggest one such strategy for Kernel Regularized Least Squares estimates by minimizing out-of-sample prediction errors over a vector  $\lambda \in \Lambda$  based on leave-one-out predictions, noting that it performs well in practice. While practical performance and the removal of external influence on the results provide intuition to set penalties in this way, it does not focus on the question whether  $\|\hat{h}_\lambda - h_0\|$  is in fact minimized, which is key to ensure that the marginal coefficients converge to the correct values, e.g.  $\|\hat{h}'_\lambda - h'_0\| \rightarrow 0$ . Here we provide additional discussion on the role of the penalty in ensuring identifiable uniqueness and establishing the consistency and normality results. We discuss that the strategy to set penalties by minimizing an out-of-sample criterion naturally pulls the estimator toward the weight vector that induces the *true* function in the limit, such that inference can be applied as usual. This is because for a given penalty  $\lambda$ , the estimated function conditional on that penalty  $\hat{h}|\lambda$  provides the optimal density across all functions  $h \in \mathcal{H}_\Theta|\lambda$  induced under that penalty, so choosing the estimate from a set of results found using different penalties  $\hat{h}|\lambda \in \Lambda$  that provides the optimal out-of-sample density, also minimizes  $\|\hat{h}_\lambda - h_0\|_{\lambda \in \Lambda}$  in the limit since  $h_0$  is the function that by definition provides the best out-of-sample density. In other words, estimating eq. (5.11) while setting  $\lambda$  based on out-of-sample prediction error minimization yields an estimated function that minimizes  $\|\hat{h} - h_0\|$  in the limit across the entire family of models generated under all weight vectors and all penalties, which is similar to the standard case in which the criterion converges to the parameter that induces the best conditional

density across the entire parameter space (White et al., 1980; White, 1994).

### Fixed effects and out-of-sample shrinkage

Linear effects can be included by using difference estimators as detailed in Hainmueller and Hazlett (2014). Nonlinear effects can be modeled by supplying group-specific trend variables and group identifiers through  $\mathbf{X}_t$ . In this case all coefficients may depend on time, and similarly across similar groups in the data. Nonlinear fixed effects approaches combined with non-parametric parts around the economic variables may result in models with an enormous size while often the amount of observations locally in the time dimension remains relatively small in environmental economic panels. Model size not only relates to the complexity of functions around the economic variables, but also to the number of fixed effects in the model. In-sample selection strategies to decide on the right number of effects are complicated in the regularized non-parametric context. While in standard regressions additional variables always improve fit, this is not the case in the current context. Adding fixed effects results in different complexity of the local weights vector. Therefore, the effect of the complexity penalty in the criterion may increase such that the penalized estimator adjusts the weighting vector to achieve lower complexity. While this reduces the penalty value, it may possibly lower the in-sample  $R^2$ . Comparing models with and without fixed effects is therefore a comparison between functional forms with different complexity and nonlinearities. This is a comparison of non-nested models with an unknown, possibly real valued, difference in degrees of freedom.<sup>17</sup>

To decide on the right number of effects, we start by estimating a model that includes all fixed effects. We then remove the least significant

---

<sup>17</sup>Degrees of freedom is a parametric concept whose translation to the non-parametric setting is complex. One can approximate the degrees of freedom empirically, which may result in numbers that are real-valued.

dummy, and obtain new results. We repeatedly evaluate the out-of-sample prediction performance while shrinking the effects, and select the model with the optimal out-of-sample density across all fixed effect models. Intuitively, this approach starts with a model similar to a linear fixed effects model, as the penalty heavily discounts the thresholds introduced by the effects resulting in flattened marginal effects, and gradually allows fixed heterogeneity to be explained by nonlinearities across covariates instead. As a result, our final estimates are guaranteed to be preferred over the standard linear fixed effects model, as judged by the out-of-sample criterion.

### 5.7.3 The role of out-of-sample performance in the interpretation

Non-parametric approaches are capable of producing parametrization mappings that approximate nonlinearities arbitrarily well, but do not necessarily also produce uniquely identifiable solutions to the criterion function if the hypothesis space produces universal approximations that fit the data arbitrarily well for any sample size. Estimation is therefore problematic without additional structure to the estimator, which in our case comes in the form of a penalty to the criterion, but in other settings may relate to bandwidths or other tuning parameters. It is a challenge in its own right to understand how this complexity-penalized estimator is positioned relative to the classical least squares approximation context as considered by White et al. (1980).<sup>18</sup> In the standard context, the best approximation is produced by a unique point in the entire parameter space, while in the penalty context a best approximation exists for every given penalty. Hence, the divergence between the *true* functional form and the *pseudo*-true approximation is not driven by boundaries to the

---

<sup>18</sup>White et al. (1980) discuss convergence toward the unique least squares approximator that may differ from the *true* parameter in the presence of misspecification bias.

parameter space as in the parametric case, but rather it is driven by the penalty. Ultimately penalization confines the hypothesis space to simple spaces and the use of excessive penalization must reflect a prior belief that the *true* functional form would not result in a large penalty. That prior belief carries over to the limit result if the effect of the penalty does not vanish. This produces a bias, or may even render the result completely arbitrary if the penalty is set without caution.

In the current setting, our penalty arises as a function of an out-of-sample criterion. As a result, the space of functions that are viewed as acceptable solutions to the criterion is generated by the data itself, and the penalized non-parametric method is able to obtain approximations of increasing complexity as the data size tends toward infinity. In the finite sample case, this estimator is appropriate given that the relationship between environmental degradation and indicators of economic development is not dominated by high-frequency components that would result in strong complexity.

### Identifiability in nonlinear models

Closely related to regulating the size of non-parametric models is the ill-posedness of unregulated non-parametric models. Before discussing the relationship between penalization and identifiability of the criterion of non-parametric estimators, we provide a simplified discussion on identifiable uniqueness and its relation to inference in the context of finite dimensional nonlinear models.<sup>19</sup>

Hypothesis testing in a framework of finite parameter nonlinear models is often plagued by the problem that verification of the assumptions required

---

<sup>19</sup>Identifiable uniqueness is a difficult concept, more elaborate general discussion can be found here (Pötscher and Prucha, 1991); formal definitions and discussion at a deeper level regarding strongly unique best approximation in Banach spaces can be found here (Smarzewski, 1986); and discussion on regulated  $M$ -estimation can be found here (Kent and Tyler, 2001); and an overview of concepts is written in (Blasques, 2010).

for identifiability, relies itself on the outcome of a hypothesis that may be difficult to test. This is problematic as identifiable uniqueness plays a key role in establishing consistency and normality of test statistics. This is illustrated by a model of the form:

$$y = \delta \exp \left( \frac{-(x - c)^2}{\gamma} \right) + \varepsilon, \quad (5.12)$$

in which the postulated relationship between  $y$  and  $x$  is assumed to follow a hyperbolic curve across levels of  $x$ . In this model the linearity hypothesis  $H_0^1 : \gamma = 0$  relating to the non-existence of the curved functional form depends on a second hypothesis  $H_0^2 : \delta \neq 0$  being true or false. This follows from the fact that for  $\delta = 0$ ,  $\gamma$  can take any value without changing the predicted density implied by the model. In this case any form of completeness required for identifiable uniqueness of the estimator, holds at most for a subset of the parameter space in which the model would in fact produce an inverted  $U$ -shaped form. Distributions corresponding to different values of  $\gamma$  are only sufficiently distinct when  $\delta$  is sufficiently bounded away from zero. Without establishing existence and uniqueness of a consistent estimator, it is impossible to establish normality, hence the distribution of test statistics remains unknown.<sup>20</sup>

More intuition is found in the following two definitions adapted from Definition 1 and Definition 2 in (Rothenberg, 1971).

DEFINITION. 1. *Two points  $\alpha_1 \in \mathcal{A}$  and  $\alpha_2 \in \mathcal{A}$  are said to be observationally equivalent with respect to a function  $h$  evaluated over  $x$  if  $h(x; \alpha_1) \equiv h(x; \alpha_2) \forall x \in \mathbb{R}$ .*

---

<sup>20</sup>Auxiliary test statistics may still be derived, but it is sometimes difficult to ensure that Taylor expansions do not capture nonlinearities of a type not predicted by the economic theory. See for example (Dijk et al., 1999) for a discussion in the threshold framework. Researchers may also choose to rely on information criteria to compare various descriptions of the data and decide between economic theories (Granger et al., 1995). In the limit, Penalized Likelihood Criteria select the model that minimizes Kull-Back Leibler divergence with probability 1 (Sin and White, 1996), but convergence rates depend on the penalty chosen. The acceptance of an economic theory thus relies on information outside the model. In a sense, a researcher has flexibility to corroborate specific theories by designing the information criteria to support them.

DEFINITION. 2. A point  $\alpha^1 \in \mathcal{A}$  is said to be *identifiable* by a function  $h$  evaluated over  $x$  if there is no other point  $\alpha \in \mathcal{A}$  that is *observationally equivalent*.

Let  $\boldsymbol{\theta} := (\delta, c, \gamma)'$  denote a vector of parameters, with  $\boldsymbol{\theta} \in \Theta$ , and  $\boldsymbol{\theta}_0 := (\delta_0, c_0, \gamma_0)'$  be the *true* vector of parameters. For consistency toward the *true* parameter, one would not only require  $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \rightarrow 0$  to be the solution *a.s.* to the criterion  $\hat{\boldsymbol{\theta}} := \arg \min_{\boldsymbol{\theta} \in \Theta} Q(y, x; \boldsymbol{\theta})$  as  $N \rightarrow \infty$ , but it needs to be the *identifiable unique* solution. Following the definitions above, then by the definition  $\boldsymbol{\theta}_0$  as the minimizer of  $Q(y, x; \boldsymbol{\theta})$ , there needs to be assurance of some form that

$$\arg \min_{\boldsymbol{\theta} \in \Theta} Q(y, x; \boldsymbol{\theta}_0) < \arg \min_{\boldsymbol{\theta} \in \Theta} Q(y, x; \boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \Theta \setminus \boldsymbol{\theta}_0, \quad (5.13)$$

excluding

$$\arg \min_{\boldsymbol{\theta} \in \Theta} Q(y, x; \boldsymbol{\theta}_0) \leq \arg \min_{\boldsymbol{\theta} \in \Theta} Q(y, x; \boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \Theta \setminus \boldsymbol{\theta}_0. \quad (5.14)$$

as the alternative. The standard assumption is that  $\Theta$  is compact. Together with *almost sure continuity* in  $\boldsymbol{\theta} \in \Theta$ , Weierstrass' theorem implies that  $\boldsymbol{\theta}_0$  exists as a non-empty set *a.s.* Equation (5.13) can result directly from the parametrized model  $\hat{y} = h(x)$  if

$$h(x; \boldsymbol{\theta}_0) \neq h(x; \boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \Theta \setminus \boldsymbol{\theta}_0, \quad (5.15)$$

such that there is no point in  $\Theta$  other than  $\boldsymbol{\theta}_0$  that is observationally equivalent to  $\boldsymbol{\theta}_0$ . Specifically the observational equivalence definition may fail to hold if  $\Theta$  is high dimensional. If eq. (5.13) is not implied by the nature of  $h$ , it can also be provided by additional structure to the criterion  $Q(\cdot; \boldsymbol{\theta})$  conditional on regions in  $\Theta$ , or by limiting the search to remain within a subset  $\hat{\boldsymbol{\theta}} := \arg \min_{\boldsymbol{\theta} \in \Theta_s \subset \Theta} Q(y, x; \boldsymbol{\theta})$ , where  $\Theta_s$  is a compact subset of the parameter space that may possibly grow in complexity along with the sample size.



DEFINITION. 1 and DEFINITION. 2 are intuitive, but provide no testable condition to decide upon the identifiability of an estimator. One insightful definition is the following adapted from (Bates and White, 1985) that ensures that the solution to the criterion is well separated.

DEFINITION. 3. *Suppose  $\theta_0$  minimizes a real-valued criterion  $Q_\infty(\cdot; \theta)$  on a compact metric space  $\Theta$ , within a circular neighborhood  $\aleph_0(r) \subset \Theta$  with radius  $r > 0$  that has a compact complement  $\aleph_0(r)^c : \Theta \setminus \aleph_0(r)$ , then  $\theta_0$  is uniquely identified on  $\Theta$  if and only if for every  $r > 0$ ,*

$$\inf_{\theta \in \aleph_0(r)^c} [Q_\infty(\theta) - Q_\infty(\theta_0)] > 0.$$

### Identifiability in non-parametric models

Non-parametric models aim to learn from the data without assuming that  $h$  is up to finitely many parameters, and work under the axiom that the parameter space  $\Theta$  may in fact be infinitely dimensional. By allowing for that, we minimize the risk that our parametrization assumptions preclude  $\theta_0 \in \Theta$ , solving for misspecification bias that results from parametric assumptions. However, without imposing further structure to the criterion it is generally not possible to establish consistency of our estimate  $\hat{\theta} \rightarrow \theta_0$  uniformly over  $\Theta$  as the compactness assumption on  $\Theta$  does not hold in infinite dimensions.<sup>21</sup> This poses a problem in verifying DEFINITION. 3, and establishing consistency results such as those of (Domowitz and White, 1982).

One solution is to focus the arguments on establishing a compact subset of the parameter space such that over the complement of the compact subset the criterion function is eventually “large”, see (Pötscher and Prucha, 1997). This follows by first constructing a subset  $\Theta_s \subset \Theta$  such that  $\theta_0 \in \Theta_s$ , and such that all the elements outside  $\Theta_s$  are valued distinguishably different by the criterion than the elements within  $\Theta_s$ ,

<sup>21</sup>By definition a set  $\mathcal{A} \in \mathbb{R}^d$  is compact if and only if it is closed and bounded.

disregard of the structure outside of  $\Theta_s$ . The subset is then closed, as its complement is open, and it is bounded as it is contained in a ball of finite radius, which implies that  $\Theta_s$  is then compact. As a consequence, it is sufficient to show that consistency holds within  $\Theta_s$ , since any  $M$ -estimator must eventually fall within this compact subset. We can summarize identifiable uniqueness of  $\theta_0$  in an open space as follows.

**DEFINITION. 4** (Identifiability in an open space). *Suppose  $\theta_0$  minimizes a real-valued criterion  $Q_\infty(\cdot; \theta)$  on an open metric space  $\Theta$ . Suppose furthermore that  $\theta_0$  minimizes  $Q_\infty(\cdot; \theta)$  within a circular neighborhood  $\Theta_s(d) \subset \Theta$  that has finite positive radius  $d > 0$  and that uniformly over  $\Theta$  there exists some positive  $\epsilon$  for which  $[Q_\infty(\theta' \in \Theta \setminus \Theta_s) - Q_\infty(\theta \in \Theta_s)] > \epsilon$ . If furthermore, there is also a circular neighborhood  $\aleph_0(r) \subset \Theta_s$  with radius  $r < d$  that has a compact complement  $\aleph_0(r)^c : \Theta_s \setminus \aleph_0(r)$ , then  $\theta_0$  is uniquely identified on  $\Theta$  if and only if for every  $r$ ,  $0 < r < d$ ,*

$$\inf_{\theta \in \aleph_0(r)^c \subset \Theta_s(d) \subset \Theta} [Q_\infty(\theta) - Q_\infty(\theta_0)] > 0.$$

In a sense, we thus want to exert some control over the structure of  $Q_\infty(\cdot; \theta)$  on  $\Theta$  such that  $\theta_0$  is uniquely identifiable by the criterion within a neighborhood  $\Theta_s$  that is distinctly different from elements outside of it, disregard whether the criterion can distinguish differences between the elements outside  $\Theta_s$ . One such an approach can be found in the well-known kernel estimator. The solution offered by the kernel method depends on selecting an appropriate bandwidth that controls for the size of local neighborhoods in the sample space throughout which nonlinearities smoothly differ. For too small bandwidths, the kernel method creates a subspace  $\Theta_k \supset \Theta_s$  that allows overly-flexible fits to the data. This can create an ill-posed problem, in which multiple solutions to the criterion within  $\Theta_k$  may still deliver equally good fits as judged by the criterion evaluated over  $\Theta_k$ . It is obvious that DEFINITION. 4 is not applicable in such a context. For too small bandwidths, the kernel method establishes

$\Theta_k \subset \Theta_s$  that is small, and while DEFINITION. 4 may work for  $\Theta_k$ , we are not sure that in fact  $\theta_0 \in \Theta_k$  due to the parametrization assumptions used to construct  $\Theta_k$ . The role of the bandwidth is therefore extremely important, identifiable uniqueness of the estimator requires the bandwidth to be sufficiently large, while reducing miss-specification bias requires the bandwidth to be sufficiently small. In an ideal framework, both factors are balanced out and  $\Theta_k$  grows as  $N \rightarrow \infty$  at an appropriate rate.

### The role of the penalty in the estimator

The fitted nonlinearities are allowed to be of any form, but  $\lambda > 0$  implies the penalty is never removed completely. Positive penalization is key to ensuring that there exists a finite radius neighborhood  $\Theta_s(d) \subset \Theta$  in which any M-estimator must eventually fall uniformly over  $\Theta$  as  $[Q_\infty(\theta' \in \Theta \setminus \Theta_s) - Q_\infty(\theta \in \Theta_s)] > \epsilon(\theta; \lambda) > 0$ , where  $\epsilon(\theta; \lambda) > 0$  is ensured for any  $\theta$  by  $\lambda > 0$ . Penalties that vanish completely at a pre-specified rate are interesting when the researcher wishes to impose penalties only when the estimator is confronted with small sample sizes. This requires however that the criterion is uniquely identified at  $\lambda = 0$  eventually. Vanishing penalties may improve inference when using estimators that have poor small sample behavior by ensuring that the estimator is relatively inert to weakly nonlinear signals and less likely to overfit the data in local regions of the sample space. Penalties that take values in  $\mathbb{R}_{>0}$ , can improve small sample behavior, but maintain a bias towards linear solutions that persists in the limit.

Note that eq. (5.11) reveals that convergence of our estimator  $\|\hat{h}_{NT} - \hat{h}_\infty\| \rightarrow 0$  to a specific target function  $\hat{h}_\infty \in \mathcal{H}_\Theta$ , where  $\hat{h}_\infty$  is possibly the *true* function or the best approximator as judged by the penalized limit criterion, is the same as  $\|\hat{\theta}_{NT} - \hat{\theta}_\infty\| \rightarrow 0$ , which is the more common notation. Hence, we shall use the latter, but really we are interested in ensuring that  $\hat{h}_\infty$  is a uniquely identifiable point in  $\mathcal{H}_\Theta$  as close to

$h_0$  as possible. Consistency and normality theorems for eq. (5.9) are provided in Hainmueller and Hazlett (2014). The results ensure a limit convergence toward the best approximation of the conditional expectation function given penalization, hence the limit solution is conditional on the researcher's choice of  $\lambda$ . The theory provided is therefore to be understood in terms of  $\hat{\boldsymbol{\theta}}_{NT}$  converging to a *pseudo-true* parameter as  $NT \rightarrow \infty$ , that by construction minimizes the penalized criterion even if the penalty does not vanish. To understand the relationship between the *pseudo-true* parameter and the *true* parameter conditional on the penalty, it is helpful to consider precisely how the penalty influences the criterion and delivers the identifiable uniqueness property.

Let  $\hat{\boldsymbol{\theta}}_\pi$  be the point  $\hat{\boldsymbol{\theta}}_\pi := \arg \min_{\boldsymbol{\theta} \in \Theta} Q_\infty(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta})$ , that minimizes the penalized criterion, and  $\boldsymbol{\theta}_0$  be the point  $\boldsymbol{\theta}_0 := \arg \min_{\boldsymbol{\theta} \in \Theta} Q_\infty(\boldsymbol{\theta})$  that minimizes an unpenalized out-sample criterion.  $\boldsymbol{\theta}_\pi$  is the best approximator similar to the misspecification case studied in (White et al., 1980), whereas  $\boldsymbol{\theta}_0$  is the *true* parameter, that is the weights vector that induces  $h_0$  through the kernel, which is the *true* function that provides the best out-of-sample density by its definition. The function  $\hat{h}_\pi := h(\mathbf{x}_t; \hat{\boldsymbol{\theta}}_\pi)$  is the best approximator of  $h_0 := h(\mathbf{x}_t; \boldsymbol{\theta}_0)$  as judged by the penalized criterion  $Q_\infty(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta})$  for a given level of penalization  $\pi$ . The penalty does not imply that  $h(\mathbf{x}_t; \hat{\boldsymbol{\theta}}_\pi) \neq h(\mathbf{x}_t; \boldsymbol{\theta}) \forall \boldsymbol{\theta} \in \Theta \setminus \hat{\boldsymbol{\theta}}_\pi$  and any  $\mathbf{x}_T \in \mathcal{X}$  and all  $NT \in \mathbb{N} \times \mathbb{Z}$ . However, it ensures that  $\hat{\boldsymbol{\theta}}_\pi$  is identifiable unique as the minimizer of the limit criterion even in the case of two observationally equivalent parametrizations  $h(\mathbf{x}_t; \hat{\boldsymbol{\theta}}_\pi) \equiv h(\mathbf{x}_t; \boldsymbol{\theta}^*)$  for some  $\boldsymbol{\theta}^* \in \Theta \setminus \hat{\boldsymbol{\theta}}_\pi$  and any  $\mathbf{x}_T \in \mathcal{X}$  and all  $NT \in \mathbb{N} \times \mathbb{Z}$ .

PROPOSITION. 4 (Identifiable uniqueness). *The function*

$$\hat{h}_\pi := \arg \min_{h_\theta \in \mathcal{H}_\Theta} Q_\infty(h_\theta) + \pi(h_\theta)$$

*produced by  $h_\mathcal{X}$  at point  $\hat{\boldsymbol{\theta}}_\pi := \arg \min_{\boldsymbol{\theta} \in \Theta} Q_\infty(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta})$  is uniquely identified within  $\mathcal{H}_{\Theta_s}$  a simple subset in the infinite dimensional Hilbert*

space  $\mathcal{H}_\Theta$ , if  $\pi$  is a strictly positive penalty function continuous on  $\Theta$ .

Central to the result is that  $\pi(\hat{\boldsymbol{\theta}}_\pi) < \pi(\boldsymbol{\theta}^*)$ , providing that for two observationally equivalent functions, the identified result is the parameter vector that induces a less complex functional form.

So far we have treated  $\pi$  to be fixed at a pre-specified level. However, for any given level of penalization, the solution to the penalized criterion is different. We can make that more explicit by writing it as the limit estimate conditional on a penalty value ( $\hat{\boldsymbol{\theta}}_\infty|\pi$ ), and analyzing the role of  $\pi$  in the divergence  $\|(\hat{\boldsymbol{\theta}}_\infty|\pi) - \boldsymbol{\theta}_0\|$ . This displays the heavy importance on determining an appropriate penalty  $\pi$  as it is crucial to the outcome, see Blasques and Duplinskiy (2018) for some thoughts on how to choose appropriate penalty weights in a general context. Asymptotically, if the impact of  $\pi$  vanishes, for example by using penalties of an  $o((NT)^{-\frac{1}{2}})$ , consistency toward  $\boldsymbol{\theta}_0$  can still be met in the limit, again see Blasques and Duplinskiy (2018) for detail. However, in small samples, similar to the Bayesian case, a researcher can exert influence on the outcome by setting the value of  $\pi$ . In the current framework,  $\pi > 0$  prevents a generality claim as it would follow in the parametric case, however we can still focus the argument on finding an optimal penalty that minimizes  $\|(\hat{\boldsymbol{\theta}}_\infty|\pi) - \boldsymbol{\theta}_0\|$ , or equivalently the function divergence  $\|(\hat{h}_\infty|\pi) - h_0\|$ .

**PROPOSITION. 5** (Best approximation across penalties and weights). *The divergence between the best approximation as judged by the penalized limit criterion given a level of penalization and the true function is smaller than the divergence as evaluated at all other limit estimates resulting under other penalty weights*

$$\|(\hat{h}_\infty|\pi_0) - h_0\| < \|(\hat{h}_\infty|\pi) - h_0\| \quad \forall \pi \in \Pi \setminus \pi_0 \subseteq \mathbb{R}_{>\delta},$$

and results under the penalty  $\pi_0$  that minimizes an out-of-sample criterion

$$\pi_0 : \arg \min_{\pi \in \Pi} Q_\infty(\hat{h}_\infty|\pi), \Pi \subseteq \mathbb{R}_{>\delta}$$

for some small positive  $\delta$ . Hence,  $(\hat{h}_\infty|\pi_0)$  is the best approximation of  $h_0$  over  $\mathcal{H}_{\Theta_s \times \Pi} := \{\mathcal{H}_{\Theta_s}|\pi_1 \times \mathcal{H}_{\Theta_s}|\pi_2 \times \dots \times \mathcal{H}_{\Theta_s}|\pi\} \forall \pi \in \Pi$ , that is across all penalties and weights.

PROPOSITION. 5 implies that if the penalty is chosen by minimizing a criterion out-of-sample, a weights vector can be estimated that produces the function closest to the target function across all penalties and weights. Effectively, a researcher is able to identify an approximation that is arbitrarily close to the *true* curve, by solving the estimator on very large data iteratively for a sufficiently wide range of penalties and selecting the result that performs optimal as an out-of-sample predictor. This is an intuitive solution as  $\theta_0$  carries a natural interpretation as the optimal out-of-sample predictor. The key result, and with that the role of the penalty, is summarized below in fig. 5.10.

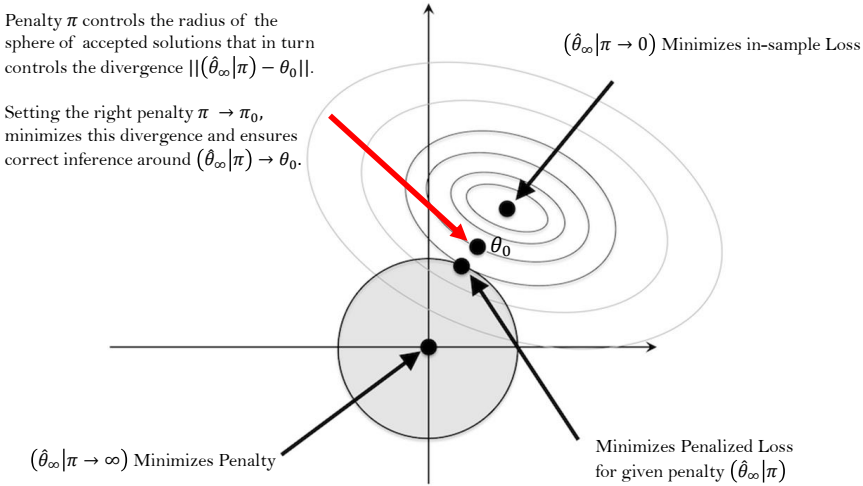


Figure 5.10: For functions  $h$  induced under parameter vectors  $\theta$ , and penalties  $\lambda$  controlled by general penalty functions  $\pi$ , the figure displays graphically the role of the penalty function in managing the closeness of the empirical result to the result that delivers the correct function  $h_0$  associated with the correct marginal coefficients  $h'_0$  of interest. The gray shaded area contains the space of accepted solutions that currently includes the result induced under an infinite penalty, but not the correct result, nor the result that would be obtained when the model fully minimizes in-sample loss. Hence the graph corresponds to the mis-specified case in which the data is under-fitted.

### 5.7.4 Conclusion

This note detailed an adaption of the Kernel Regularized Least Squares model to the panel context, paying particular focus to automated selection of fixed effects. The model was applied in our main paper to study nonlinear trends between environmental indicators and economic development. The key feature of the model that makes it attractive in this type of applied studies is that it provides a straightforward approach to nonlinearity and interpretability within one integrated framework. The difficulty with the approach is that estimation relies on externally set tuning parameters that are not part of the estimated vector of parameters that define the functional relationships in the data. This makes the interpretation of local marginal coefficients highly dependent on correctly tuning the model. The discussion provided high level arguments for optimizing the estimation criterion on validation samples in order to ensure the coefficients admit to standard interpretation.

The discussion provided some examples that highlighted that penalization in the non-parametric context differs from penalization in the GLM case, such as in the popular LASSO. While penalized GLM's require the penalty to vanish asymptotically for generality claims, positive penalization in the limit may be a necessity to ensure identifiable uniqueness for non-parametric models. Regularization or penalization, while primarily known for dealing with over-fitting, was in fact a way to flexibly establish simple subspaces in which consistency theorems hold. As a result, the consistency and normality limits are uniquely defined for every level of penalization, which makes it less straightforward to interpret the estimator, and its derivatives, with usual confidence. However, penalties may still be found that yield estimates that conform standard interpretation. Specifically, penalties that result from minimizing an out-of-sample criterion pull the consistency limit toward the result that induces the optimal conditional distribution implied by the weighted kernel across all penalties and

weights, as judged by the out-of-sample criterion. That result is similar to the standard convergence toward the parameter that delivers the best modeled density in terms of divergence with respect to the true density of the data Kullback and Leibler (1951); White (1994). Under that result, the estimator converges to the result that delivers the best approximation to the true distribution of the data.

It is important to stress that the argument is based on out-of-sample optimization of the same criterion function that was used to fit the model in the estimation sample. Particularly in the case of classification problems this may deviate from common practices. For example, classification problems are often tuned by maximizing accuracy measures that involve fractions of correctly predicted or mis-predicted classes. These widely used metrics do not satisfy the smoothness properties imposed on the in-sample criterion function to obtain consistency to a limit result at a given value of the penalty parameter. A straightforward example of one such violated assumption is the assumed continuity of the estimation criterion in all its arguments, which is needed as part of a standard Consistency proof to ensure limit preserving properties. This continuity breaks because for any level of accuracy (simply the percentage of correctly classified observations), there can exist an infinite number of parameterizations that are judged to be exactly identical by the accuracy criterion. Two simple examples are one model that is correct by predicting .49 and .51 probabilities versus one that predicts 0 versus 1. Moreover, a minute change in a parameter value may change the model's accuracy from 100% to 0%, for example by swapping the margins around probabilities close to .5., so model's with near identical parameters can also be judged as wildly different by an accuracy-based criterion.



## Proofs

Proof to PROPOSITION. 4

*Proof.*  $\hat{\boldsymbol{\theta}}_\pi := \arg \min_{\boldsymbol{\theta} \in \Theta} Q_\infty(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta})$ , is by definition the minimizer of  $Q_\infty(\cdot; \boldsymbol{\theta}) + \pi(\boldsymbol{\theta})$  that is by construction of the least squares function and the penalty function  $\pi : \Theta \rightarrow \mathbb{R}_{>0}$  a real-valued criterion on an open metric space  $\Theta$ . Furthermore there exists some positive constant  $\epsilon$  for which

$$[Q_\infty(\boldsymbol{\theta}' \in \Theta \setminus \Theta_s) + \pi(\boldsymbol{\theta}' \in \Theta \setminus \Theta_s) - Q_\infty(\boldsymbol{\theta} \in \Theta_s) + \pi(\boldsymbol{\theta} \in \Theta_s)] > \epsilon,$$

because for  $Q_\infty(\boldsymbol{\theta} \in \Theta \setminus \Theta_s) \equiv Q_\infty(\boldsymbol{\theta} \in \Theta_s)$ ,

$$[\pi(\boldsymbol{\theta}' \in \Theta \setminus \Theta_s) - \pi(\boldsymbol{\theta} \in \Theta_s)] > \epsilon,$$

by the monotonicity of  $\pi$  on  $\Theta$ . This implies that  $\hat{\boldsymbol{\theta}}_\pi$  minimizes  $Q_\infty(\cdot; \boldsymbol{\theta}) + \pi(\boldsymbol{\theta})$  within a neighborhood  $\Theta_s \subset \Theta$ . Furthermore  $\Theta_s(d)$  has finite radius  $d < \infty$  because

$$[\pi(\boldsymbol{\theta} \in \Theta_s(d)) - \pi(\boldsymbol{\theta} \in \Theta \setminus \Theta_s(d))] \leq \epsilon$$

implies  $d < \infty$ , by finiteness of  $\epsilon$  in turn implied by continuity of the penalty. Finally,  $\Theta_s(d) \subset \Theta$  is compact as it closed because its radius is finite, and its complement  $\Theta \setminus \Theta_s$  is open.

We have now established that uniformly over  $\Theta$ , the estimator must fall eventually inside  $\Theta_s$ . The rest of the argument follows from standard identifiability arguments in compact parameter spaces as in (Bates and White, 1985; Domowitz and White, 1982) focused on  $\Theta_s$ . That is, define a circular neighborhood  $\aleph_k(r) \subset \Theta_s$  with nonnegative radius  $r < d$  that has a compact complement  $\aleph_k(r)^c : \Theta_s \setminus \aleph_k(r)$ .  $\boldsymbol{\theta}_k$  is uniquely identified

on  $\Theta$  as by  $0 < r < d < \infty$ , for every  $(r, d)$ ,

$$\inf_{\boldsymbol{\theta} \in \mathbb{N}_k(r)^c \cap \Theta_s(d) \subset \Theta} \left[ Q_\infty(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta}) - Q_\infty(\hat{\boldsymbol{\theta}}_\pi) + \pi(\hat{\boldsymbol{\theta}}_\pi) \right] > 0.$$

In our case, this is implied by continuity of the criterion and additionally by the fact that for any observationally equivalent point  $\pi(\boldsymbol{\theta}^*)$  such that  $Q_\infty(\boldsymbol{\theta}^*) \equiv Q_\infty(\hat{\boldsymbol{\theta}}_\pi)$ , by definition of  $\hat{\boldsymbol{\theta}}_\pi$  as the minimizer of  $\min_{\boldsymbol{\theta} \in \Theta} Q_\infty(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta})$  the continuity of  $\pi$  implies

$$\pi(\hat{\boldsymbol{\theta}}_\pi) < \pi(\boldsymbol{\theta}^*).$$

□

Proof to PROPOSITION. 5

*Proof.* Let  $\hat{\boldsymbol{\theta}}_\pi = \hat{\boldsymbol{\theta}}_\infty|_\pi := \arg \min_{\boldsymbol{\theta} \in \Theta} Q_\infty(\boldsymbol{\theta}) + \pi(\boldsymbol{\theta})$  be the minimizer of the penalized criterion for a certain level of penalization and  $\boldsymbol{\theta}_0 := \arg \min_{\boldsymbol{\theta} \in \Theta} Q_\infty(\boldsymbol{\theta})$  the minimizer of an unpenalized out-of-sample criterion. When plugging the *true* parameter in the penalized criterion, then taking  $\|\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}_0\| \rightarrow 0$  implies that similarly  $|[Q_\infty(\hat{\boldsymbol{\theta}}_\pi) + \pi(\hat{\boldsymbol{\theta}}_\pi)] - [Q_\infty(\boldsymbol{\theta}_0) + \pi(\boldsymbol{\theta}_0)]| \rightarrow 0$ . This minimization is solved if the in-sample criterion evaluates  $|Q_\infty(\hat{\boldsymbol{\theta}}_\pi) - Q_\infty(\boldsymbol{\theta}_0)| \rightarrow 0$  equivalently, as then immediately also  $|\pi(\hat{\boldsymbol{\theta}}_\pi) - \pi(\boldsymbol{\theta}_0)| \rightarrow 0$ . Hence, either  $|\pi(\hat{\boldsymbol{\theta}}_\pi) - \pi(\boldsymbol{\theta}_0)| \rightarrow 0$  or  $|Q_\infty(\hat{\boldsymbol{\theta}}_\pi) - Q_\infty(\boldsymbol{\theta}_0)| \rightarrow 0$ , is sufficient for  $\|\hat{\boldsymbol{\theta}}_\pi - \boldsymbol{\theta}_0\| \rightarrow 0$ .

Any such result following from taking both penalties  $\pi(\hat{\boldsymbol{\theta}}_\pi)$  and  $\pi(\boldsymbol{\theta}_0)$  to zero simultaneously as  $N \rightarrow \infty$  is prohibited by the fact that  $\pi : \Theta \rightarrow \mathbb{R}_{>0}$ . However  $|Q_\infty(\hat{\boldsymbol{\theta}}_\pi) + \pi(\hat{\boldsymbol{\theta}}_\pi)| - |Q_\infty(\boldsymbol{\theta}_0) + \pi(\boldsymbol{\theta}_0)|$  attains a minimum when setting the penalty to minimize the criterion defined on out-of-sample errors. Specifically since  $\boldsymbol{\theta}_0$  is by construction the minimum of the out-of-sample criterion in the limit, setting  $\pi_0$  to minimize

$\arg \min_{\pi \in \Pi \ \forall \ \subseteq \mathbb{R}_{\geq 0}} \mathcal{Q}_{\infty}(\hat{\boldsymbol{\theta}}_{\pi})$  gives

$$\pi_0 : \arg \min_{\pi \in \Pi} |\mathcal{Q}_{\infty}(\hat{\boldsymbol{\theta}}_{\pi}) - \mathcal{Q}_{\infty}(\boldsymbol{\theta}_0)|.$$

and if  $|\mathcal{Q}_{\infty}(\hat{\boldsymbol{\theta}}_{\pi}) - \mathcal{Q}_{\infty}(\boldsymbol{\theta}_0)| \rightarrow 0$ , it must follow that

$$|\mathcal{Q}_{\infty}(\hat{\boldsymbol{\theta}}_{\pi}) - \mathcal{Q}_{\infty}(\boldsymbol{\theta}_0)| \rightarrow 0,$$

and,

$$|\pi(\hat{\boldsymbol{\theta}}_{\pi}) - \pi(\boldsymbol{\theta}_0)| \rightarrow 0.$$

If  $\Pi \subseteq \mathbb{R}_{\geq 0}$  is constructed such that  $\pi_0 \in \Pi$  for which the  $\arg \min$ 's above reach 0, then  $\|(\hat{h}_{\infty}|\pi_0) - h_0\| = 0$  would follow and we reach the *true* target function. Now  $\Pi \subseteq \mathbb{R}_{\geq 0}$  can contain penalties infinitely close to 0, in practice one must work with finite sets for a grid search across  $\Pi$  and construct instead a set  $\Pi \subseteq \mathbb{R}_{\geq \delta}$  being the set of all possible parameters bounded away from zero by some arbitrarily small positive constant  $\delta$ . If  $\pi_0 \notin \Pi$  for which  $\|(\hat{h}_{\infty}|\pi_0) - h_0\| = 0$ , then still

$$|\mathcal{Q}_{\infty}(\hat{\boldsymbol{\theta}}_{\pi_0}) - \mathcal{Q}_{\infty}(\boldsymbol{\theta}_0)| < |\mathcal{Q}_{\infty}(\hat{\boldsymbol{\theta}}_{\pi}) - \mathcal{Q}_{\infty}(\boldsymbol{\theta}_0)| \ \forall \ \pi \in \Pi \setminus \pi_0 \subseteq \mathbb{R}_{>\delta},$$

thus also

$$|\mathcal{Q}_{\infty}(\hat{\boldsymbol{\theta}}_{\pi_0}) - \mathcal{Q}_{\infty}(\boldsymbol{\theta}_0)| < |\mathcal{Q}_{\infty}(\hat{\boldsymbol{\theta}}_{\pi}) - \mathcal{Q}_{\infty}(\boldsymbol{\theta}_0)| \ \forall \ \pi \in \Pi \setminus \pi_0 \subseteq \mathbb{R}_{>\delta}$$

and therefore

$$\|(\hat{\boldsymbol{\theta}}_{\infty}|\pi_0) - \boldsymbol{\theta}_0\| < \|(\hat{\boldsymbol{\theta}}_{\infty}|\pi) - \boldsymbol{\theta}_0\| \ \forall \ \pi \in \Pi \setminus \pi_0 \subseteq \mathbb{R}_{>\delta},$$

which induces through the definition of  $h$  as the weighted kernel also

$$\|(\hat{h}_{\infty}|\pi_0) - h_0\| < \|(\hat{h}_{\infty}|\pi) - h_0\| \ \forall \ \pi \in \Pi \setminus \pi_0 \subseteq \mathbb{R}_{>\delta},$$

implying that  $(\hat{h}_{\infty}|\pi_0)$  turns out to be the best approximation of  $h_0$  for all penalties in  $\Pi$  that each result itself as a best approximator within

the subset  $\mathcal{H}_{\Theta_s}|\pi$  within the penalized criterion necessarily falls given the level of penalization. In this case  $\pi_0$  simply plays the role of a pseudo-true penalty that delivers a pseudo-true results, which can be detected when the penalty is at the boundary of the grid  $\Pi$  or is expected when the resolution of the grid is such that  $\Pi$  is not approximately continuous.  $\square$



# Chapter 6

## Vector Spatial Time Series

### Chapter Summary

This paper introduces a Spatial Vector Autoregressive Moving Average (SVARMA) model in which multiple cross-sectional time series are modeled as multivariate, possibly fat-tailed, spatial autoregressive ARMA processes. The estimation requires specifying the cross-sectional spillover channels through spatial weights matrices. The paper explores a kernel method to estimate the network topology based on similarities in the data. This method is able to capture interesting network structures that transmit effects based on geographic proximity, but also over far distances based on economic similarity. The paper discusses the model's properties and its estimation using a penalized Maximum Likelihood criterion. The empirical performance of the estimator is explored in a simulation study. The model is used to study a spatial time series of pollution and household expenditure data in Indonesia. The analysis finds that the new model improves in terms of implied density, and better neutralizes residual correlations than the VARMA, using fewer parameters. The results suggest that growth in household expenditures precedes pollution reduction, particularly after the expenditures of poorer households increase; that increasing pollution is followed by reduced growth in expenditures, particularly reducing the growth of poorer households; and that there are significant spillovers from bottom-up growth in expenditures. The paper does not find evidence for top-down growth spillovers. Feedback between the identified mechanisms may contribute to pollution-poverty traps and the results imply that pollution damages are economically significant.<sup>1</sup>

---

<sup>1</sup>This chapter is based on “*Pollution and Expenditures in a Penalized Vector Spatial Autoregressive Time Series Model with Data-Driven Networks*” published by the *World Bank*, the full reference is Andree et al. (2019).

## 6.1 Introduction

Environmental and economic systems are deeply tied with one another, but consensus on the causal pathways is even in the most isolated settings seldom achieved. For instance: Does economic growth lead to environmental degradation or improvement? At the same time, to what extent does pollution take its toll on growth? The answers to both questions – and their interrelation – might tell us how places end up in pollution-poverty traps, or succeed in cleaning up the environment. The scope of these questions clearly calls for a holistic framework around the environmental-economic domain with both space and time dimensions. In this paper we introduce a framework that allows the researcher to model multiple interacting spatial time series.

Time series offers invaluable insights to trace the arrow of causality. Univariate autoregressive moving average (ARMA) models are among the most fundamental statistical models to explore dynamics in observations that are collected sequentially over time. As we are interested in interactions between variables, we focus on their multivariate counterparts, known as vector autoregressive moving averages (VARMA). Moving averages are characterized by a cutoff in the auto-covariance functions. This implies that the effects represent parts in a model with short memory, while autoregressive parts represent long-memory effects. Short memory effects may relate to unobservable factors that slowly assimilate into the model, e.g. effects for which it takes time to be completely absorbed by a system. This is realistic for policy interventions in the context of economic systems, but it may also be realistic for natural phenomena. The ability to model effects that decay or remain free from feedback provides a framework to differentiate between long and short run causality as in Dufour and Renault (1998); Dufour et al. (2006) and Dufour and Taamouti (2010). This has added value when one is specifically interested in testing economic theories about the timing and duration of responses.

The VARMA constitutes the backbone of many studies on causality due to the strong relationship between invertibility and Granger-causality, and the ability to test for the direction of effects (Sims, 1972). Estimation of VARMA models is discussed for example by Roy et al. (2014), but also in textbooks by Brockwell and Davis (2002), Reinsel (2003), and Lütkepohl (2005). In this paper we work around the concept of Granger-causality (Granger, 1969, 1980; Covey and Bessler, 1992).<sup>2</sup> This concept involves eliminating the history of variables from the joint distribution of all variables. There is no Granger-causality from the eliminated variables if the conditional density of the model did not improve significantly. To avoid problems related to repeated testing, discussed for example by Hendry (2017), we follow Granger et al. (1995) in using Information Criteria (IC) to decide between economic theories. Minimization of IC, guarantees the selection of the model that attains the lower average Kullback-Leibler bound in the limit, see Sin and White (1996) for detail. IC methods favor parsimony, hence also work when some parameters may be unidentified under the null. They offer a general solution when models are strictly nested, overlapping or non-nested, linear or nonlinear, and well-specified or miss-specified. In the miss-specified case, minimizing IC results in a *pseudo-true* model that still delivers the best possible hypothesis about Granger-causality as judged by the criterion function across all possible hypotheses generated under the model and the parameter space.

Consistent estimation of VARMA models is closely related to the ability to identify it uniquely. In particular, stationary and invertible VARMA models have both VAR and VMA representations. Standard approaches in the VARMA literature that deal with non-uniqueness focus on final equations or echelon forms (see Lütkepohl (2005)). We follow a penalization approach to ensure a unique VARMA solution to the estimation criterion. This approach can be seen as a Ridge or Lasso regression for

---

<sup>2</sup>We say that one variable does not cause the other, if adding past observations of the former to the information set with which we predict future observations of the latter does not improve the conditional density.



VARMA models. By penalizing either the VAR or VMA coefficients in the criterion function, we rule out the multiplicity of solutions where both components essentially cancel each other out.

While the VARMA treatment takes care of the feedback over time, it does not incorporate the possibility of contemporaneous feedback. To illustrate the latter, a shock can affect an area both directly as well as indirectly through its neighbors. The spatial structure therefore acts as a multiplier of the initial shock. If we neglect this multiplier, the VARMA will likely overestimate the direct effects of interest. Hence, it is crucial to filter out the spatial dependence at each point in time. Extending the VARMA framework with spatial effects yields the spatial vector autoregressive moving average (SVARMA) model. The SVARMA can be thought of as the MA extension to the spatial-VAR discussed in (Beenstock and Felsenstein, 2007). To model spatial dependence, we need to specify the underlying spatial structure. Spatial weights are designed around a concept of distance, which may not necessarily be geographic. In this paper we build networks based on economic similarity rather than geographic proximity. Under this notion, areas are more likely to share dynamics when they have similar economic fundamentals. At the same time, they are not likely to share spillovers, if they are dissimilar. We propose a flexible method that allows to integrate estimation of the spatial structure using kernels. In this context, the kernel bandwidth controls the neighborhood size that in turn determines similarity. Large bandwidths lead to many far and weak connections and small bandwidths yield strong local clusters.

We use the penalized SVARMA framework with integrated estimation of networks to study interactions between pollution and household expenditures in Indonesia between 1999-2014. We focus particularly on the effect of economic growth on pollution levels, the effect that pollution in turn has on economic growth, and the dynamic feedback that arises as both

channels spill over into each other. Additionally, we seek to disentangle how the different households are affected by – and affect – pollution change. In turn, this strongly depends on the presence of bottom-up and top-down growth spillovers. Finally, we explore the differential in these relationships between average urban areas and highly polluted areas. We use the estimated parameters in an Impulse Response framework. Our methods and data suggest several interesting feedback mechanisms.

The remaining part of this paper is as follows. Section 6.2 introduces the model. Specifically, we detail the process equations, and our approach to build connectivity up from the data using kernels. Section 6.3 discusses the properties of the model, specifically stability, invertibility, non-uniqueness, and the IRF. Section 6.4 provides the tools needed for estimation. Our appendix provides simulation results on the empirical distributions of all the parameters in sample sizes relevant to our empirical application. The framework is applied in section 6.5 to study dynamics in a multivariate cross-sectional time-series of pollution and household expenditures. We study the IRF and discuss policy implications of the results. Section 6.6 concludes.

## **6.2 Spatial Vector Autoregressive Moving Average model**

This section details VARMA approaches for multiple panels that exhibit spatial feedback. Figure 6.1 summarizes the components of the SVARMA and its relation to other widely used models. SVARMA allows instantaneous effects between observations within cross-sections, and long and short run effects in the time-dimension between and within panels. This provides a dynamic framework to study causation and feedback between spatially autocorrelated time-series. Our use of the spatial framework is intended to filter out dependencies and improve

estimation of the underlying cross-sectional ARMA structures. This is important because contemporaneous, cross-sectional feedback works as a multiplier. Without distinguishing this feedback from the impulse mechanisms, the direct impacts may be severely overestimated. This is similar to the contemporaneous case in which instruments are used to isolate effect from feedback.

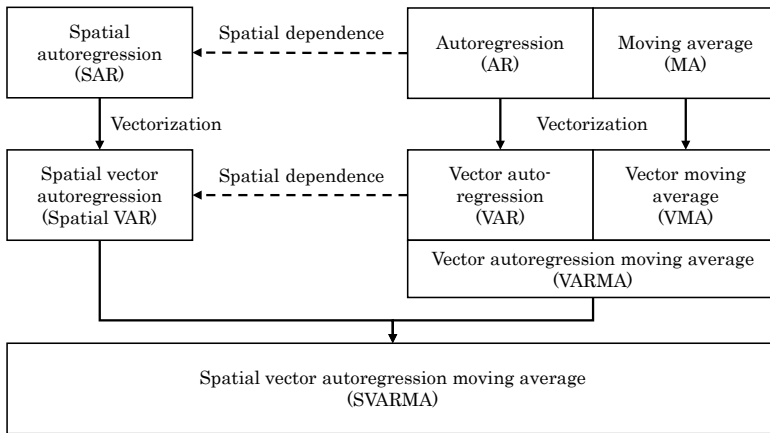


Figure 6.1: This chart presents an overview of the constituents of the Spatial vector autoregressive moving average (SVARMA) model described in this section. Note that AR and MA processes may also be defined on single cross-sections resulting in spatial-time series, or cross-sectional ARMA models – not depicted in this diagram.

The SVARMA model can improve inference compared to VAR or spatial VARs. The distinction between autoregressive and residual properties is useful for forecasting and for distinguishing between short and long effects, but moreover it plays a role in deriving consistent model statistics.<sup>3</sup> If the autoregressive parameter is correct in the sense that the response at the *true* parameter confirms to the mean of the endogenous variable conditional on partial information, then the score vector is generally not a martingale difference sequence as the disturbance vector in the *true* model is still autocorrelated. While the AR structure of the model is

<sup>3</sup>Neutralizing serial dependence is required to satisfy the martingale property of the score needed to apply a standard CLT.

correct, the objective function does not correspond to the *true* objective function. The random variables that compose the score are therefore not guaranteed to be martingale difference sequences. While the AR structure produces correct responses, it will generally not be possible to assign correct probability to the possibility that those responses are in fact zero.<sup>4</sup> As an effect, the statistical framework used to asses validity of the causal claims is invalidated.

We use the following notation,  $a$  is a scalar,  $\mathbf{a}$  is a vector,  $A$  is a matrix, and  $\mathbf{A}$  is a matrix that arises from stacking multiple blocks of  $A$  together.  $\mathcal{A}$  is the collection of matrices  $\{A_0, A_1, \dots, A_p\}$ ,  $\mathcal{A}$  collects  $\{\mathbf{A}_0, \mathbf{A}_1, \dots, \mathbf{A}_p\}$ . Finally,  $A_{i:j}$  and  $\mathbf{A}_{i:j}$  respectively select elements  $i$  to  $j$  from those sets. We reserve  $\mathbf{w} := (\mathbf{x}, \mathbf{y})$  for the joint sequence of two vector processes  $\mathbf{x}$  and  $\mathbf{y}$ . While we admit that in the case of two univariate sequences, the joint sequence is a vector, we use  $w := (x, y)$  for the joint process in this isolated case. To avoid confusion between  $\mathbf{w} \in \mathcal{W}$ , we divert from most spatial literature by using  $C$  as a connectivity matrix.

### 6.2.1 Vector Autoregressive Moving Average model

In the multiple univariate sequence case,  $w := (x, y)$ ,  $\varepsilon := (\varepsilon^x, \varepsilon^y)$ , a VARMA is a process

$$A_0 w_t + A_1 w_{t-1} + \dots + A_p w_{t-p} = M_0 \varepsilon_t + M_1 \varepsilon_{t-1} + \dots + M_q \varepsilon_{t-q} \quad \forall t \in \mathbb{Z}, \quad (6.1)$$

with parameter matrices structured as

$$A := \begin{bmatrix} a^{xx} & a^{xy} \\ a^{yx} & a^{yy} \end{bmatrix}, M := \begin{bmatrix} m^{xx} & m^{xy} \\ m^{yx} & m^{yy} \end{bmatrix}. \quad (6.2)$$

---

<sup>4</sup>Corrections to the CLT are available if the score vector exhibits a suitable form of weak dependence, see for example Pötscher and Prucha (1997). In practice it is not straightforward to judge whether the score adheres a suitable form of weak dependence. This suggests that a researcher is always better neutralizing the residuals when possible.

In the multiple cross-section case  $\mathbf{w} := (\mathbf{x}, \mathbf{y})$ ,  $\boldsymbol{\varepsilon} := (\boldsymbol{\varepsilon}^x, \boldsymbol{\varepsilon}^y)$  stacked  $n_x$  and  $n_y$  vectors for every  $t$ , we can work by defining the parameter matrices as  $A^{ij} := a^{ij}I_{n_i}$  and  $M^{ij} := m^{ij}I_{n_i}$ , structured as

$$\mathbf{A}_{0:p} := \begin{bmatrix} A_{0:p}^{xx} & A_{0:p}^{xy} \\ A_{0:p}^{yx} & A_{0:p}^{yy} \end{bmatrix}, \mathbf{M}_{0:p} := \begin{bmatrix} M_{0:p}^{xx} & M_{0:p}^{xy} \\ M_{0:p}^{yx} & M_{0:p}^{yy} \end{bmatrix}, \mathbf{I} := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}, \quad (6.3)$$

in which  $O$  is a matrix of zeros, to write the cross-sectional VARMA as

$$\mathbf{A}_0 \mathbf{w}_t + \mathbf{A}_1 \mathbf{w}_{t-1} + \dots + \mathbf{A}_p \mathbf{w}_{t-p} = \mathbf{M}_0 \boldsymbol{\varepsilon}_t + \mathbf{M}_1 \boldsymbol{\varepsilon}_{t-1} + \dots + \mathbf{M}_q \boldsymbol{\varepsilon}_{t-q} \quad \forall t \in \mathbb{Z}, \quad (6.4)$$

in which  $\{\mathbf{A}_0, \mathbf{A}_1, \dots, \mathbf{A}_p\} \in \mathcal{A}$  and  $\{\mathbf{M}_0, \mathbf{M}_1, \dots, \mathbf{M}_p\} \in \mathcal{M}$  are thus  $n_w \times n_w$  parameter matrices induced by scalar coefficients, and  $\boldsymbol{\varepsilon}_t \sim p_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}_t, \boldsymbol{\Sigma}; \boldsymbol{\nu})$  is a disturbance vector that has  $n_x$  elements drawn from a distribution with an unknown scale matrix  $\Sigma^x$  and possibly other parameters contained in  $\boldsymbol{\nu}^x$  and the next  $n_y$  elements drawn from a distribution with an unknown scale matrix  $\Sigma^y$  and possibly other parameters contained in  $\boldsymbol{\nu}^y$ . This allows  $\Sigma^x \neq \Sigma^y$  and  $\boldsymbol{\nu}^x \neq \boldsymbol{\nu}^y$ , but also  $\Sigma^x = \Sigma^y$  and  $\boldsymbol{\nu}^x = \boldsymbol{\nu}^y$ , or any combination thereof. The parametric distributions however are of the same family, and controlled by a same function  $p_{\boldsymbol{\varepsilon}}$ .

It is standard that eq. (6.4) is linear in all its components, and does not allow for any simultaneous feedback. Following standard normalization rules,  $\mathbf{A}_0$  and  $\mathbf{M}_0$  have unit diagonals, i.e.  $\mathbf{A}_0 = \mathbf{M}_0 = \mathbf{I}$ , but this is not necessarily the case. In the multiple cross-section case eq. (6.4) no longer involves multiple one-dimensional sequences, and  $\mathbf{A}_0 = \mathbf{M}_0 = \mathbf{I}$  is severely restrictive, especially as  $n$  grows. If observations within the cross-section influence each other over time with an interval  $\tau$ , while cross-sections are observed at an interval  $t$  that is a multiple of  $\tau$ , then the interactions between cross-sectional observations seem instantaneous from the observer's perspective, see also the examples in Granger (1980). The SVARMA is intended to explain part of the values of elements in

$\mathbf{w}$  in terms of the remaining contemporaneous elements of  $\mathbf{w}_t$ . We work with  $\mathbf{A}_0$  as a matrix that allows for instantaneous spillovers. We focus on the specific case in which elements in  $n_y$  and elements in  $n_x$  are cross-sectionally dependent.

### 6.2.2 Spatial Vector Autoregressive Moving Average model

We can write SVARMA using  $\mathbf{M} = \mathbf{I}$  by defining  $\mathbf{A}_0$  in eq. (6.4) as a matrix consisting of a unit diagonal and a non-unit-diagonal component  $\mathbf{C}$  that structures the contemporaneous feedback across the elements of  $n_w$ ,  $\mathbf{A}_0 = \mathbf{I} + \mathbf{A}_\mathbf{C}$ , with  $\mathbf{A}_\mathbf{C} = -\boldsymbol{\rho} \circ \mathbf{C}$  in which  $\boldsymbol{\rho}$  is a vector with the first  $n_x$  elements consisting out of  $\rho^x$  and the subsequent  $n_y$  elements equal to  $\rho^y$ .  $\boldsymbol{\rho}$  multiplies element-wise, or “weighs” the connectivity matrix  $\mathbf{C}$  that has diagonal blocks  $C_{n_x}, C_{n_y}$  and zeros on the off diagonal blocks,

$$(\mathbf{I} + \mathbf{A}_\mathbf{C})\mathbf{w}_t + \mathbf{A}_1\mathbf{w}_{t-1} + \dots + \mathbf{A}_p\mathbf{w}_{t-p} = \boldsymbol{\varepsilon}_t + \mathbf{M}_1\boldsymbol{\varepsilon}_{t-1} + \dots + \mathbf{M}_q\boldsymbol{\varepsilon}_{t-q} \quad \forall t \in \mathbb{Z}. \quad (6.5)$$

Alternatively, we can work with  $\mathbf{A}_0 = \mathbf{I}$ , after multiplying all the autoregressive filters and moving average parameters with the appropriate spatial multipliers:

$$\mathbf{w}_t + \mathbf{S}\mathbf{A}_1\mathbf{w}_{t-1} + \dots + \mathbf{S}\mathbf{A}_p\mathbf{w}_{t-p} = \mathbf{S}\boldsymbol{\varepsilon}_t + \mathbf{S}\mathbf{M}_1\boldsymbol{\varepsilon}_{t-1} + \dots + \mathbf{S}\mathbf{M}_q\boldsymbol{\varepsilon}_{t-q} \quad \forall t \in \mathbb{Z}, \quad (6.6)$$

with  $\mathbf{S} = (\mathbf{I} + \mathbf{A}_\mathbf{C})^{-1}$ . We refer to eq. (6.6) as the structural representation of the SVARMA. Finally, we can also work with spatial errors, and spatially multiplied autoregressive coefficients by introducing  $\boldsymbol{\epsilon}_t = \mathbf{S}\boldsymbol{\varepsilon}_t$  and  $\mathbf{H} = \mathbf{S}\mathbf{A}$ , such that for  $\mathbf{A}_0 = \mathbf{I} + \mathbf{A}_\mathbf{C} = \mathbf{S}^{-1}$ ,  $\mathbf{H}_0 = \mathbf{S}\mathbf{S}^{-1} = \mathbf{I}$  we have

$$\mathbf{w}_t + \mathbf{H}_1\mathbf{w}_{t-1} + \dots + \mathbf{H}_p\mathbf{w}_{t-p} = \boldsymbol{\epsilon}_t + \mathbf{M}_1\boldsymbol{\epsilon}_{t-1} + \dots + \mathbf{M}_q\boldsymbol{\epsilon}_{t-q} \quad \forall t \in \mathbb{Z}. \quad (6.7)$$

This is the normalized VARMA representation of the SVARMA, and differs from the non-spatial model by the fact that while we parameterize the time dynamics at the cross-sectional level, a heterogeneous

dependence structure at the observational level arises through the spatial network matrices. This is a powerful way of modeling high-dimensional dependencies at the observational level as it allows for a large number of correlation channels using relatively few parameters. We will keep the model in this form unless stated otherwise.

### 6.3 Model properties

We can define two operators that respectively filter the (spatial) autoregressive effects and produce the moving averages, and summarize the SVARMA as

$$\mathbf{H}(L)\mathbf{w}_t = \mathbf{M}(L)\boldsymbol{\epsilon}_t \quad \forall t \in \mathbb{Z}, \quad (6.8)$$

by defining  $L$  as a lag operator that has the effect that  $L\mathbf{w}_t = \mathbf{w}_{t-1}$ , and where  $\mathbf{H}(L) = \mathbf{H}_0 + \mathbf{H}_1L + \dots + \mathbf{H}_pL^p$  and  $\mathbf{M}(L) = \mathbf{M}_0 + \mathbf{M}_1L + \dots + \mathbf{M}_qL^q$  are full rank matrix-valued polynomials.

Equation (6.8) is convenient notation for the SVARMA because it allows us to condition theory directly on components similar to the standard case of eq. (6.4), and understand standard results for invertibility, stability, and Granger-causality simply as high-level conditions on the spatially multiplied autoregressive and moving average components. In the general case of misspecification, model invertibility and process invertibility are not the same.<sup>5</sup> Though non-stationary processes may be invertible, they are generally not causal in the control theoretical sense (Boudjellaba et al., 1992). Analysis should therefore focus on invertible stationary processes under an axiom of correct specification. This complicates matters with respect to the more commonly excepted axiom of misspecification that provides descriptions in terms of pseudo-true correlations in the data. When the model is correct, fading memory properties and process invertibility cannot simply be assumed to be properties of the data. Instead,

---

<sup>5</sup>See for example Blasques et al. (2018) for results on the relation between filters and DGPs.

these properties are directly related to the properties of the model itself and the range of parameter values considered.<sup>6</sup> Below, we will highlight relevant parameter regions and discuss invertibility, and stability results for SVARMA models following theory for standard VARMA models found in Lütkepohl (2005) or Brockwell and Davis (2002). The results will also show how multiple representations may equally well describe the data, which is why we shall discuss a penalized estimation criterion.

### 6.3.1 Causal SVAR and it's SMA representation

An important aspect of stationary SVARMA models is that under regularity conditions the SVAR(1) part is causal (in the control theoretical sense that it is a nonanticipative system) and has an infinite-order SMA representation. Say an SVAR(1) is written as

$$\mathbf{w}_t = \Phi \mathbf{w}_{t-1} + \epsilon_t \quad \forall t \in \mathbb{Z}, \quad (6.9)$$

with  $\Phi z = -\mathbf{H}_1 Lz - \dots - \mathbf{H}_p L^p z$ . Assuming some form of fading memory, eq. (6.9) may be expanded by a process of infinite back-substitution, giving rise to an infinite-order multivariate spatial autoregressive moving average:

$$\mathbf{w}_t = \{\epsilon_t + \Phi \epsilon_{t-1} + \Phi^2 \epsilon_{t-2} + \dots + \Phi^\infty \epsilon_{t-\infty}\} \quad \forall t \in \mathbb{Z}. \quad (6.10)$$

For the sequence  $\{\Phi, \Phi^1, \Phi^2, \dots, \Phi^\infty\}$  to converge, it is necessary and sufficient that all the moduli of the eigenvalues of  $\Phi$  remain within the unit circle, see section 6.3.3. Stationarity and invertibility conditions that apply to eq. (6.8) are naturally an extension of this first order autoregressive case, which is itself a generalization of the scalar ARMA case.

---

<sup>6</sup>Proofs for Stationarity and Ergodicity of data generated by VARMA models are widespread and can be found for example in (Nsiri and Roy, 1993). Stelzer (2008) treat multivariate Generalized ARMA models including non-identity links, (Zheng et al., 2015) treat nonlinear theory for Multivariate Markov-switching ARMA processes, finally Andree et al. (2017a) show that multivariate ARMA structures can generate geometrically Ergodic data even when a nonlinear observation-driven spatial dependence process is considered.



This high-level condition is the same as the one for VARMA models, the difference is that in the case of the SVARMA, the autoregressive properties are partly determined also by the spatial multiplier. Specifically, if  $\det(\mathbf{H}(z)) \neq 0 \forall z \in \mathbb{C}, |z| < 1$ , then there exists an infinite order representation

$$\mathbf{w}_t = \Psi(L)\epsilon_t = \{\Psi_0\epsilon_t + \Psi_1\epsilon_{t-1} + \Psi_2\epsilon_{t-2} + \dots + \Psi_\infty\epsilon_{t-\infty}\} \forall t \in \mathbb{Z}. \quad (6.11)$$

with the matrices  $\Psi_k$  generated by

$$\mathbf{H}(z)\Psi(z) = \mathbf{M}(z). \quad (6.12)$$

The conditions

$$\mathbf{H}_0 := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}, \mathbf{M}_0 := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}, \text{ imply that } \Psi_0 := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}. \quad (6.13)$$

### 6.3.2 Invertible SMA as a SVAR

If and only if  $\det(\mathbf{M})(z) \neq 0$  for all  $z$  such that  $|z| < 1$ , the process is invertible and the spatial disturbance vector can also be written as

$$\epsilon_t = \Pi(L)\mathbf{w}_t = \{\Pi_0\mathbf{w}_{t-1} + \Pi_1\mathbf{w}_{t-1} + \Pi_2\mathbf{w}_{t-2} + \dots + \Pi_\infty\mathbf{w}_{t-\infty}\} \forall t \in \mathbb{Z}. \quad (6.14)$$

The matrices  $\Pi_k$  are generated by

$$\mathbf{M}(z)\Pi(z) = \mathbf{H}(z). \quad (6.15)$$

The conditions eq. (6.13) imply that

$$\Pi_0 := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}. \quad (6.16)$$

### 6.3.3 Stability in canonical state space

The stability and invertibility conditions may alternatively be understood in a state-space context. Consider a controllable canonical state-space representation:

$$\mathbf{w}_t = \mathbf{H}^{-1}(L)\{\mathbf{M}(L)\boldsymbol{\epsilon}_t\} = \mathbf{M}(L)\boldsymbol{\Xi}_t \quad \forall t \in \mathbb{Z}, \quad (6.17)$$

where  $\boldsymbol{\Xi}_t = \mathbf{H}^{-1}(L)\boldsymbol{\epsilon}_t$ .

Equation (6.17) is defined through a transition equation that corresponds to a first-order Markov process. It is commonly known that multivariate linear stationary processes that have coefficients that are absolutely summable are invertible if and only if its spectral density is regular everywhere. One can work with eq. (6.17) to derive the companion matrix, and see that stability follows if the eigenvalues of  $\boldsymbol{\Phi}$  lie inside the unit circle. Additional details are provided in section 6.7.2.

### 6.3.4 Uniqueness

Since an invertible SVARMA process has both SVAR and SMA representations by rewriting either part, uniqueness is not ensured. In order to ensure uniqueness of the SVARMA, restrictions on the AR and MA operators are required to ensure that there is only a single pair of  $\mathbf{H}(L)$  and  $\mathbf{M}(L)$  that satisfy eq. (6.8). The first source of non-uniqueness relates to the fact that multiple combinations for  $\mathbf{H}(L)$  and  $\mathbf{M}(L)$  can be found for different values of the operators at  $t = 0$ . This is ruled out by a suitable form of normalization. It is usually ruled out that the operators cancel each other out by the assumption that the AR and MA operators have no common factors. However, even if restrictions are in place that ensure this in an estimation algorithm, it does not rule out that SVAR and SMA representations of the SVARMA can be found that fit the data equally well. Lütkepohl (2005) discusses the

so-called final equations and echelon forms that are unique. Additional restrictions on the structure of both  $\mathbf{H}$  and  $\mathbf{M}$  can be found, but we propose to penalize the MA parts in the criterion. The penalty ensures that the criterion always prefers setting both AR and MA parts to zero rather than having them cancel each other out at any arbitrary value. Furthermore, if both an SVAR representation can be found and an SMA representation, the SVAR representation will be favored over the SMA in order to minimize the penalty. In principle, the penalization approach works if either the AR or the MA parts are penalized. Penalizing the AR part involves a prior belief that the sequences do not feedback, and that the impulse responses are of a short-memory type. Penalizing the MA parts can intuitively be understood as prioring on the belief that the *true* process exhibits endogenous feedback, which reconciles better with the endogeneity concerns that lead many micro-economists to promote the use of IV approaches in contemporaneous regressions, and the general goal of having a parsimonious description of the data to reduce regression uncertainty.

### 6.3.5 Impulse Response Functions

Given an SVARMA system, it may be insightful to know precisely how idiosyncratic impulses on the input side affect the output variables. By considering an isolated impulse in  $\boldsymbol{\varepsilon}$ , for example a positive shock in  $\boldsymbol{\varepsilon}^x$  while holding all other disturbances at zero for all times, one can isolate the effect of an exogenous change in  $\mathbf{x}_t$  as it moves through the entire SVARMA system. Specifically, consider a mechanism activated at a certain  $t$  that produces a pulse sequence

$$\mathbf{p}(t) = \begin{cases} \boldsymbol{\zeta}, & t = 0, \\ 0, & t \neq 0. \end{cases} \quad \forall t \in \mathbb{Z}.$$

$\zeta$  is the magnitude of the value of the considered impact. If  $\mathbf{e}$  is the vector with a unit in the positions where a shock occurs, the response by the system is represented by

$$\mathbf{w}_t = \Psi(L)\{\mathbf{p}(t)\mathbf{e}\} \quad \forall t \in \mathbb{Z}. \quad (6.18)$$

This system is inactive until  $t = 0$ , after which it generates the sequence  $\{\Psi_0\mathbf{e}, \Psi_1\mathbf{e}, \dots, \Psi_\infty\mathbf{e}, \dots\}$ . The impulse travels through the entire SVARMA structure with speed depending on the spatial autoregressive and time autoregressive parameters. It is possible to trace all the routes by taking into account how the spatial autoregressive polynomial  $\mathbf{H}(z)$  is structured. Finally, confidence bands around the response can be obtained by repeating an experiment of identical impact, and drawing different parameters for the SVARMA structure randomly from their confidence bands. Trivially, the sequence eq. (6.18) converges to zero exponentially fast *a.s.*, for a stationary and ergodic model. Hence, even when the aggregate behavior of all parameters is not directly of interest, the IRF provides a useful tool to explore stability of the estimated model, which is important also for Granger-causal inference on the individual parameters.

## 6.4 Estimation

### 6.4.1 Parameterizing spatial weight matrices using Gaussian kernels

Key to the estimation of contemporaneous spatial effects is specifying a network structure that defines the spill-over channels between cross-sectional observations. In spatial literature, the weights matrix is based on geographical distances (Anselin, 1988), but it is equally possible to define networks based on economic distances (see for example the application of Blasques et al. (2016)). Furthermore, spatial relationships in the environmental domain may occur both over short and far distances

(Hewitt et al., 2018). In our case, physical transmission of pollution through the air can be expected to lead to spillovers that are transmitted over short geographical distances. However, it may also be the case that pollution in the sort run is driven by economic activities that spill over across a cluster of urban environments that are close in an economic sense, implying that linkages across further geographical distances may be equally relevant to describe the process.

To allow for network structures that can transmit effects at short geographical distances, as well as over economic distances, we propose a flexible approach based on Gaussian kernels that can produce weights matrices based on distances within any specified set of exogenous variables  $\mathbf{v}$ . Specifically, spatial weights, or more generally, the connectivity matrices  $C$  can be constructed by first computing a Gaussian kernel

$$G = k(\mathbf{v}_i, \mathbf{v}_j; b) = \exp \left( \frac{-\|\mathbf{v}_i - \mathbf{v}_j\|^2}{b} \right), \quad (6.19)$$

with  $\|\mathbf{v}_i - \mathbf{v}_j\|$  being the Euclidean distance, and  $b$  being a bandwidth parameter that determines the network smoothness. After the kernel is computed, one can design a matrix  $D$ :

$$D = G - I = k(\mathbf{v}_i, \mathbf{v}_j; b) = \exp \left( \frac{-\|\mathbf{v}_i - \mathbf{v}_j\|^2}{b} \right) - I, \quad (6.20)$$

that sets the diagonal to zero. Note that the diagonal of the Gaussian kernel is 1, so one can simply subtract the identity matrix. The spatial weight matrix  $C$  can subsequently be constructed by row-normalizing  $D$ .

To better understand the role between distances in the exogenous variables  $\mathbf{v}$ , and the type of network structures that this procedure produces, a closer look at the properties of the Gaussian kernel is helpful. For  $b > 0$ , the kernel  $k$  can be understood as a measure of similarity, which is seen by applying a Cauchy-Schwarz inequality

$$k(\mathbf{v}_i, \mathbf{v}_j; b)^2 \leq k(\mathbf{v}_i, \mathbf{v}_i; b)k(\mathbf{v}_j, \mathbf{v}_j; b) \quad \forall (\mathbf{v}_i, \mathbf{v}_j; b > 0) \in \mathcal{X} \times \mathcal{X} \times \mathcal{B}.$$

This reveals that when two points  $\mathbf{v}_i$  and  $\mathbf{v}_j$  are similar, then the kernel  $k(\mathbf{v}_i, \mathbf{v}_j; b)_{b>0}$  will return a value close to 1. On the other hand, when  $\mathbf{v}_i$  and  $\mathbf{v}_j$  are dissimilar, it will reach a value close to 0. This immediately suggests that geographic weights matrices can be constructed using this approach if  $\mathbf{v}$  describes the physical locations of observations, for example by using coordinates.

While  $\mathbf{v}$  plays the crucial role of describing the possible similarities between locations,  $b$  controls the type of network connections that result based on these similarities. For a positive but small  $b$ , few but strong network links arise. For larger values of  $b$ , a large number of positive, but weaker, connections result. The bandwidth can in principle also become negative. In this case the relationship between closeness in between two data points and the strength of their connection inverts. In particular, negative bandwidths produce positive network connectivities based on dissimilarities in  $\mathbf{v}$ . This is seen by the following. When  $b$  is negative and  $\mathbf{v}_i$  and  $\mathbf{v}_j$  are similar, then  $k(\mathbf{v}_i, \mathbf{v}_j; b)_{b<0}$  will be close to 1, but the kernel will attain values larger than 1 when  $\mathbf{v}_i$  and  $\mathbf{v}_j$  are dissimilar

$$k(\mathbf{v}_i, \mathbf{v}_j; b)^2 \geq k(\mathbf{v}_i, \mathbf{v}_i; b)k(\mathbf{v}_j, \mathbf{v}_j; b) \quad \forall (\mathbf{v}_i, \mathbf{v}_j; b < 0) \in \mathcal{X} \times \mathcal{X} \times \mathcal{B}.$$

This type of clustering based on dissimilarities may not make sense when considering clustering in a geographical context, but in some equilibrating processes, intensification of contraction can in fact be the result of divergences. Both have empirical relevance. For example, when the kernel is drawn around the level series of a cross-sectional time-series, the resulting contraction between dissimilar observations is similar to the error-correction effect that is commonly modeled using Vector Error Correction Models. For positive bandwidths, on the other hand, the similarity view of the kernel approach carries a similar interpretation as that of Tobler's law, that underlies the intuition of the SAR.

Figure 6.2, summarizes the various possibilities visually. In particular, it plots the connectivity matrix for different bandwidth values using a single vector of values  $\mathbf{v} = N/25$ ,  $N \in \{1, 2, \dots, 25\}$ . One can see that disregard of the sign of  $b$  the surfaces are smooth when the bandwidth is large in magnitude. We also see that the connection between the values  $\mathbf{v}_1$  and  $\mathbf{v}_{25}$  is closer to zero when  $b$  is positive, but closer to 1 when  $b$  is negative. Section 6.4 discusses how to find an appropriate value empirically.

### 6.4.2 Penalized Maximum Likelihood Estimator

To relax the Gaussian assumption that may not hold for data that exhibits extreme tail movement with high probability, often the case in the environmental-economic data, we discuss estimation in the context of the Students'  $t$ -estimation. In line with our discussion on uniqueness, we apply  $L^2$  (Euclidean distance) penalties set on the moving average components that vanish with a weight of  $1/\sqrt{NT}$ . Penalizing the  $L^1$  norm (absolute sum), as in popularized Ridge estimations, encourages parameter vectors with many elements set to zero, which results in an unidentified problem for  $\mathbf{b}$ .  $L^2$  penalization, like in the LASSO framework, encourages solutions where parameters are small, and in fact the penalty effect reduces in strength as parameters become close to zero. To reduce dimensionality, we suggest to evaluate the  $AICc$  around the PMLE, and apply zero restrictions following minimization of information loss.  $L^2$  penalization of  $\mathbf{b}$  increases exponentially in strength for  $\|\mathbf{b}\| > 1$  while weakening in strength as  $\|\mathbf{b}\| \rightarrow 0$ , and favors networks with fewer, but stronger links. This prior is justified by the improved small sample behavior of the MLE of spatial auto-regressions with higher degree of sparseness of the weights matrix (Bao and Ullah, 2007). Our penalized Students'  $t$ -criterion with vanishing penalties maintains generality in the limit and naturally generalizes the standard Gaussian case, while imposing less strict assumptions regarding thin-tailedness of the moving

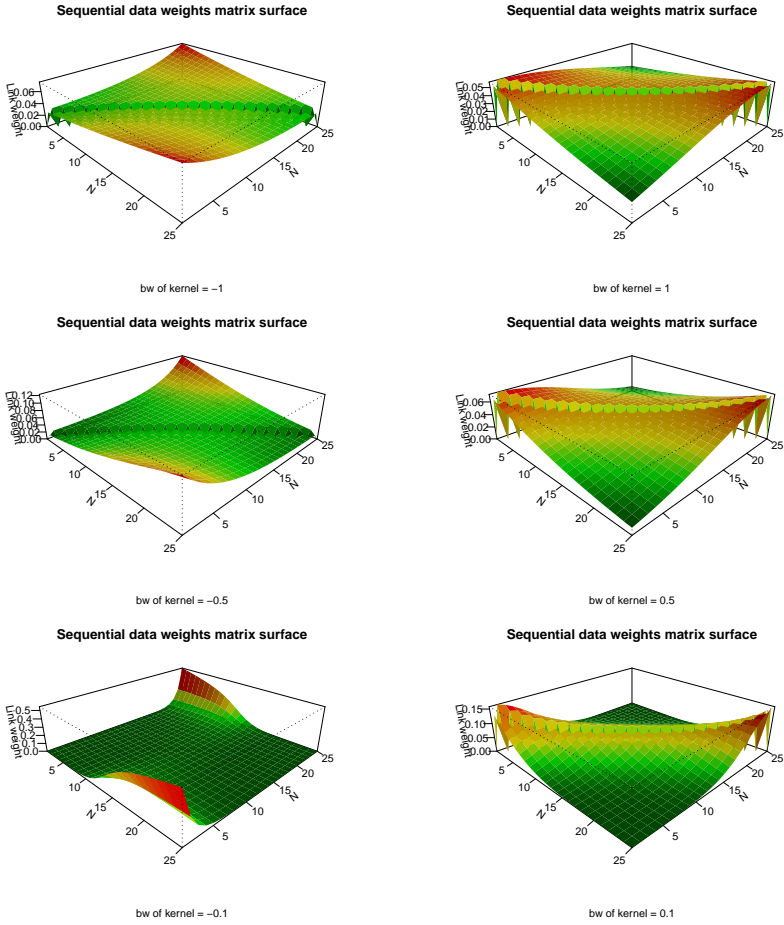


Figure 6.2: Surfaces of spatial weights produced using the kernel approach for different bandwidth values, on identical data produced with  $N/25$ ,  $N \in \{1, 2, \dots, 25\}$ .



averages thereby allowing for large exogenous impacts to occur with high probability.

Let  $\boldsymbol{\theta}$  denote the collection of parameters of the SVARMA model,  $\boldsymbol{\theta} := (\mathbf{H}, \mathbf{M})$ , of which  $\boldsymbol{\theta}^{\mathbf{S}} := (\boldsymbol{\rho}, \mathbf{b})$  is a subset of spatial parameters. We define the PMLE as:

$$\hat{\boldsymbol{\theta}}_T := \arg \min_{\boldsymbol{\theta} \in \Theta} Q_T(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta}) + \lambda \gamma(\boldsymbol{\theta}), \quad (6.21)$$

with the ML criterion defined as

$$Q_T := \ell_T(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta}) = \sum_t^T \ell_t(\mathbf{v}, \mathbf{w}_t; \boldsymbol{\theta}), \quad (6.22)$$

$$\ell_t(\mathbf{v}, \mathbf{w}_t; \boldsymbol{\theta}) = \ln p_{\varepsilon}(\mathbf{w}_t - f(\mathbf{v}, \mathbf{w}_t; \boldsymbol{\theta}), \boldsymbol{\Sigma}; \boldsymbol{\nu}),$$

with  $f(\mathbf{v}, \mathbf{w}_t; \boldsymbol{\theta})$  shorthand for the data modeled by the SVARMA with spatial matrices conditional on a vector of data  $\mathbf{v}$ , and the penalty defined as

$$\lambda \gamma(\boldsymbol{\theta}) = 1/\sqrt{NT} \sum |\mathbf{M}|^2. \quad (6.23)$$

Using the standard expression for the multivariate  $t$ -distribution with  $\boldsymbol{\nu} = \nu^w = (\nu^x, \nu^y)$  degrees of freedom for each channel, and variance  $\boldsymbol{\Sigma} = \Sigma^w = (\Sigma^x, \Sigma^y)$  for each channel, we obtain

$$\ell_t(\mathbf{v}, \mathbf{w}_t; \boldsymbol{\theta}) = D(\boldsymbol{\theta}^{\mathbf{S}}, \mathbf{v}) + K(\boldsymbol{\theta}) + E(\boldsymbol{\theta}, \mathbf{v}, \mathbf{w}_t), \quad (6.24)$$

where  $D(\boldsymbol{\theta}^{\mathbf{S}}, \mathbf{v})$  is the log determinant of

$$D(\boldsymbol{\theta}^{\mathbf{S}}, \mathbf{v}) := \ln \det \mathbf{S}(\boldsymbol{\theta}^{\rho}, \mathbf{C}(\mathbf{v}; \mathbf{b})), \quad (6.25)$$

with  $\mathbf{S}(\boldsymbol{\theta}^{\rho}, \mathbf{C}(\mathbf{v}; \mathbf{b}))$  as the spatial multiplier matrix conditional on data  $\mathbf{v}$  and bandwidth parameters  $\mathbf{b}$  that we defined as

$$\mathbf{S}(\boldsymbol{\theta}^{\rho}, \mathbf{C}(\mathbf{v}; \mathbf{b})) = (\mathbf{I} - \boldsymbol{\rho} \circ \mathbf{C}(\mathbf{v}; \mathbf{b}))^{-1}, \quad (6.26)$$

with  $\mathbf{C}(\mathbf{v}; \mathbf{b})$  constructed as detailed in section 6.4.1. Importantly, the

log determinant equals the sum of the log determinants of its diagonal blocks, as the off-diagonal blocks are zero

$$\begin{aligned} D(\boldsymbol{\theta}^S, \mathbf{v}) &= \ln \det \mathbf{S}(\boldsymbol{\theta}^\rho, \mathbf{C}(\mathbf{v}; \mathbf{b})) = \ln \det S^x(\rho^x, C_{n_x}(\mathbf{v}; b^x)) \\ &\quad + \ln \det S^y(\rho^y, C_{n_y}(\mathbf{v}; b^y)) \end{aligned} \quad (6.27)$$

and each determinant is evaluated over  $S(\rho, C(\mathbf{v}; b)) = (I - \rho C(\mathbf{v}; b))^{-1}$  with  $\rho C(\mathbf{v}; b)$  as the diagonal blocks of

$$\rho \circ \mathbf{C}(\mathbf{v}; \mathbf{b}) = \begin{bmatrix} \rho^x C_{n_x}(\mathbf{v}; b^x) & O_{n_x} \\ O_{n_y} & \rho^y C_{n_y}(\mathbf{v}; b^y) \end{bmatrix}. \quad (6.28)$$

$K(\boldsymbol{\theta})$  is a constant, that can be similarly expressed as a sum

$$K(\boldsymbol{\theta}) := \ln \Gamma((\nu + N)/2) \left[ \det \Sigma^{\frac{1}{2}} (\nu \pi)^{\frac{N}{2}} \Gamma(\nu/2) \right]^{-1}, \quad (6.29)$$

for each  $(\nu, \Sigma) \in ((\nu^x, \Sigma^x), (\nu^y, \Sigma^y))$ . Finally, the random element  $E(\boldsymbol{\theta}, \mathbf{v}, \mathbf{w}_t)$  can naturally be defined as the sum

$$\begin{aligned} E(\boldsymbol{\theta}, \mathbf{v}, \mathbf{w}_t) &:= \\ &-\frac{1}{2}(\nu^x + N) \ln \left( 1 + \nu^{x-1} (\mathbf{x}_t - f^x(\mathbf{v}, \mathbf{w}_t; \boldsymbol{\theta}^x))' \Sigma^{x-1} (\mathbf{x}_t - f^x(\mathbf{v}, \mathbf{w}_t; \boldsymbol{\theta}^x)) \right) \\ &-\frac{1}{2}(\nu^y + N) \ln \left( 1 + \nu^{y-1} (\mathbf{y}_t - f^y(\mathbf{v}, \mathbf{w}_t; \boldsymbol{\theta}^y))' \Sigma^{y-1} (\mathbf{y}_t - f^y(\mathbf{v}, \mathbf{w}_t; \boldsymbol{\theta}^y)) \right). \end{aligned} \quad (6.30)$$

The channel-wise summing of the likelihood is possible as long as feedback stays within each cross-section, and contemporaneous spillovers between  $\mathbf{x}$  and  $\mathbf{x}$  are not modeled. This channel-wise computation allows parallelization for each  $\ell_t(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta})$ , which reduces computation time of each evaluation of  $\ell_t(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta})$  tremendously. Since  $f(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta})$  depends on the moving averages that in turn result as difference combinations of  $\mathbf{w}_t - f(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta})$ , the components of eq. (6.30) can only be computed simultaneously for identical  $t$ . In the Appendix we discuss restrictions that are advantageous in terms of reducing the computational cost, and

detail how this trades with flexibility of the implied density.

Limit properties of  $\mathbf{b}$  are not developed in the literature to our knowledge, but we do not regard it as an interesting parameter for inference. For Granger-causal inference we are interested in  $\hat{\boldsymbol{\theta}}_T \setminus \mathbf{b}_T$ , and  $\mathbf{b}$  has the sole purpose of improving  $\hat{\boldsymbol{\theta}}_T \setminus \mathbf{b}_T$  by reducing misspecification bias of  $\mathbf{C}(\mathbf{v}; \mathbf{b})$  that may result in bias in  $\boldsymbol{\theta}^\rho$ . This can be diagnosed by comparing against non-spatial VARMA using standard diagnostics. To explore the small sample behavior, we perform a simulation study. It turns out that the small sample distribution of the penalized bandwidth is reasonable, while the distribution of the unpenalized bandwidth is heavily distorted in our small  $T$  study. In both cases however, we see that  $\hat{\boldsymbol{\theta}}_T \setminus \mathbf{b}_T$  behaves well. We also provide results that highlight the significant bias in the ARMA parts when no spatial dynamics are modeled.

Finally, due to the dependence on moving averages that are not available as difference combinations for the first  $q$  periods are unavailable, the estimation algorithm requires an initialization of  $\hat{\boldsymbol{\varepsilon}}_t$  for  $t \leq q$ . As  $T \rightarrow \infty$ , the impact of the initialization on the filter fades exponentially fast almost surely for a stationary process, see for example Straumann and Mikosch (2006). For small  $T$  however, the impact remains. We focus our simulations on the small  $T$  case to investigate this.

### 6.4.3 Small sample distribution of the (P)MLE

To explore the adequacy of the S(V)ARMA in filtering out space-time-dynamics, we conduct a simulation study. We investigate both the MLE that arises by setting  $\lambda = 0$  and the PMLE with  $\lambda = 1/\sqrt{NT}$ . Remember that this penalty vanishes as the data grows, ensuring consistency in the limit while penalizing only in small sample regions. For this reason we explore simulations across growing data dimensions. In particular, because our application covers two sets of estimation results that are identical in time dimension but different in the cross-sectional dimension

( $T - p = 12$ ,  $N = 60$  and  $N = 113$ ), we explore the (P)MLE across growing  $N = (10, 25, 75, 125)$  while keeping  $T$  fixed to the dimension of the application. We set the parameters to realistic values given the empirical application.

Apart from the behavior of the ARMA components we are interested in the adequacy of the (P)MLE in dynamically estimating appropriate values of the bandwidth parameter that produces alternative spatial structures. We also explore explicitly whether the spatial structure improves the ARMA estimates, and explore robustness to over-fitting under the null of a non-spatial ARMA process. The *DGP* is

$$\mathbf{y}_t = 0.6C(\mathbf{x}; b)\mathbf{y}_t - 0.35\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t + 0.25\boldsymbol{\varepsilon}_{t-1}, \quad (6.31)$$

where  $\mathbf{x}$  is drawn uniquely in every experiment from a Student's- $t$  distribution with  $\nu = 120$ ,  $\boldsymbol{\varepsilon}_t$  is drawn from a Student's- $t$  distribution with  $\nu = 5$ . We explore both a spatial structure with few but strong links with  $b = .15$  and a smoother network with  $b = 2$ . The decision to focus on the heavy tail case is guided by our empirical results.

As we can see in fig. 6.3 the PMLE performs reasonably well already in small samples, but even in the largest samples we do not obtain the limit result for the individual parameters. This is not surprising given the small  $T$ . The initialization of the moving averages at zero cannot fade, leading to a downward bias of MA parameters and an upward bias of AR parameters. In fact, by increasing  $N$  and fixing  $T$ , the bias increases further as the ratio of distorted information due to zero-initialization of innovations grows along with the ratio of  $N/T$ . Nonetheless, the ARMA parameters are jointly well behaved, even when both  $N$  and  $T$  are small. We conclude that inference on the joint parameters (such as when simulating the IRF using all the model's parameters) is therefore valid in our application, while statements that involve differentiation between short- and long-term effects should be made with caution.

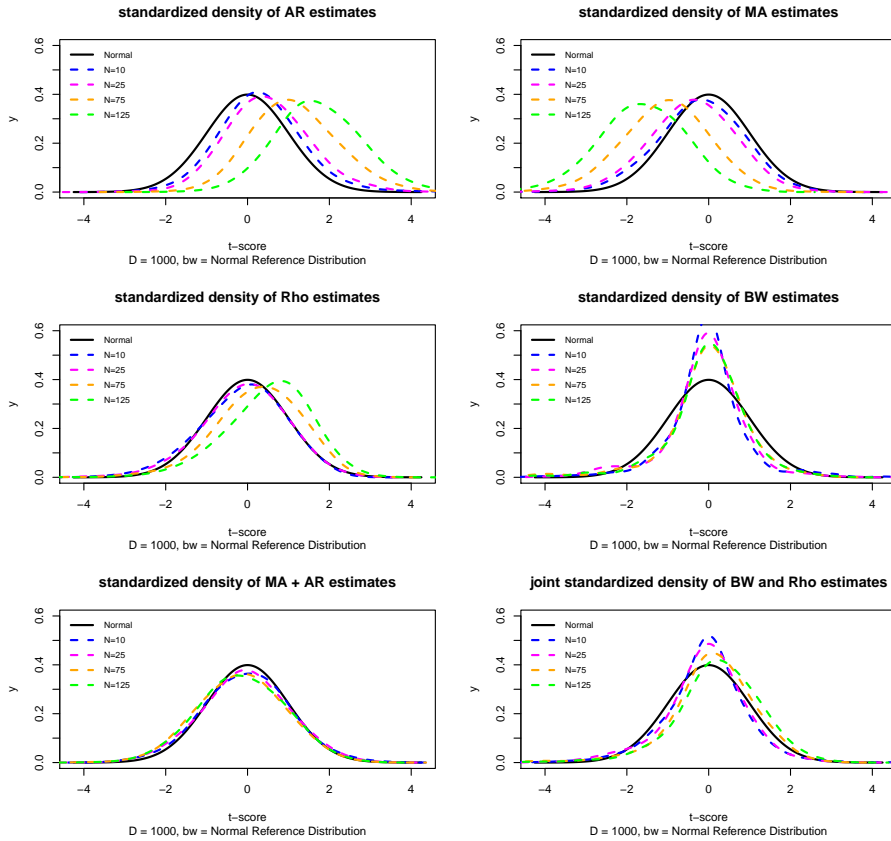


Figure 6.3: Penalized small sample distributions of the correctly specified SARMA, bandwidth of the spatial kernel matrix in the DGP set to .15

Figure 6.5 in the appendix shows the results for the MLE. It is clear that the penalization improves the empirical distribution of the bandwidth parameter substantially. Note that the MLE is not identified because both AR and MA distributions could potentially fit the data equally well. The PMLE was designed to ensure identification, and the simulations confirm that the distribution of individual AR and MA parameters of the PMLE are slightly better. Figure 6.6 and fig. 6.7, also in the appendix, document results for  $b = 2$ . This experiment shows that our conclusions are insensitive to the value of  $b$ .

Figure 6.8 in the appendix shows results for misspecified non-spatial ARMA estimation. This reveals that when the cross-sectional process exhibits spatial effects, and these spatial effects are not modeled, then the ARMA parameters become severely biased. This highlights that estimating a conventional non-spatial VARMA when the cross-sectional time series processes are in fact spatial, leads to bad inference as the temporal parameters capture a share of the unmodeled spatial correlations.

Finally, to investigate the behavior of the SVARMA with the Gaussian kernel structure as spatial weights matrix when the data is in fact non-spatial, we present results in Figure 6.9 (appendix). The bandwidth density of the (P)MLE is centered around 0, with the PMLE having a notably nicer distribution. Note that the kernel structure is not identified when the bandwidth is zero, and it could take on any value potentially allowing the structure to find some (dis)similarities that over-fit the data. The results show that the penalization technique is useful, and the non-penalized MLE has a long tail of incorrect high bandwidth values. The spatial dependence parameter remains, however, well-behaved in both cases. This suggests that the researcher can decide between SVARMA and VARMA mechanics by focusing on Wald-type test around the spatial dependence parameter.

Combined, all the simulation results not only confirm that the SVARMA model performs well in empirically relevant situations, but also that not specifying the spatial effects results in biased results. The SVARMA remains a robust analysis tool also when the data is non-spatial.

## 6.5 Application to subnational pollution and household expenditure data in Indonesia

In this application we study interactions between household level expenditures and pollution. It has long been theorized that as economies

develop, pollution initially increases at an exponential rate. However at some point on the development path, parts in the economy start to adopt cleaner technologies and acceleration in pollution slows down till pollution levels reach a maximum after which the entire economy enters into a state characterized by a decline in pollution. We do not aim to provide a large survey of the literature, for a progression of the debate, see (World Bank, 1992; Grossman and Krueger, 1995; Stern et al., 1996; Stern, 1998, 2004; Andree et al., 2019). For many, the central question is whether increases in wealth and income result in increasing pressure on the environment, or whether economic development provides the basis for environmental improvement. In turn, environmental degradation may negatively interact with growth and contribute to the creation of urban pollution traps. In this application we revisit the empirical issue and focus on the question whether pollution increases or decreases after income. Furthermore, we are interested in the order of effects, the presence of feedback, and distributional impacts of effects. We therefore focus our study on air pollution, average per capita household expenditures, and bottom quintile per capita household expenditures and explore the interactions in the context of multiple spatial time series in Indonesia over the period 1999-2014. We seek to distinguish between the effects of average household growth and bottom household growth on pollution and see if there is differential in potential impacts of pollution on the two different income groups.

### 6.5.1 Data

Our analysis relies on two longitudinal data sets. First, as a proxy for air pollution, we use the global estimates of fine particulate matter developed by van Donkelaar et al. (2016). Second, we are interested in distinguishing between the economic development of average households and poor households. As a proxy, we use annual averages of monthly

household expenditures for the average households and for the bottom quintile households as defined in the Indonesia Database for Policy and Economic Research (INDO-DAPOER, World Bank Group).<sup>7</sup> The expenditure data are available from 1999 to 2014. The data set also contains several other economic, social and demographic indicators at the district-level, primarily sourced from various surveys and the Indonesia Central Bureau of Statistics (BPS), but the coverage of other potential proxies for local poverty and average economic growth is sparse.

The air pollution data set contains estimates on mean annual (1999 to 2015) concentrations of fine particulate matter ( $PM_{2.5}$ ), coarse dust particles of 2.5 micrometers in diameter, that proxy a wider range of air pollutants. The data points are available at a 0.01-degree resolution and have been derived from a combination of satellite-, simulation- and monitor-based sources. The authors address several inconsistencies in satellite-derived  $PM_{2.5}$  data by calibrating their estimates with ground-based observations and reducing the noise of seasonal anomalies.

We are primarily interested in the environmental-economic interactions in urban environments. To narrow the focus, we used a gridded population data set (Gridded Population of the World, v4 at 30 arc-seconds resolution) to distinguish urban from rural districts. We defined urban areas as a contiguous patch of pixels with population density higher than 300 per square kilometer and a population count higher than 5,000. This is similar to the approach followed by OECD and EC-DG Regio to define global Functional Urban Areas, scaling down the population counts to be relevant in a subnational context. Our approach identifies 219 areas with urban clusters. To establish a link between urban air pollution and the INDO-DAPOER database, we summarized the  $PM_{2.5}$  annual

---

<sup>7</sup><https://data.worldbank.org/data-catalog/indonesia-database-for-policy-and-economic-research>. Some district names and borders have changed over time. To construct the time series, we used the database's "District Proliferation Crosswalk" file to match observations in the data to the district definition provided by the Global Administrative Areas repository (GADM) available at <http://www.gadm.org/>. Indonesia's latest district configuration covers 497 districts of which 427 were successfully matched.



grids to the district-level using the mean value for pollution grids sensed over urban patches in each district. This captures output directly from urban activity, and reduces the outside influence of fires and agricultural activity. Figure 6.10 contains kernel densities of the pollution levels, and changes, in each year for all the 219 urban clusters.

Since we are particularly interested in the effects of considerable pollution, we drop any areas that at one or more points in time have a concentration below  $6 \text{ mcg/m}^3$ . To ensure that the sample is relatively homogeneous and not too impacted by outliers, we also removed several regions in which pollution briefly spiked to values over  $40 \text{ mcg/m}^3$  in 2006, during which a particularly strong fire season occurred. After removing the relatively unpolluted areas and these extreme pollution outliers, we are left with a final number of 113 areas that meet our criteria of being a polluted urban cluster. Apart from the 113 areas that we defined as polluted urban areas, we perform an additional estimation focusing specifically on 60 heavily polluted areas that exceed the WHO air quality guidelines in all years.

### 6.5.2 Estimation approach

We use percentage changes, and work with demeaned series that are cleared from both the time-invariant and cross-sectionally invariant impacts similarly to a Fixed Effects approach, to remove any trending behavior or strongly dependent co-movements, and control for heterogeneity. We find nonzero medians after removing all average effects, indicative of heavy tail action. This strengthens justification for our  $t$ -approach against the Gaussian alternative. Plotted distributions of levels and returns are included in the Appendix, section 6.7.4.

We base our spatial weights matrix on Gaussian kernels around features computed from the local distributions in returns (prior to demeaning). Specifically, we use the first, second and fourth moments (excess), together

with 25 and 75 quantiles of the local returns to describe the sample distributions, and cumulative returns to describe the total effect of moving through that distribution. The similarity approach around these local statistics informs the model on similarities in the behavior and direction of the local time-series. The cross-sectional spillover channels thus arise as functions of similarities in the local temporal patterns, which suggest that those regions share commonalities such as co-integrating forces or common latent factors. We estimate VARMA and SVARMA models with both  $p, q$  equal to three, such that if Granger-causal effects follow after one lag, variables can potentially influence each other indirectly through another channel while direct effects may in fact be zero. We minimize the  $AICc$  evaluated at the PMLE, to minimize divergence w.r.t. the *true* probability measure.

### 6.5.3 Results

Table 6.1 presents the estimation results for the SVARMA( $AICc$ ) for all observations, table 6.5 in the appendix contains the additional estimation results for the more polluted ( $PM_{2.5} > 10$ ) samples. For comparison, VARMA( $AICc$ ) results are contained in tables 6.3 and 6.4 in the appendix. The parameter results suggest that the processes are fat-tailed, Gaussian estimation would be overwhelmingly rejected both in the VARMA and SVARMA frameworks. Second, the  $AICc$  drops with 494.341 points at  $PM_{2.5} > 6$  and by 277.103 points at  $PM_{2.5} > 10$ , indicating that the SVARMA improves the conditional density implied by the model significantly over the VARMA. Our  $\hat{R}^2$  estimates<sup>8</sup> suggest that we explain

---

<sup>8</sup>We use a pseudo- $R^2$  using the  $SSR$  of residuals evaluated at the PMLE versus the residuals evaluated at all parameters equal to 0 (and bandwidths at any value),

$$\hat{R}^2 = 1 - \frac{\sum_1^T |\mathbf{w}_T - f(\mathbf{w}_T; \hat{\boldsymbol{\theta}})|^2}{\sum_1^T |\mathbf{w}_T - f(\mathbf{w}_T; \boldsymbol{\theta}_{\boldsymbol{\theta}=0})|^2}, \quad (6.32)$$

in which  $\boldsymbol{\theta}_{\boldsymbol{\theta}=0}$  implies that all the structural parameters are set to zero – not to be confused with  $\boldsymbol{\theta}_0$  as the *true* values.

roughly more than 70% of the variance in the data, confirming slightly higher explanatory power using the SVARMA specifications (0.737 versus 0.705 at  $\text{PM}_{2.5} > 6$ , 0.732 versus 0.722 at  $\text{PM}_{2.5} > 10$ ). In both cases the SVARMA, however, uses less ARMA parameters (29 versus 34 at  $\text{PM}_{2.5} > 6$ , 27 versus 31 at  $\text{PM}_{2.5} > 10$ .) implying that the improvements from spatial filtering are significant.

We can see that the bandwidths that control the network smoothness are different in each channel of the model. Figure 6.11 in the appendix plots the network surfaces, we have ordered the link weights from high to low. This reveals that the bandwidths at  $\text{PM}_{2.5} > 6$  produce smooth network structures in both expenditures equations with many weak links, which implies that economic spillovers are weakly shared across many observations with many indirect spillovers. Observations in the pollution cross-section are more often linked to only a few other observations, but share strong direct spillovers. As there are many near-zero links, this implies that feedback effects in the pollution equation remain relatively centered in local pollution clusters. Average expenditures have a higher bandwidth value than bottom expenditures, hence the results indicate that bottom expenditures spill over in smaller but stronger clusters than average expenditures.

To assess how well the estimated structure fits the data, we also estimate a cross-sectional AR model on the residuals on an equation-by-equation basis. Under the null, the models are estimated on random data and we should expect 1 out of 10 lags to be significant at .10 purely out of chance. We compute  $1, \dots, r$  individual  $LR$  ratios for AR models with up to  $r$  lags against a zero lag model, and correcting the  $p$ -values using a Bonferroni-correction. The smallest  $p$ -value out of  $r$  Bonferroni-corrected  $p$ -values is reported. These residual correlation tests also favor the SVARMA representation (the VARMA at  $\text{PM}_{2.5} > 6$  retains significant residual correlations). The rejections of residual correlations, and reasonable  $\hat{R}^2$ ,

Table 6.1: SVARMA(AICc) results at  $PM_{2.5} > 6$ ,  $\hat{R}^2 = 0.737$ , 41 estimated parameters on  $(N - \max(p, q) \times T) \times 3 = 4068$  data points with 372 fixed demeaning components.  $AICc = -7390.091$ .

	Pollution	Bottom Expenditures	Expenditures
$\phi\ pol_{t-1}$	-0.068** (-2.57)	-0.092*** (-2.866)	-0.047*** (-2.391)
$\phi\ pol_{t-2}$	-0.070*** (-4.381)	-0.063*** (-2.969)	
$\phi\ pol_{t-3}$	0.026* (1.652)		0.054** (2.507)
$\phi\ bot_{t-1}$		-0.108* (-1.94)	0.089*** (2.577)
$\phi\ bot_{t-2}$		-0.139*** (-4.736)	-0.129*** (-2.791)
$\phi\ bot_{t-3}$	-0.039** (-2.119)	-0.099*** (-3.544)	-0.141*** (-3.534)
$\phi\ exp_{t-1}$		0.071*** (2.964)	-0.374*** (-12.805)
$\phi\ exp_{t-2}$		0.158*** (4.453)	
$\phi\ exp_{t-3}$		0.074*** (3.14)	
$M\ pol_{t-1}$	-0.515*** (-15.292)	0.126*** (3.233)	
$M\ pol_{t-2}$			
$M\ pol_{t-3}$		-0.052* (-1.814)	-0.082** (-2.131)
$M\ bot_{t-1}$	-0.038* (-1.94)	-0.253*** (-4.296)	
$M\ bot_{t-2}$			0.252*** (4.313)
$M\ bot_{t-3}$			0.128** (2.203)
$M\ exp_{t-1}$			
$M\ exp_{t-2}$		-0.134*** (-3.621)	-0.387*** (-10.705)
$M\ exp_{t-3}$			-0.147*** (-4.273)
$\rho$	0.812*** (27.765)	0.305*** (3.177)	0.327*** (2.976)
$b$	0.088	0.18	0.229
$\sigma$	0.119	0.129	0.176
$\nu$	2.004	7.313	4.703
4-lag white-noise $p$	1.000	0.129	0.176

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01  
Constant omitted,  $t$ -statistics in parenthesis for the SARMA components.

Table 6.2: Cumulative effects after 15 years following an initial 10% increase in the impulse variable. Based on 10.000 simulations from the model, drawing parameters randomly from the estimated parameter distributions and discarding 50 initialization steps before applying the impulse.

Percentiles:	PM <sub>2.5</sub> > 6			PM <sub>2.5</sub> > 10		
	25%	50%	75%	25%	50%	75%
Impulse: Pollution						
Pollution	-28.203%	-14.470%	-6.403%	-50.507%	-17.071%	0.334%
Bottom expenditures	-3.066%	-2.227%	-1.534%	-8.312%	-5.742%	-3.876%
Average expenditures	-1.399%	-0.854%	-0.409%	-4.162%	-2.645%	-1.520%
Impulse: Bottom expenditures						
Pollution	-3.143%	-2.416%	-1.794%	-5.966%	-4.171%	-2.761%
Bottom expenditures	6.504%	7.192%	7.940%	4.389%	5.132%	5.928%
Average expenditures	1.435%	2.089%	2.773%	0.668%	1.221%	1.747%
Impulse: Average expenditures						
Pollution	-0.043%	-0.003%	0.031%	-0.407%	-0.235%	-0.106%
Bottom expenditures	-0.329%	-0.027%	0.268%	-0.919%	-0.634%	-0.363%
Average expenditures	2.522%	2.954%	3.330%	1.525%	2.146%	2.772%

lead us to conclude that no major components are missing in either of the SVARMA specifications, hence we interpret the parameters and standard errors in their usual context.

### Impulse Response analysis

To explore the dynamics implied by the estimated results, we use the parameters to simulate IRF's. We perform 3 experiments. First we trace the effect after an isolated impact of 10% increase in pollution across all areas, we consider a similar impact to the bottom expenditures, and finally we repeat the experiment for average expenditures. The impact vectors are not designed to mimic a plausible event, our foremost goal is to track the direct and indirect Granger-causality channels implied by the estimated model. However, 10% is roughly in line with one standard deviation of the residuals for each variable. Confidence bandwidths are constructed by simulating from the models, randomly drawing parameters from their empirical distributions. The first 50 time steps are discarded before the impact vector is activated to prevent dependence of the dynamics on the initialization. Table 6.2 summarizes the results.

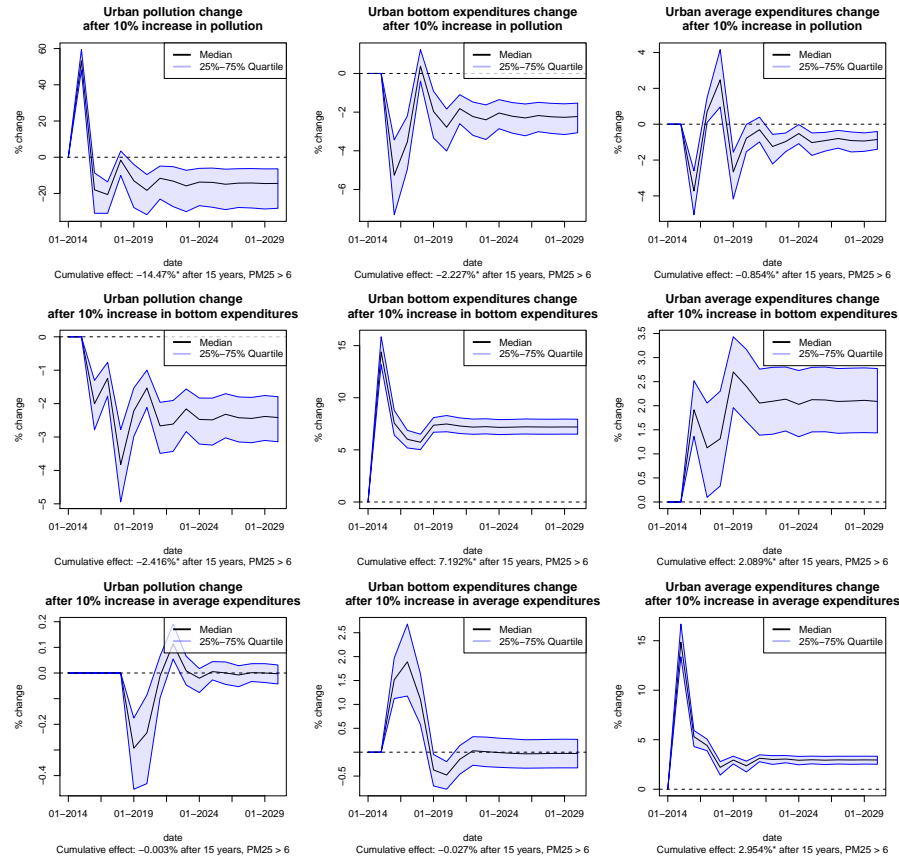


Figure 6.4: IRF plots for exogenous shocks in pollution, bottom household expenditures and average household expenditures  $PM_{2.5} > 6$ . Effects that exclude zero in the final year, are marked by \*.

Figure 6.4 shows the results for the model estimated at  $PM_{2.5} > 6$ , and fig. 6.12 in the appendix shows the results from the model estimated at  $PM_{2.5} > 10$ . The figures are produced by 10,000 random draws and show the cumulative effects resulting from compounding the percentage changes including spatial feedback effects. Table

We find that across all districts with  $PM_{2.5} > 6$ , average expenditure growth has no long-term effect on pollution. Growth in the bottom expenditures, however, reduces pollution by -2.416%. At higher pollution concentrations we find that the effect of bottom expenditure growth on

pollution is even stronger (-4.171%). Growth in average expenditures in these highly polluted areas is also found to reduce pollution, albeit with smaller impact (-.235%). Exogenous pollution effects in both models have a short-term multiplier effect due to feedback, with the effect peaking briefly over 50%. The long-term impacts, however, produce a wide range of outcomes that are mostly negative or include zero. Therefore, our results suggest that ongoing effects of exogenous pollution, such as increasing populations and changes in urban structure, contribute to pollution build up by constantly keeping the short-run effects positive. This suggests that a region remains polluted as long as exogenous effects continue to enter the system, while the highest pollution levels will eventually dissipate as these structural contributions stabilize, and further decline as continued income growth takes over as a predominant driver of pollution decline.

Another result is that at both  $PM_{2.5} > 6$  and  $PM_{2.5} > 10$ , average growth is non-inclusive. At  $PM_{2.5} > 6$ , an increase in average household expenditures does not significantly spill over to bottom households in the long-run, and at  $PM_{2.5} > 10$  the long-run impact is  $-0.634\%$ . Growth in bottom expenditures, on the other hand, boosts the average ( $7.192\%$  at  $PM_{2.5} > 6$  and  $5.132\%$  at  $PM_{2.5} > 10$ ). Pollution is additionally identified as a negative effect on bottom growth,  $-2.227\%$  at  $PM_{2.5} > 6$ . The effect intensifies at higher pollution concentrations,  $-5.742\%$  on average across all districts with  $PM_{2.5} > 10$ . Average household expenditures are relatively more resilient, but are also negatively impacted by pollution ( $-0.854\%$  at  $PM_{2.5} > 6$ ), especially at higher pollution levels ( $-2.645\%$  at  $PM_{2.5} > 10$ ).

The results suggest several feedback mechanisms. First, average growth is non-inclusive. Second, pollution lowers primarily after bottom expenditures increase, while average growth is less effective in reducing pollution. Third, average household expenditures are more resilient to pollution

effects. Taken together, these three effects compound in downwards pressure on bottom growth, subsequently also slowing pollution clean-up, and creating an environment in which heavily polluted urban poverty traps may potentially arise if pollution and poverty are not addressed. Pollution impacts also block part of the potential multiplier effect that bottom-up growth would produce. Growth spillovers from the bottom to the average are strong however, suggesting that pollution-poverty environments may have strong negative impacts on the wider urban economy. Jointly, these inferred mechanisms suggest that a bottom-up approach to growth can help reduce the likelihood of pollution-poverty trap scenarios and even later on remains a no-regret strategy for growth as it induces positive spillovers.

### **Economic significance**

The results from the impulse response analysis indicates that pollution damages account for significant economic losses. Using the converged IRF impacts, and using a 2017 dollar conversion rate, we can draft the following crude economic costs associated with the analyzed 10% country wide pollution increases by using the average expenditure levels of the distinguished household groups. We use the income 2014 values, and extrapolate to 2017 to match our conversion rate, by compounding the average growth rate observed per household group. Table 6.6 in the appendix summarizes the per capita expenditures used for our calculations.

Using 2014 population estimates from INDO-DAPOER, together with the average local population growth rates, we would see approximately 83,104,069 people living in heavily polluted areas in 2017. Another 47,463,131 people live in the 6 to 10  $PM_{2.5}$  range.<sup>9</sup> By population weighting the effect of the analyzed increase in pollution levels, an estimated

---

<sup>9</sup>As a reference, the United Nations put the total Indonesian population at 261,115,456 in 2016.



total economic loss to household expenditures reaches over 3 billion dollars. Poor households account for approximately half a billion dollars of those losses. Various factors can further add to this number in the future, including migration toward areas with higher pollution concentration and overall continued growth in urban populations, growth in income and increasing pollution levels in areas that are now still relatively clean. The average pollution level in 2014 in heavily polluted areas was 21.75 according to our aggregated sensor estimates, and the .95 percentile is at 26.98, showing that a 25% increase in the average urban area can still occur. In addition, we look at household expenditures that constitute only part of GDP, and thus capture only part of the potential economic damages. We do not model the potential direct and indirect impacts on other components of GDP. Opportunity costs related to diverting government expenditures to health-related issues while social returns to investment might be higher elsewhere in an unpolluted economy may be another hidden cost. Without intervention the damages would run into the multi-billions over the course of only a few years.

## 6.6 Conclusion

This paper discussed and estimated a fat-tailed Spatial Vector Autoregressive Moving Average (SVARMA) model in which multiple spatial autoregressive time series are modeled together. The model was used to study Granger-causal interactions between spatial autoregressive time series of subnational pollution and household expenditure data. The application used data that was not spatially contiguous in all cases and explored the use of a Gaussian kernel to estimate the spatial weights based on similarities in covariates. The analysis found that the model improved over the non-spatial VARMA and highlighted interesting dynamics between poverty and pollution.

Our economic findings are summarized in three main points: first, expenditure growth reduces pollution, particularly growth of poor households; second, pollution reduces growth in expenditures, particularly of poor households; third, growth is non-exclusive, there are significant spillovers from bottom-up growth but not from top-down growth. This imbalance in growth spillovers aligns with a body of literature debunking so-called “trickle-down” economics (see, for example, Quiggin (2009); Ranieri and Almeida Ramos (2013)), and suggests instead that investment in the poor is more effective than raising average incomes. Non-inclusive growth, lower resilience of the poor to pollution damages, and the importance of growth in bottom households to reduce pollution, together lay the basis for polluted poverty traps.

We find that damages from pollution in Indonesia are considerable, over 3 billion annually for a 10% increase in particulate matter concentrations. This is in line with earlier research that has indicated that considerable economic impacts of air pollution stem from health effects that decrease length and quality of life, increases in health expenditures, and reductions in labor supply and productivity Preker et al. (2016); Levinson (2012); Hanna and Oliva (2015); Zivin and Neidell (2012). In 2013, one-tenth of deaths worldwide were attributable to air pollution, resulting in about \$225 billion annually in lost labor income (World Bank and Institute for Health Metrics and Evaluation, 2016).

While these results point toward an economic failure, our analysis also suggests potentials for enhanced growth. Policy targeted at exogenous pollution can have positive growth effects by reducing the harmful effects of pollution. Positive economic effects, specifically on the poor, in turn help combat air pollution. Bottom-up growth spills over positively to average growth while reducing pollution, and can therefore be seen both as an effective component in pollution reduction strategies as well as in general economic growth programs. Health policies for the poor that

reduce the economic impact on these households, may similarly have economic benefits for the broader economy by leveraging growth spillovers and pollution reduction effects. Optimal pollution policies have both a positive effect on expenditures, specifically for the poor, while reducing exogenous pollution. Simple examples may include distributing cleaner gas stoves such as under the Clean Stove Initiative of the World Bank. This type of initiative reduces particulate matter emissions by reducing the amount of wood, agricultural residues, dung, and coal burned, while having a positive effect directly on bottom household wealth. Wealth increase in the bottom, then has the potential to spill over through the entire economy. In a different fashion, a pollution tax such as under Chile's Green Tax Strategy, may in fact well be a less optimal way of pollution control, specifically if it is not sufficiently progressive.<sup>10</sup> In these cases, impacting household income and expenditures interferes with the overall effectiveness. Tax-based policies may possibly be made more effective if the tax revenues are in turn invested in the poor.

The analysis also found that the economic impacts of pollution are higher in more severely polluted areas. Combined, the evidence points toward a pro-active stance towards both poverty reduction and pollution abatement as early in the development process as possible. A "grow first, solve later" attitude in either case leads to the lesser effective growth strategy. Letting pollution increase, results in increasingly higher damages. Both in a cumulative, but also in a marginal sense. Slowed growth of the poor prolongs poverty, which in turn slows down a potential pollution decline. The narrative of pollution naturally reducing as development occurs is a decades-old concept, and has been surrounded by controversy and debate related to its implications for development (see Stagl (1999) and Soumyananda (2004) for examples). The so-called "clean-up phase" that historically accompanied middle- and late-stage income growth has long

---

<sup>10</sup>This does not imply that pollution taxes are not effective. In fact, multiple studies have shown the effectiveness of tax-based approaches in curbing pollution (Deschenes et al., 2012; Shapiro and Walker, 2016).

been misinterpreted as a justification for knowingly developing through “dirty” means and neglecting to establish policy interventions that would curb early-stage pollution. We hope our evidence contributes to an ending of this unjustified and harmful interpretation that can only lead to bad economic outcomes. This conclusion has been put forward also by others, already in earlier literature (Panayatou, 1997; Lee, 2012).

## 6.7 Appendix

### 6.7.1 Restrictions

#### Restricted SVARMA 1

A model in which the joint process has autoregressive forces that feedback in the time-dimension between the sequences, while variables feedback simultaneously within the cross-sections, could be written as

$$\begin{bmatrix} \mathbf{x}_t + H_1^{xx} \mathbf{x}_{t-1} + H_1^{xy} \mathbf{y}_{t-1} + \dots + H_p^{xx} \mathbf{x}_{t-p} + H_p^{xy} \mathbf{y}_{t-p} \\ \mathbf{y}_t + H_1^{yx} \mathbf{y}_{t-1} + H_1^{yy} \mathbf{x}_{t-1} + \dots + H_p^{yx} \mathbf{y}_{t-p} + H_p^{yy} \mathbf{x}_{t-p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_t^x + M_1^{xx} \boldsymbol{\epsilon}_{t-1}^x + \dots + M_q^{xx} \boldsymbol{\epsilon}_{t-q}^x \\ \boldsymbol{\epsilon}_t^y + M_1^{yy} \boldsymbol{\epsilon}_{t-1}^y + \dots + M_q^{yy} \boldsymbol{\epsilon}_{t-q}^y \end{bmatrix} \quad \forall t \in \mathbb{Z}. \quad (6.33)$$

This model constrains  $M_{0:p}^{xy}$  and  $M_{0:p}^{yx}$  to zero, implying that residuals and lagged residuals enter only in one cross-section, while the observations may still depend on the observations in both cross-sections. We can write this efficiently by working with parameter matrices

$$\begin{aligned} \mathbf{H}_0 &:= \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}, \quad \mathbf{H}_{1:p} := \begin{bmatrix} H_{1:p}^{xx} & H_{1:p}^{xy} \\ H_{1:p}^{yx} & H_{1:p}^{yy} \end{bmatrix}, \\ \mathbf{M}_0 &:= \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}, \quad \mathbf{M}_{1:p} := \begin{bmatrix} M_{1:p}^{xx} & O_{n_x} \\ O_{n_y} & M_{1:p}^{yy} \end{bmatrix}. \end{aligned} \quad (6.34)$$

### Restricted SVARMA 2

Alternatively, we can work with moving averages that enter both equations directly, e.g., the second part of the equality in eq. (6.33) is of the form:

$$\begin{bmatrix} \epsilon_t^x + M_1^{xx} \epsilon_{t-1}^x + M_1^{xy} \epsilon_{t-1}^y + \dots + M_q^{xx} \epsilon_{t-q}^x + M_q^{xy} \epsilon_{t-q}^y \\ \epsilon_t^y + M_1^{yx} \epsilon_{t-1}^x + M_1^{yy} \epsilon_{t-1}^y + \dots + M_q^{yx} \epsilon_{t-q}^x + M_q^{yy} \epsilon_{t-q}^y \end{bmatrix}. \quad (6.35)$$

The matrix representation results from

$$\begin{aligned} \mathbf{H}_0 &:= \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}, \quad \mathbf{H}_{1:p} := \begin{bmatrix} H_{1:p}^{xx} & H_{1:p}^{xy} \\ H_{1:p}^{yx} & H_{1:p}^{yy} \end{bmatrix}, \\ \mathbf{M}_0 &:= \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}, \quad \mathbf{M}_{1:p} := \begin{bmatrix} M_{1:p}^{xx} & M_{1:p}^{xy} \\ M_{1:p}^{yx} & M_{1:p}^{yy} \end{bmatrix}. \end{aligned} \quad (6.36)$$

This model allows that each effect goes through a spatial multiplier that may differ in structure and strength for each panel variable.

We make a clear distinction between the two cases because the equations in the first model can be computed without the moving averages of other variables being available. Therefore, the criterion functions can be evaluated on an equation-by-equation basis which allows better parallelization of tasks. In the second model, the impulse generating mechanisms may cross-interact, and all equations have to be evaluated simultaneously or in matrix form. This becomes computationally demanding even for a small number of variables and moderate  $n_w$  and  $T$ . It is still possible to invert the contemporaneous spillovers on an equation by equation basis, which means that parts of the computation can still be parallelized. The second model is a restricted version of the case in which both observations and residuals have contemporaneous effects between variables.<sup>11</sup> From a

---

<sup>11</sup>The unrestricted model with contemporaneous effects between variables results from

$$\mathbf{H}_{0:p} := \begin{bmatrix} H_{0:p}^{xx} & H_{0:p}^{xy} \\ H_{0:p}^{yx} & H_{0:p}^{yy} \end{bmatrix}, \quad \mathbf{M}_{0:p} := \begin{bmatrix} M_{0:p}^{xx} & M_{0:p}^{xy} \\ M_{0:p}^{yx} & M_{0:p}^{yy} \end{bmatrix},$$

practical aspect it is useful to first consider models of the type eq. (6.34) first, and use the results to feed numerical algorithms to estimate models of the eq. (6.36) type.

### 6.7.2 Stability in terms of the companion matrix

Consider the Markov Chain,

$$\mathbf{w}_t = \mathbf{M}(L)\{\mathbf{H}^{-1}(L)\boldsymbol{\epsilon}_t\} = \mathbf{M}(L)\boldsymbol{\Xi}_t \quad \forall t \in \mathbb{Z},$$

with identity normalization of the spatially multiplied autoregressive matrix at  $t = 0$ , and  $p = q$  for simplicity. After generating the spatially correlated residuals  $\boldsymbol{\epsilon}_t$  from  $\boldsymbol{\varepsilon}_t$ , the values of  $\mathbf{w}_t$  can be generated in two stages. First,

$$\boldsymbol{\Xi}_t = \boldsymbol{\epsilon}_t - \{\mathbf{H}_1\boldsymbol{\Xi}_{t-1} + \dots + \mathbf{H}_p\boldsymbol{\Xi}_{t-p}\},$$

then,

$$\mathbf{w}_t = \mathbf{M}_0\boldsymbol{\Xi}_t + \mathbf{M}_1\boldsymbol{\Xi}_{1t-1} + \dots + \mathbf{M}_{p-1}\boldsymbol{\Xi}_{t-p+1}.$$

By defining the set of  $p$  state variables:

$$\begin{aligned} \boldsymbol{\Xi}_{1t} &= \boldsymbol{\Xi}_t, \\ \boldsymbol{\Xi}_{2t} &= \boldsymbol{\Xi}_{t-1}, \\ &\vdots \\ \boldsymbol{\Xi}_{pt} &= \boldsymbol{\Xi}_{t-p+1}. \end{aligned}$$

and rewriting the Markov Chain in terms of the left hand side variables:

$$\mathbf{w}_{1t} = \boldsymbol{\epsilon}_t - \{\mathbf{H}_1\boldsymbol{\Xi}_{1t-1} + \dots + \mathbf{H}_p\boldsymbol{\Xi}_{pt-1}\}.$$

we can use the state vector  $\boldsymbol{\Xi}_t = [\boldsymbol{\Xi}_{1t}, \boldsymbol{\Xi}_{2t}, \dots, \boldsymbol{\Xi}_{pt}]'$  to write the system

---

in which the connectivity matrices that generate the off-diagonal blocks  $H_{0:p}^{xy}$  and  $H_{0:p}^{yx}$  may be designed to have non-zero diagonals. While interesting from a theoretical perspective, we were not able to design algorithms for estimation that carried value in a practical context.

after defining  $\mathbf{O} = \mathbf{0} \circ \mathbf{I}$ :

$$\begin{bmatrix} \Xi_{1t} \\ \Xi_{2t} \\ \vdots \\ \Xi_{pt} \end{bmatrix} = \begin{bmatrix} -\mathbf{H}_1 & \dots & -\mathbf{H}_{p-1} & -\mathbf{H}_p \\ \mathbf{I} & \dots & -\mathbf{O} & \mathbf{O} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{O} & \dots & \mathbf{I} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \Xi_{1t-1} \\ \Xi_{2t-1} \\ \vdots \\ \Xi_{pt-1} \end{bmatrix} + \begin{bmatrix} \mathbf{I} \\ \mathbf{O} \\ \vdots \\ \mathbf{O} \end{bmatrix} \epsilon(t),$$

with the sparse matrix on the right side of the equality being the companion matrix that has the following accompanying measurement equation

$$\mathbf{w}_t = \mathbf{M}_0 \Xi_{1t} + \dots + \mathbf{M}_{p-1} \Xi_{pt} \quad \forall t \in \mathbb{Z}.$$

Stability then be expressed in terms of the companion matrix  $\Phi$ . Remember that its elements correspond to the inverted autoregressive components  $\mathbf{H}$ , hence it is straightforward that this yields the conditions that the eigenvalues of  $\Phi$  must lie within the unit circle:

$$\det(\mathbf{I} - \Phi(z)) = \det(\mathbf{H}(z)) = \det(\mathbf{H}_0 + \mathbf{H}_1 + \dots + \mathbf{I} + \mathbf{H}_p z^p) \neq 0 \quad \forall |z| \leq 1.$$

Note that if  $\rho = 0$ ,  $\mathbf{S} = (\mathbf{I} + \mathbf{O})^{-1} = \mathbf{I}$ , as an effect  $\mathbf{H} = \mathbf{A}$ , which gives

$$\det(\mathbf{I} - \Phi(z)) = \det(\mathbf{A}(z)) = \det(\mathbf{A}_0 + \mathbf{A}_1 + \dots + \mathbf{I} + \mathbf{A}_p z^p) \neq 0 \quad \forall |z| \leq 1.$$

Finally, this only differs from the standard condition cited in VARMA literature that

$$\det(I - \Phi(z)) = \det(A(z)) = \det(A_0 + A_1 + \dots + I + A_p z^p) \neq 0 \quad \forall |z| \leq 1,$$

by construction of our parameter matrices that link the scalar coefficients to the cross-sectional observations. However, since there is no parameter heterogeneity left, the two conditions are identical. Finally, to better understand the relationship between the spatial multiplier for nonzero  $\rho$  and the autoregressive parameter in determining stability, the additional results in (Andree et al., 2017a) are of help. While the stability conditions

of SVARMA or straightforward in terms of high-level conditions, they involve many parameters and in practice it may be less straightforward to calculate them for testing purposes. We suggest that for practical purposes, it may be less cumbersome to simulate from the model under impulses, and see if the responses converge as the researcher should be interested in this either way.

### 6.7.3 Small sample distribution of the (P)MLE

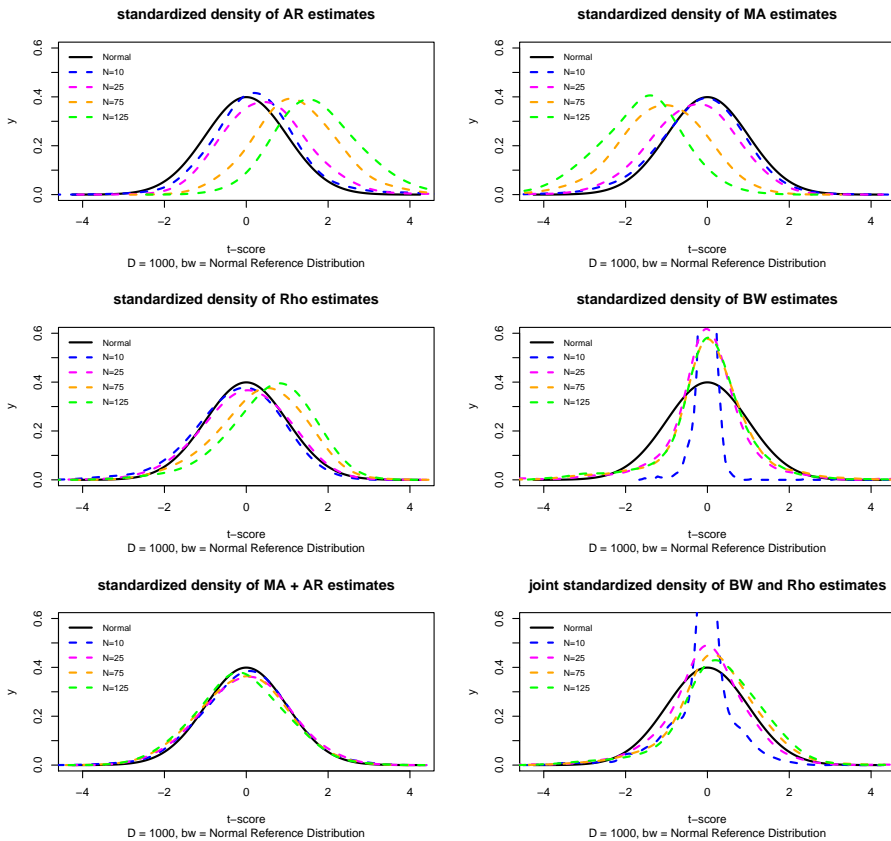


Figure 6.5: Unpenalized small sample distributions of the correctly specified SARMA, bandwidth of the spatial kernel matrix in the DGP set to .15.



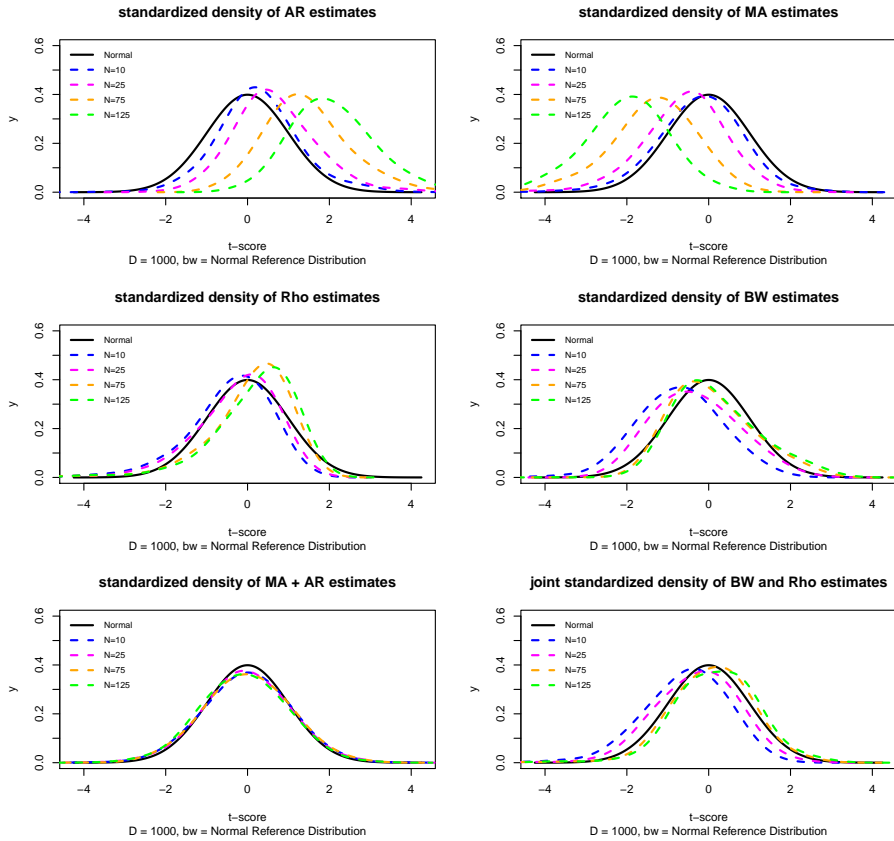


Figure 6.6: Penalized small sample distributions of the correctly specified SARMA, bandwidth of the spatial kernel matrix in the DGP set to 2.

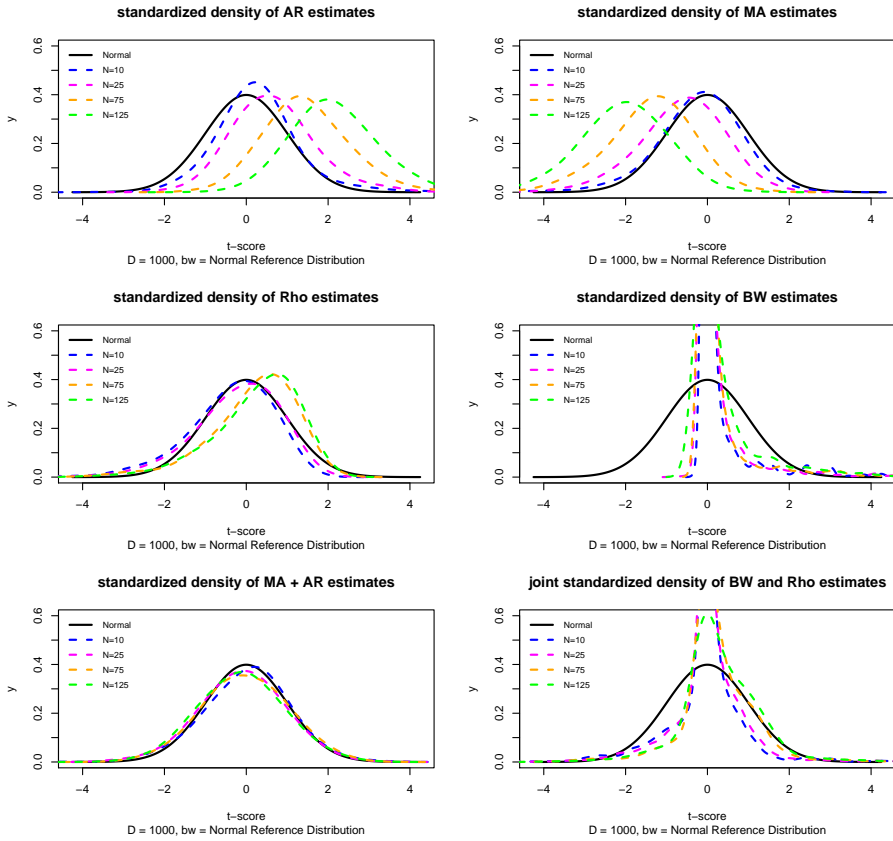


Figure 6.7: Unpenalized small sample distributions of the correctly specified SARMA, bandwidth of the spatial kernel matrix in the DGP set to 2.

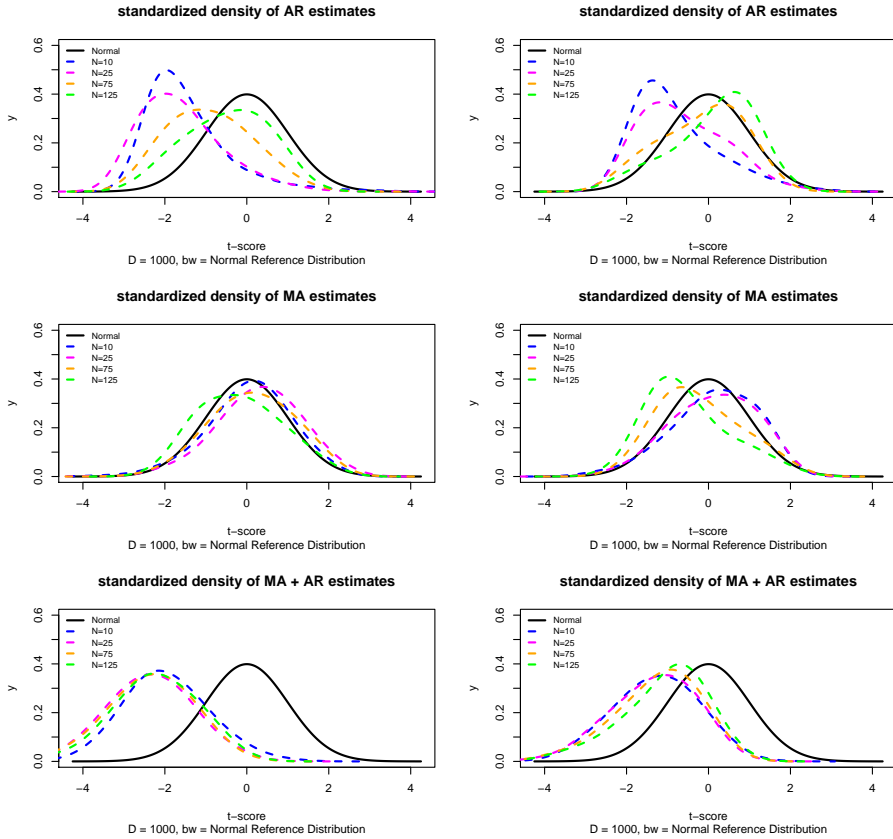


Figure 6.8: Unpenalized small sample distributions of the misspecified ARMA, when the *true* process is an SARMA with bandwidth of the spatial kernel matrix set to .15 (left) and 2 (right).

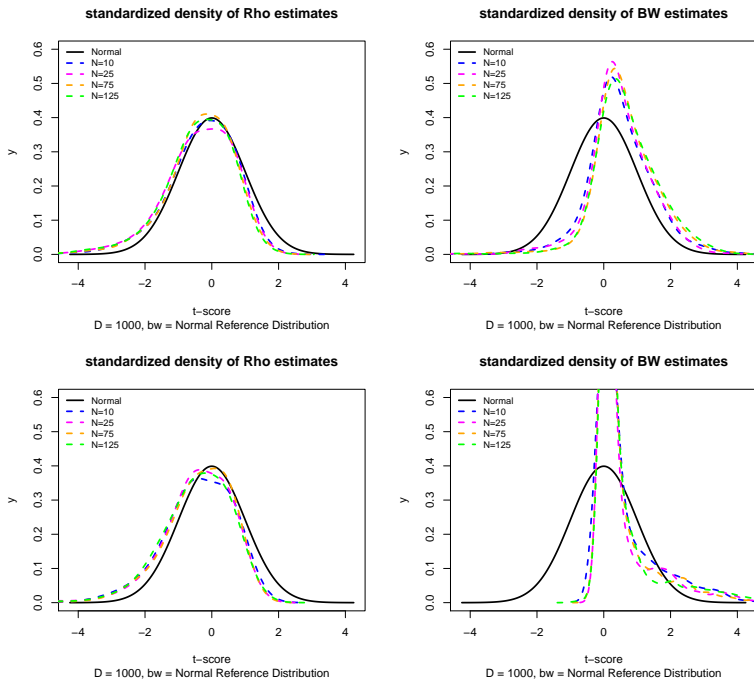


Figure 6.9: Penalized (upper) small sample distributions of bandwidth and spatial parameter in the SARMA, when the *true* process is a cross-sectional ARMA with zero spatial effects. The bandwidth density is centered around 0, note that the kernel structure is not identified at this value.

6.7.4 Pollution data

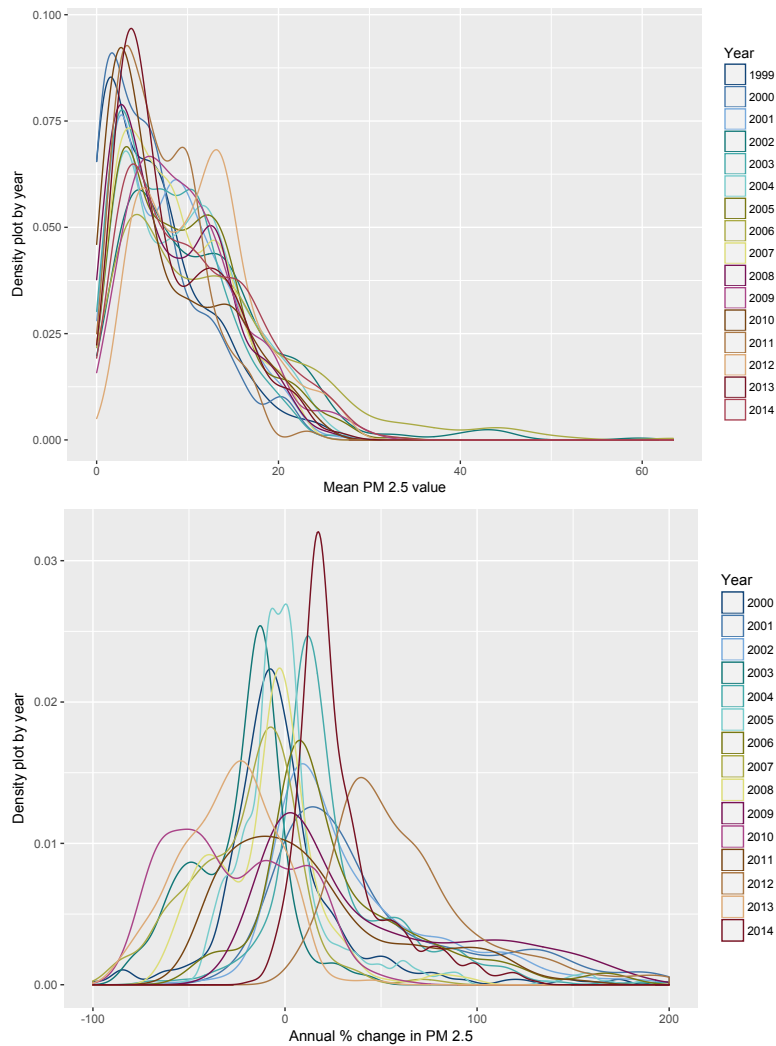


Figure 6.10: Densities of pollution levels (left) and changes in pollution (right) for 219 areas with an urban patch of over 5,000 people and densities of 300 per square kilometer or higher.

## 6.7.5 Additional regression results

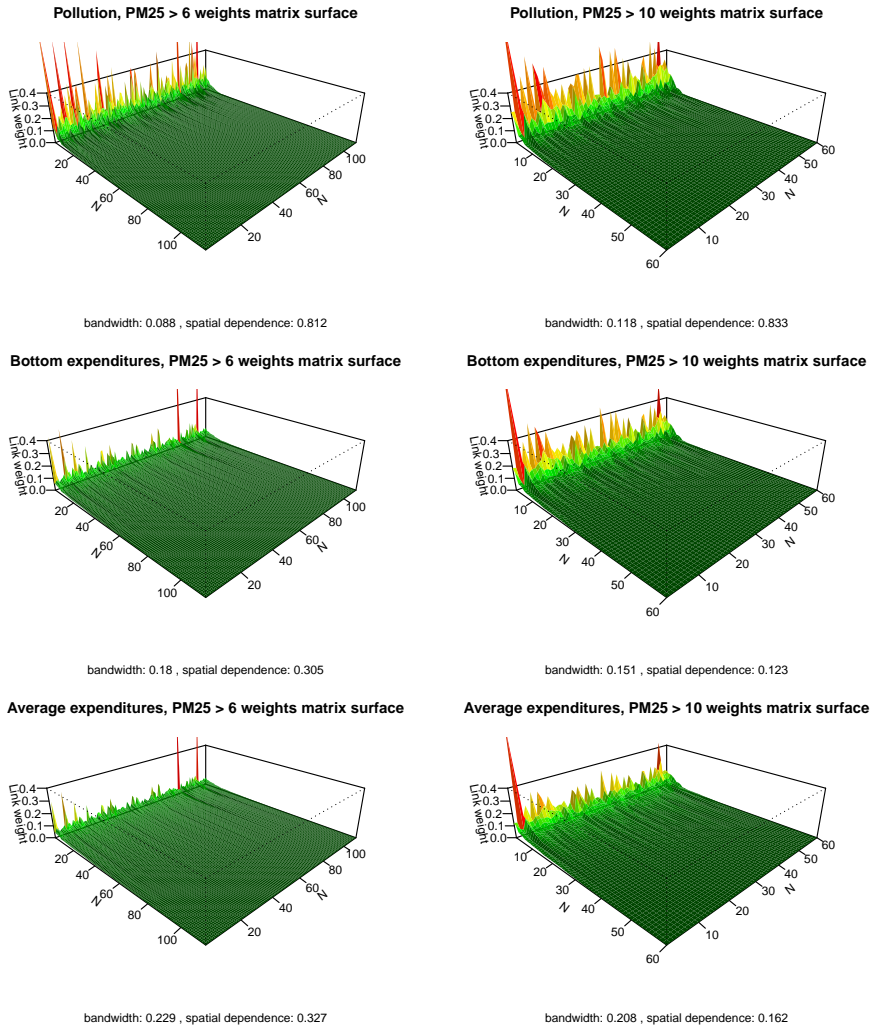


Figure 6.11: Surfaces of estimated spatial weights, ordered by link strengths (observations in no particular order), revealing the different links and links strengths across the different channels of the SVARMA structure.

Table 6.3: VARMA(AICc) results at  $PM_{2.5} > 6$ ,  $\hat{R}^2 = 0.705$ , 42 estimated parameters on  $(N - \max(p, q) \times T) \times 3 = 4068$  data points with 372 fixed demeaning components.  $AICc = -6895.750$ .

	Pollution	Bottom Expenditures	Expenditures
$\phi pol_{t-1}$	-0.456*** (-6.407)	-0.155*** (-2.745)	0.124* (1.862)
$\phi pol_{t-2}$	-0.201*** (-6.171)	-0.101*** (-3.215)	
$\phi pol_{t-3}$			0.078*** (3.081)
$\phi bot_{t-1}$	0.101** (2.019)	-0.119** (-2.02)	
$\phi bot_{t-2}$		-0.138*** (-4.645)	-0.123*** (-2.655)
$\phi bot_{t-3}$	-0.056** (-2.362)	-0.094*** (-3.333)	-0.156*** (-3.825)
$\phi exp_{t-1}$		0.069*** (2.884)	-0.277*** (-4.884)
$\phi exp_{t-2}$		0.159*** (4.557)	
$\phi exp_{t-3}$		0.072*** (3.099)	
$M pol_{t-1}$	-0.142* (-1.853)	0.145** (2.444)	-0.196*** (-2.7)
$M pol_{t-2}$	-0.100** (-2.236)		0.129*** (2.788)
$M pol_{t-3}$	0.068* (1.806)	-0.085*** (-2.877)	-0.088** (-2.225)
$M bot_{t-1}$	-0.148*** (-2.585)	-0.235*** (-3.855)	0.095** (2.538)
$M bot_{t-2}$			0.221*** (3.692)
$M bot_{t-3}$			0.141** (2.354)
$M exp_{t-1}$			-0.119* (-1.827)
$M exp_{t-2}$		-0.135*** (-3.658)	-0.356*** (-8.198)
$M exp_{t-3}$			-0.137*** (-3.75)
$\sigma$	0.109	0.087	0.100
$\nu$	3.797	5.031	5.721
4-lag white-noise $p$	1.000	0.085*	0.025**

Note:

\* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$

Constant omitted,  $t$ -statistics in parenthesis for the ARMA components.

Table 6.4: VARMA(AICc) results at  $PM_{2.5} > 10$ ,  $\hat{R}^2 = 0.722$ , 37 estimated parameters on  $(N - \max(p, q) \times T) \times 3 = 2160$  data points with 213 fixed demeaning components.  $AICc = -3876.735$ .

	Pollution	Bottom Expenditures	Expenditures
$\phi\ pol_{t-1}$	-0.578*** (-17.043)	-0.542*** (-4.875)	-0.413*** (-2.811)
$\phi\ pol_{t-2}$	-0.234*** (-4.764)	-0.320*** (-4.866)	-0.204** (-2.459)
$\phi\ pol_{t-3}$	0.064* (1.883)		
$\phi\ bot_{t-1}$	0.138* (2.327)	-0.440*** (-11.281)	-0.248** (-2.558)
$\phi\ bot_{t-2}$	0.083** (2.374)		
$\phi\ bot_{t-3}$		-0.061* (-1.85)	
$\phi\ exp_{t-1}$		0.151*** (3.049)	-0.172** (-2.654)
$\phi\ exp_{t-2}$		0.056* (1.837)	-0.219*** (-5.579)
$\phi\ exp_{t-3}$			
$Mpol_{t-1}$		0.569*** (4.903)	0.342** (2.244)
$Mpol_{t-2}$	-0.184*** (-3.11)		
$Mpol_{t-3}$	-0.134*** (-2.594)	-0.294*** (-4.529)	-0.208*** (-2.751)
$Mbot_{t-1}$	-0.188*** (-2.644)		0.348*** (3.247)
$Mbot_{t-2}$		-0.344*** (-6.453)	-0.154** (-2.138)
$Mbot_{t-3}$			
$Mexp_{t-1}$		-0.106** (-2.021)	-0.303*** (-4.569)
$Mexp_{t-2}$	-0.079*** (-2.579)		
$Mexp_{t-3}$	0.054* (1.789)		-0.286*** (-6.83)
$\rho$			
$b$			
$\sigma$	0.094	0.079	0.101
$\nu$	4.107	9.474	4.573
p white-noise	1.000	0.311	0.498

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Constant omitted,  $t$ -statistics in parenthesis for the ARMA components.



Table 6.5: SVARMA(AICc) results at  $PM_{2.5} > 10$ ,  $\hat{R}^2 = 0.732$ , 39 estimated parameters on  $(N - \max(p, q) \times T) \times 3 = 2160$  data points with 213 fixed demeaning components.  $AICc = -4153.838$ .

	Pollution	Bottom Expenditures	Expenditures
$\phi\ pol_{t-1}$	-0.097** (-2.503)	-0.150*** (-2.803)	-0.085** (-2.367)
$\phi\ pol_{t-2}$		-0.073** (-2.048)	-0.049 (-1.482)
$\phi\ pol_{t-3}$	0.041* (1.773)	-0.045 (-1.446)	
$\phi\ bot_{t-1}$	0.092** (2.446)		
$\phi\ bot_{t-2}$		-0.271*** (-4.573)	-0.200*** (-3.518)
$\phi\ bot_{t-3}$			-0.093** (-2.326)
$\phi\ exp_{t-1}$			-0.324*** (-5.421)
$\phi\ exp_{t-2}$		0.133*** (2.995)	
$\phi\ exp_{t-3}$		0.052* (1.964)	
$M\ pol_{t-1}$	-0.362*** (-6.212)	0.198*** (3.19)	
$M\ pol_{t-2}$	-0.105** (-2.33)		
$M\ pol_{t-3}$	0.120*** (2.96)		
$M\ bot_{t-1}$	-0.146*** (-3.43)	-0.447*** (-12.057)	
$M\ bot_{t-2}$		0.194*** (2.864)	0.260*** (3.825)
$M\ bot_{t-3}$		-0.216*** (-4.225)	
$M\ exp_{t-1}$			-0.106 (-1.496)
$M\ exp_{t-2}$		-0.139*** (-2.911)	-0.400*** (-8.092)
$M\ exp_{t-3}$			-0.159*** (-3.934)
$\rho$	0.833*** (24.272)	0.123 (1.374)	0.162 (1.46)
$b$	0.118	0.151	0.208
$\sigma$	0.906	0.080	0.101
$\nu$	2.004	7.313	4.703
p white-noise	1.000	0.187	0.864

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01  
Constant omitted, *t*-statistics in parenthesis for the SARMA components.

### 6.7.6 Additional Impulse Response analysis results

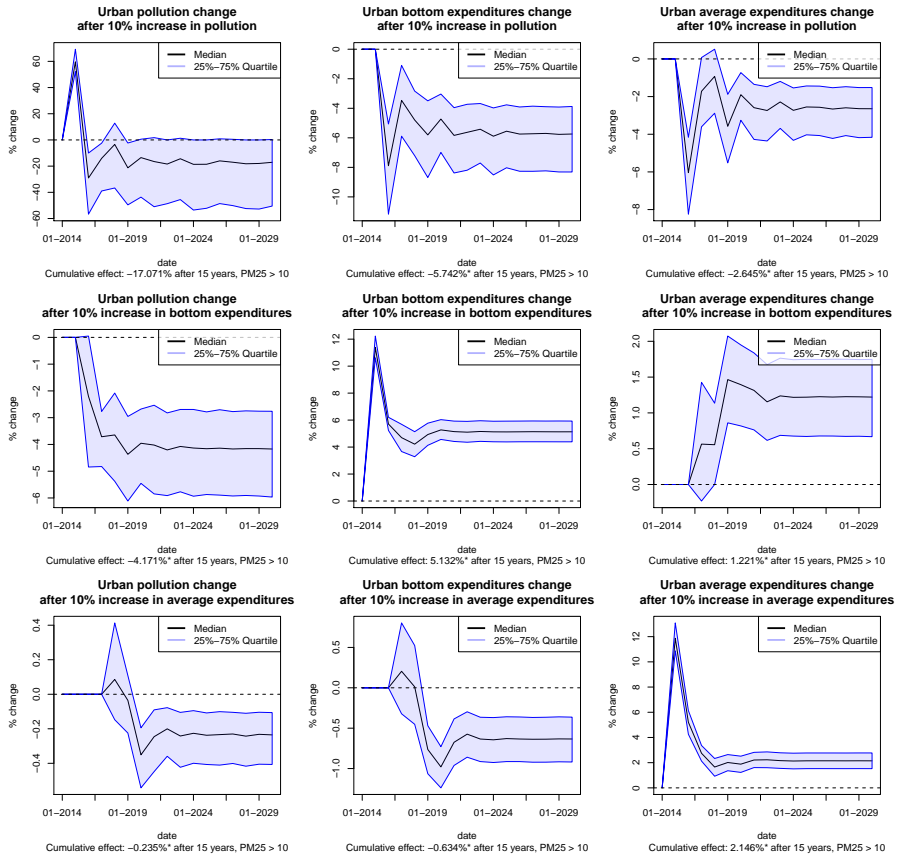


Figure 6.12: IRF plots for exogenous shocks in pollution, bottom household expenditures and average household expenditures  $PM_{2.5} > 10$ . Effects that exclude zero in the final year, are marked by \*.

Table 6.6: Economic pollution costs based on a conversion rate from IDR to dollars of 100,000 IDR to 7.410 USD – Pulled from Google Finance on 15 October, 2017.

	Annual expenditures in USD per capita	Average annual loss in USD for 10% $PM_{2.5}$ increase
Bottom household $PM_{2.5}^{6+}$	397.132	8.844
Average household $PM_{2.5}^{6+}$	1074.54	9.177
Bottom household $PM_{2.5}^{10+}$	420.50	24.145
Average household $PM_{2.5}^{10+}$	1183.96	31.316



# Chapter 7

## Probability and Causality in Spatial Time Series

### Chapter Summary

The current paper discusses approximating a correct theory of cause and effect by minimizing distance to its associated probability measure in a space of measures in which each element is associated with a stochastic representation of a candidate theory. The discussion encourages researchers to use flexible dynamical models to model and discover the *true* quantitative relationships that may be hidden in interrelated stochastic data. The argument is based on the use of a decision criterion that scales to a metric that measures distance between any given measure. When this is the case, a metric space can be considered in which equivalences can be established by partitioning into classes of zero-distance points. Equivalence to the *true* measure, that is associated with the *true* frequencies in Markov chains of iterated causes and effects, is established by reaching zero distance in that space. When the hypothesis space is incorrectly constructed, equivalence is established with respect to a pseudo-*true* measure that by definition is closest to the correct hypothesis across all considered hypotheses. The specific case of Maximum Likelihood is further discussed. In particular, squared Hellinger distance marks a lower bound of Kullback-Leibler divergence. This implies that maximizing complexity penalized likelihood minimizes distance toward the *true* probability measure. As such, it is an objective that approximates the correct causal structure from interrelated stochastic data that are observed and modeled sequentially over time.<sup>1</sup>

---

<sup>1</sup>This chapter is based on “*Probability, Causality and Stochastic Formulations of Economic Theory*”, available on the Social Science Research Network. The reference is (Andree, 2019).

## 7.1 Introduction

The 20th century has seen much work done on establishing the statistical properties of estimators widely used in econometrics. Notably, Kolmogorov (1933) laid out the axiomatic foundations of modern probability theory and, one year later, Doob (1934) proved the law of large numbers using a probabilistic interpretation of Birkhoff's ergodic theorem. Doob then used this to prove theorems of Fisher (1922, 1925) and Hotelling (1930) on estimating a parameter of a distribution by method of Maximum Likelihood, establishing both the Consistency and Normality of the MLE. Wald (1949) provided a proof for the multi-parameter case with greater generality. Earlier assertions on the efficiency of the MLE by Fischer, and importantly Cramer (1946), were eventually substantiated by rigorous proof by Rao (1962) resulting in what is now known as the Cramer-Rao bound. Generalizations that cover the nonlinear case were developed, initially with difficult to verify conditions (Le Cam, 1953) and (Kraft, 1955), but later for the general case of stationary Markov processes (Roussas, 1965) which in fact was a result that extended the theory by Wald (1949). It took several decades, but eventually the nonlinear Least Squares case was tackled (Jennrich, 1969; Malinvaud, 1970) which set the basis to a general asymptotic theory of extremum estimators. In the decades that followed, asymptotic properties of extremum estimators have covered multivariate dynamic settings, miss-specified models, heterogeneity, and dependence of the data. A good modern review is Pötscher and Prucha (1997).

While the important early statisticians Pearson and Fischer were primarily biometricians, their statistical methods for data analysis were eagerly integrated into economics. Wald, who played a crucial role in developing the Consistency and Normality results, spent much of his time with econometricians and he produced economic theories of his own. Arguably, the most notable contribution in integrating probabil-

ity into econometrics is, however, by Haavelmo (1944) “*The Probability Approach in Econometrics*”. In a less well-known paper that was published one year before, Haavelmo (1943) already provided the basis for stochastic formulation of economic theories and the integration of error terms into regressions. Good historical accounts on how Haavelmo’s work shaped modern econometrics are by Spanos (1989) and Bjerkholt (2007), and a more general reconstruction of the interaction between early econometricians and statisticians is provided by Aldrich (2010).

Though the probabilistic view laid out by Haavelmo to model economic theories has largely been embraced by many applied economists, it seems that more mechanic definitions of causality are largely preferred over probabilistic ones. Particularly Rubins’ viewpoint (Rubin, 1974), which originated from studies on human psychology, has been widely embraced as a model for causality. The core idea behind the Rubins’ approach to identification is that treatment groups of populations with otherwise equal properties can be used to isolate a treatment effect, as no other factor can otherwise be attributed to account for the differential in an observed outcome. This view on causality implies that treatments must, in a deterministic manner, cause outcomes to occur. In fact, Pearl (2000) states that cause and effect relations are fundamentally deterministic, explicitly excluding quantum mechanical phenomena from his concept of cause and effect but mentioning that causal analysis involves probability language (see also the review by Neuberger et al. (2003)). The probabilistic approaches to causality, such as laid out by Granger (1969, 1980); Covey and Bessler (1992) that involve contrasting the probabilistic forecasting performance of a univariate and bivariate specification, are done away by Pearl (2000). In particular, Pearl (2000) makes explicit mention that this is not causality, and that the concepts of “strong exogeneity” (Engle et al., 1983) and Granger-causality are only statistical concepts. His view on causality is purely mechanical. At the same time, there are many examples in physics, the study that was born out of classical mechanics,

that approach causality from the probabilistic angle. On one hand this may relate to the fact that branches in both physics and economics evolve around models of dynamical systems in which the control-theoretic concept of a non-anticipative system is the basis for causal relationships, see for example Liang (2016); Harnack et al. (2017); Krakovska et al. (2018) for examples of recent causal studies in physics that work around estimating the time dependencies in dynamical systems. On the other hand, it may be related to developments in quantum mechanics that suggest that reality itself is probabilistic in nature, a profound notion that arguably still has not been fully clarified since the Bohr-Einstein debates. This would, at some level of abstraction, turn the universe non-causal under Pearl's view, which may be a philosophically difficult proposition. For example, in *"Causality and Chance in Modern Physics"*, Bohm (1999), one year before Pearl, argues that any theory about reality that embraces either one of causality and chance, to the exclusion of the other, is inherently incomplete.

The conflicting views might seem an inconsistency, and the mechanical approach to causality that is widely used in economics seems in stark contrast to the viewpoint presented in Haavelmo (1943) that turned econometrics into a probabilistic study. Particularly, Haavelmo's core argument was that it is the very nature of economic behavior itself, that implies the necessity of stochastic formulations of economic theory and the inclusion of error terms in otherwise exact relationships to make simplifications of reality elastic enough for application. Moreover, Kalman (1983) definitively argues that the classical model of reality developed in mechanical physics is simply inapplicable to the problems of economics. It is certainly interesting that, after a century of probability work by statisticians and econometricians that led statistics to be accepted as the leading model for inference, the working definition of causality used by many economists is deterministic in nature, while physicists are open to work with Granger's definition.

Efforts to reunite the conflicting views have recently begun to produce interesting results. New developments began by noting that questions about important concepts in economics, such as choice and uncertainty, can even in very simplistic settings not be answered within Pearl's framework. White and Chalak (2009); White et al. (2014) extend Pearl's causal model to include optimization, equilibrium, learning concepts, and choice that are integral parts of economics and game theory, or social systems in which agents act and react under uncertainty. Under the extended causal framework, White and Lu (2010) forge the previously missing link between Granger causality and structural causality by showing that, given a corresponding conditional form of exogeneity, Granger causality holds *if and only if* a corresponding form of structural causality holds, and Eichler and Didelez (2010) provide conditions under which Granger non-causality implies that an intervention has had no effect. White et al. (2011) show that tests for Granger causality can be used to test for direct causality in sequential systems, and Lu et al. (2017) produce tests for cross-section and panel data valid in a general case that does not assume linearity, monotonicity in observables or unobservables, or separability between observed and unobserved variables in the structural relations. White and Pettenuzzo (2014) show that instead of relying on exogeneity (weak, strong, or super) conditional on the model or *Data Generating Process* (DGP), causal effects can also be consistently estimated by relying on correct specification of the conditional mean sequence. This highlights the importance of knowledge regarding the important features of the DGP. Specifically, economic theory may suggest which variables are meaningful, while the functional form (numbers of lags, cointegration, or structural shifts), may be resolved directly from the data.

In this paper, we continue the debate focusing on the application strategy to estimate causal relationships, taking a general, data-driven, stand. Specifically, we assume that a theorized causal relationship between two economic variables in the possible presence of unobserved factors leads to



a probability law that regulates the transitions from one phase to another in Markov chains of iterated processes of causes and effects. This is particularly relevant given the arguments of Haavelmo (1943) and others discussed, that are in favor of formulating economic theories stochastically. In this probabilistic setting, the properties of the extremum estimators introduced earlier provide a natural interpretation of an estimated result regardless of whether correct specification is assumed. Specifically, under simple conditions that are often guaranteed by the design of standard estimation problems, the limit result is the closest to the correct hypothesis about causality out of all considered hypotheses. Assuming correct specification, ensures naturally that minimal distance is zero, which corresponds to the setting of White and Pettenuzzo (2014). When the hypothesis space is incorrectly constructed, equivalence is established with respect to the pseudo-*true* measure that by definition, again, is closest to the correct hypothesis out of all considered hypotheses. If the hypothesis space is sufficiently large to ensure a small divergence between the *true* causal probability law and the closest possible modeled measure, then the limit result should naturally capture important aspects of the *true* causal probability law even under miss-specification. This suggests that in the absence of clear economic theories to guide model specification, a researcher can still focus on ensuring that the parameter space is able to produce as much hypotheses about causality as possible and proceed with a general estimation method that penalizes model complexity.

The core of the argument is based on the use of a decision criterion that scales to a metric measuring distance between any two probability measures. When this is the case, a metric space can be considered in which equivalences can be established by partitioning into classes of zero-distance points. Equivalence to the *true* measure, that correctly describes the *true* frequencies in Markov chains of iterated processes of causes and effects, is established by reaching zero distance in that metric space. These type of decision criteria are common in econometrics and an example is

provided in the context of Maximum Likelihood. We show that squared Hellinger distance marks the lower bound of Kullback-Leibler divergence, implying that minimizing information loss using the AIC, or another suitable penalized Likelihood variation that ranks hypotheses according to their Kullback-Leibler divergence, minimizes a distance metric toward the *true* probability measure. As such, minimizing AIC is a theoretically sound objective to uncover the correct causal structure from interrelated stochastic data that are observed and modeled sequentially over time.

The remainder of this paper is structured as follows. Section 7.2 introduces definitions of causality in terms of probability measures, section 7.3 discusses the divergence between modeled measures and the causal measure, section 7.4 discusses the particular case of the Maximum Likelihood estimator and squared Hellinger distance. Section 7.5 ends with concluding remarks.

## 7.2 Causality and probability

Cause and effect in deterministic settings involve propositions along the lines of “*if  $X$  occurs then  $Y$  must occur*”. That deterministic definition of causality is difficult to reconcile with probability. Causality statements in a statistical context often spur a great deal of discussion among researchers. In fact, while many researchers meet the concept of causality early in their career, few eventually agree on what it truly means and how it should be approached in an empirical context. To introduce a concept of causality appropriate in a probabilistic setting, let us consider a simple game of chance; a dice. Throwing a dice does not cause a certain outcome. In that sense, one cannot say that “*if  $X$  occurs then  $Y$  must occur*” with  $X$  being a throw, and  $Y$  being the outcome of a throw. In fact, the outcome is one of seven, six being one to six eyes, and seven being no outcome at all. Each outcome occurs with a certain probability,

the latter being zero. The measure that assigns probability to each of the outcomes describes the *true* probabilistic property of the dice. In that sense, a faulty dice may *cause* a certain outcome to occur with higher probability *truly*. One can say that a faulty dice is characterized by a probability measure that leads to outcomes with a certain outcome having a higher probability assigned to it, such that the expected value of a throw minus the expected value of a throw of a non-faulty dice is nonzero. In this sense, one can describe the causal effect of being faulty, in terms of probability. Specifically, “ $\nabla Y$  must occur with probability  $P \geq 0$  if  $X$  has occurred” with  $\nabla Y$  being a non-zero difference value between throws of the faulty dice and a correct dice.

In this probabilistic view, a theory about causality is a statement about the properties of the *true* measure that describes a process stochastically. Specifically, a causal relationship can be described in terms of whether the *true* probability measure produces a non-empty stochastic sequence describing the directly caused effects from one variable to the other. Or, equivalently, whether the *true* probability measure is associated with a non-empty stochastic sequence of differences between the process that is driven by causes that produce real-valued effects from one variable to the other and the process that does not react to the causes. This is somewhat different than attributing the presence of causal relationships directly to the values of the parameters in a mathematical model of reality, though, as shall be discussed, the definition based on the probability measure equivalently produces statements about parameters or functions. Drawing on Approximation Theory, one can transfer the measure theoretical definitions of *true* causality, to a modeled probability measure in the limit based on an equivalence argument. The modeled measure, in turn, is naturally associated with parameters that determine the functional behavior. Due to well-known results on consistency for approximate extremum estimates, the approximation of the *true* measure eventually thus provides valid descriptions of causality based on empirically modeled

data when a sufficient amount of observations has been collected.

In an empirical sense, stating that a dice is faulty, or equivalently saying that a certain outcome occurs with higher probability, is a statement about the *true* probabilistic property of that dice and such conclusions may result from a modeled probability measure that best confirms to many observed outcomes. Similarly, saying that a modeled dice is faulty, or equivalently saying that a modeled outcome has higher probability than assigned by the *true* probability measure, is a statement about the probabilistic properties of the modeled dice. Such conclusions may result from observing many modeled dices, and many outcomes, and comparing observed outcomes to modeled outcomes repeatedly and selecting the model that best resembles reality. Most estimators by design select from a set of hypothetical realities by some process of divergence minimization w.r.t. the *true* measure. If that decision process is exhaustive across all divergences between possible measures and the *true* measure, then the closest possible measure will be chosen. If the *true* measure is included in all considered measures, then the decision process will end by selecting that measure. When the axiom of correct specification is abandoned, and the correct probability measure is not included in the set of modeled measures, the *true* measure is replaced by a *pseudo-true* measure. This measure by definition still minimizes divergence w.r.t. the *true* measure. The interpretation that a *pseudo-true* measure carries is that, after observing the data and considering all the measures that are induced under all the possible parameter vectors, the *pseudo-measure* probability measure best confirms to the *true* probability measure. In this case, the decision process thus ends with accepting the best approximation of the correct hypothesis as the one from which to derive causal claims, as no better hypothesis about reality can be constructed until a larger set of hypotheses formally comes under review. This is a stronger result than the common statement that  $X$  only helps predicting  $Y$  with the arrow of time as the indicator of the direction of effects.

Improved predictability can be a local result within the space of potential hypotheses. This reveals the intrinsic relationship between the size of the parameter space that is set when the regression is specified, and the empirical claim that results from estimating that regression, suggesting that the quality of causal inference depends on the flexibility of the model to produce a wide variety of potentially (in)correct structures.

Let us now more formally express these thoughts. Notation is as follows,  $\mathbb{N}$ ,  $\mathbb{Z}$  and  $\mathbb{R}$ , respectively denote the sets of natural, integer, and real numbers. If  $\mathcal{A}$  is a set,  $\mathfrak{B}(\mathcal{A})$  denotes the Borel- $\sigma$  algebra over  $\mathcal{A}$ , and  $\times_{t=1}^{t=T} \mathcal{A}$ , alternatively denoted as  $\mathcal{A}_T$ , is the Cartesian product of  $T$  copies of  $\mathcal{A}$ . Definitional equivalence is denoted  $:=$ , which is to be distinguished from  $\equiv$  denoting equivalence, for example in the functional sense. For two maps  $f$  and  $g$ , their composition arises from their point-wise application and is denoted  $f \circ g := f(g)$ . The tensor product is denoted  $\otimes$ . Finally, the empty set  $\emptyset$  is also used in the context of an empty sequence, that sometimes would be notated as  $()$  in literature.

Directional causality is interesting when at least two sequences are considered. Specifically, when the focus is on a  $T$ -period sequence  $\{\mathbf{x}_t(\omega)\}_{t=1}^T$ , that is a subset of the realized path of the  $n_{\mathbf{x}}$ -variate stochastic sequence  $\mathbf{x}(\omega) := \{\mathbf{x}_t(\omega)\}_{t \in \mathbb{Z}}$  for events in the event space  $\omega \in \Omega$ . That is,  $\mathbf{x}_t(\omega) \in \mathcal{X} \subseteq \mathbb{R}^{n_{\mathbf{x}}} \forall (\omega, t) \in \Omega \times \mathbb{Z}$ . The random sequence  $\mathbf{x}(\omega)$  is a Borel- $\sigma$   $\mathcal{F}/\mathfrak{B}(\mathcal{X}_{\infty})$ -measurable map  $\mathbf{x} : \Omega \rightarrow \mathcal{X}_{\infty} \subseteq \mathbb{R}_{\infty}^{n_{\mathbf{x}}}$ . In this,  $\mathbb{R}_{\infty}^{n_{\mathbf{x}}} := \times_{t=-\infty}^{t=\infty} \mathbb{R}^{n_{\mathbf{x}}}$  denotes the Cartesian product of infinite copies of  $\mathbb{R}^{n_{\mathbf{x}}}$  and  $\mathcal{X}_{\infty} = \times_{t=-\infty}^{t=\infty} \mathcal{X}$  with  $\mathfrak{B}(\mathcal{X}_{\infty}) := \mathfrak{B}(\mathbb{R}_{\infty}^{n_{\mathbf{x}}}) \cap \mathcal{X}_{\infty}$ , and  $\mathfrak{B}(\mathbb{R}_{\infty}^{n_{\mathbf{x}}})$  denotes the Borel sigma algebra on the finite dimensional cylinder set of  $\mathbb{R}_{\infty}^{n_{\mathbf{x}}}$ , see Billingsley (1995), p.159. As always, the complete probability space of interest is described by a triplet  $(\Omega, \mathcal{F}, \mathbb{P})$ , with  $\mathcal{F}$  as the  $\sigma$ -field defined on the event space.  $\mathbb{P}$  is used here as a placeholder as we shall introduce probability measures of interest shortly.

If  $\mathbf{x}$  was considered as a univariate sequence free from exogenous drivers,

then for every event  $\omega \in \Omega$ , the stochastic sequence  $\mathbf{x}_t(\omega)$  would live on the probability space  $(\mathcal{X}_\infty, \mathfrak{B}(\mathcal{X}_\infty), P^{\mathbf{x}})$  where  $P^{\mathbf{x}}$  is defined over elements of  $\mathfrak{B}(\mathcal{X}_\infty)$ . In a similar fashion, one can consider  $\{\mathbf{y}_t(\omega)\}_{t=1}^T$  as the subset of the realized path of the  $n_{\mathbf{y}}$ -variate stochastic sequence  $\mathbf{y}(\omega) := \{\mathbf{y}_t(\omega)\}_{t \in \mathbb{Z}}$  indexed by identical  $t$  for events  $\omega \in \Omega$ . If  $\mathbf{y}$  would live similarly isolated from outside influence, then for every  $\omega \in \Omega$ , the stochastic sequence  $\mathbf{y}_t(\omega)$  would operate on a space  $(\mathcal{Y}_\infty, \mathfrak{B}(\mathcal{Y}_\infty), P^{\mathbf{y}})$  where  $P^{\mathbf{y}}$  assigns probability to all the elements of  $\mathfrak{B}(\mathcal{Y}_\infty)$ . We have a system of two unrelated sequences,<sup>2</sup>

$$\begin{aligned} \mathbf{x} &:= \{\mathbf{x}_t = f^{\mathbf{x}\mathbf{x}}(\mathbf{x}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y} &:= \{\mathbf{y}_t = f^{\mathbf{y}\mathbf{y}}(\mathbf{y}_{t-1}), t \in \mathbb{Z}\} \end{aligned} \quad (7.1)$$

The structure reveals that  $P^{\mathbf{x}}$  is simply induced by the function  $f^{\mathbf{x}\mathbf{x}}$  on  $\mathfrak{B}(\mathcal{X})$  according to  $P^{\mathbf{x}}(B_{\mathbf{x}}) = P^{\mathbf{x}} \circ (f^{\mathbf{x}\mathbf{x}})^{-1}(B_{\mathbf{x}}) \forall B_{\mathbf{x}} \in \mathfrak{B}(\mathcal{X}_\infty)$  and  $P^{\mathbf{y}}$  is induced by the function  $f^{\mathbf{y}\mathbf{y}}$  on  $\mathfrak{B}(\mathcal{Y})$  in a similar way, see Dudley (2002) p.118 and Davidson (1994) p.115. The notion is important to the extent that it has been argued (see Hendry (2017) for discussion) that probabilistic definitions of causality are not strictly causal in the sense that they do not provide insight in the origin of the probability law that regulates the process of interest, and that a (correct) time-series model only describes correctly the probabilistic behavior as the outcome of that unknown causal origin. The notation here shows, however, explicitly the relation between the functional behavior of a system and it's induced probability measure that assigns probability to all possible outcomes. This suggests that such critiquing views rather relate to disagreements around the level of detail in the structure of a model that in turn would be guided by the research question of interest and the availability of data. Particularly, dynamical systems in economics are often modeled using aggregate macro-economic data that does not have the same granularity as micro-economic data that contains information about behavior of

---

<sup>2</sup>This naturally covers to most common auto-regression case (only stated for  $\mathbf{y}_t$  here )  $\mathbf{y}_t = f^{\mathbf{y}\mathbf{y}}(\mathbf{y}_{t-1}) + \varepsilon_t$ , where  $\varepsilon_t$  is unobserved. The linear auto-regression case is obtained when  $f^{\mathbf{y}\mathbf{y}}$  is a scaled identify function.

individual economic agents.

If interrelated stochastic sequences are at the center of inference, the building blocks required for describing the processes are more complicated. This increases the potential complexity of  $P^{\mathbf{x}}$  and  $P^{\mathbf{y}}$  tremendously, but it also allows to conclude decisively between causality, non-causality and feedback. Consider a simple stochastic system:

$$\begin{aligned}\mathbf{x} &:= \{\mathbf{x}_t = f^{\mathbf{xx}}(\mathbf{x}_{t-1}) + f^{\mathbf{xy}}(\mathbf{y}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y} &:= \{\mathbf{y}_t = f^{\mathbf{yx}}(\mathbf{x}_{t-1}) + f^{\mathbf{yy}}(\mathbf{y}_{t-1}), t \in \mathbb{Z}\}.\end{aligned}\tag{7.2}$$

In this multivariate context,  $f^{\mathbf{xy}}$  and  $f^{\mathbf{yx}}$  will be referred to as the direct causal maps, while  $f^{\mathbf{xx}}$  and  $f^{\mathbf{yy}}$  control the memory properties within each channel. When  $\mathbf{x}$  and  $\mathbf{y}$  are analyzed individually, the properties of  $f^{\mathbf{xx}}$  and  $f^{\mathbf{yy}}$  are of key interest, they carry information on the future positions of  $\mathbf{x}_{t+1}$  and  $\mathbf{y}_{t+1}$ , and provide predictability without considering outside influence directly. However, correct causal inference around the interdependencies of  $\mathbf{x}$  and  $\mathbf{y}$  may be preferred over developing predictive capabilities that can result from many configurations within the parameter space that are associated with untrue probability measures. The properties of  $f^{\mathbf{xy}}$  and  $f^{\mathbf{yx}}$  determine the direction in which effects move, and verifying their properties is central to causality studies, while  $f^{\mathbf{xx}}$  and  $f^{\mathbf{yy}}$ , on the other hand, play a central role in the system's responses to external impulses by shaping memory of the causal initial impact of a sequence of interventions, even if that sequence turns inactive immediately after impact. The functions that control memory properties within channels in some sense determine the reflex of the future onto the past, and specifying correct empirical equivalents to  $f^{\mathbf{xx}}$  and  $f^{\mathbf{yy}}$  is just as crucial to the inference about the causal interdependencies as specifying mechanisms for the action of interest is. To understand directional cause,

and the role that  $f^{\mathbf{x}\mathbf{x}}$  and  $f^{\mathbf{y}\mathbf{y}}$  play, it is useful to consider the following:

$$\begin{aligned}\mathbf{x}^0 &:= \{\mathbf{x}_t^0 = f^{\mathbf{x}\mathbf{y}}(\mathbf{y}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y}^0 &:= \{\mathbf{y}_t^0 = f^{\mathbf{y}\mathbf{x}}(\mathbf{x}_{t-1}), t \in \mathbb{Z}\} \end{aligned} \quad (7.3)$$

with  $\mathbf{x}^0$  and  $\mathbf{y}^0$  defined as  $\mathbf{x}_t^0 = \mathbf{x}_t - f^{\mathbf{x}\mathbf{x}}(\mathbf{x}_{t-1})$  and  $\mathbf{y}_t^0 = \mathbf{y}_t - f^{\mathbf{y}\mathbf{y}}(\mathbf{y}_{t-1})$ . Given the realized sequences  $\mathbf{y}(\omega)$  and  $\mathbf{x}(\omega)$  generated by eq. (7.2), the sequential system eq. (7.3) moves forward in time as the one-step ahead directly caused parts of  $\mathbf{y}$  and  $\mathbf{x}$  that are filtered from the reverberating effects of  $f^{\mathbf{x}\mathbf{x}}$  and  $f^{\mathbf{y}\mathbf{y}}$ . More specifically, while  $\mathbf{y}$  partially consists out of memory, there is a part  $\mathbf{y}^0$  that at any point is directly mapped from the previous state of  $\mathbf{x}$ , while at the same time  $\mathbf{x}$  consists partially out of memory and a part  $\mathbf{x}^0$  directly generated from the last position of  $\mathbf{y}$ . In this view, directional causality can be stated in terms of whether eq. (7.3) produces any values.

Importantly, the system also reveals that by the definitions of  $\mathbf{x}_t^0$  and  $\mathbf{y}_t^0$ , obtaining appropriate estimates for  $f^{\mathbf{x}\mathbf{y}}$  and  $f^{\mathbf{y}\mathbf{x}}$  involves  $f^{\mathbf{x}\mathbf{x}}$  and  $f^{\mathbf{y}\mathbf{y}}$  being modeled correctly as  $\mathbf{x}_t^0$  and  $\mathbf{y}_t^0$  are not observed and only result as functions from the observable processes  $\mathbf{y}$  and  $\mathbf{x}$ . Moreover, if  $\mathbf{y}(\omega)$  and  $\mathbf{x}(\omega)$  are triggered by an event, then it is possible by process of infinite backward substitution to write eq. (7.3) as an infinite chain initialized in the infinite past. Plugging in the equalities  $\mathbf{x}_t = \mathbf{x}_t^0 + f^{\mathbf{x}\mathbf{x}}(\mathbf{x}_{t-1})$  and  $\mathbf{y}_t = \mathbf{y}_t^0 + f^{\mathbf{y}\mathbf{y}}(\mathbf{y}_{t-1})$  and defining the random functions  $f_{\mathbf{y}}^0(\mathbf{y}_t^0, \mathbf{y}_{t-1}^0) = f^{\mathbf{x}\mathbf{y}}(\mathbf{y}_t^0 + f^{\mathbf{y}\mathbf{y}}(\mathbf{y}_{t-1}^0))$  and  $f_{\mathbf{x}}^0(\mathbf{x}_t^0, \mathbf{x}_{t-1}^0) = f^{\mathbf{y}\mathbf{x}}(\mathbf{x}_t^0 + f^{\mathbf{x}\mathbf{x}}(\mathbf{x}_{t-1}^0))$ , one can write

$$\begin{aligned}\mathbf{x}^0 &:= \{\mathbf{x}_t^0 = f_{\mathbf{y}}^0(\mathbf{y}_{t-1}^0, \mathbf{y}_{t-2}^0), t \in \mathbb{Z}\} \\ \mathbf{y}^0 &:= \{\mathbf{y}_t^0 = f_{\mathbf{x}}^0(\mathbf{x}_{t-1}^0, \mathbf{x}_{t-2}^0), t \in \mathbb{Z}\} \end{aligned} \quad (7.4)$$

Repeating infinitely, and extending infinitely in the direction  $T \rightarrow \infty$ ,

$$\begin{aligned}\mathbf{x}^0 &:= \{\mathbf{x}_{\infty}^0 = (f_{\mathbf{y}}^0)^{\infty}(\mathbf{y}_1^0, \mathbf{y}_1), t \in \mathbb{Z}\} \\ \mathbf{y}^0 &:= \{\mathbf{y}_{\infty}^0 = (f_{\mathbf{x}}^0)^{\infty}(\mathbf{x}_1^0, \mathbf{x}_1), t \in \mathbb{Z}\} \end{aligned} \quad (7.5)$$

$(f_{\mathbf{y}}^0)^{\infty}$  and  $(f_{\mathbf{x}}^0)^{\infty}$  are the maps that generate  $\mathbf{y}^0$  and  $\mathbf{x}^0$  infinitely after  $\mathbf{y}$



and  $\mathbf{x}$  have been generated into infinity. Subscript  $_1$  has been used here to mark the initialization points. This shows that  $\mathbf{x}^0$  can be written as a sequence of iterating random functions that are all defined on  $\mathbf{y}$ , and  $\mathbf{y}^0$  defined on  $\mathbf{x}$  in a similar way.<sup>3</sup> For ease of notation, let us write

$$\begin{aligned}\mathbf{x}^0 &:= \{\mathbf{x}_t^0 = \mathbf{f}_\mathbf{y}^0(\mathbf{y}_{-\infty:t}), t \in \mathbb{Z}\} \\ \mathbf{y}^0 &:= \{\mathbf{y}_t^0 = \mathbf{f}_\mathbf{x}^0(\mathbf{x}_{-\infty:t}), t \in \mathbb{Z}\}.\end{aligned}\tag{7.6}$$

where bold-faced  $\mathbf{f}^0$  is used to refer to the entire sequence of functions  $f^0$  up to  $t$ , starting in the infinite past  $t = -\infty$ . This highlights that generating the unobserved quantities,  $\mathbf{x}^0$  and  $\mathbf{y}^0$  from the observed quantities  $\mathbf{x}$  and  $\mathbf{y}$  by back substitution, eventually involves the unobserved quantities  $\mathbf{x}_1$  and  $\mathbf{y}_1$ . This means that some feasible form of approximation is needed.

Note first that  $\mathbf{f}_\mathbf{y}^0 : \mathcal{Y} \rightarrow \mathcal{X} \subseteq \mathbb{R}$  is a  $\mathfrak{B}(\mathcal{Y})/\mathfrak{B}(\mathcal{X})$ -measurable mapping, and  $\mathbf{f}_\mathbf{x}^0 : \mathcal{X} \rightarrow \mathcal{Y} \subseteq \mathbb{R}$  is a  $\mathfrak{B}(\mathcal{X})/\mathfrak{B}(\mathcal{Y})$ -measurable mapping. The sequence  $\mathbf{x}^0$  thus lives on  $(\mathcal{X}_\infty, \mathfrak{B}(\mathcal{X}_\infty), P_0^\mathbf{x})$ , where  $P_0^\mathbf{x}$  is induced according to  $P_0^\mathbf{x}(B_\mathbf{x}) = P_0^\mathbf{y} \circ (\mathbf{f}_\mathbf{y}^0)^{-1}(B_\mathbf{x}) \ \forall B_\mathbf{x} \in \mathfrak{B}(\mathcal{X}_\infty)$ , and  $\mathbf{y}^0$  lives on  $(\mathcal{Y}_\infty, \mathfrak{B}(\mathcal{Y}_\infty), P_0^\mathbf{y})$ , where  $P_0^\mathbf{y}$  is induced according to  $P_0^\mathbf{y}(B_\mathbf{y}) = P_0^\mathbf{x} \circ (\mathbf{f}_\mathbf{x}^0)^{-1}(B_\mathbf{y}) \ \forall B_\mathbf{y} \in \mathfrak{B}(\mathcal{Y}_\infty)$ . The notation shows that the probability measures underlying the stochastic causal sequences result from the functional behavior of the entire system. In particular, the causal sequences can be written as recursive direct effects, and the probability measures underlying the causal sequences are induced by the functional relationships that describe these dynamical dependencies.

In many cases, a researcher is not able to observe all the relevant variables. When a third, possibly unobserved external variable  $\mathbf{z}$  with effect  $f^\mathbf{z}(\mathbf{z})$ ,

---

<sup>3</sup>Equation (7.5) reveals an important implication for causality studies. The sequences that constitute the directly caused parts of  $\mathbf{x}$  and  $\mathbf{y}$  are ultimately dependent on the values at which the observable process has been initialized. That is, the entire causal pathway depends on the initial impact. In practice one cannot observe all impacts including those that occurred in the infinite past, and assurances are required that the initialization effect on the causal pathway must eventually not matter given sufficient observations. This is central to contraction studies.

is considered, the researcher is confronted with the situation that

$$\begin{aligned} \mathbf{x} &:= \{\mathbf{x}_t = f^{\mathbf{x}\mathbf{x}}(\mathbf{x}_{t-1}) + f^{\mathbf{x}\mathbf{y}}(\mathbf{y}_{t-1}) + f^{\mathbf{x}\mathbf{z}}(\mathbf{z}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y} &:= \{\mathbf{y}_t = f^{\mathbf{y}\mathbf{x}}(\mathbf{x}_{t-1}) + f^{\mathbf{y}\mathbf{y}}(\mathbf{y}_{t-1}) + f^{\mathbf{y}\mathbf{z}}(\mathbf{z}_{t-1}), t \in \mathbb{Z}\} \end{aligned} \quad (7.7)$$

If  $\mathbf{z}$  is unobserved, it can still be approximated as a difference combination of  $\mathbf{x}$  and  $\mathbf{y}$ . To obtain an approximated sequence of the *true*  $\mathbf{z}$  sequence to condition empirical counterparts for  $f^{\mathbf{x}\mathbf{z}}$  and  $f^{\mathbf{y}\mathbf{z}}$  on, one can work with:

$$\begin{aligned} \mathbf{z} &:= \{\mathbf{z}_t = (f^{\mathbf{x}\mathbf{z}})^{-1}(\mathbf{x}_{t+1} - (f^{\mathbf{x}\mathbf{x}}(\mathbf{x}_t) + f^{\mathbf{x}\mathbf{y}}(\mathbf{y}_t))), t \in \mathbb{Z}\} \\ \mathbf{z} &:= \{\mathbf{z}_t = (f^{\mathbf{y}\mathbf{z}})^{-1}(\mathbf{y}_{t+1} - (f^{\mathbf{y}\mathbf{x}}(\mathbf{x}_t) + f^{\mathbf{y}\mathbf{y}}(\mathbf{y}_t))), t \in \mathbb{Z}\} \end{aligned} \quad (7.8)$$

Equation (7.8) suggests to write eq. (7.7) in terms of  $\mathbf{y}$  and  $\mathbf{x}$  only by defining  $\mathbf{z}$  as a difference combination of  $\mathbf{x}$  and  $\mathbf{y}$ .<sup>4</sup> This allows us to define the spaces and measures on which the multivariate process operates in terms of  $\mathbf{x}$  and  $\mathbf{y}$  only even in the presence of  $\mathbf{z}$ . If the process is invertible, one can simply write:<sup>5</sup>

$$\begin{aligned} \mathbf{x} &:= \{\mathbf{x}_t = f^{\mathbf{x}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y} &:= \{\mathbf{y}_t = f^{\mathbf{y}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}), t \in \mathbb{Z}\} \end{aligned} \quad (7.9)$$

For every  $t \in \mathbb{Z}$ , the map  $f^{\mathbf{x}} \circ (\mathbf{y}_{t-1}, \mathbf{x}_{t-1}) : \Omega \rightarrow \mathcal{Y}$  is  $\mathcal{F}/\mathfrak{B}(\mathcal{Y} \times \mathcal{X})$ -measurable and  $\mathbf{y}(\omega)$  lives on the space  $(\mathcal{Y}_\infty, \mathfrak{B}(\mathcal{Y}_\infty \times \mathcal{X}_\infty), P^{\mathbf{y}})$  where the probability measure  $P^{\mathbf{y}}$  is induced by  $f^{\mathbf{x}}$  on  $\mathfrak{B}(\mathcal{Y}_\infty \times \mathcal{X}_\infty)$  according to the point-wise application of  $P^{\mathbf{x}}$  and the inverse of  $f^{\mathbf{x}}$ .<sup>6</sup> Similar arguments follow for  $f^{\mathbf{y}}$ . This tells us that in the multivariate case with possibly unobserved variables, the probability measures underlying the individual

<sup>4</sup>Apart from stability conditions on the endogenous process, one requires also that the exogenous impacts enter the system in some suitable manner such that  $(f^{\mathbf{y}\mathbf{z}})^{-1}$  and  $(f^{\mathbf{x}\mathbf{z}})^{-1}$  are absolute summable. Following the same arguments that resulted in eq. (7.5), the initialization of the exogenous impacts  $\mathbf{z}_1$  should similarly not carry information influential in the empirical estimates of  $f^{\mathbf{x}\mathbf{y}}$  and  $f^{\mathbf{y}\mathbf{x}}$  conditional on partial information.

<sup>5</sup>By aggregating the functions

$$\begin{aligned} \mathbf{x} &:= \{\mathbf{x}_t = f^{\mathbf{x}\mathbf{x}}(\mathbf{x}_{t-1}) + f^{\mathbf{x}\mathbf{y}}(\mathbf{y}_{t-1}) + f^{\mathbf{x}\mathbf{z}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}), t \in \mathbb{Z}\} \\ \mathbf{y} &:= \{\mathbf{y}_t = f^{\mathbf{y}\mathbf{x}}(\mathbf{x}_{t-1}) + f^{\mathbf{y}\mathbf{y}}(\mathbf{y}_{t-1}) + f^{\mathbf{y}\mathbf{z}}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}), t \in \mathbb{Z}\} \end{aligned}$$

<sup>6</sup> $P^{\mathbf{y}}(B_{\mathbf{y}} \times B_{\mathbf{x}}) = P^{\mathbf{x}} \circ (f^{\mathbf{x}})^{-1}(B_{\mathbf{y}} \times B_{\mathbf{x}}) \forall (B_{\mathbf{y}} \times B_{\mathbf{x}}) \in \mathfrak{B}(\mathcal{Y}_\infty \times \mathcal{X}_\infty)$ .

sequences are possibly intertwined with those of the other sequences. This strongly complicates candidates and studies for the probability measure  $P^{\mathbf{w}}$  that underlies the joint process  $\mathbf{w} := \{\mathbf{w}_t = (\mathbf{y}_t, \mathbf{x}_t), t \in \mathbb{Z}\}$  operating on  $(\mathcal{W}_\infty, \mathfrak{B}(\mathcal{W}_\infty), P^{\mathbf{w}})$ .<sup>7</sup>

Nevertheless, when the correct invertible filters for all the time dynamics of the observed part of the system are specified, one can still rewrite general systems of the form eq. (7.7) into a representation that follows eq. (7.6). One can thus always state causality conditions relevant for correct inference, based on the subsystems that produce the directly caused effects eq. (7.6). In particular, one can keep the focus on  $P_0^{\mathbf{x}}$  and  $P_0^{\mathbf{y}}$ , bearing in mind that they are lower-level components of  $P^{\mathbf{w}}$  that defines the complete estimation objective.

DEFINITION. 5 (Non-causality). *The stochastic sequences  $\mathbf{x}(\omega)$  and  $\mathbf{y}(\omega)$  are not causality related if  $P_0^{\mathbf{x}}$  and  $P_0^{\mathbf{y}}$  are null measures, such that  $\mathbf{x}^0(\omega) \in \emptyset \forall (\omega, t) \in \Omega \times \mathbb{Z}$  and  $\mathbf{y}^0(\omega) \in \emptyset \forall (\omega, t) \in \Omega \times \mathbb{Z}$ .*

DEFINITION. 6 (Uni-directional Causality). *Causality runs uni-directionally from the stochastic sequence  $\mathbf{x}(\omega)$  to another stochastic sequence  $\mathbf{y}(\omega)$  (visa versa), if  $P_0^{\mathbf{x}}$  is a null measure, and  $P_0^{\mathbf{y}}$  is a non-null measure, such that  $\mathbf{x}^0(\omega) \in \emptyset \forall (\omega, t) \in \Omega \times \mathbb{Z}$  and  $\mathbf{y}^0(\omega) \in \mathcal{Y} \forall (\omega, t) \in \Omega \times \mathbb{Z}$  (visa versa).*

DEFINITION. 7 (Bi-directional Causality). *The stochastic sequence  $\mathbf{x}(\omega)$  is causal with respect to  $\mathbf{y}(\omega)$  and  $\mathbf{y}(\omega)$  is causal with respect to  $\mathbf{x}(\omega)$ , if  $P_0^{\mathbf{x}}$  and  $P_0^{\mathbf{y}}$  are both non-null measures, such that  $\mathbf{x}^0(\omega) \in \mathcal{X} \forall (\omega, t) \in \Omega \times \mathbb{Z}$  and  $\mathbf{y}^0(\omega) \in \mathcal{Y} \forall (\omega, t) \in \Omega \times \mathbb{Z}$ .*

With null-measures, it is meant that the stochastic sequence describing the directly caused effects from one variable to the other takes values in the emptyset with probability 1. This is because the functions that induce the probability measure cancel out, hence they can be removed from the

<sup>7</sup>The sequence is more complicated, and realizes under the events  $\omega \in \Omega$ ,  $\mathbf{w}_t(\omega) \in \mathcal{W}$ , where  $\mathcal{W} := \mathcal{Y} \times \mathcal{X}$  and  $\mathbf{w}(\omega) \in \mathcal{W}_\infty$ , with  $\mathcal{W}_\infty := \mathcal{Y}_\infty \times \mathcal{X}_\infty \subseteq \mathbb{R}_\infty^{n_{\mathbf{x}}+n_{\mathbf{y}}} := \times_{t=-\infty}^{t=\infty} \mathbb{R}^{n_{\mathbf{x}}+n_{\mathbf{y}}}$ , and the probability measure of the joint process  $P^{\mathbf{w}}$  is thus defined on the product  $\sigma$ -algebra  $\mathfrak{B}(\mathcal{W}_\infty) = \mathfrak{B}(\mathcal{X}_\infty \times \mathcal{Y}_\infty) = \mathfrak{B}(\mathcal{X}_\infty) \otimes \mathfrak{B}(\mathcal{Y}_\infty) := \mathcal{W}_\infty \cap \mathfrak{B}(\mathbb{R}_\infty^{n_{\mathbf{x}}+n_{\mathbf{y}}})$  (see, Dudley (2002) p119.).

equations resulting in a probability measure that is not induced by any remaining rule or relationship. Respectively, conditioning on impacts in  $\mathbf{x}$ , these probabilistic causality definitions can thus be understood as:

1. Whenever an intervention in  $\mathbf{x}$  occurs, there is no chance that  $\mathbf{y}$  reacts as a result of that.
2. Whenever an intervention in  $\mathbf{x}$  occurs, there is positive chance that  $\mathbf{y}$  reacts as a result of that.
3. Whenever an intervention in  $\mathbf{x}$  occurs, there is positive chance that  $\mathbf{y}$  reacts as a result of that. Subsequently there is positive chance that  $\mathbf{x}$  reacts to this initial reaction, a probabilistic process that repeats recursively.

### 7.3 Limit divergence on the space of modeled probability measures

The definitions of causality in terms of the lower-level components of  $P^{\mathbf{w}}$ , suggest that correct causal statements can be obtained empirically by extracting relevant counterparts to  $P_0^{\mathbf{x}}$  and  $P_0^{\mathbf{y}}$  from a relevant counterpart to  $P^{\mathbf{w}}$ , and investigating the stochastic sequences produced by these modeled measures. For such an approach to be of relevance in an empirical context, one must ensure that the concepts introduced, adequately transfer over from the *true* measure  $P^{\mathbf{w}}$  to a modeled measure  $P^{\hat{\mathbf{w}}}$ . The focus is therefore shifted towards detailing how  $P^{\hat{\mathbf{w}}}$  can be approximated as a minimally divergent measure relative to  $P^{\mathbf{w}}$ , and draw on Approximation Theory to construct equivalence around the *true* measure under an axiom of correct specification.

For some event  $\omega \in \Omega$ , a realized  $T$ -period sequence  $\mathbf{w}_T(\omega) := (\mathbf{y}_T(\omega), \mathbf{x}_T(\omega))$  consisting of sequences  $\{\mathbf{y}_t(\omega)\}_{t=1}^{t=T}$  and  $\{\mathbf{x}_t(\omega)\}_{t=1}^{t=T}$  can be observed. The *true* function  $f^{\mathbf{w}}$ , consists of our main functions of

interest  $f^{\mathbf{x}}$  and  $f^{\mathbf{y}}$  that in turn are composed of  $f^{\mathbf{xy}}$  and  $f^{\mathbf{yx}}$  that are of particular interest to the researcher focused on causality, but possibly also nonzero functions  $f^{\mathbf{xx}}$  and  $f^{\mathbf{yy}}$  that shape the responses of an initial causal effect. The exact properties are generally unknown to the observer, but one can design a parametrization mapping that learns the behavior of  $f^{\mathbf{x}}$  and  $f^{\mathbf{y}}$  when exposed to sufficient data. To learn from the data an approximation of  $f^{\mathbf{x}}$  and  $f^{\mathbf{y}}$ , one can postulate a model

$$\hat{\mathbf{w}} := \{\hat{\mathbf{w}}_t = f(\mathbf{w}_{t-1}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta, t \in \mathbb{Z}\}, \quad (7.10)$$

with  $f : \mathcal{W} \times \Theta \rightarrow \mathcal{W}$  as our postulated model function and  $\hat{\mathbf{w}}$  as the modeled data. In the context of parametric inference, the parameter space  $\Theta$  is trivially of finite dimensionality, but also in the nonparametric case, the vector  $\boldsymbol{\theta} \in \Theta$  indexes parametric models nested by the nonparametric model, each inducing its own probability measure, and  $\Theta$  indexes families of parametric models each inducing a space of parametric functions generated under  $\Theta$ . In this discussion the focus remains limited to parametric inference, hence a compact set of potential hypotheses is considered. The arguments are trivially extended to the nonparametric case, by focusing on a compact subset  $\Theta_s \subset \Theta$  of solutions.<sup>8</sup> For example, by using priors or penalties that discard  $\Theta \setminus \Theta_s$  such that any solution of the criterion necessarily falls within a compact subset space. Let  $f$  be  $\mathfrak{B}(\mathcal{W})$ -measurable  $\forall \boldsymbol{\theta} \in \Theta$  so that  $f(\mathbf{w}_t; \boldsymbol{\theta}) : \Omega \rightarrow \mathcal{W}$  is  $\mathcal{F}/\mathfrak{B}(\mathcal{W})$ -measurable  $\forall \boldsymbol{\theta} \in \Theta$  and  $t \in \mathbb{Z}$ .  $F_\Theta := \{f(\cdot; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$  is our space of parametric functions defined on  $\mathcal{W}$  generated under  $\Theta$  under the injective  $f_\mathcal{W} : \Theta \rightarrow F_\Theta(\mathcal{W})$  where  $f_\mathcal{W}(\boldsymbol{\theta}) := f(\cdot; \boldsymbol{\theta}) \in F_\Theta(\mathcal{W}) \forall \boldsymbol{\theta} \in \Theta$ . Under any *true* probability measure  $P^\mathbf{w}$ , every potential parameter vector included in the parameter space  $\boldsymbol{\theta} \in \Theta$  induces a probability measure  $P_\boldsymbol{\theta}^{\hat{\mathbf{w}}}$  indexed by  $\boldsymbol{\theta}$  on  $\mathfrak{B}(\mathcal{W}_\infty)$ , according to  $P_\boldsymbol{\theta}^{\hat{\mathbf{w}}}(B_\mathbf{w}) = P^\mathbf{w} \circ f^{-1}(B_\mathbf{w}, \boldsymbol{\theta}) \forall (B_\mathbf{w}, \boldsymbol{\theta}) \in \mathfrak{B}(\mathcal{W}_\infty \times \Theta)$ . Thus, for every potential parameter vector included in

---

<sup>8</sup>For example, by letting  $\Theta_s$  grow as  $T \rightarrow \infty$ , hence focusing on the case  $\Theta_{s1} \subset \Theta_{s2} \dots \subset \Theta_{s\infty} \subseteq \Theta$ , see for example Geman, Stuart; Hwang (1982).

the parameter space  $\boldsymbol{\theta} \in \Theta$ , there is a triplet  $(\mathcal{W}_\infty, \mathfrak{B}(\mathcal{W}_\infty), P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$  that describes the probability space of modeled data under  $\boldsymbol{\theta}$ . The triplet  $(\mathcal{W}_\infty, \mathfrak{B}(\mathcal{W}_\infty), P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$  is thus itself an element of the measure spaces indexed by  $\boldsymbol{\theta}$  across all  $\Theta$ . Given the *true* probability measure  $P^{\mathbf{w}}$  on  $\mathfrak{B}(\mathcal{W})$ , this process is summarized by a functional  $\mathfrak{P} : F_\Theta(\mathcal{W}) \rightarrow \mathcal{P}_\Theta^{\hat{\mathbf{w}}}$ , that maps elements from the space of parametric functions generated by the entire parameter space  $F_\Theta(\mathcal{W})$ , onto the space  $\mathcal{P}_\Theta^{\hat{\mathbf{w}}}$  of probability measures defined on the sets of  $\mathfrak{B}(\mathcal{W}_\infty)$  generated by  $\Theta$  through  $f(\cdot; \boldsymbol{\theta})$ .

Now,  $f^{\mathbf{w}}$  is generally not only unknown, but for a finite  $\Theta$  there is no guarantee that  $\exists \boldsymbol{\theta}_0 \in \Theta : P \circ f_{\mathcal{W}}(\boldsymbol{\theta}_0) = P^{\mathbf{w}}$ , implying that in many empirical applications one is concerned with the situation where  $P^{\mathbf{w}} \notin \mathcal{P}_\Theta^{\hat{\mathbf{w}}}$ . However, if  $\exists P^{\mathbf{w}} \in \mathcal{P}_\Theta^{\hat{\mathbf{w}}}$ , one can learn all about  $P^{\mathbf{w}}$ , by uncovering the properties of  $f$ , given a sufficient amount of observations is available.<sup>9</sup> Let

$$\hat{\boldsymbol{\theta}}_T := \arg \min_{\boldsymbol{\theta} \in \Theta} Q_T(\mathbf{w}_T; \boldsymbol{\theta}), \quad (7.11)$$

$\hat{\boldsymbol{\theta}}_T : \Omega \rightarrow \Theta$ , be the extremum estimate for  $\boldsymbol{\theta}_0$  as judged by the criterion  $Q_T : \mathcal{W}_T \times \Theta \rightarrow \mathbb{R}$ . Trivially,  $\mathcal{W}_T := \mathcal{Y}_T \times \mathcal{X}_T$  and  $\mathbf{w}_T(\omega) \in \mathcal{W}_T$ . To see that under correct specification it is possible to approximate the *true* function  $f^{\mathbf{w}}$  in terms of equivalence (in the sense of function equivalence Kolmogorov and Fomin (1975) p.288), one can write the criterion function also as a function of the *true* function and the postulated model  $Q_T(f^{\mathbf{w}}(\mathbf{w}_T), f(\mathbf{w}_T; \boldsymbol{\theta}))$  in which it is made use of the fact that  $f^{\mathbf{w}}(\mathbf{w}_T) := \{f^{\mathbf{w}}(\mathbf{w}_t)\}_{t=1}^T := \mathbf{w}_T$  and  $f(\mathbf{w}_T; \boldsymbol{\theta}) := \{f(\mathbf{w}_t; \boldsymbol{\theta})\}_{t=1}^T := \hat{\mathbf{w}}_T$ .

The discussion further evolves toward showing that the element in  $\mathcal{P}_\Theta^{\hat{\mathbf{w}}}$  that is closest to  $P^{\mathbf{w}}$ , minimizes a divergence metric that results from a transformation of the limit criterion that measures the divergence between the *true* density and the density implied by the model. It is important to again note that  $\mathcal{P}_\Theta^{\hat{\mathbf{w}}}$  is induced by the proposed candidates for  $P^{\mathbf{w}}$ .

<sup>9</sup>As discussed in literature on miss-specification, even when the axiom of correct specification is abandoned,  $f$  may converge to a function that produces the optimal conditional a density which may reveal important properties of  $f^{\mathbf{w}}$ .

Studies on causality thus rely on flexible model design as the researcher determines which hypotheses are considered in a study by exerting control over  $\Theta$ . Naturally if  $\Theta_1 \subset \Theta_2$ , then  $\Theta_2$  produces a larger  $\mathcal{P}_{\Theta_2}^{\hat{\mathbf{w}}} \supset \mathcal{P}_{\Theta_1}^{\hat{\mathbf{w}}}$ . This suggests that minimizing this divergence metric over a large as possible  $\mathcal{P}_{\Theta}^{\hat{\mathbf{w}}}$  results in selecting  $P^{\hat{\mathbf{w}}}$  at a point in  $\mathcal{P}_{\Theta}^{\hat{\mathbf{w}}}$  that attains equivalence to  $P^{\mathbf{w}}$  only when  $\Theta$  is large enough to produce a correctly specified hypothesis set. Note that the definition of  $F_{\Theta} := \{f(\cdot; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$  as our space of parametric functions generated under  $\Theta$ , under the injective  $f_{\mathcal{W}} : \Theta \rightarrow F_{\Theta}(\mathcal{W})$  and the functional  $\mathfrak{P} : F_{\Theta}(\mathcal{W}) \rightarrow \mathcal{P}_{\Theta}^{\hat{\mathbf{w}}}$  that induces the space of probability measures, is defined on the sample space  $\mathcal{W}$ . This highlights that the correct specification argument  $P^{\mathbf{w}} \in \mathcal{P}_{\Theta}^{\hat{\mathbf{w}}}$ , not only stresses flexible parametrization in the sense that parameterized dependencies can take on many values, but also in the sense of using correct data.<sup>10</sup> When little is known about  $f$ , one is thus not only concerned with flexibility in terms of the type of parametric functions generated under  $\Theta$ , but also the variables on which the modeled measures are defined. When these concerns are appropriately addressed, testing for causality is deciding based on the approximation  $P^{\hat{\mathbf{w}}}$  whether the best approximation of the *true* model suggests 1) that  $\mathbf{x}$  and  $\mathbf{y}$  live in isolation, 2) unidirectional causality, or 3) that  $P^{\mathbf{w}}$  produces feedback.

To turn this problem into a selection problem that can be solved by divergence minimization w.r.t. the *true* measure, first introduce the limit criterion by taking  $T \rightarrow \infty$  and working with the modeled data as the minimizer of the criterion. Specifically, let the limit criterion be  $Q_{\infty}(\boldsymbol{\theta}) := Q_T(f^{\mathbf{w}}(\mathbf{w}_T), f(\mathbf{w}_T; \arg \min_{\boldsymbol{\theta} \in \Theta} Q_T(\mathbf{w}_T; \boldsymbol{\theta})))$  evaluated at  $T \rightarrow \infty$  with  $Q_{\infty} : \Theta \rightarrow \mathbb{R}$  and  $Q_{\infty}(\boldsymbol{\theta}) = Q_{\infty}^{\mathcal{P}}(P^{\mathbf{w}}; P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}) \forall \boldsymbol{\theta} \in \Theta$  with the criterion  $Q_{\infty}(\boldsymbol{\theta}) = Q_{\infty}^{\mathcal{P}}$  as a measure of divergence  $d_{\mathcal{P}}$  on the

---

<sup>10</sup>Indeed, the potential parameters that would interact with data that is not used, are essentially treated as zero, so the focus on using correct data is implicitly already contained in the standard statements of correct specification that focus directly on the dimensions of  $\Theta$ . The distinction is nevertheless useful because nonparametric models are often popularized as methods to reduce misspecification bias as  $\Theta$  becomes infinite dimensional, but this does not imply that  $P^{\mathbf{w}} \in \mathcal{P}_{\Theta}^{\hat{\mathbf{w}}}$  if important data is missing.

*true* probability measure and the modeled measure. More specifically,  $d_{\mathcal{P}} \equiv Q_{\infty}^{\mathcal{P}} : \mathcal{P}_{\Theta}^{\mathbf{w}} \times \mathcal{P}_{\Theta}^{\mathbf{w}} \rightarrow \mathbb{R}_{\geq 0}$ . By definition of  $Q_{\infty}^{\mathcal{P}}$  as a divergence on the space that contains  $P^{\mathbf{w}}$  and  $P_{\theta}^{\mathbf{w}} \forall \theta \in \Theta$ , the element  $\theta_0$  is thus the minimizer of that divergence.

Moreover,  $\arg \min$  in the parameter sense,  $\arg \min$  in the function sense in terms of a divergence metric on the *true* function, and  $\arg \min$  in the measure sense in terms of a divergence metric on the *true* probability measure, are equivalent limits under the same consistency result. To see this, it is convenient to focus once more on the target and write  $\theta_0 = \arg \min_{\theta \in \Theta} Q_{\infty}^{\mathcal{P}} \equiv \arg \min_{\theta \in \Theta} Q_{\infty}^F(f^{\mathbf{w}}, f_{\mathcal{W}}(\theta))$ , with  $Q_{\infty}^F : F(\mathcal{W}) \times F(\mathcal{W}) \rightarrow \mathbb{R}_{\geq 0}$ , to make clear that the criterion establishes a divergence  $d_F$  on  $F(\mathcal{W}) \times F(\mathcal{W})$ , which is in turn induced by  $d_{\mathcal{P}}$  through  $\mathfrak{P}$  according to  $d_F(f^1, f^2) = d_{\mathcal{P}}(P(f^1), P(f^2)) \forall (f^1, f^2) \in F(\mathcal{W}) \times F(\mathcal{W})$ . This ensures that our statement on the probability measure is relevant under standard consistency results that are focused on the convergence of an estimated parameter vector toward  $\theta_0$ , while equivalently the Impulse Response Functions converge to the *true* IRF at  $\theta_0$ . This implies that deciding between DEFINITION. 5-DEFINITION. 7 can be read from the responses produced by the IRF that minimizes divergence w.r.t. the *true* IRF

Not necessarily, but convenient for a proof that holds easily in practical situations, is to assume existence of a strictly increasing function  $r : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$  that ensures existence of a transformation of the limit criterion into a metric,  $d_{\mathcal{P}}^* \equiv r \circ d_{\mathcal{P}}$ , with  $r$  being a continuously strictly increasing function. Under these assumptions a simple result follows. For convenience all assumptions are summarized in ASSUMPTION. 13.

ASSUMPTION. 13. *For a limit criterion  $Q_{\infty} : \Theta \rightarrow \mathbb{R}$  of the form  $Q_{\infty}(\theta) \equiv Q_{\infty}^{\mathcal{P}}(P^{\mathbf{w}}, P_{\theta}^{\mathbf{w}}) \forall \theta \in \Theta$ ,  $d_{\mathcal{P}} \equiv Q_{\infty}^{\mathcal{P}} : \mathcal{P}^{\mathbf{w}} \times \mathcal{P}^{\mathbf{w}} \rightarrow \mathbb{R}_{\geq 0}$  is a divergence. Assume there exists a continuous strictly increasing function  $r : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$  such that  $d_{\mathcal{P}}^* \equiv r \circ d_{\mathcal{P}}$  is a metric. The functional  $f_{\mathcal{W}} : \Theta \rightarrow F_{\Theta}(\mathcal{W})$  is injective and  $\theta_0 \in \Theta$ .*

PROPOSITION. 6. *Assume ASSUMPTION. 13, then the following are equiv-*



alent limits:

1.  $\boldsymbol{\theta}_0$ ,
2.  $\arg \min_{\boldsymbol{\theta} \in \Theta} Q_\infty(\boldsymbol{\theta})$ ,
3.  $\arg \min_{\boldsymbol{\theta} \in \Theta} d_F^*(f^{\mathbf{w}}, f^{\hat{\mathbf{w}}}(\cdot, \boldsymbol{\theta}))$ ,
4.  $\arg \min_{\boldsymbol{\theta} \in \Theta} Q_\infty^{\mathcal{P}}(P^{\mathbf{w}}, P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$ ,
5.  $\arg \min_{\boldsymbol{\theta} \in \Theta} d_{\mathcal{P}}^*(P^{\mathbf{w}}, P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$ .

REMARK. 6. *Dropping the axiom of correct specification implies  $\hat{\boldsymbol{\theta}}_\infty \neq \boldsymbol{\theta}_0$ , hence the equivalences of 3-5 are now w.r.t. item 2.*

The equivalences in PROPOSITION. 6 not only ensure that for a correctly specified model  $\exists \boldsymbol{\theta}_0 \in \Theta$ , the element  $\boldsymbol{\theta}_0$  results in functional equivalence between the model and the *true* model (item 3), but also in zero divergence between the probability measures  $P^{\mathbf{w}}$  and  $P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}$  (item 4). Moreover, it follows that at  $\boldsymbol{\theta}_0$ , the empirically estimated probability measure  $P^{\hat{\mathbf{w}}}$  is equivalent to  $P^{\mathbf{w}}$  in the sense that there is zero distance between the two (item 5).

REMARK. 7. *PROPOSITION. 6 is applicable to a large class of extremum estimators, even those not initially conceived as minimizers of distance. In particular it is often possible to find a divergence on the space of probability measures. For example, Method of Moments estimators are naturally defined in terms of features of the underlying probability measures. In section 7.4 we also shall give an example using Kullback-Leibler divergence for which penalized Likelihood is an estimator. In this case squared Hellinger distance can be shown to be a lower bound.*

COROLLARY. 6 now delivers that our definitions set on the *true* measures, transfer to modeled probability measures in the limit for correctly specified cases. It is well-known that standard consistency proofs apply also to approximate extremum estimators, therefore assuming additionally that  $\sup_{\boldsymbol{\theta} \in \Theta} |Q_T(\mathbf{w}_T; \boldsymbol{\theta}) - Q_\infty(\boldsymbol{\theta})| \rightarrow 0$  a.s., is sufficient for a consistency result together with uniqueness of  $\boldsymbol{\theta}_0$  within the compact hypothesis space  $\Theta$ .

This implies that our causality conditions on the *true* measures do not only transfer to the approximate in the limit, but also for large  $T$  under standard regularity conditions. Essentially this is the setting considered by White and Pettenuzzo (2014). Summarized:

COROLLARY. 6. *Given a true probability measure  $P^{\mathbf{w}}$ , and an equivalent modeled probability measure  $P^{\hat{\mathbf{w}}}$  in the sense that  $d_{P^{\hat{\mathbf{w}}}}^* = r \circ d_P(P^{\mathbf{w}}, P_{\theta}^{\hat{\mathbf{w}}}) \sim 0$ , there are four possibilities for causality:*

1. *There is no causation if  $P_0^{\hat{\mathbf{x}}}$  and  $P_0^{\hat{\mathbf{y}}}$  adhere to DEFINITION. 5.*
2.  *$\mathbf{x}$  causes  $\mathbf{y}$  if the probability measure  $P_0^{\hat{\mathbf{y}}}$  adheres to DEFINITION. 6.*
3.  *$\mathbf{y}$  causes  $\mathbf{x}$  if the probability measure  $P_0^{\hat{\mathbf{x}}}$  adheres to DEFINITION. 6.*
4. *There is bi-directional causality if  $P_0^{\hat{\mathbf{x}}}$  and  $P_0^{\hat{\mathbf{y}}}$  adhere to DEFINITION. 7.*

Finally, in the case of a miss-specified model, REMARK. 6 implies that the divergence between the optimal probability measure as judged by the criterion and the *true* probability measure attains a minimum at a strictly positive value  $d_{P^{\mathbf{w}}}^* = r \circ d_P(P^{\mathbf{w}}, \arg \min_{\theta \in \Theta} Q_{\infty}^P(P^{\mathbf{w}}, P_{\theta}^{\hat{\mathbf{w}}})) > 0$ . In this case, the quantity  $d_{P^{\mathbf{w}}}^*$  determines how “close” the empirical claim is to the *true* hypothesis about causality. While it is difficult to make strong claims about this quantity, it is evident that minimizing  $d_{P^{\mathbf{w}}}^*$  may involve widening  $\mathcal{P}_{\Theta}^{\hat{\mathbf{w}}}$  in the direction of  $P^{\mathbf{w}}$  by increasing the dimensionality of  $\Theta$  by allowing flexibility and investigating a wide range of data. Disregard the value of  $d_{P^{\hat{\mathbf{w}}}}^*$ , the following holds.

PROPOSITION. 7. *If  $\theta_0 \notin \Theta$ , then  $P^{\mathbf{w}} \notin \mathcal{P}_{\Theta}^{\hat{\mathbf{w}}}$ . However,  $\hat{\theta}_{\infty}$  is still the pseudo-true parameter that minimizes  $r \circ d_P(P^{\mathbf{w}}, P_{\theta}^{\hat{\mathbf{w}}})$  over  $\Theta$ . Therefore  $P^{\hat{\mathbf{w}}}$  is the probability measure minimally divergent from  $P^{\mathbf{w}}$  within  $\mathcal{P}_{\Theta}^{\hat{\mathbf{w}}}$ . As such it follows that from all the potential probability measures in  $\mathcal{P}_{\Theta}^{\hat{\mathbf{w}}}$ , the measure closest to  $P^{\mathbf{w}}$  is supportive of one out of 1 – 4 in COROLLARY. 6 based on the properties of  $P_0^{\hat{\mathbf{x}}}$  and  $P_0^{\hat{\mathbf{y}}}$  as the best approximations.  $P^{\hat{\mathbf{w}}}$  provides the best approximation of the true causal measure across all the hypotheses considered.*

This leads to the following collection of results.

COROLLARY. 7. *Given a true probability measure  $P^{\mathbf{w}}$ , and a non-equivalent, but pseudo-true modeled probability measure,  $P^{\hat{\mathbf{w}}}$ , in the sense that  $d_{P^{\mathbf{w}}}^* = r \circ d_P(P^{\mathbf{w}}, P_{\theta}^{\hat{\mathbf{w}}})$  has attained a non-zero minimum, there are four possible optimal hypotheses about causality as judged by the criterion:*

1. *There is no causation if  $P_0^{\hat{\mathbf{x}}}$  and  $P_0^{\hat{\mathbf{y}}}$  adhere to DEFINITION. 5.*
2.  *$\mathbf{x}$  causes  $\mathbf{y}$  if the probability measure  $P_0^{\hat{\mathbf{y}}}$  adheres to DEFINITION. 6.*
3.  *$\mathbf{y}$  causes  $\mathbf{x}$  if the probability measure  $P_0^{\hat{\mathbf{x}}}$  adheres to DEFINITION. 6.*
4. *There is bi-directional causality if  $P_0^{\hat{\mathbf{x}}}$  and  $P_0^{\hat{\mathbf{y}}}$  adhere to DEFINITION. 7.*

Respectively, conditioning on interventions in  $\mathbf{x}$ , the results can be understood as:

1. Whenever an intervention in  $\mathbf{x}$  occurs, our best hypothesis is that there is no chance that  $\mathbf{y}$  reacts as a result of that.
2. Whenever an intervention in  $\mathbf{x}$  occurs, our best hypothesis is that there is positive chance that  $\mathbf{y}$  reacts as a result of that.
3. Whenever an intervention in  $\mathbf{x}$  occurs, our best hypothesis is that there is positive chance that  $\mathbf{y}$  reacts as a result of that, and these interactions continue to repeat with positive probability.

## 7.4 Limit Squared Hellinger distance

Both COROLLARY. 6 and COROLLARY. 7 assume that an appropriate transformation of the limit criterion exists that provides us with a metric or norm. This assumption allows us to make use of the classical theorems on existence and uniqueness of best approximations that have been naturally obtained for metric, normed and inner product spaces (Cheney and Respass, 1982). While this retains simplicity of the argument, it also shows that a direct interpretation of COROLLARY. 6 and COROLLARY. 7

can be obtained within the framework of Maximum Likelihood. Let us first define our criterion as the Maximum Likelihood Estimator:

$$\arg \min_{\theta \in \Theta} Q_T(\mathbf{w}_T; \theta) := \arg \max_{\theta \in \Theta} \sum_{t=1}^T \ln p_t(\mathbf{w}_t | \theta). \quad (7.12)$$

Note that this is conform to the form

$$Q_\infty(\theta) := Q_T(f^{\mathbf{w}}(\mathbf{w}_T), f(\mathbf{w}_T; \arg \min_{\theta \in \Theta} Q_T(\mathbf{w}_T; \theta)))$$

with  $T \rightarrow \infty$  and  $Q_\infty : \Theta \rightarrow \mathbb{R}$ . It can be shown that under this definition with  $Q_\infty(\theta) = Q_\infty^{\mathcal{P}}(P^{\mathbf{w}}; P_{\hat{\theta}}^{\hat{\mathbf{w}}}) \forall \theta \in \Theta$  that the criterion  $Q_\infty(\theta) = Q_\infty^{\mathcal{P}}$  is a measure of divergence  $d_{\mathcal{P}}$  on the *true* probability measure and the modeled measure. Specifically, we can introduce a divergence  $d_{\mathcal{P}} \equiv Q_\infty^{\mathcal{P}} : \mathcal{P}^{\mathbf{w}} \times \mathcal{P}^{\hat{\mathbf{w}}} \rightarrow \mathbb{R}_{\geq 0}$  as follows. Let  $p^{\mathbf{w}}(\mathbf{w}_t | \theta_{\mathbf{w}})$  and  $p^{\hat{\mathbf{w}}}(\mathbf{w}_t | \theta_{\hat{\mathbf{w}}})$  be respectively the *true* density evaluated under the *true* parameter and a modeled density at  $\hat{\theta}$  evaluated under the estimated parameter, both at time  $t$ , with respect to the Lebesgue measure (such that they are simply probability density functions), then the following is a divergence from the true probability measure to the modeled probability measure (Kullback-Leibler divergence, see Kullback and Leibler (1951)):

$$\begin{aligned} & KL(P^{\mathbf{w}}(\mathbf{w} | \theta_{\mathbf{w}}) || P^{\hat{\mathbf{w}}}(\mathbf{w} | \theta_{\hat{\mathbf{w}}})) \\ &= \begin{cases} \int_{-\infty}^{\infty} p^{\mathbf{w}}(\mathbf{w} | \theta_{\mathbf{w}}) \ln \frac{p^{\mathbf{w}}(\mathbf{w} | \theta_{\mathbf{w}})}{p^{\hat{\mathbf{w}}}(\mathbf{w} | \theta_{\hat{\mathbf{w}}})} d\mathbf{w} & \forall p^{\mathbf{w}}(\mathbf{w} | \theta_{\mathbf{w}}) << p^{\hat{\mathbf{w}}}(\mathbf{w} | \theta_{\hat{\mathbf{w}}}) \\ \infty & \forall p^{\hat{\mathbf{w}}}(\mathbf{w} | \theta_{\hat{\mathbf{w}}}) >> p^{\mathbf{w}}(\mathbf{w} | \theta_{\mathbf{w}}) \end{cases}. \end{aligned} \quad (7.13)$$

Naturally,  $KL(P^{\mathbf{w}}(\mathbf{w} | \theta_{\mathbf{w}}) || P^{\hat{\mathbf{w}}}(\mathbf{w} | \theta_{\hat{\mathbf{w}}})) \geq 0$  with equality if and only if  $p^{\mathbf{w}}(\mathbf{w} | \theta_{\mathbf{w}}) = p^{\hat{\mathbf{w}}}(\mathbf{w} | \theta_{\hat{\mathbf{w}}})$  almost everywhere, i.e. when the probability measures are the same (this is known as Gibb's inequality and can be verified by applying Jensen's Inequality).

Kullback-Leibler divergence is not a distance metric as was used in COROLLARY. 6 and COROLLARY. 7 to establish equivalences by partition-

ing into classes of zero-distance points. In particular, it is asymmetric

$$KL(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})) \neq KL(P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})||P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})), \quad (7.14)$$

and the triangle inequality is also not satisfied. However, it has the product-density property

$$KL(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})) = \sum_t^T \ln KL(p_t^{\mathbf{w}}(\mathbf{w}_t|\boldsymbol{\theta}_{\mathbf{w}})||p_t^{\hat{\mathbf{w}}}(\mathbf{w}_t|\boldsymbol{\theta}_{\hat{\mathbf{w}}})), \quad (7.15)$$

for  $p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) = p_1^{\mathbf{w}}(\mathbf{w}_1|\boldsymbol{\theta}_{\mathbf{w}}) \cdot p_2^{\mathbf{w}}(\mathbf{w}_2|\boldsymbol{\theta}_{\mathbf{w}}) \dots p_T^{\mathbf{w}}(\mathbf{w}_T|\boldsymbol{\theta}_{\mathbf{w}})$ , and  $p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})$  defined similarly. Hence the MLE is an unbiased estimator of minimized Kullback-Leibler divergence:

$$\begin{aligned} \arg \min_{\boldsymbol{\theta} \in \Theta} Q_T(\mathbf{w}_T; \boldsymbol{\theta}) &:= \arg \max_{\boldsymbol{\theta} \in \Theta} \sum_{t=1}^T \ln \frac{p^{\mathbf{w}}(\mathbf{w}_t|\boldsymbol{\theta}_{\mathbf{w}})}{p^{\hat{\mathbf{w}}}(\mathbf{w}_t|\boldsymbol{\theta}_{\hat{\mathbf{w}}})} \\ &= \arg \min_{\boldsymbol{\theta} \in \Theta} KL(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})). \end{aligned} \quad (7.16)$$

Note that under standard assumptions, a Law of Large Numbers can be applied to obtain the convergence, hence by maximizing likelihood, we minimize Kullback-Leibler divergence. Now, we need to either find a continuously scaling function  $r$  to ensure that it also minimizes distance between the *true* measure and the modeled measure so that we may reach zero at  $d_{P^{\hat{\mathbf{w}}}}^* = r \circ d_P(P^{\mathbf{w}}, P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}) \sim 0$ . Alternatively, we find the distance metric directly. We argued above that Kullback-Leibler divergence is not a proper distance (in particular it is not symmetric and does not satisfy the triangle inequality). However, notably useful is specifying  $d_{P^{\hat{\mathbf{w}}}}^*$  directly as the Hellinger distance between a modeled probability measure and the true probability measure (Hellinger, 1909):

$$H(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}), P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})) = \sqrt{\frac{1}{2} \int \left( \sqrt{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})} - \sqrt{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})} \right)^2 d\mathbf{w}}. \quad (7.17)$$

Specifically, the squared Hellinger distance provides a lower bound for

the Kullback-Leibler divergence. Therefore, maximizing likelihood implies minimizing Kullback-Leibler divergence which implies minimizing Hellinger distance. This is easily seen by the following:

PROPOSITION. 8. *Squared Hellinger distance provides a lower bound to Kullback-Leibler divergence:*

$$\left( H(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})) \right)^2 \leq KL(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})||P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})).$$

We end this sections with some notes on practical considerations. Let  $L_T(\boldsymbol{\theta})$  denote the sample Log likelihood at  $\boldsymbol{\theta} \in \Theta$ . Naturally, if  $\Theta_s \subset \Theta$ , it follows that  $\mathcal{P}_{\Theta}^{\hat{\mathbf{w}}} \supset \mathcal{P}_{\Theta_s}^{\hat{\mathbf{w}}}$ . In the limit, this means that maximizing Likelihood, minimizes Hellinger distance over both  $\mathcal{P}_{\Theta}^{\hat{\mathbf{w}}}$  and  $\mathcal{P}_{\Theta_s}^{\hat{\mathbf{w}}}$ . Following COROLLARY. 6, if  $\boldsymbol{\theta} \in \Theta_s$ , this results in selecting  $P^{\hat{\mathbf{w}}}$  at a point in  $\mathcal{P}_{\Theta_s}^{\hat{\mathbf{w}}}$  that attains equivalence to  $P^{\mathbf{w}}$ . In practice, when finite data is used, two different points, one in  $\mathcal{P}_{\Theta}^{\hat{\mathbf{w}}} \setminus \mathcal{P}_{\Theta_s}^{\hat{\mathbf{w}}}$  and one in  $\mathcal{P}_{\Theta_s}^{\hat{\mathbf{w}}}$ , may be obtained because the finite sample Log Likelihoods  $L_T(\hat{\boldsymbol{\theta}}_{sT})$  and  $L_T(\hat{\boldsymbol{\theta}}_T)$  that are available are both asymptotically biased estimators of the expected Log Likelihood  $\mathbb{E}L_T(\boldsymbol{\theta}_0)$ . This is easily shown by using a simple quadratic expansion

$$\begin{aligned} & \lim_{T \rightarrow \infty} \mathbb{E} \left( L_T(\hat{\boldsymbol{\theta}}_T) - \mathbb{E}L_T(\boldsymbol{\theta}_0) \right) \\ &= \lim_{T \rightarrow \infty} \mathbb{E} \sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)' \frac{1}{T} L_T''(\boldsymbol{\theta}_T) \sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \neq 0. \end{aligned} \quad (7.18)$$

Under considerably restrictive conditions original work by Akaike (1973, 1974) showed that the right hand-side approaches the dimension of  $\hat{\boldsymbol{\theta}}_T$  and hence, an asymptotically unbiased estimator of  $\mathbb{E}\ell_t(\boldsymbol{\theta}_0)$  is given by  $\frac{1}{T} \sum_{t=2}^T \ell_t(\hat{\boldsymbol{\theta}}_T) - k$ . Akaike also proposed the well known AIC given by  $\text{AIC} = 2T(k - \frac{1}{T} \sum_{t=2}^T \ell_t(\hat{\boldsymbol{\theta}}_T))$ . Several authors have shown that the AIC can be used to consistently rank models according to Kullback-Leibler divergence in considerably more general settings including the mis-specified case and have suggested further finite sample improvements Hurvich and Tsai (1989, 1991); Sin and White (1996). The AIC is also valid to decide between economic theories for which no test statistics can

be found Granger et al. (1995).

This means that while maximizing Log Likelihood over  $\Theta$  is not the same objective as minimizing Kullback-Leibler divergence in finite samples, working with a complexity penalized Log Likelihood, i.e. minimizing the AIC, does select the model that attains the lowest  $KL$  bound of all considered models generated under  $\Theta$ . Hence, in practice, a researcher can minimize the AIC as the practical objective to minimize Hellinger distance, and use correct specification tests to decide whether COROLLARY. 6 or COROLLARY. 7 is relevant.

## 7.5 Concluding remarks

During the 20th century, probability theory and economic theory have been closely developed together. While empirical studies in economics rely heavily on probabilistic concepts for inference, definitions for causality are often viewed through a deterministic lens. This paper discussed a probabilistic view on causality. In this view, a theory about causality is seen as a statement about the properties of the *true* measure that describes an observed process stochastically. The correct economic theory thus concerns the *true* frequencies in Markov chains of iterated processes of causes and effects, in which the transitions from one phase to another are regulated by the *true* probability law. This *true* probability law has been used to define causality in terms of stochastic sequences of caused effects.

Some argue that similar system theoretic definitions of causality, most notably the one from Granger, are not causal in the sense that they do not provide economic insight in the origin of the *true* probability law, but rather describe (correctly) the probabilistic behavior of the outcome of a causal origin. Clearly, these definitional discussions lie outside the scope of the statistical framework used in an empirical setting and relate to the

structure of the research question itself. In fact, we have seen that the relation between the functional behavior of a system and the probability measure that regulates its transitions from one phase to another, can be made explicit such that the direct relationship between theorized functional behavior and the stochastic properties of data produced under that functional behavior, is easily established. This thus suggests that the critiquing views rather relate to disagreements around whether the functional behavior that is looked at in an application, is critically of interest to policy.

Apart from definitional issues, the distinction between “good” predictors and causal effects is another central part of discussion. In many cases, researchers do not accept an empirical result to be causal, but settle by agreeing that the relationship that is found constitutes a good predictor. From the point of view currently presented, it is not acceptable that a suboptimal predictor could in fact be a better candidate for the causal description of the mechanisms that produced the data. An empirical model of reality found by a distance-minimization process, attains the status of the one closest to the *true* model. Proofs that sample averages approach their infinite counterparts, are among the most fundamental results in probability theory. In practice there may be various violations to the required regularity conditions for the convergence of a criterion function, and attention must be paid to ensure that empirical models are constructed in an appropriate manner. The *true* probability measure, however, is by definition the optimal description of observed data sequences when tested infinitely many times against other ones, and doing away the result that is closest to this description as merely predictive, and not as one that is close to the causal origin of the observed data, seems therefore a flawed attack.

Still, economics has been criticized to not deliver on a number of important prediction problems, even though economists, disregard their differences



in views on causality, have paid important attention to uncover causal relationships in their analyses of economic systems. Some examples include not being able to accurately predict a downturn in markets or find a definitive answer to the relationship between employment and government expenditures. The argument that not observing an outcome that was predicted by a supposedly causal model, invalidates the causal claim, is naturally flawed as well. A prediction made with the correct probability measure of a dice is only correct in the frequency domain – e.g., one out of six for an ordinary dice. In a similar manner, we would say that stress and bad lifestyle habits cause increased risk of a heart attack, which is similarly a probabilistic statement that provides accurate predictions only in the frequency domain. The optimal, causal, predictor must hence always be understood as the predictor that minimizes distance between predicted probability of occurrence and the *true* future probabilistic occurrence, and those laws will only ever be correct within the frequency domain.

## Proofs

### Proof for Proposition 1.

*Proof.* By construction of the criterion as stated in ASSUMPTION. 13,  $\arg \min_{\theta \in \Theta} Q_{\infty}(\theta)$  is its minimizer, and by assuming  $\theta_0 \in \Theta$ , it is also equal to  $\theta_0$ . Hence, item 2 is equivalent to item 1 by definition under correct specification.

The equivalence of the deterministic limit criterion (item 2) as a function describing the divergence of the underlying probability measures of  $\mathbf{w}$  and  $\hat{\mathbf{w}}$  (item 4) is assumed, however, given a limit criterion function  $Q_{\infty} : \Theta \rightarrow \mathbb{R}$  and a flexible definition of divergence (e.g. a pre-metric such as the *KL*-divergence), it is often possible to find a divergence  $d_{\mathcal{P}} : \mathcal{P}_{\Theta} \times \mathcal{P}_{\Theta} \rightarrow \mathbb{R}_{\geq 0}$  on the space of probability measures satisfying  $\arg \min_{\theta \in \Theta} d_{\mathcal{P}}(\mathcal{P}^{\mathbf{w}}, \mathcal{P}_{\theta}^{\hat{\mathbf{w}}}) = \arg \min_{\theta \in \Theta} Q_{\infty}(\theta)$ . The *KL*-divergence example is provided in this paper in the context of the Maximum Likelihood criterion.

By the assumption that  $r$  exists, the deterministic limit criterion that minimizes divergence, is also the minimizer of a distance metric  $d_{\mathcal{P}}^*(P^{\mathbf{w}}, P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}})$ , hence item 4 is also equivalent to item 2.

Finally, since  $f_{\mathcal{W}} : \Theta \rightarrow F_{\Theta}(\mathcal{W})$  is injective,  $(P^{\mathbf{w}}, P_{\boldsymbol{\theta}}^{\hat{\mathbf{w}}}) \equiv d_F^*(f^{\mathbf{w}}, f(\cdot, \boldsymbol{\theta})) \forall \boldsymbol{\theta} \in \Theta$  and  $d_F^*$  is a metric on  $F_{\Theta}(\mathcal{W})$ ,  $\boldsymbol{\theta}_0$  is also the minimizer of  $d_F^*(f^{\mathbf{w}}, f(\cdot, \boldsymbol{\theta})) \forall \boldsymbol{\theta} \in \Theta$  providing that item 3 is equivalent to item 2.

□

### Proof for Proposition 2.

*Proof.* The result follows immediately by the arguments used in PROPOSITION. 6 dropping only the first equivalence. □

### Proof for proposition 3.

*Proof.* First, Hellinger distance is

$$H(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}), P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})) = \sqrt{\frac{1}{2} \int \left( \sqrt{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})} - \sqrt{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})} \right)^2 d\mathbf{w}},$$

hence,

$$\left( H(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}), P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})) \right)^2 = \frac{1}{2} \int \left( \sqrt{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})} - \sqrt{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})} \right)^2 d\mathbf{w}.$$

Now, the R.H.S. can be written as

$$\frac{1}{2} \int p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) d\mathbf{w} + \frac{1}{2} \int p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}}) d\mathbf{w} - \int \sqrt{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})} d\mathbf{w}.$$

The integral of a probability density over its domain equals 1, hence the sum of the first two terms is 1, hence this can be rewritten as

$$1 - \int \sqrt{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})} d\mathbf{w}.$$

This has an upper bound, provided by the inequality

$$1 - \int \sqrt{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})} d\mathbf{w} \leq -\ln \int \sqrt{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})} d\mathbf{w}.$$

Write R.H.S. as  $-\ln \int \left[ \sqrt{\frac{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}} p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) \right]$  and take expectations to get

$$\mathbb{E} - \ln \int \left[ \sqrt{\frac{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}} p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) \right] = -\ln \mathbb{E} \left[ \sqrt{\frac{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}} p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) \right].$$

Note that

$$-\ln \mathbb{E} \left[ \sqrt{\frac{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}} p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) \right] < -\mathbb{E} \left[ \ln \sqrt{\frac{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}} p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) \right],$$

by Jensen's inequality.

Finally,  $\mathbb{E} \left[ \ln \sqrt{\frac{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})}{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}} p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) \right]$  can be written as

$$E \left[ \ln \frac{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})} p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) \right],$$

where the last expression is equivalent to Kullback-Leibler divergence by an elementary row operation

$$E \left[ \ln \frac{p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}})}{p^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})} p^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) \right] \equiv KL(P^{\mathbf{w}}(\mathbf{w}|\boldsymbol{\theta}_{\mathbf{w}}) || P^{\hat{\mathbf{w}}}(\mathbf{w}|\boldsymbol{\theta}_{\hat{\mathbf{w}}})) .$$

□

# Chapter 8

## Conclusion

The models that researchers estimate are necessarily an idealization of a complex reality. Advances in our capacity to compute, along with continued increases in the dimensions of datasets, have enormously increased both the complexity of what we attempt to achieve in analysis and the models that we use to pursue those goals. The aim of the basic theory with which we opened the introduction of this thesis was to provide clearly formulated and generalizable interpretation to standard empirical results. Given the advances in data and complexity, it is clear that analysis must acknowledge that the models ideally estimated aim at achieving a greater degree of idealization than was held possible when the theory of linear estimation of a parameter from a modest numbers of observations was first developed. With the general Consistency and Normality results for  $M$ -estimators that were introduced, there was much more freedom to think about more complex models that might provide a better description of reality. This thesis was devoted to exploring dynamic spatial time series models that can provide a better fit to the data using minimal complexity.

Chapter 3 first characterized spatial heterogeneity. This was done from the perspective of the data generating process itself. Specifically, we used a spatial model based on an economic rationale and parametrized it based on estimates from the literature. This was used to simulate

likely economic outcomes at a grid-cell level. While inherently not a problem related to statistical inference in the way it was discussed in the introduction of this thesis, the analysis produced several useful insights. Specifically, we saw that by imposing simple linear relationships at a high resolution, aggregate system behavior tended to follow nonlinear patterns. This is important as, in reality, we tend to observe economic outcomes at a coarse scale while processes are arguably driven by the total sum of interactions between a large number of individual economic actors. Furthermore, we saw that the geophysical nature of our landscape plays an important role in economic processes. In particular, the natural organization in geological factors tends to contribute to spatial clustering, even when spatial interdependencies across various distances are not explicitly parameterized in the data generating process. This is also important, as it is easy to miss out on one or several unobserved common factors, that may follow this type of spatial organization, in empirical applications. This immediately implies that the residuals in simple cross-sectional regressions are likely to be spatially correlated and may follow structural patterns that vary by types of regimes. In the introduction of this thesis we had already emphasized the crucial role that neutralizing residuals plays in rendering the parameter distributions approximately normal.

In Chapter 4, we tackled the problem of spatial dependence in time series. Specifically, we specified the spatial autoregressive time series model discussed in the introduction of the thesis and studied it in more detail. Building on our notion that the linearity assumption may be too restrictive, especially as the spatial dimensions grow, we extended the model to allow the parameter that determines dependence between neighbors to vary across time and space in an idiosyncratic manner. This allows dependence to vary over different regimes that may be covered by the cross-sectional data. The model allowed each observation in the cross-section to have a different history of attraction to its neighbors and the

magnitude of the induced feedback effects to vary continuously over time. This type of dynamic behavior could not be understood under standard dynamic time series theory provided in the introduction. We therefore extended the theory to allow for dynamic multivariate time series and provided a general theory that allowed the nonlinear dynamics to become spatial. We applied the model to a short spatial time series of urban densities and saw that the linear spatial model was not able to handle both the urban and rural dynamics in a single framework, causing the model to severely underestimate urban densities and overestimate rural densities. These regime-specific dependencies could, however, correctly be captured by the nonlinear model, allowing to analyze transitory effects across both the urban, rural, and urban gradients in one single framework. We also applied the nonlinear model to a long financial time series, and saw that it was able to fit both periods of financial stability during which spatial dependence was flat and periods of financial unrest in which there was substantially stronger idiosyncratic behavior.

In Chapter 5 we dropped the parametric assumption, and worked in a non-parametric framework in which the exact form of the nonlinearities did not have to be assumed. Instead of modeling the dependence between spatial observations to describe clustering in the data endogenously, we allowed for the flexibility to let dependence on exogenous variables vary nonlinearly across levels in the data. This resulted in rich dependence structures in which individual observations are part of different spatial and temporal regimes, each having possibly unique relationships with the outcome variable. We learned that there are methods that can approximate any type of nonlinearities arbitrarily well, while the estimation problem could still be solved linearly. In particular, the Kernel model mapped the input to a higher dimensional feature space, from where linear relationships could be established with the outcome variable. The growing number of local parameters used in those type of approximation strategies, however, violate the standard compactness assumption intro-

duced in the introduction of the thesis that was used to obtain existence and measurability of the estimator. Hence, the uniform convergence that was obtained from point-wise convergence and stochastic equicontinuity on a compact parameter space was also lost. We saw that to estimate these models, it was necessary to regulate the size of the parameter space appropriately which ensured that there was sufficient data to support the degrees of freedom. The regularization method effectively ensured that the parameter space grew at an appropriate rate as the data grew. This delivered a type of consistency that had a different interpretation than what was discussed in the introduction of the thesis. In particular, the limit result depended on the user-defined tolerance for complexity, which was determined by a hyper-parameter that was not estimated by the criterion function itself. The appendix of this chapter discussed the implication of this external influence on the interpretation of the result and concluded that standard interpretation to the results is supported as long as the hyper-parameter was tuned by optimizing the criterion out-of-sample.

Chapter 6 moved away from the nonlinear world, and moved back into the linear one. In this chapter we focused on multivariate interactions between multiple spatial time-series. Naturally, once the asymptotic results for multivariate nonlinear time series models put forward in Chapter 3 and the penalization from Chapter 4 are understood, it is straightforward to apply these ideas together to the setting of multiple nonlinear spatial time-series. From a practical standpoint we, unfortunately, are still quite constrained by modern computing capacity to work with such complex descriptions of reality. Interesting linear dynamics between multiple spatial time series could still be modeled though, which admittedly already results in detailed dynamics at the observational level. In particular, the spatial spillover effects implied heterogeneous relationships at the local level, and the multiple variable setting thus allowed us to explore cause and effect between interrelated cross-sectional time series while taking into account

that different cross-sectional variables themselves exhibit spatial feedback between observations that result in heterogeneous local impacts after shocks occur. We saw that models that do not factor in the cross-sectional dependences were likely to over-estimate the temporal effects and provided a generally poorer fit to the data that violated the martingale difference sequence assumption imposed on the score. Finally, the chapter explored the kernel trick from Chapter 4 as a mechanism to generate data-driven spatial weight matrices. The analysis showed that appropriate network structures could be estimated using Maximum Likelihood. This allowed generalizing the spatial dependencies discussed in this thesis and apply them to settings in which cross-sectional dependencies arise because of economic similarities or through other non-geographic channels.

Finally, in Chapter 7 we moved back to our starting discussion around estimators, and to the notion of correct specification specifically. Only this time, we approached the topic from a more general angle. We reconsidered the basic idea of inference and considered why flexible models, such as the ones introduced in this thesis, are desirable tools for inference in the first place. While the assumption of correct specification surfaced many times in parts of this thesis, it is easy to admit that this is possibly the most difficult assumption of all. In Chapter 4 and 6 we made use of different strategies to verify whether our estimated models provide an appropriate fit to the data. Nevertheless, when formulating empirical models we naturally abstract from reality and work with a description that is only an approximation to a complex reality. While mis-specification is often accepted in practice, it should not be a reason to opt for simple approximations merely because it is difficult to describe reality in fullness and easy to acknowledge that a simple model does not appropriately reflect that fullness. Particularly, when a result is taken as causal and representative of the real world, then that statement must reflect a belief that reality could be produced by a model that is reasonably similar to the estimated one. This means that if one is



interested in making causal statements, then the estimated model used to build the arguments should at least be able to produce dynamics that we believe are relevant in the real world. In particular, the Stationarity and Ergodicity of the data introduced as an assumption in the introduction if this thesis must come from the model itself. If one is willing to verify all the stability conditions of the possibly complex analyzed dynamics, as we did in Chapter 3, then one must also be ensured that the empirical strategy that is followed inherently ensures that the estimator finds the correct causal structure. Critical here is that increasing model complexity leads to a higher number of parameters, hence an increased overall model uncertainty. We discussed approximation of causal structures in more detail and provided an argument that minimizing complexity penalized criteria such as the AIC, as we did in Chapters 4 and 6, is the right objective in empirical settings.

## 8.1 Final remarks

With the theory and methods introduced in this thesis, researchers can now estimate a wide range of flexible models that take into account possible heterogeneity in dependencies across time and space. While there are many thoroughly developed options for analysis of spatial time series data, there are still many possible other research methodologies left to cover. A few directions for future research are the following.

First, the applications in thesis focused primarily on modeling conditional mean sequences, possibly with observation-driven nonlinear dynamics. The notions put forward in this work can easily be extended to higher moments. For example, the nonlinear dynamics explored in the context of the smooth transition spatial autoregressive model could be extended to allow for nonlinear cross-sectional dependence in multivariate GARCH models to allow instantaneous transmission of volatility spillovers in an

asymmetric way. This is particularly relevant when one is interested in understanding risk by means of numerically calculating Value-at-Risk or Expected Shortfall for a collection of interrelated investments using stochastic simulations. Basic univariate threshold GARCH models have already been developed to incorporate simple regime switching behavior into volatility regressions, but the standard application is one of instantaneous switching between linear autoregressive regimes. Spatial GARCH models have also been developed to allow for linear instantaneous dependence in processes that share AR and GARCH parameters. The obvious drawback is that, while financial assets may exhibit feedback, particularly when markets crash or surge, they may be assumed to follow individual temporal dynamics. From that perspective, Generalized Orthogonal GARCH is a useful model as it allows one to parameterize interactions in the conditional mean sequence using a VAR structure, while also allowing for volatility spillovers in a multivariate GARCH equation. The GO-GARCH spillovers are, however, not instantaneous. Instead, they lag over time. Given that these various models are already available, a generalization of multivariate GARCH, spatial GARCH and the threshold dynamics, seems within reach of the practitioner. The resulting nonlinear spatial dependence in conditional mean and conditional variance, together with VAR parameters, would provide a framework in which one can analyze shocks that travel through a system, both in regimes that are dominated by commonalities or idiosyncrasies. Second, not all the world's phenomena can be described with continuous data. Future research may focus on extensions relevant to model categorical, ordinal and count data that are collected sequentially over time at possibly dependent locations. This may require assuming distributions of a different type than those assumed in the theory developed here. For example the Poisson distribution would be the starting point for basic count series, and a Poisson mixture like the negative binomial distribution could be the starting point to tackle zero-inflation. Mixture models that

involve multiple distributions can also be used to combine both the characteristics of continuous process and those of count process jointly in a time series. For example, the jump-diffusion model combines continuous Brownian motion paths from Gaussian log returns with discontinuities, or jumps, that are drawn from a compound Poisson process. Generalization of jump-diffusion to the spatial time series setting may be interesting, but possibly they will have to wait until spatial multivariate volatility models are better understood. The development could be particularly challenging because jumps may occur simultaneously in a spatial time series, but the magnitude of jumps may differ over the cross-section while the assimilation of these jumps into the series may also happen partly in an idiosyncratic manner.

Third, the state-space framework, in particular the Kalman filter, has been extremely important in time series analysis and much work can be done to integrate the idea of cross-sectional nonlinearity and spatial dependence into this framework. This may be a particularly interesting direction for further advancement when one deals with processes that are only partially observed or measured with possible error. The smoothing framework could be particularly helpful to develop nonlinear interpolations for spatial time series that are intermittently observed. Ultimately, this seems to be an unavoidable problem for which tools will be needed. If we assume that local data gathering processes operate and report back information independently from one another, then logically it becomes likely that there will be local series in close proximity of one another that overlap mildly at best when one starts to track more regions in an economic system. A basic example would be a survey program in which households in different areas report back on local market prices whenever they buy goods. The challenge of constructing a continuous spatial time series will then have to deal with missing observations in space and time. While this seems an advanced application, the problems are relevant to key policy indicators that have been gathered for a long time already. Currently,

typical large survey programs such as those carried out by institutions like the World Bank, carry on for weeks or possibly months. During that time, seasons and economic circumstances may change. While the surveys thus actually represent a partially complete spatial time series, key statistics are often derived from them in the form of a single complete cross-section of data. The standard approach that many follow is to simply ignore away temporal changes assuming that they are randomly distributed over the survey program, and use the surveys to construct a single figure relevant for, say, the year. Often, one can find footnotes in reports and papers acknowledging that the underlying micro-data may have been gathered at different times. Performing a proper spatial time series interpolation before collapsing the data to a certain point in time would likely result in much more accurate estimates.

As we continue to develop theory for those complex settings, our datasets continue to grow increasingly rich, and the advances in our capacity to compute continue to accelerate, we may be able to model real-world processes in an increasingly accurate manner. The models we may use to approximate complex realities then become increasingly complex as well. We must therefore never forget the foundation on which we built. While we may achieve a greater degree of idealization than was ever held possible, the elegance of simple models was that they dealt with a modest numbers of parameters to summarize a complex world in a clearly formulated, tractable, an generalized fashion. Sometimes this is enough.



# Bibliography

Agriculture and Agri-Food Canada (2003). The Greenhouse Gas Mitigation Program for Canadian agriculture (GHGMP). Technical report, Agriculture and Agri-Food Canada.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Technical report.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Aldrich, J. (2010). The Econometricians’ Statisticians, 1895-1945. *History of Political Economy*, 42(1):111–154.

Alig, R., Stewart, S., Wear, D., Stein, S., and Nowak, D. (2010). Forest Land Conversion and Recent Trends. *Advances in Threat Assessment and Their Application to Forest and Rangeland Management*.

Alig, R. J. and Plantinga, A. J. (2004). Future forestland area: impacts from population growth and other factors that affect land values. *Journal of Forestry*. 102(8): 19-24.

Amisano, G. and Giacomini, R. (2007). Comparing Density Forecasts via Weighted Likelihood Ratio Tests. *Journal of Business & Economic Statistics*, 25(2):177–190.

Andree, B. P. J. (2019). Probability, Causality and Stochastic Formulations of Economic Theory. *SSRN Electronic Journal*.

- Andree, B. P. J., Blasques, F., and Koomen, E. (2017a). Smooth Transition Spatial Autoregressive Models. *Tinbergen Institute Discussion Papers*.
- Andree, B. P. J., Chamorro, A., Spencer, P., and Dogo, H. (2019). Environment and Development. *World Bank Policy Research Working Papers*, WPS8756.
- Andrée, B. P. J., Chamorro, A., Spencer, P., Koomen, E., and Dogo, H. (2019). Revisiting the relation between economic growth and the environment; a global assessment of deforestation, pollution and carbon emission. *Renewable and Sustainable Energy Reviews*, 114:109221.
- Andree, B. P. J., Diogo, V., and Koomen, E. (2017b). Efficiency of second-generation biofuel crop subsidy schemes: Spatial heterogeneity and policy design. *Renewable and Sustainable Energy Reviews*, 67:848–862.
- Andree, B. P. J., Spencer, P., Chamorro, A., Wang, D., Azari, S. F., and Dogo, H. (2019). Pollution and Expenditures in a Penalized Vector Spatial Autoregressive Time Series Model with Data-Driven Networks. *World Bank Policy Research Working Papers*, WPS8757.
- Anselin, L. (1988). *Spatial Econometrics: Methods and Models*.
- Anselin, L. (1995). Local Indicators of Spatial Association-LISA. *Geographical Analysis*, 27(2):93–115.
- Antle, J., Capalbo, S., Mooney, S., Elliott, E., and Paustian, K. (2003). Spatial heterogeneity, contract design, and the efficiency of carbon sequestration policies for agriculture. *Journal of Environmental Economics and Management*, 46(2):231–250.
- Apergis, N. (2016). Environmental Kuznets curves: New evidence on both panel and country-level CO2 emissions. *Energy Economics*, 54:263–271.
- Apergis, N. and Ozturk, I. (2015). Testing Environmental Kuznets Curve hypothesis in Asian countries. *Ecological Indicators*, 52:16–22.

- Audigier, V., White, I. R., Jolani, S., Debray, T. P. A., Quartagno, M., Carpenter, J., van Buuren, S., and Resche-Rigon, M. (2018). Multiple Imputation for Multilevel Data with Continuous and Binary Variables. *Statistical Science*, 33(2):160–183.
- Awaworyi Churchill, S., Inekwe, J., Ivanovski, K., and Smyth, R. (2018). The Environmental Kuznets Curve in the OECD: 1870–2014. *Energy Economics*, 75:389–399.
- Babcock, B. a., Lakshminarayan, P. G., and Wu, J. (1996). The Economics of a Public Fund for Environmental Amenities: A Study of CRP Contracts. *American Journal of Agricultural Economics*, 78(November):961–971.
- Bakam, I., Balana, B. B., and Matthews, R. (2012). Cost-effectiveness analysis of policy instruments for greenhouse gas emission mitigation in the agricultural sector. *Journal of Environmental Management*, 112:33–44.
- Baltagi, B. H., Fingleton, B., and Pirotte, A. (2014). Spatial lag models with nested random effects: An instrumental variable procedure with an application to English house prices. *Journal of Urban Economics*, 80:76–86.
- Bao, Y., Lee, T.-H., and Saltoglu, B. (2007). Comparing density forecast models. *Journal of Forecasting*, 26(3):203–225.
- Bao, Y. and Ullah, A. (2007). Finite sample properties of maximum likelihood estimator in spatial models. *Journal of Econometrics*, 137(2):396–413.
- Basile, R., Durbán, M., Mínguez, R., María Montero, J., and Mur, J. (2014). Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and nonlinearities. *Journal of Economic Dynamics and Control*, 48:229–245.



- Bates, C. and White, H. (1985). A Unified Theory of Consistent Estimation for Parametric Models. *Econometric Theory*, 1(02):151–178.
- Batidzirai, B., Faaij, A. P. C., and Smeets, E. (2006). Biomass and bioenergy supply from Mozambique. *Energy for Sustainable Development*, 10(1):54–81.
- Beenstock, M. and Felsenstein, D. (2007). Spatial Economic Analysis Spatial Vector Autoregressions Spatial Vector Autoregressions. *Spatial Economic Analysis*, 2(2):167–196.
- Beenstock, M. and Felsenstein, D. (2019). *The Econometric Analysis of Non-Stationary Spatial Panel Data*. Advances in Spatial Science. Springer International Publishing, Cham.
- Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83.
- Billingsley, P. (1961). The Lindeberg-Levy Theorem for Martingales. *Proceedings of the American Mathematical Society*, 12(5):788.
- Billingsley, P. (1995). *Probability and Measure*. 3rd edition.
- Bjerkholt, O. (2007). Writing “The Probability Approach” With Nowhere To Go: Haavelmo in the United States, 1939–1944. *Econometric Theory*, 23(05):775.
- Blasques, F. (2010). Identifiable Uniqueness Conditions for a Large Class of Extremum Estimators. In *ETA International Symposium on Econometric Theory and Applications*, Singapore.
- Blasques, F. and Duplinskiy, A. (2018). Penalized indirect inference. *Journal of Econometrics*, 205(1):34–54.
- Blasques, F., Gorgi, P., Koopman, S. J., and Wintenberger, O. (2018). Feasible invertibility conditions and maximum likelihood estimation for

- observation-driven models. *Electronic Journal of Statistics*, 12(1):1019–1052 Blasques, F., Gorgi, P., Koopman, S. J.,.
- Blasques, F., Koopman, S. J., Lucas, A., and Schaumburg, J. (2016). Spillover dynamics for systemic risk measurement using spatial financial time series models. *Journal of Econometrics*, 195(2):211–223.
- Bo, S. (2011). A Literature Survey on Environmental Kuznets Curve. *Energy Procedia*, 5:1322–1325.
- Bohm, D. (1999). *Causality and chance in modern physics*. University of Pennsylvania Press.
- Bonhomme, S. and Manresa, E. (2015). Grouped Patterns of Heterogeneity in Panel Data. *Econometrica*, 83(3):1147–1184.
- Boudjellaba, H., Dufour, J.-M., and Roy, R. (1992). Testing Causality Between Two Vectors in Multivariate Autoregressive Moving Average Models. *Journal of the American Statistical Association*, 87(420):1082.
- Bouwman, A. F., van der Hoek, K. W., Eickhout, B., and Soenario, I. (2005). Exploring changes in world ruminant production systems. *Agricultural Systems*, 84(2):121–153.
- Brock, W. A. and Taylor, M. S. (2010). The Green Solow model. *Journal of Economic Growth*, 15(2):127–153.
- Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer.
- Brockwell, P. J. and Davis, R. A. (2002). *Time series : theory and methods*.
- Brosse, N., Dufour, A., Meng, X., Sun, Q., and Ragauskas, A. (2012). Miscanthus: A fast-growing crop for biofuels and chemicals production.

- Brown, L. D. and Levine, M. (2007). Variance Estimation In Nonparametric Regression Via The Difference Sequence Method. *The Annals of Statistics*, 35(5):2219–2232.
- Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33:261–304.
- Burton, E. and Sanjour, W. (1967). An Economic Analysis of the Control of Sulphur Oxides Air Pollution. Technical report, Washington, DC.
- Caceres, C., Carrière-Swallow, Y., Demir, I., and Gruss, B. (2016). U.S. Monetary Policy Normalization and Global Interest Rates. *IMF Working Paper*.
- Carvalho, D. and Fidora, M. (2015). Capital inflows and euro area long-term interest rates. *Journal of International Money and Finance*, 54:186–204.
- Caslin, B., Finnan, J., and Johnston, C. (2015). Miscanthus Best Practice Guidelines. Technical report.
- Cheney, E. and Respass, J. (1982). *Best approximation problems in tensor-product spaces*, volume 102. Pacific Journal of Mathematics, etc.].
- Cho, S.-H., Lambert, D. M., and Chen, Z. (2010). Geographically weighted regression bandwidth selection and spatial autocorrelation: an empirical example using Chinese agriculture data. *Applied Economics Letters*, 17(8):767–772.
- Clements, M. P., Franses, P. H., Smith, J., and van Dijk, D. (2003). On SETAR non-linearity and forecasting. *Journal of Forecasting*, 22(5):359–375.

- Cliff, A. and Ord, J. (1969). The Problem of Spatial Autocorrelation. In A. J. Scott, editor, *In London Papers in Regional Science*, pages 25–55. Pion, London.
- Cliff, A. and Ord, K. (1972). Testing for Spatial Autocorrelation Among Regression Residuals. *Geographical Analysis*.
- Cline, D. B. H. and Pu, H.-M. H. (1998). Verifying irreducibility and continuity of a nonlinear time series. *Statistics & Probability Letters*, 40:139–148.
- Cline, D. B. H. and Pu, H.-M. H. (1999). Geometric Ergodicity of Nonlinear Time Series. *Statistica Sinica*, 9:1103–1118.
- College of Agriculture Food and Rural Enterprise (2005). Grass Challenge for Dairy Farmers, Challenge Note 2A – Grass Budgeting. Technical report, Department of Agriculture and Rural Development, Northern Ireland.
- Costanza, R., Wainger, L., Folke, C., and Mäler, K.-G. (1993). Modeling Complex Ecological Economic Systems. *BioScience*, 43(8):545–555.
- Covey, T. and Bessler, D. A. (1992). Testing for Granger’s Full Causality. *The Review of Economics and Statistics*, 74(1):146.
- Coyle, D. (2014). *GDP: a brief but affectionate history*. Princeton University Press.
- Cramer, H. (1946). *Mathematical methods of statistics*. Princeton University Press.
- Cruse, B., Liedloff, A. C., and Wintle, B. A. (2012). A new method for dealing with residual spatial autocorrelation in species distribution models. *Ecography*, 35(10):879–888.

- Crespo Cuaresma, J., Danylo, O., Fritz, S., McCallum, I., Obersteiner, M., See, L., and Walsh, B. (2017). Economic Development and Forest Cover: Evidence from Satellite Data. *Scientific Reports*, 7:40678.
- Das, D., Kelejian, H. H., and Prucha, I. R. (2003). Finite sample properties of estimators of spatial autoregressive models with autoregressive disturbances. *Papers in Regional Science*, 82(1):1–26.
- Dasgupta, S., Laplante, B., Wang, H., and Wheeler, D. (2002). Confronting the Environmental Kuznets Curve. *Journal of Economic Perspectives*, 16(1):147–168.
- Davidson, J. (1994). *Stochastic Limit Theory*. Oxford University Press.
- Davies, J. B., Sandström, S., Shorrocks, A., and Wolff, E. N. (2011). The Level and Distribution of Global Household Wealth. *The Economic Journal*, 121(551):223–254.
- Davies, R. B. (1977). Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternative. *Biometrika*, 64(2):247.
- Davies, R. B. (1987). Hypothesis Testing when a Nuisance Parameter is Present Only Under the Alternatives. *Biometrika*, 74(1):33.
- de Jong, K. (2013). Resultaten Melkveehouderij Bedrijfsanalyse 2013. Technical report, PPP Agro Advies, Ilpendam.
- de Wit, M. and Faaij, A. (2010). European biomass resource potential and costs. *Biomass and Bioenergy*, 34(2):188–202.
- Debarsy, N., Jin, F., and Lee, L.-f. (2015). Large sample properties of the matrix exponential spatial specification with an application to FDI. *Journal of Econometrics*, 188(1):1–21.
- Deschenes, O., Greenstone, M., and Shapiro, J. S. (2012). Defensive Investments and the Demand for Air Quality: Evidence from the NOx Budget Program. *NBER Working Paper No. 18267*.

- Diebold, F. X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests. *Journal of Business & Economic Statistics*, 33(1):1–1.
- Diebold, F. X. and Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.
- Dijk, D., Teräsvirta, T., and Franses, P. (2002). Smooth transition autoregressive models—a survey of recent developments. *Econometric Reviews*, (August 2013):37–41.
- Dijk, D. V., Franses, P. H., Lucas, A., and Lucas, A. (1999). Testing for Smooth Transition Nonlinearity in the Presence of Outliers. *Journal of Business & Economic Statistics*, 17(2):217.
- Diks, C., Panchenko, V., and van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163(2):215–230.
- Diogo, V., Koomen, E., and Kuhlman, T. (2015). An economic theory-based explanatory model of agricultural land-use patterns: The Netherlands as a case study. *Agricultural Systems*, 139:1–16.
- Diogo, V., Koomen, E., and van der Hilst, F. (2012). Second generation biofuel production in the Netherlands. Technical report, Amsterdam.
- Diogo, V., Koomen, E., Witte, F., and Schaap, B. (2013). Understanding the spatial distribution of agricultural land use in view of climate-driven hydrological changes - Expert Pool Report. Technical report, Knowledge for Climate, Utrecht.
- Diogo, V., van der Hilst, F., van Eijck, J., Verstegen, J. A., Hilbert, J., Carballo, S., Volante, J., and Faaij, A. (2014). Combining empirical and theory-based land-use modelling approaches to assess economic

- potential of biofuel production avoiding iLUC: Argentina as a case study. *Renewable and Sustainable Energy Reviews*, 34:208–224.
- Dixit, A. K. and Pindyck, R. S. (1995). The Options Approach to Capital Investment. *Harvard Business Review*, 73(3):105–115.
- Dogo, H., Brandon, C., Heger, M., Chonabayashi, S., Gaskell, J., Brown, T., Bangalore, M., Norman, T., Lee, J. J., Spencer, P., Chamorro, A., Andree, B. P. J., de Azevedo, J. P. W., Nguyen, M., Deen, S., and Bagstad, K. (2019). *Hidden Dimension Of Poverty: Natural Resources and the Environment*. World Bank Publications, Washington, DC.
- Domowitz, I. and White, H. (1982). Misspecified models with dependent observations. *Journal of Econometrics*, 20(1):35–58.
- Doob, J. L. (1934). Probability and Statistics. *Transactions of the American Mathematical Society*, 36(4):759–775.
- Dornburg, V., Faaij, A., Verweij, P., Langeveld, H., van de Ven, G., Wester, F., van Keulen, H., van Diepen, K., Meeusen, M., Banse, M., Ros, J., van Vuuren, D., van den Born, G., van Oorschot, M., Smout, F., van Vliet, J., Aiking, H., Londo, M., and Mozzaf, K. (2008). Biomass Assessment: Assessment of global biomass potentials and their links to food , water , biodiversity , energy demand and economy Main report. *Climate Change Scientific Assessment and Policy Analysis*, 500114009(January 2008):0.
- Dudley, R. M. (2002). *Real analysis and probability*. Cambridge University Press.
- Dufour, J.-M., Pelletier, D., and Renault, É. (2006). Short run and long run causality in time series: inference. *Journal of Econometrics*, 132(2):337–362.
- Dufour, J.-M. and Renault, E. (1998). Short Run and Long Run Causality in Time Series: Theory. *Econometrica*, 66(5):1099.

- Dufour, J.-M. and Taamouti, A. (2010). Short and long run causality measures: Theory and inference. *Journal of Econometrics*, 154(1):42–58.
- Dutch Emission Authority (2013). Naleving jaarverplichting 2012 hernieuwbare energie vervoer en verplichting brandstoffen luchtverontreiniging. Technical report.
- Eichler, M. and Didelez, V. (2010). On Granger causality and the effect of interventions in time series. *Lifetime Data Analysis*, 16(1):3–32.
- Elbersen, H. W., Bakker, R. R., and Elbersen, B. S. (2005). A simple method to estimate practical field yields of biomass grasses in Europe. In *14th European Biomass Conference, 17–21 October 2005*, Paris. Wageningen University.
- Elhorst, J. P. (2010a). Applied Spatial Econometrics: Raising the Bar. *Spatial Economic Analysis*.
- Elhorst, J. P. (2010b). *Spatial Econometrics: From Cross Sectional Data to Spatial Panels*. Springer.
- Engle, R. F., Hendry, D. F., and Richard, J.-F. (1983). Exogeneity. *Econometrica*, 51(2):277.
- Epple, D. and Sieg, H. (1999). Estimating Equilibrium Models of Local Jurisdictions. *The Journal of Political Economy*, 10(2):130–153.
- European Commission (2007). Biofuels: aid per hectare of energy crops reduced as the area exceeds 2 million hectares.
- European Commission (2013). Political agreement on new direction for common agricultural policy.
- Eurostat - Statistical Office of the European Communities (2009). Panorama of energy: Energy statistics to support EU policies and solutions. Technical report, Luxembourg.



- Evers, A., de Haan, M., Blanken, K., Hemmer, J., Hollander, C., Holshof, G., and Ouweltjes, W. (2007). Results low-cost farm 2006. Technical report, Animal Sciences Group, Department of Livestock Research, Lelystad.
- Farber, S. and Páez, A. (2007). A systematic investigation of cross-validation in GWR model estimation: Empirical analysis and Monte Carlo simulations. *Journal of Geographical Systems*, 9(4):371–396.
- Farrell, A. E., Plevin, R. J., Turner, B. T., Jones, A. D., O'Hare, M., and Kammen, D. M. (2006). Ethanol can contribute to energy and environmental goals. *Science*, 311(5760):506–508.
- Fell, H. and Linn, J. (2013). Renewable electricity policies, heterogeneity, and cost effectiveness. *Journal of Environmental Economics and Management*, 66(3):688–707.
- Fischer, G., Prieler, S., van Velthuizen, H., Berndes, G., Faaij, A., Londo, M., and de Wit, M. (2010a). Biofuel production potentials in Europe: Sustainable use of cultivated land and pastures, Part II: Land use scenarios. *Biomass and Bioenergy*, 34(2):173–187.
- Fischer, G., Prieler, S., van Velthuizen, H., Lensink, S. M., Londo, M., and de Wit, M. (2010b). Biofuel production potentials in Europe: Sustainable use of cultivated land and pastures. Part I: Land productivity potentials. *Biomass and Bioenergy*, 34(2):159–172.
- Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 222(594-604):309–368.
- Fisher, R. A. (1925). Theory of Statistical Estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(05):700.
- Flach, B., Lieberz, S., Rondon, M., Williams, B., and Teiken, C. (2015).

- EU Biofuels Annual 2015. Technical Report GAIN Report Number: NL5028.
- Fleming, R. a. and Adams, R. M. (1997). The Importance of Site-Specific Information in the Design of Policies to Control Pollution. *Journal of Environmental Economics and Management*, 33(3):347–358.
- Folland, G. B. (2009). *A Guide to Advanced Real Analysis*. Dolciani Mathematical Expositions. Cambridge University Press.
- Fotheringham, A. S. (1981). Spatial structure and distance-decay parameters. *Annals of the Association of American Geographers*, 71(3):425–436.
- Fotheringham, A. S. (2009). “The Problem of Spatial Autocorrelation” and Local Spatial Statistics. *Geographical Analysis*, 41(4):398–403.
- Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley.
- Fowlie, M. L. and Muller, N. Z. (2013). Market-Based Emissions Regulation when Damages Vary Across Sources: What Are The Gains From Differentiation? *NBER Working Paper Series*.
- Frankel, J., Schmukler, S. L., and Servén, L. (2004). Global transmission of interest rates: monetary independence and currency regime. *Journal of International Money and Finance*, 23:701–733.
- Freedman, D. and Diaconis, P. (1982). De Finetti’s Theorem for Symmetric Location Families. *The Annals of Statistics*, 10(1):184–189.
- Frías, M. and Ruiz-Medina, M. (2016). Wavelet nonparametric estimation from strong spatial correlated high-dimensional data. *Spatial Statistics*, 18:363–385.
- Gallant, R. and White, H. (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Cambridge University Press.

- Gao, Y., Zhang, X., Wang, S., and Zou, G. (2016). Model averaging based on leave-subject-out cross-validation. *Journal of Econometrics*, 192(1):139–151.
- Gasser, T., Guivarch, C., Tachiiri, K., Jones, C. D., and Ciais, P. (2015). Negative emissions physically needed to keep global warming below 2C. *Nature Communications*, 6:7958.
- Geman, Stuart; Hwang, C.-R. (1982). Nonparametric Maximum Likelihood Estimation by the Method of Sieves. *The Annals of Statistics*, 10(2):401–414.
- Glass, A. J., Kenjegalieva, K., and Sickles, R. C. (2016). A spatial autoregressive stochastic frontier model for panel data with asymmetric efficiency spillovers. *Journal of Econometrics*, 190(2):289–300.
- Gomiero, T., Paoletti, M. G., and Pimentel, D. (2010). Biofuels: Efficiency, Ethics, and Limits to Human Appropriation of Ecosystem Services. *Journal of Agricultural and Environmental Ethics*, 23(5):403–434.
- Gosselink, J., Bos, B., Bokma, S., and Koerkamp, P. (2008). Oudere koeien voor een duurzame veehouderij. Technical report, Wageningen.
- Graham, J. W. (2012). *Missing Data*. Springer New York, New York, NY.
- Granger, C., King, M. L., and White, H. (1995). Comments on testing economic theories and the use of model selection criteria. *Journal of Econometrics*, 67(1):173–187.
- Granger, C. and Teräsvirta, T. (1993). Modelling Nonlinear Economic Relationships. *Oxford University Press*, 61(4):1241–1243.
- Granger, C. W. (1980). Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352.

- Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424.
- Grieg-Gran, M., Porras, I., and Wunder, S. (2005). How can market mechanisms for forest environmental services help the poor? Preliminary lessons from Latin America. *World Development*, 33(9 SPEC. ISS.):1511–1527.
- Grossman, G. M. and Krueger, A. B. (1991). Environmental Impacts of a North American Free Trade Agreement. *NBER Working Papers*.
- Grossman, G. M. and Krueger, A. B. (1995). Economic Growth and the Environment. *The Quarterly Journal of Economics*, 110(2):353–377.
- Haavelmo, T. (1943). The Statistical Implications of a System of Simultaneous Equations. *Econometrica*, 11:1–12.
- Haavelmo, T. (1944). The Probability Approach in Econometrics. *Econometrica*, 12((Suppl.)):1–115.
- Hainmueller, J. and Hazlett, C. (2014). Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. *Political Analysis*, 22(2):143–168.
- Hamelinck, C. N., van Hooijdonk, G., and Faaij, A. P. C. (2005). Ethanol from lignocellulosic biomass: Techno-economic performance in short-, middle- and long-term.
- Hanna, R. and Oliva, P. (2015). The effect of pollution on labor supply: Evidence from a natural experiment in Mexico City. *Journal of Public Economics*, 122:68–79.
- Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., and Townshend, J. R. G. (2013). High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, 342(6160):850–853.

- Härdle, W., Werwatz, A., Müller, M., and Sperlich, S. (2004). *Nonparametric and Semiparametric Models*. Springer Berlin Heidelberg.
- Harnack, D., Laminski, E., Schünemann, M., and Pawelzik, K. R. (2017). Topological Causality in Dynamical Systems. *Physical Review Letters*, 119(9):098301.
- Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291.
- Helfand, G. E. and House, B. W. (1995). Regulating Nonpoint Source Pollution under Heterogeneous Conditions. *American Journal of Agricultural Economics*, 77(4):1024–1032.
- Hellinger, E. (1909). Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die reine und angewandte Mathematik (Crelle's Journal)*, 1909(136):210–271.
- Hendry, D. F. (2017). *European journal of pure and applied mathematics.*, volume 10.
- Hewitt, J., Hoeting, J. A., Done, J. M., and Towler, E. (2018). Remote effects spatial process models for modeling teleconnections. *Environmetrics*, 29(8):e2523.
- Hoogwijk, M., Faaij, A., Eickhout, B., De Vries, B., and Turkenburg, W. (2005). Potential of biomass energy out to 2100, for four IPCC SRES land-use scenarios.
- Hordijk, L. (1974). Spatial correlation in the disturbances of a linear interregional model. *Regional and Urban Economics*.
- Hordijk, L. and Paelinck, J. (1976). Some principles and results in spatial econometrics. *Recherches Économiques de Louvain*.

- Horowitz, J. L. (2011). Applied Nonparametric Instrumental Variables Estimation. *Econometrica*, 79(2):347–394.
- Hoshino, T. (2016). Semiparametric Spatial Autoregressive Models with Endogenous Regressors: With an Application to Crime Data. *Journal of Business & Economic Statistics*, pages 1–51.
- Hotelling, H. (1930). The Consistency and Ultimate Distribution of Optimum Statistics. *Transactions of the American Mathematical Society*, 32:847–59.
- Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.
- Hurvich, C. M. and Tsai, C.-L. (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika*, 78(3):499–509.
- Hyde, W. F., Amacher, G. S., and Magrath, W. (1996). Deforestation and Forest Land Use: Theory, Evidence, and Policy Implications. *The World Bank Research Observer*, 11(2):223–248.
- International Energy Agency (2016). Energy and Air Pollution. *World Energy Outlook - Special Report*, page 266.
- Jennrich, R. I. (1969). Asymptotic Properties of Non-Linear Least Squares Estimators. *The Annals of Mathematical Statistics*, 40(2):633–643.
- Just, R. E. and Antle, J. M. (1990). Interactions Between Agricultural and Environmental Policies : A Conceptual Framework. *American Economic Review*, 80(2):197–202.
- Kabaila, P. (1983). On The Asymptotic Efficiency Of Estimators Of The Parameters Of An ARMA Process. *Journal of Time Series Analysis*, 4(1):37–47.

- Kalman, R. (1983). Identifiability and Modeling in Econometrics. *Developments in Statistics*, 4:97–136.
- Kapetanios, G. (2001). Model Selection in Threshold Models. *Journal of Time Series Analysis*, 22(6):733–754.
- Keene, A. and Deller, S. C. (2015). Evidence of the Environmental Kuznets’ Curve among US Counties and the Impact of Social Capital. *International Regional Science Review*, 38(4):358–387.
- Kelejian, H. H. and Prucha, I. R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, 157(1):53–67.
- Kent, J. T. and Tyler, D. E. (2001). Regularity and Uniqueness for Constrained M -Estimates and Redescending M -Estimates. *The Annals of Statistics*, 29(1):252–265.
- Kharroubi, E., Zampolli, F., Kharroubi, E., and Zampolli, F. (2016). Monetary independence in a financially integrated world: what do measures of interest rate co-movement tell us? *BIS Papers Series*, 88:193–205.
- Koçar, G. and Civas, N. (2013). An overview of biofuels from energy crops: Current status and future prospects. *Renewable and Sustainable Energy Reviews*, 28:900–916.
- Kolmogorov, A. N. A. N. (1933). *Foundations of the theory of probability*.
- Kolmogorov, A. N. A. N. and Fomin, S. V. S. V. (1975). *Introductory real analysis*. Dover Publications.
- Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer Series in Statistics. Springer, New York.

- Koomen, E., Dekkers, J., and van Dijk, T. (2008). Open-space preservation in the Netherlands: Planning, practice and prospects. *Land Use Policy*, 25(3):361–377.
- Koomen, E., Kuhlman, T., Groen, J., and Bouwman, A. (2005). Simulating the future of agricultural land use in The Netherlands. *Tijdschrift Voor Economische En Sociale Geografie*, 96(2):218–224.
- Kostov, P. (2009). A Spatial Quantile Regression Hedonic Model of Agricultural Land Prices. *Spatial Economic Analysis*, 4(1):53–72.
- Kowarik, A. and Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software*, 74(7):1–16.
- Kraft, C. (1955). Some conditions for consistency and uniform consistency of statistical procedures. *University of California Publications in Statistics*, 2:125–242.
- Krakovska, A., Jakubík, J., Chvosteková, M., Coufal, D., Jajcay, N., and Paluš, M. (2018). Comparison of six methods for the detection of causality in a bivariate time series. *Physical Review E*, 97(4):042207.
- Krengel, U. (1985). *Ergodic theorems*. Walter de Gruyter.
- Kuhlman, T., Diogo, V., and Koomen, E. (2013). Exploring the potential of reed as a bioenergy crop in the Netherlands. *Biomass and Bioenergy*, 55:41–52.
- Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5):1–26.
- Kuhn, M. and Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York, New York, NY.
- Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.



- Landell-Mills, N. and Porras, I. T. (2002). Silver bullet or fools' gold? A global review of markets for forest environmental services and their impact on the poor. *Development*, 100(March):249.
- Lange, K. (1999). *Numerical analysis for statisticians*. Springer.
- Le Cam, L. (1953). On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes' Estimates. *University of California Publications in Statistics*, 2:23–52.
- Lee, L. (2012). Environmental poverty, a decomposed environmental Kuznets curve, and alternatives: Sustainability lessons from China. *Ecological Economics*, 73(1):86–92.
- Lee, L. F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, 72(6):1899–1925.
- LEI (2012). Land- en tuinbouwcijfers 2012. Technical report, The Hague.
- Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., and Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525(7569):367–371.
- LeSage, J. P. (2008). An Introduction to Spatial Econometrics. *Revue d'économie industrielle*, 123(123):19–44.
- LeSage, J. P. and Fischer, M. M. (2008). Spatial growth regressions: Model specification, estimation and interpretation. *Spatial Economic Analysis*, 3(3):275–304.
- LeSage, J. P. and Pace, R. K. (2009). *Introduction to spatial econometrics*. CRC Press.
- Leung, Y., Mei, C. L., and Zhang, W. X. (2000). Testing for spatial autocorrelation among the residuals of the geographically weighted regression. *Environment and Planning A*, 32(5):871–890.

- Levinson, A. (2001). The Ups and Downs of the Environmental Kuznets Curve. *Georgetown University Working Papers*.
- Levinson, A. (2012). Valuing public goods using happiness data: The case of air quality. *Journal of Public Economics*, 96(9-10):869–880.
- Li, W. K. (1988). The akaike information criterion in threshold modelling: Some empirical evidences. In *Nonlinear Time Series and Signal Processing*, pages 88–96. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Liang, X. S. (2016). Information flow and causality as rigorous notions ab initio. *Physical Review E*, 94(5):052201.
- List, J. A. and Gallet, C. A. (1999). The environmental Kuznets curve: does one size fit all? *Ecological Economics*, 31(3):409–423.
- Little, R. J. A. and Rubin, D. B. (2012). *Statistical analysis with missing data*.
- Lorenz, H.-W. (1993). Nonlinearities and Economic Dynamics. In *Nonlinear Dynamical Economics and Chaotic Motion*, pages 26–79. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Lu, X., Su, L., and White, H. (2017). Granger Causality and Structural Causality in Cross-section and Panel Data. *Econometric Theory*, 33(02):263–291.
- Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Malinvaud, E. (1970). The Consistency of Nonlinear Regressions. *The Annals of Mathematical Statistics*, 41(3):956–969.
- Managi, S. and Jena, P. R. (2008). Environmental productivity and Kuznets curve in India. *Ecological Economics*, 65(2):432–440.
- McQuarrie, A. D. R. and Tsai, C.-L. (1998). *Regression and Time Series Model Selection*. World Scientific.

- Mérel, P. and Wimberger, E. (2012). Improving air quality in California's San Joaquin Valley: The role of vehicle heterogeneity in optimal emissions abatement. *Journal of Environmental Economics and Management*, 63(2):169–186.
- Micchelli, C. A., Xu, Y., and Zhang, H. (2006). Universal Kernels. *Journal of Machine Learning Research*, 7:2651–2667.
- Ministerie van Economische Zaken Landbouw en Innovatie (2013). Basisregistratie Percelen 2012. Technical report, The Hague.
- Moltedo, A., Troubat, N., Lokshin, M., and Sajaia, Z., editors (2014). *Analyzing Food Security Using Household Survey Data: Streamlined Analysis with ADePT Software*. The World Bank.
- Murray, J. S. (2018). Multiple Imputation: A Review of Practical and Theoretical Findings. *Statistical Science*, 33(2):142–159.
- Neuberg, L. G., Neuberg, and Gerson, L. (2003). Causality: Models, Reasoning, and Inference, by Judea Pearl, Cambridge University Press, 2000. *Econometric Theory*, 19(04):675–685.
- Nsiri, S. and Roy, R. (1993). On the Invertibility on Multivariate Linear Processes. *Journal of Time Series Analysis*, 14(3):305–316.
- OECD and FAO (2015). OECD-FAO Agricultural Outlook 2015-2024. Technical report, Paris.
- Ord, K. (1975). Estimation Methods for Models of Spatial Interaction. *Journal of the American Statistical Association*, 70(349):120–126.
- Organisation for Economic Cooperation and Development (2016). *The economic consequences of outdoor air pollution*. OECD Publishing, Paris.

- Özokcu, S. and Özdemir, Ö. (2017). Economic growth, energy, and environmental Kuznets curve. *Renewable and Sustainable Energy Reviews*, 72:639–647.
- Paelinck, J. H. and Klaasen, L. H. (1979). Spatial econometrics. *Spatial econometrics*.
- Paez, A., Uchida, T., and Miyamoto, K. (2002). A general framework for estimation and inference of geographically weighted regression models: 1. Location-specific kernel bandwidths and a test for locational heterogeneity. *Environment and Planning A*, 34(4):733–754.
- Panayatou, T. (1997). Demystifying the environmental Kuznets curve: turning a black box into a policy tool. *Environment and Development Economics*, 2(4):S1355770X97000259.
- Panayotou, T. (1993). Empirical tests and policy analysis of environmental degradation at different stages of economic development. *ILO Working Papers*.
- Parks, P. J. and Hardie, I. W. (1995). Least-Cost Forest Carbon Reserves - Cost-Effective Subsidies to Convert Marginal Agricultural Land to Forests. *Land Economics*, 71(1):122–136.
- Pautsch, G. R., Kurkalova, L. a., Babcock, B. a., and Kling, C. L. (2001). the Efficiency of Sequestering Carbon in Agricultural Soils. *Contemporary Economic Policy*, 19(2):123–134.
- Pearl, J. (2000). *Causality : models, reasoning, and inference*. Cambridge University Press.
- Perman, R. and Stern, D. I. (2003). Evidence from panel unit root and cointegration tests that the Environmental Kuznets Curve does not exist. *Australian Journal of Agricultural and Resource Economics*, 47(3):325–347.

- Peters, G. P., Andrew, R. M., Boden, T., Canadell, J. G., Ciais, P., Quéré, C. L., Marland, G., Raupach, M. R., and Wilson, C. (2013). The challenge to keep global warming below 2C. *Nature Climate Change*, 3(January):4–6.
- Peters, G. P., Marland, G., Le Quéré, C., Boden, T., Canadell, J. G., and Raupach, M. R. (2012). Rapid growth in CO<sub>2</sub> emissions after the 2008-2009 global financial crisis. *Nature Climate Change*, 2(1):2–4.
- Pina, G. (2017). International reserves and global interest rates. *Journal of International Money and Finance*, 74:371–385.
- Pindyck, R. (1991). Irreversibility, Uncertainty, and Investment. *Journal of Economic Literature*, 29(3):1110–1148.
- Pindyck, R. S. (2007). Uncertainty in environmental economics. *Review of Environmental Economics and Policy*, 1(1):45–65.
- Pötscher, B. M. and Prucha, I. R. (1991). Basic structure of the asymptotic theory in dynamic nonlinear econometric models, part i: consistency and approximation concepts. *Econometric Reviews*, 10(2):125–216.
- Pötscher, B. M. and Prucha, I. R. (1997). *Dynamic Nonlinear Econometric Models*. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Pozzer, A., Zimmermann, P., Doering, U. M., Van Aardenne, J., Tost, H., Dentener, F., Janssens-Maenhout, G., and Lelieveld, J. (2012). Effects of business-as-usual anthropogenic emissions on air quality. *Atmospheric Chemistry and Physics*, 12(15):6915–6937.
- Preker, A. S., Adeyi, O. O., Lapetra, M. G., Simon, D. C., and Keuffel, E. (2016). Health Care Expenditures Associated With Pollution: Exploratory Methods and Findings. *Annals of Global Health*, 82(5):711–721.

- Psaradakis, Z., Sola, M., Spagnolo, F., and Spagnolo, N. (2009). Selecting nonlinear time series models using information criteria. *Journal of Time Series Analysis*, 30(4):369–394.
- Puu, T. (1991). Nonlinear Economic Dynamics. In *Nonlinear Economic Dynamics*, pages 1–7. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Quiggin, J. (2009). Six Refuted Doctrines. *Economic Papers*, 28(3):239–248.
- Ragauskas, A. J., Williams, C. K., Davison, B. H., Britovsek, G., Cairney, J., Eckert, C. A., Jr, W. J. F., Hallett, J. P., Leak, D. J., Liotta, C. L., Mielenz, J. R., Murphy, R., Templer, R., and Tschaplinski, T. (2006). The Path Forward for Biofuels. *Science*, 311(January):484–489.
- Ranieri, R. and Almeida Ramos, R. (2013). Inclusive growth: Building up a concept.
- Rao, R. R. (1962). Relations between Weak and Uniform Convergence of Measures with Applications. *The Annals of Mathematical Statistics*, 33(2):659–680.
- Regan, C. M., Bryan, B. A., Connor, J. D., Meyer, W. S., Ostendorf, B., Zhu, Z., and Bao, C. (2015). Real options analysis for land use management: Methods, application, and implications for policy.
- Reinsel, G. C. (2003). *Elements of multivariate time series analysis*. Springer.
- REN21 (2015). Renewables 2015 Global Status Report. Technical report, Paris.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pages 1135–1144.

- Rosenthal, S. S. and Strange, W. C. (2003). Geography, Industrial Organization, and Agglomeration. *Review of Economics and Statistics*, 85(2):377–393.
- Rothenberg, T. J. (1971). Identification in Parametric Models. *Econometrica*, 39(3):577–591.
- Roussas, G. G. (1965). Extension to Markov processes of a result by A. Wald about the consistency of the maximum likelihood estimate. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 4(1):69–73.
- Roy, A., McElroy, T. S., and Linton, P. (2014). Estimation of Causal Invertible VARMA Models.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.
- Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91(434):473.
- Salgado, C. M., Azevedo, C., Proença, H., and Vieira, S. M. (2016). Missing Data. In *Secondary Analysis of Electronic Health Records*, pages 143–162. Springer International Publishing, Cham.
- Schölkopf, B. and Smola, A. J. (2001). *Learning with kernel: Support Vector Machines, Regularization, Optimization and Beyond*.
- Shafik, N. and Bandyopadhyay, S. (1992). Economic growth and environmental quality : time series and cross-country evidence. *World Bank Policy Research Papers*.

- Shahbaz, M., Shafiullah, M., Papavassiliou, V. G., and Hammoudeh, S. (2017). The CO<sub>2</sub>–growth nexus revisited: A nonparametric analysis for the G7 economies over nearly two centuries. *Energy Economics*, 65:183–193.
- Shapiro, J. S. and Walker, R. (2016). Why is Pollution from U.S. Manufacturing Declining? The Roles of Environmental Regulation, Productivity, and Trade. *Cowles Foundation Discussion Paper No. 1982R*.
- Sims, C. A. (1972). Money, Income, and Causality. *The American Economic Review*, 62(4):540–552.
- Sin, C.-Y. and White, H. (1996). Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics*, 71(1):207–225.
- Singh, A., Smyth, B. M., and Murphy, J. D. (2010). A biofuel strategy for Ireland with an emphasis on production of biomethane and minimization of land-take. *Renewable and Sustainable Energy Reviews*, 14(1):277–288.
- Sleeter, B. M., Sohl, T. L., Wilson, T. S., Sleeter, R. R., Soulard, C. E., Bouchard, M. A., Sayler, K. L., Reker, R. R., and Griffith, G. E. (2012). Projected Land-Use and Land-Cover Change in the Western United States. *Baseline and Projected Future Carbon Storage and Greenhouse-Gas Fluxes in Ecosystems of the Western United States*.
- Smarzewski, R. (1986). Strongly unique best approximation in Banach spaces. *Journal of Approximation Theory*, 47(3):184–194.
- Smeets, E. M. W. and Faaij, A. P. C. (2007). Bioenergy potentials from forestry in 2050: An assessment of the drivers that determine the potentials. *Climatic Change*, 81(3-4):353–390.
- Soumyananda, D. (2004). Environmental Kuznets Curve Hypothesis: A Survey. *Ecological Economics*, 49(4):431–455.



- Spanos, A. (1989). On Rereading Haavelmo: A Retrospective View of Econometric Modeling. *Econometric Theory*, 5(03):405–429.
- Stagl, S. (1999). Delinking Economic Growth from Environmental Degradation? A Literature Survey on the Environmental Kuznets Curve Hypothesis. *Wirtschafts Universitat Wien Working Paper No. 6*.
- Stanton, T.L.; LeValley, S. (2010). Feed Composition for Cattle and Sheep. Technical report, Colorado.
- Statistical Office of the European Communities and Organisation for Economic Co-operation and Development (2012). *Eurostat-OECD methodological manual on purchasing power parities*. OECD.
- Stavins, R. N. (1998). What Can We Learn from the Grand Policy Experiment? Lessons from SO<sub>2</sub> Allowance Trading. *Journal of Economic Perspectives*, 12(3):69–88.
- Stavins, R. N. (1999). The costs of carbon sequestration: a revealed-preference approach. *American Economic Review*, 89(4):994–1009.
- Stelzer, R. (2008). Multivariate Markov-switching ARMA processes with regularly varying noise. *Journal of Multivariate Analysis*, 99(6):1177–1190.
- Stern, D. I. (1998). Progress on the Environmental Kuznets Curve? *Environment and Development Economics*, 3(2):173–196.
- Stern, D. I. (2004). The Rise and Fall of the Environmental Kuznets Curve. *World Development*, 32(8):1419–1439.
- Stern, D. I. and Common, M. S. (2001). Is There an Environmental Kuznets Curve for Sulfur? *Journal of Environmental Economics and Management*, 41(2):162–178.

- Stern, D. I., Common, M. S., and Barbier, E. B. (1996). Economic growth and environmental degradation: The environmental Kuznets curve and sustainable development. *World Development*, 24(7):1151–1160.
- Straumann, D. and Mikosch, T. (2006). Quasi-Maximum-Likelihood Estimation in Conditionally Heteroskedastic Time Series: A Stochastic Recurrence Equations Approach. *The Annals of Statistics*, 34(5):2449–2495.
- Styles, D. and Jones, M. B. (2007). Current and future financial competitiveness of electricity and heat from energy crops: A case study from Ireland. *Energy Policy*, 35(8):4355–4367.
- Su, S., Xiao, R., and Zhang, Y. (2012). Multi-scale analysis of spatially varying relationships between agricultural landscape patterns and urbanization using geographically weighted regression. *Applied Geography*, 32(2):360–375.
- Tait, J. (2011). The ethics of biofuels. *GCB Bioenergy*, 3(3):271–275.
- Tallis, H. M., Hawthorne, P. L., Polasky, S., Reid, J., Beck, M. W., Brauman, K., Bielicki, J. M., Binder, S., Burgess, M. G., Cassidy, E., Clark, A., Fargione, J., Game, E. T., Gerber, J., Isbell, F., Kiesecker, J., McDonald, R., Metian, M., Molnar, J. L., Mueller, N. D., O’Connell, C., Ovando, D., Troell, M., Boucher, T. M., and McPeck, B. (2018). An attainable global vision for conservation and human well-being. *Frontiers in Ecology and the Environment*, 16(10):563–570.
- Teräsvirta, T. (1994). Specification, Estimation and Evaluation of Smooth Transition Autoregressive Models. *Journal of the American Statistical Association*, 89(425):208–218.
- Teräsvirta, T. and Anderson, H. M. (1992). Characterizing Nonlinearities in Business Cycles Using Smooth Transition Autoregressive Models. *Journal of Applied Econometrics*, 7(S):S119–36.

- Teräsvirta, T., Tjøstheim, D., and Granger, C. W. J. (2010). *Modelling nonlinear economic time series*. Oxford University Press.
- The World Bank (2011). *Purchasing power parities and the real size of world economies : a comprehensive report of the 2011 International Comparison Program*.
- The World Bank (2013). *Measuring the Real Size of the World Economy*. The World Bank.
- Tibshirani, R. (1996). Regression selection and shrinkage via the lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288.
- Tietenberg, T. H. (1990). Economic instruments for environmental regulation. *Oxford Review of Economic Policy*, 6(1):17–33.
- Tong, H. (1983). *Threshold models in non-linear time series analysis*. Springer-Verlag.
- Tong, H. (1990). *Non-linear time series : a dynamical system approach*. Clarendon Press.
- Tong, H. (2015). Threshold models in time series analysis—Some reflections. *Journal of Econometrics*, 189(2):485–491.
- van Bakel, P., van der Waal, B., de Haan, M., Spruyt, J., and Evers, A. (2007). HELP-2006; Uitbreiding en actualisering van de HELP-2005-tabellen ten behoeve van het Waternood-instrumentarium. Technical report.
- van Dam, J., Faaij, A. P. C., Lewandowski, I., and Fischer, G. (2007). Biomass production potentials in Central and Eastern Europe under different scenarios. *Biomass and Bioenergy*, 31(6):345–366.
- van den Broek, R., Teeuwisse, S., Healion, K., Kent, T., van Wijk, A., Faaij, A., and Turkenburg, W. (2001). Potentials for electricity production from wood in Ireland. *Energy*, 26(11):991–1013.

- van der Hilst, F., Dornburg, V., Sanders, J. P. M., Elbersen, B., Graves, A., Turkenburg, W. C., Elbersen, H. W., van Dam, J. M. C., and Faaij, A. P. C. (2010). Potential, spatial distribution and economic performance of regional biomass chains: The North of the Netherlands as example. *Agricultural Systems*, 103(7):403–417.
- van der Vaart, A. W. (2000). *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press.
- van der Voort, M., Timmer, R., Geel, W., Runia, W., and Corré, W. (2008). *Economie van Energiegewassen*. Technical report, Wageningen.
- van der Wolf, M. de; Klooster, A. (2006). *Kwantitatieve Informatie Akkerbouw en Vollegrondsgroenteteelt*. Technical report, Wageningen.
- van Donkelaar, A., Martin, R. V., Brauer, M., Hsu, N. C., Kahn, R. A., Levy, R. C., Lyapustin, A., Sayer, A. M., and Winker, D. M. (2016). Global Estimates of Fine Particulate Matter using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors. *Environmental Science & Technology*, 50(7):3762–3772.
- Voeks, R. (1997). Real Options: Managerial Flexibility and Strategy in Resource Allocation. *Journal of Banking & Finance*, 21(2):285–288.
- Vollebergh, H. R. J., Dijkgraaf, E., and Melenberg, B. (2005). Environmental Kuznets curves for CO<sub>2</sub>: heterogeneity versus homogeneity. *SSRN*, (November):39.
- von Borstel, J., Eickmeier, S., and Krippner, L. (2016). The interest rate pass-through in the euro area during the sovereign debt crisis. *Journal of International Money and Finance*, 68:386–402.
- Wagner, M. (2015). The Environmental Kuznets Curve, Cointegration and Nonlinearity. *Journal of Applied Econometrics*, 30(6):948–967.

- Wald, A. (1949). Note on the Consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics*, 20(4):595–601.
- Westreich, D. (2012). Berkson’s bias, selection bias, and missing data. *Epidemiology (Cambridge, Mass.)*, 23(1):159–64.
- Wheeler, D. and Tiefelsdorf, M. (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, 7(2):161–187.
- Wheeler, D. C. (2007). Diagnostic tools and a remedial method for collinearity in geographically weighted regression. *Environment and Planning A*, 39(10):2464–2481.
- White, H. (1994). *Estimation, inference, and specification analysis*. Cambridge University Press.
- White, H. and Chalak, K. (2009). Settable Systems: An Extension of Pearl’s Causal Model with Optimization, Equilibrium, and Learning. *Journal of Machine Learning Research*, 10(Aug):1759–1799.
- White, H., Chalak, K., and Lu, X. (2011). Causality in Time Series Linking Granger Causality and the Pearl Causal Model with Settable Systems. *JMRL: Workshop and Conference Proceedings 12*, pages 1–29.
- White, H. and Lu, X. (2010). Granger Causality and Dynamic Structural Systems. *Journal of Financial Econometrics*, 8(2):193–243.
- White, H. and Pettenuzzo, D. (2014). Granger causality, exogeneity, cointegration, and economic policy analysis. *Journal of Econometrics*, 178:316–330.
- White, H., White, and Halbert (1980). Using Least Squares to Approximate Unknown Regression Functions. *International Economic Review*, 21(1):149–70.

- White, H., Xu, H., and Chalak, K. (2014). Causal discourse in a game of incomplete information. *Journal of Econometrics*, 182(1):45–58.
- Wiesenthal, T., Leduc, G., Christidis, P., Schade, B., Pelkmans, L., Govaerts, L., and Georgopoulos, P. (2009). Biofuel support policies in Europe: Lessons learnt for the long way ahead. *Renewable and Sustainable Energy Reviews*, 13(4):789–800.
- Wong, C. S. and Li, W. K. (1998). A note on the corrected Akaike information criterion for threshold autoregressive models. *Journal of Time Series Analysis*, 19(1):113–124.
- World Bank (1992). World Development Report 1992. Development and the Environment. Technical report, World Bank.
- World Bank and Institute for Health Metrics and Evaluation (2016). The cost of air pollution : strengthening the economic case for action.
- World Health Organization (2016). Ambient Air Pollution: A global assessment of exposure and burden of disease. *World Health Organization*, pages 1–131.
- Young, L. (2003). Carbon sequestration in agriculture: the U.S. policy context. *American Journal of Agricultural Economics*, 85(5):1164–1170.
- Zheng, T., Xiao, H., and Chen, R. (2015). Generalized ARMA models with martingale difference errors. *Journal of Econometrics*, 189(2):492–506.
- Zivin, J. G. and Neidell, M. (2012). The impact of pollution on worker productivity. *American Economic Review*, 102(7):3652–3673.
- Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429.



The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

- 712. Y. KANTOR, *Urban Form and the Labor Market*
- 713. R.M. TEULINGS, *Untangling Gravity*
- 714. K.J.VAN WILGENBURG, *Beliefs, Preferences and Health Insurance Behavior*
- 715. L. SWART, *Less Now or More Later? Essays on the Measurement of Time Preferences in Economic Experiments*
- 716. D. NIBBERING, *The Gains from Dimensionality*
- 717. V. HOORNWEG, *A Tradeoff in Econometrics*
- 718. S. KUCINSKAS, *Essays in Financial Economics*
- 719. O. FURTUNA, *Fiscal Austerity and Risk Sharing in Advanced Economies*
- 720. E. JAKUCIONYTE, *The Macroeconomic Consequences of Carry Trade Gone Wrong and Borrower Protection*
- 721. M. LI, *Essays on Time Series Models with Unobserved Components and Their Applications*
- 722. N. CIURILĂ, *Risk Sharing Properties and Labor Supply Disincentives of Pay-As-You-Go Pension Systems*
- 723. N.M. BOSCH, *Empirical Studies on Tax Incentives and Labour Market Behaviour*




724. S.D. JAGAU, *Listen to the Sirens: Understanding Psychological Mechanisms with Theory and Experimental Tests*
725. S. ALBRECHT, *Empirical Studies in Labour and Migration Economics*
726. Y.ZHU, *On the Effects of CEO Compensation*
727. S. XIA, *Essays on Markets for CEOs and Financial Analysts*
728. I. SAKALAUSKAITE, *Essays on Malpractice in Finance*
729. M.M. GARDBERG, *Financial Integration and Global Imbalances.*
730. U. THÜMMEL, *Of Machines and Men: Optimal Redistributive Policies under Technological Change*
731. B.J.L. KEIJERS, *Essays in Applied Time Series Analysis*
732. G. CIMINELLI, *Essays on Macroeconomic Policies after the Crisis*
733. Z.M. LI, *Econometric Analysis of High-frequency Market Microstructure*
734. C.M. OOSTERVEEN, *Education Design Matters*
735. S.C. BARENDSE, *In and Outside the Tails: Making and Evaluating Forecasts*
736. S. SÓVÁGÓ, *Where to Go Next? Essays on the Economics of School Choice*
737. M. HENNEQUIN, *Expectations and Bubbles in Asset Market Experiments*
738. M.W. ADLER, *The Economics of Roads: Congestion, Public Transit and Accident Management*
739. R.J. DÖTTLING, *Essays in Financial Economics*
740. E.S. ZWIERS, *About Family and Fate: Childhood Circumstances and Human Capital Formation*
741. Y.M. KUTLUAY, *The Value of (Avoiding) Malaria*

742. A. BOROWSKA, *Methods for Accurate and Efficient Bayesian Analysis of Time Series*
743. B. HU, *The Amazon Business Model, the Platform Economy and Executive Compensation: Three Essays in Search Theory*
744. R.C. SPERNA WEILAND, *Essays on Macro-Financial Risks*
745. P.M. GOLEC, *Essays in Financial Economics*
746. M.N. SOUVERIJN, *Incentives at work*
747. M.H. COVENEY, *Modern Imperatives: Essays on Education and Health Policy*
748. P. VAN BRUGGEN, *On Measuring Preferences*
749. M.H.C. NIENTKER, *On the Stability of Stochastic Dynamic Systems and their use in Econometrics*
750. S. GARCIA MANDICÓ, *Social Insurance, Labor Supply and Intra-Household Spillovers*
751. Y. SUN, *Consumer Search and Quality*
752. I. KERKEMEZOS, *On the Dynamics of (Anti) Competitive Behaviour in the Airline Industry*
753. G.W. GOY, *Modern Challenges to Monetary Policy*
754. A.C. VAN VLODROP, *Essays on Modeling Time-Varying Parameters*
755. J. SUN, *Tell Me How To Vote, Understanding the Role of Media in Modern Elections*
756. J.H. THIEL, *Competition, Dynamic Pricing and Advice in Frictional Markets: Theory and Evidence from the Dutch Market for Mortgages*
757. A. NEGRIU, *On the Economics of Institutions and Technology: a Computational Approach*
758. F. GRESNIGT, *Identifying and Predicting Financial Earth Quakes using Hawkes Processes*

759. A. EMIRMAHMUTOGLU, *Misperceptions of Uncertainty and Their Applications to Prevention*
760. A. RUSU, *Essays in Public Economics*
761. M.A. COTOFAN, *Essays in Applied Microeconomics: Non-Monetary Incentives, Skill Formation, and Work Preferences*





Stochastic economic processes are often characterized by dynamic interactions between variables that are dependent in both space and time. Analyzing these processes raises a number of questions about the econometric methods used that are both practically and theoretically interesting. This work studies econometric approaches to analyze spatial data that evolves dynamically over time.

The book provides a background on least squares and maximum likelihood estimators, and discusses some of the limits of basic econometric theory. It then discusses the importance of addressing spatial heterogeneity in policies. The next chapters cover parametric modeling of linear and nonlinear spatial time series, non-parametric modeling of nonlinearities in panel data, modeling of multiple spatial time series variables that exhibit long and short memory, and probabilistic causality in spatial time series settings.

Bo P.J. Andrée holds a BSc in Geology and Economics and an MSc in Spatial, Transport and Environmental Economics from the Free University Amsterdam. He performed studies for the World Bank, United Nations, OECD, European Commission, Asian Development Bank and the Dutch Government. His most recent project was on food crisis prediction with the Chief Economist of the World Bank. He currently lives in Amsterdam with his wife.

