Theory and Application of Dynamic Spatial Time Series Models

Andree, B.P.J.

2020

document version
Publisher's PDF, also known as Version of record

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.[1]

[1]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).[2] 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

[2]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 discus 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.

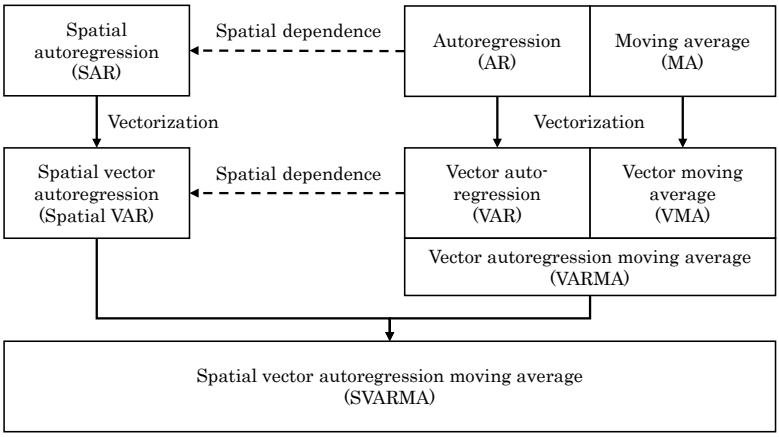

Figure 6.1: This chart presents an overview of the constituents of the Spatial vector autoregressive moving average (SVARMA) model described in this section. Note that AR and MA processes may also be defined on single cross-sections resulting in spatial-time series, or cross-sectional ARMA models – not 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.[3] 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

[3]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.[4] 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 $\{A_0, A_1, ..., A_p\}$, $\boldsymbol{\mathcal{A}}$ collects $\{\mathbf{A}_0, \mathbf{A}_1, ..., \mathbf{A}_p\}$. 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

$$A_0 w_t + A_1 w_{t-1} + ... + A_p w_{t-p} = M_0 \varepsilon_t + M_1 \varepsilon_{t-1} + ... + M_q \varepsilon_{t-q} \; \forall \, t \in \mathbb{Z}, \quad (6.1)$$

with parameter matrices structured as

$$A := \begin{bmatrix} a^{xx} & a^{xy} \\ a^{yx} & a^{yy} \end{bmatrix}, M := \begin{bmatrix} m^{xx} & m^{xy} \\ m^{yx} & m^{yy} \end{bmatrix}. \quad (6.2)$$

[4] 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

$$\mathbf{A}_{0:p} := \begin{bmatrix} A^{xx}_{0:p} & A^{xy}_{0:p} \\ A^{yx}_{0:p} & A^{yy}_{0:p} \end{bmatrix}, \mathbf{M}_{0:p} := \begin{bmatrix} M^{xx}_{0:p} & M^{xy}_{0:p} \\ M^{yx}_{0:p} & M^{yy}_{0:p} \end{bmatrix}, \mathbf{I} := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}, \quad (6.3)$$

in which O is a matrix of zeros, to write the cross-sectional VARMA as

$$\mathbf{A}_0 \mathbf{w}_t + \mathbf{A}_1 \mathbf{w}_{t-1} + ... + \mathbf{A}_p \mathbf{w}_{t-p} = \mathbf{M}_0 \boldsymbol{\varepsilon}_t + \mathbf{M}_1 \boldsymbol{\varepsilon}_{t-1} + ... + \mathbf{M}_q \boldsymbol{\varepsilon}_{t-q} \; \forall \; t \in \mathbb{Z}, \quad (6.4)$$

in which $\{\mathbf{A}_0, \mathbf{A}_1, ..., \mathbf{A}_p\} \in \boldsymbol{\mathcal{A}}$ and $\{\mathbf{M}_0, \mathbf{M}_1, ..., \mathbf{M}_p\} \in \boldsymbol{\mathcal{M}}$ 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 Σ^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 Σ^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 τ, while cross-sections are observed at an interval t that is a multiple of τ, 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 ρ^x and the subsequent n_y elements equal to ρ^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,

$$(\mathbf{I}+\mathbf{A}_\mathbf{C})\mathbf{w}_t + \mathbf{A}_1\mathbf{w}_{t-1} + ... + \mathbf{A}_p\mathbf{w}_{t-p} = \boldsymbol{\varepsilon}_t + \mathbf{M}_1\boldsymbol{\varepsilon}_{t-1} + ... + \mathbf{M}_q\boldsymbol{\varepsilon}_{t-q} \; \forall \; t \in \mathbb{Z}. \tag{6.5}$$

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:

$$\mathbf{w}_t + \mathbf{S}\mathbf{A}_1\mathbf{w}_{t-1} + ... + \mathbf{S}\mathbf{A}_p\mathbf{w}_{t-p} = \mathbf{S}\boldsymbol{\varepsilon}_t + \mathbf{S}\mathbf{M}_1\boldsymbol{\varepsilon}_{t-1} + ... + \mathbf{S}\mathbf{M}_q\boldsymbol{\varepsilon}_{t-q} \; \forall \; t \in \mathbb{Z}, \tag{6.6}$$

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

$$\mathbf{w}_t + \mathbf{H}_1\mathbf{w}_{t-1} + ... + \mathbf{H}_p\mathbf{w}_{t-p} = \boldsymbol{\epsilon}_t + \mathbf{M}_1\boldsymbol{\epsilon}_{t-1} + ... + \mathbf{M}_q\boldsymbol{\epsilon}_{t-q} \; \forall \; t \in \mathbb{Z}. \tag{6.7}$$

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

$$\mathbf{H}(L)\mathbf{w}_t = \mathbf{M}(L)\boldsymbol{\epsilon}_t \; \forall \; t \in \mathbb{Z}, \tag{6.8}$$

by defining L as a lag operator that has the effect that $L\mathbf{w}_t = \mathbf{w}_{t-1}$, and where $\mathbf{H}(L) = \mathbf{H}_0 + \mathbf{H}_1 L + ... + \mathbf{H}_p L^p$ and $\mathbf{M}(L) = \mathbf{M}_0 + \mathbf{M}_1 L + ... + \mathbf{M}_q L^q$ 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.[5] 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,

[5] See for example Blasques et al. (2018) for results on the relation between filters and *DGP*s.

these properties are directly related to the properties of the model itself and the range of parameter values considered.[6] 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

$$\mathbf{w}_t = \mathbf{\Phi}\mathbf{w}_{t-1} + \boldsymbol{\epsilon}_t \ \forall \ t \in \mathbb{Z}, \tag{6.9}$$

with $\mathbf{\Phi}z = -\mathbf{H}_1 Lz - ... - \mathbf{H}_p L^p z$. 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:

$$\mathbf{w}_t = \{\boldsymbol{\epsilon}_t + \mathbf{\Phi}\boldsymbol{\epsilon}_{t-1} + \mathbf{\Phi}^2\boldsymbol{\epsilon}_{t-2} + ... + \mathbf{\Phi}^\infty\boldsymbol{\epsilon}_{t-\infty}\} \ \forall \ t \in \mathbb{Z}. \tag{6.10}$$

For the sequence $\{\mathbf{\Phi}, \mathbf{\Phi}^1, \mathbf{\Phi}^2, ..., \mathbf{\Phi}^\infty\}$ to converge, it is necessary and sufficient that all the moduli of the eigenvalues of $\mathbf{\Phi}$ 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.

[6] 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

$$\mathbf{w}_t = \boldsymbol{\Psi}(L)\boldsymbol{\epsilon}_t = \{\boldsymbol{\Psi}_0\boldsymbol{\epsilon}_t + \boldsymbol{\Psi}_1\boldsymbol{\epsilon}_{t-1} + \boldsymbol{\Psi}_2\boldsymbol{\epsilon}_{t-2} + ... + \boldsymbol{\Psi}_\infty\boldsymbol{\epsilon}_{t-\infty}\} \ \forall \ t \in \mathbb{Z}. \quad (6.11)$$

with the matrices $\boldsymbol{\Psi}_k$ generated by

$$\mathbf{H}(z)\boldsymbol{\Psi}(z) = \mathbf{M}(z). \quad (6.12)$$

The conditions

$$\mathbf{H}_0 := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}, \ \mathbf{M}_0 := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}, \text{ imply that } \boldsymbol{\Psi}_0 := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}. \quad (6.13)$$

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

$$\boldsymbol{\epsilon}_t = \boldsymbol{\Pi}(L)\mathbf{w}_t = \{\boldsymbol{\Pi}_0\mathbf{w}_{t-1} + \boldsymbol{\Pi}_1\mathbf{w}_{t-1} + \boldsymbol{\Pi}_2\mathbf{w}_{t-2} + ... + \boldsymbol{\Pi}_\infty\mathbf{w}_{t-\infty}\} \ \forall \ t \in \mathbb{Z}. \quad (6.14)$$

The matrices $\boldsymbol{\Pi}_k$ are generated by

$$\mathbf{M}(z)\boldsymbol{\Pi}(z) = \mathbf{H}(z). \quad (6.15)$$

The conditions eq. (6.13) imply that

$$\boldsymbol{\Pi}_0 := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}. \quad (6.16)$$

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:

$$\mathbf{w}_t = \mathbf{H}^{-1}(L)\{\mathbf{M}(L)\boldsymbol{\epsilon}_t\} = \mathbf{M}(L)\boldsymbol{\Xi}_t \; \forall \; t \in \mathbb{Z}, \tag{6.17}$$

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

$$\mathbf{p}(t) = \begin{cases} \zeta, t = 0, \\ 0, t \neq 0. \end{cases} \quad \forall\, t \in \mathbb{Z}.$$

ζ 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

$$\mathbf{w}_t = \mathbf{\Psi}(L)\{\mathbf{p}(t)\mathbf{e}\} \; \forall \; t \in \mathbb{Z}. \tag{6.18}$$

This system is inactive until $t = 0$, after which it generates the sequence $\{\mathbf{\Psi}_0\mathbf{e}, \mathbf{\Psi}_1\mathbf{e}, ..., \mathbf{\Psi}_\infty\mathbf{e}, \}$. 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

$$G = k(\mathbf{v}_i, \mathbf{v}_j; b) = \exp\left(\frac{-\|\mathbf{v}_i - \mathbf{v}_j\|^2}{b}\right), \tag{6.19}$$

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:

$$D = G - I = k(\mathbf{v}_i, \mathbf{v}_j; b) = \exp\left(\frac{-\|\mathbf{v}_i - \mathbf{v}_j\|^2}{b}\right) - I, \tag{6.20}$$

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

$$k(\mathbf{v}_i, \mathbf{v}_j; b)^2 \leq k(\mathbf{v}_i, \mathbf{v}_i; b)k(\mathbf{v}_j, \mathbf{v}_j; b) \; \forall \; (\mathbf{v}_i, \mathbf{v}_j; b > 0) \in \mathcal{X} \times \mathcal{X} \times \mathcal{B}.$$

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

$$k(\mathbf{v}_i, \mathbf{v}_j; b)^2 \geq k(\mathbf{v}_i, \mathbf{v}_i; b)k(\mathbf{v}_j, \mathbf{v}_j; b) \; \forall \; (\mathbf{v}_i, \mathbf{v}_j; b < 0) \in \mathcal{X} \times \mathcal{X} \times \mathcal{B}.$$

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 caries 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 $\mathbf{v} = N/25, N \in \{1, 2, ..., 25\}$. 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 $\|\mathbf{b}\| \to 0$, 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

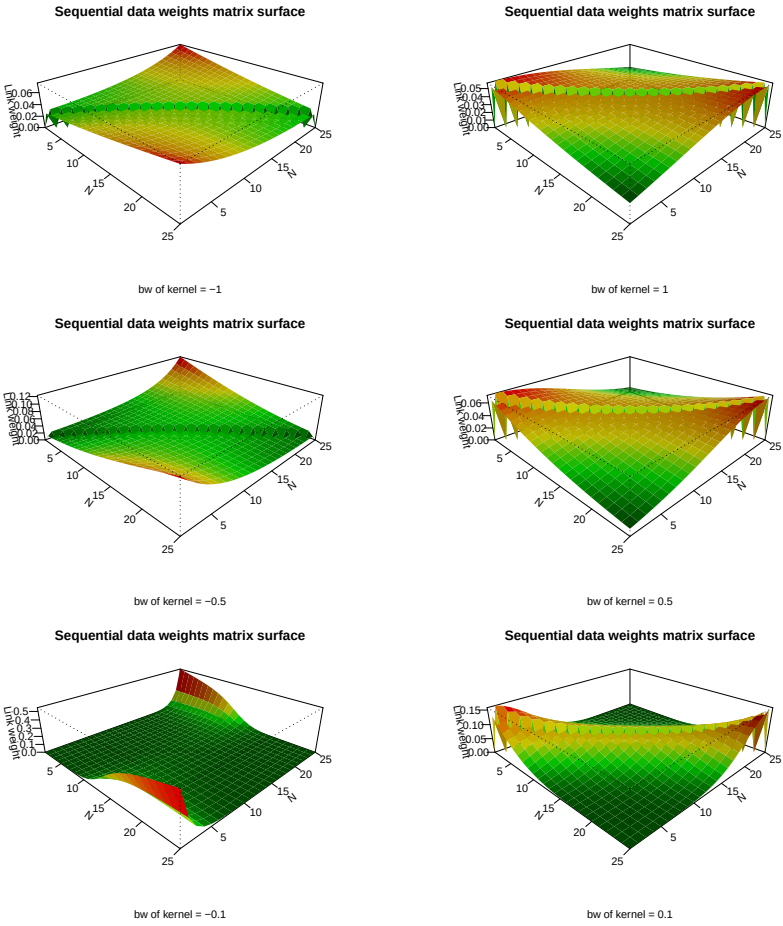

Figure 6.2: Surfaces of spatial weights produced using the kernel approach for different bandwidth values, on identical data produced with $N/25, N \in \{1, 2, ..., 25\}$.

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:

$$\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta}) + \lambda\gamma(\boldsymbol{\theta}), \tag{6.21}$$

with the ML criterion defined as

$$\begin{aligned} Q_T := \ell_T(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta}) &= \sum_t^T \ell_t(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta}), \\ \ell_t(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta}) &= \ln p_\varepsilon(\mathbf{w}_t - f(\mathbf{v}, \mathbf{w}_t; \boldsymbol{\theta}), \boldsymbol{\Sigma}; \boldsymbol{\nu}), \end{aligned} \tag{6.22}$$

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

$$\lambda\gamma(\boldsymbol{\theta}) = 1/\sqrt{NT}\sum|\mathbf{M}|^2. \tag{6.23}$$

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

$$\ell_t(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta}) = D(\boldsymbol{\theta}^{\mathbf{S}}, \mathbf{v}) + K(\boldsymbol{\theta}) + E(\boldsymbol{\theta}, \mathbf{v}, \mathbf{w}_t), \tag{6.24}$$

where $D(\boldsymbol{\theta}^{\mathbf{S}}, \mathbf{v})$ is the log determinant of

$$D(\boldsymbol{\theta}^{\mathbf{S}}, \mathbf{v}) := \ln\det\mathbf{S}\Big(\boldsymbol{\theta}^{\rho}, \mathbf{C}(\mathbf{v}; \mathbf{b})\Big), \tag{6.25}$$

with $\mathbf{S}\Big(\boldsymbol{\theta}^{\rho}, \mathbf{C}(\mathbf{v}; \mathbf{b})\Big)$ as the spatial multiplier matrix conditional on data $\mathbf{v}$ and bandwidth parameters $\mathbf{b}$ that we defined as

$$\mathbf{S}(\boldsymbol{\theta}^{\rho}, \mathbf{C}(\mathbf{v}; \mathbf{b})) = \Big(\mathbf{I} - \boldsymbol{\rho}\circ\mathbf{C}(\mathbf{v}; \mathbf{b})\Big)^{-1}, \tag{6.26}$$

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

$$D(\boldsymbol{\theta}^{\mathbf{S}}, \mathbf{v}) = \ln\det \mathbf{S}\Big(\boldsymbol{\theta}^{\boldsymbol{\rho}}, \mathbf{C}(\mathbf{v};\mathbf{b})\Big) = \ln\det S^{x}\Big(\rho^{x}, C_{n_x}(\mathbf{v}; b^{x})\Big) + \ln\det S^{y}\Big(\rho^{y}, C_{n_y}(\mathbf{v}; b^{y})\Big) \tag{6.27}$$

and each determinant is evaluated over $S\Big(\rho, C(\mathbf{v};b)\Big) = \Big(I - \rho C(\mathbf{v};b)\Big)^{-1}$ with $\rho C(\mathbf{v};b)$ as the diagonal blocks of

$$\boldsymbol{\rho} \circ \mathbf{C}(\mathbf{v};\mathbf{b}) = \begin{bmatrix} \rho^{x} C_{n_x}(\mathbf{v};b^{x}) & O_{n_x} \\ O_{n_y} & \rho^{y} C_{n_y}(\mathbf{v};b^{y}) \end{bmatrix}. \tag{6.28}$$

$K(\boldsymbol{\theta})$ is a constant, that can be similarly expressed as a sum

$$K(\boldsymbol{\theta}) := \ln\Gamma\Big((\nu+N)/2\Big)\Big[\det\Sigma^{\frac{1}{2}}(\nu\pi)^{\frac{N}{2}}\Gamma\left(\nu/2\right)\Big]^{-1}, \tag{6.29}$$

for each $(\nu,\Sigma) \in ((\nu^{\mathbf{x}},\Sigma^{\mathbf{x}}),(\nu^{\mathbf{y}},\Sigma^{\mathbf{y}}))$. Finally, the random element $E(\boldsymbol{\theta},\mathbf{v},\mathbf{w}_t)$ can naturally be defined as the sum

$$E(\boldsymbol{\theta},\mathbf{v},\mathbf{w}_t) := -\frac{1}{2}(\nu^{\mathbf{x}}+N)\ln\Big(1+\nu^{\mathbf{x}-1}(\mathbf{x}_t - f^{\mathbf{x}}(\mathbf{v},\mathbf{w}_t;\boldsymbol{\theta}^{\mathbf{x}}))'\Sigma^{\mathbf{x}-1}\left(\mathbf{x}_t - f^{\mathbf{x}}(\mathbf{v},\mathbf{w}_t;\boldsymbol{\theta}^{\mathbf{x}})\right)\Big) -\frac{1}{2}(\nu^{\mathbf{y}}+N)\ln\Big(1+\nu^{\mathbf{y}-1}(\mathbf{y}_t - f^{\mathbf{y}}(\mathbf{v},\mathbf{w}_t;\boldsymbol{\theta}^{\mathbf{y}}))'\Sigma^{\mathbf{y}-1}\left(\mathbf{y}_t - f^{\mathbf{y}}(\mathbf{v},\mathbf{w}_t;\boldsymbol{\theta}^{\mathbf{y}})\right)\Big). \tag{6.30}$$

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 $\boldsymbol{\theta}^{\boldsymbol{\rho}}$. 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 cased 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 $T \to \infty$, 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

$$\mathbf{y}_t = 0.6C(\mathbf{x}; b)\mathbf{y}_t - 0.35\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t + 0.25\boldsymbol{\varepsilon}_{t-1}, \tag{6.31}$$

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.

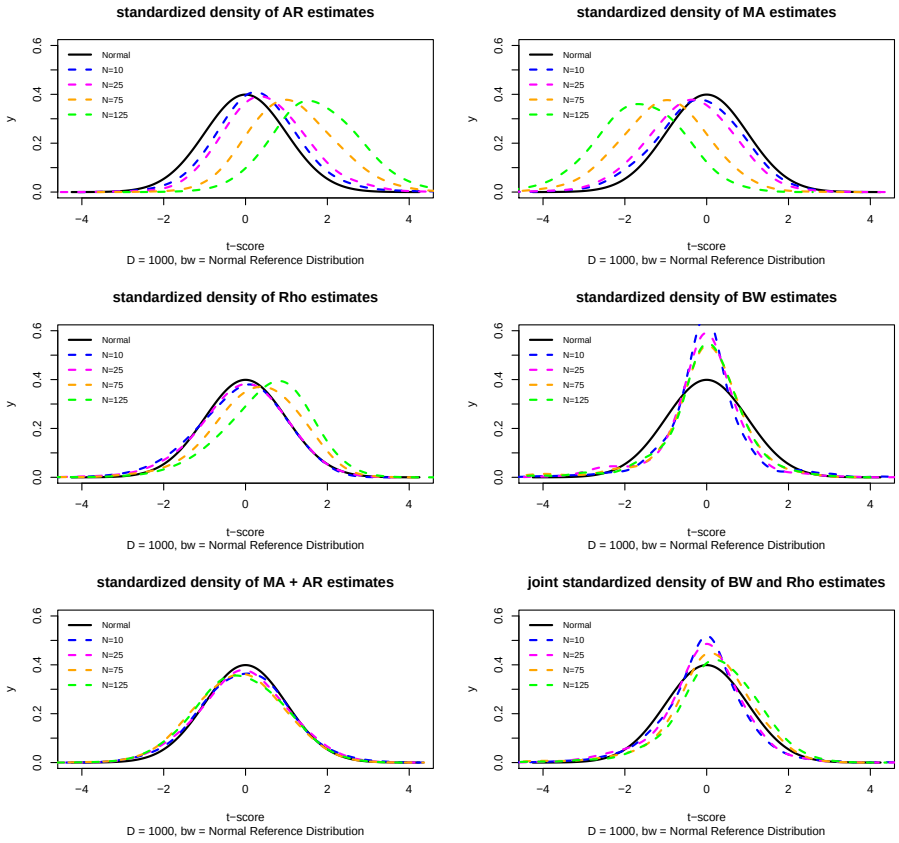

Figure 6.3: Penalized small sample distributions of the correctly specified SARMA, bandwidth of the spatial kernel matrix in the DGP set to .15

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).[7] 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

[7] https://data.worldbank.org/data-catalog/indonesia-database-for-policy-and-economic-research. Some district names and boarders 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 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 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[8] suggest that we explain

[8]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),

$$\hat{R}^2 = 1 - \frac{\sum_1^T |\mathbf{w}_T - f(\mathbf{w}_T; \hat{\boldsymbol{\theta}})|^2}{\sum_1^T |\mathbf{w}_T - f(\mathbf{w}_T; \boldsymbol{\theta}_{\boldsymbol{\theta}=0})|^2}, \tag{6.32}$$

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 $PM_{2.5} > 6$, 0.732 versus 0.722 at $PM_{2.5} > 10$). In both cases the SVARMA, however, uses less ARMA parameters (29 versus 34 at $PM_{2.5} > 6$, 27 versus 31 at $PM_{2.5} > 10$.) 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 $PM_{2.5} > 6$ 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 $1, ..., r$ 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 $PM_{2.5} > 6$ 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**	-0.092***	-0.047***
	(-2.57)	(-2.866)	(-2.391)
$\phi\ pol_{t-2}$	-0.070***	-0.063***	
	(-4.381)	(-2.969)	
$\phi\ pol_{t-3}$	0.026*		0.054**
	(1.652)		(2.507)
$\phi\ bot_{t-1}$		-0.108*	0.089***
		(-1.94)	(2.577)
$\phi\ bot_{t-2}$		-0.139***	-0.129***
		(-4.736)	(-2.791)
$\phi\ bot_{t-3}$	-0.039**	-0.099***	-0.141***
	(-2.119)	(-3.544)	(-3.534)
$\phi\ exp_{t-1}$		0.071***	-0.374***
		(2.964)	(-12.805)
$\phi\ exp_{t-2}$		0.158***	
		(4.453)	
$\phi\ exp_{t-3}$		0.074***	
		(3.14)	
$M\ pol_{t-1}$	-0.515***	0.126***	
	(-15.292)	(3.233)	
$M\ pol_{t-2}$			
$M\ pol_{t-3}$		-0.052*	-0.082**
		(-1.814)	(-2.131)
$M\ bot_{t-1}$	-0.038*	-0.253***	
	(-1.94)	(-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***	-0.387***
		(-3.621)	(-10.705)
$M\ exp_{t-3}$			-0.147***
			(-4.273)
ρ	0.812***	0.305***	0.327***
	(27.765)	(3.177)	(2.976)
b	0.088	0.18	0.229
σ	0.119	0.129	0.176
ν	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.

	$PM_{2.5} > 6$			$PM_{2.5} > 10$		
Percentiles:	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.

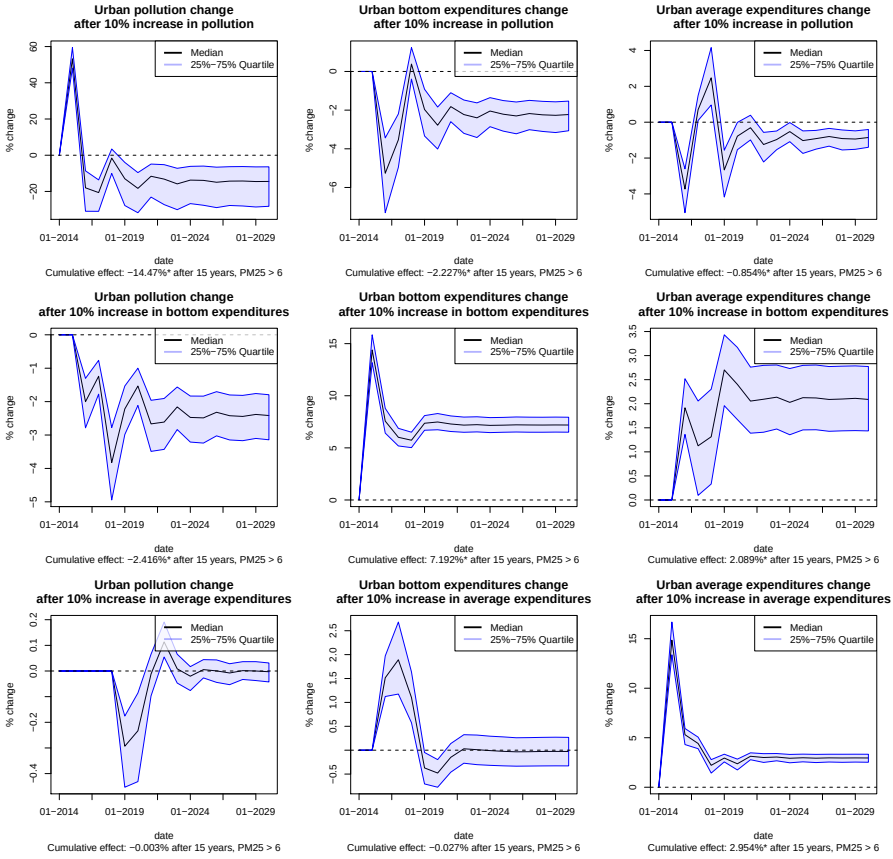

Figure 6.4: IRF plots for exogenous shocks in pollution, bottom household expenditures and average household expenditures $PM_{2.5}$ >6. Effects that exclude zero in the final year, are marked by *.

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 tp 10 $PM_{2.5}$ range.[9] By population weighting the effect of the analyzed increase in pollution levels, an estimated

[9] 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.[10] 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

[10]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

$$\begin{bmatrix} \mathbf{x}_t + H_1^{xx}\mathbf{x}_{t-1} + H_1^{xy}\mathbf{y}_{t-1} + ... + H_p^{xx}\mathbf{x}_{t-p} + H_p^{xy}\mathbf{y}_{t-p} \\ \mathbf{y}_t + H_1^{yx}\mathbf{y}_{t-1} + H_1^{yy}\mathbf{x}_{t-1} + ... + H_p^{yx}\mathbf{y}_{t-p} + H_p^{yy}\mathbf{x}_{t-p} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\epsilon}_t^x + M_1^{xx}\boldsymbol{\epsilon}_{t-1}^x + ... + M_q^{xx}\boldsymbol{\epsilon}_{t-q}^x \\ \boldsymbol{\epsilon}_t^y + M_1^{yy}\boldsymbol{\epsilon}_{t-1}^y + ... + M_q^{yy}\boldsymbol{\epsilon}_{t-1}^y \end{bmatrix} \forall\ t \in \mathbb{Z}. \tag{6.33}$$

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

$$\mathbf{H}_0 := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix},\ \mathbf{H}_{1:p} := \begin{bmatrix} H_{1:p}^{xx} & H_{1:p}^{xy} \\ H_{1:p}^{yx} & H_{1:p}^{yy} \end{bmatrix}, \tag{6.34}$$

$$\mathbf{M}_0 := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix},\ \mathbf{M}_{1:p} := \begin{bmatrix} M_{1:p}^{xx} & O_{n_x} \\ O_{n_y} & M_{1:p}^{yy} \end{bmatrix}.$$

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:

$$\left[\begin{array}{l} \boldsymbol{\epsilon}_t^x + M_1^{xx}\boldsymbol{\epsilon}_{t-1}^x + M_1^{xy}\boldsymbol{\epsilon}_{t-1}^x + ... + M_q^{xx}\boldsymbol{\epsilon}_{t-q}^x + M_q^{xy}\boldsymbol{\epsilon}_{t-q}^x \\ \boldsymbol{\epsilon}_t^y + M_1^{yx}\boldsymbol{\epsilon}_{t-1}^y + M_1^{yy}\boldsymbol{\epsilon}_{t-1}^y + ... + M_q^{yx}\boldsymbol{\epsilon}_{t-q}^y + M_q^{yy}\boldsymbol{\epsilon}_{t-q}^y \end{array} \right]. \tag{6.35}$$

The matrix representation results from

$$\mathbf{H}_0 := \left[\begin{array}{cc} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{array} \right], \ \mathbf{H}_{1:p} := \left[\begin{array}{cc} H_{1:p}^{xx} & H_{1:p}^{xy} \\ H_{1:p}^{yx} & H_{1:p}^{yy} \end{array} \right], \tag{6.36}$$

$$\mathbf{M}_0 := \left[\begin{array}{cc} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{array} \right], \ \mathbf{M}_{1:p} := \left[\begin{array}{cc} M_{1:p}^{xx} & M_{1:p}^{xy} \\ M_{1:p}^{yx} & M_{1:p}^{yy} \end{array} \right].$$

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.[11] From a

[11] The unrestricted model with contemporaneous effects between variables results from

$$\mathbf{H}_{0:p} := \left[\begin{array}{cc} H_{0:p}^{xx} & H_{0:p}^{xy} \\ H_{0:p}^{yx} & H_{0:p}^{yy} \end{array} \right], \ \mathbf{M}_{0:p} := \left[\begin{array}{cc} M_{0:p}^{xx} & M_{0:p}^{xy} \\ M_{0:p}^{yx} & M_{0:p}^{yy} \end{array} \right],$$

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,

$$\mathbf{w}_t = \mathbf{M}(L)\{\mathbf{H}^{-1}(L)\boldsymbol{\epsilon}_t\} = \mathbf{M}(L)\boldsymbol{\Xi}_t \ \forall \ t \in \mathbb{Z},$$

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,

$$\boldsymbol{\Xi}_t = \boldsymbol{\epsilon}_t - \{\mathbf{H}_1\boldsymbol{\Xi}_{t-1} + ... + \mathbf{H}_p\boldsymbol{\Xi}_{t-p}\},$$

then,

$$\mathbf{w}_t = \mathbf{M}_0\boldsymbol{\Xi}_t + \mathbf{M}_1\boldsymbol{\Xi}_{1t-1} + ... + \mathbf{M}_{p-1}\boldsymbol{\Xi}_{t-p+1}.$$

By defining the set of p state variables:

$$\begin{aligned}
\boldsymbol{\Xi}_{1t} &= \boldsymbol{\Xi}_t, \\
\boldsymbol{\Xi}_{2t} &= \boldsymbol{\Xi}_{t-1}, \\
&\vdots \\
\boldsymbol{\Xi}_{p_t} &= \boldsymbol{\Xi}_{t-p+1}.
\end{aligned}$$

and rewriting the Markov Chain in terms of the left hand side variables:

$$\mathbf{w}_{1t} = \boldsymbol{\epsilon}_t - \{\mathbf{H}_1\boldsymbol{\Xi}_{1t-1} + ... + \mathbf{H}_p\boldsymbol{\Xi}_{p_{t-1}}\}.$$

we can use the state vector $\boldsymbol{\Xi}_t = \left[\boldsymbol{\Xi}_{1t}, \boldsymbol{\Xi}_{2t}, ..., \boldsymbol{\Xi}_{p_t}\right]'$to write the system

in which the connectivity matrices that generate the off-diagonal blocks $H^{xy}_{0:p}$ and $H^{yx}_{0:p}$ 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 $\mathbf{O} = 0 \circ \mathbf{I}$:

$$\begin{bmatrix} \boldsymbol{\Xi}_{1t} \\ \boldsymbol{\Xi}_{2t} \\ \vdots \\ \boldsymbol{\Xi}_{pt} \end{bmatrix} = \begin{bmatrix} -\mathbf{H}_1 & \dots & -\mathbf{H}_{p-1} & -\mathbf{H}_p \\ \mathbf{I} & \dots & -\mathbf{O} & \mathbf{O} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{O} & \dots & \mathbf{I} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Xi}_{1t-1} \\ \boldsymbol{\Xi}_{2t-1} \\ \vdots \\ \boldsymbol{\Xi}_{pt-1} \end{bmatrix} + \begin{bmatrix} \mathbf{I} \\ \mathbf{O} \\ \vdots \\ \mathbf{O} \end{bmatrix} \boldsymbol{\epsilon}(t),$$

with the sparse matrix on the right side of the equality being the companion matrix that has the following accompanying measurement equation

$$\mathbf{w}_t = \mathbf{M}_0 \boldsymbol{\Xi}_{1t} + ... + \mathbf{M}_{p-1} \boldsymbol{\Xi}_{pt} \; \forall \; t \in \mathbb{Z}.$$

Stability then be expressed in terms of the companion matrix $\boldsymbol{\Phi}$. 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 $\boldsymbol{\Phi}$ must lie within the unit circle:

$$\det(\mathbf{I} - \boldsymbol{\Phi}(z)) = \det(\mathbf{H}(z)) = \det(\mathbf{H}_0 + \mathbf{H}_1 + ... + \mathbf{I} + \mathbf{H}_p z^p) \neq 0 \; \forall \; |z| \leq 1.$$

Note that if $\boldsymbol{\rho} = 0$, $\mathbf{S} = (\mathbf{I} + \mathbf{O})^{-1} = \mathbf{I}$, as an effect $\mathbf{H} = \mathbf{A}$, which gives

$$\det(\mathbf{I} - \boldsymbol{\Phi}(z)) = \det(\mathbf{A}(z)) = \det(\mathbf{A}_0 + \mathbf{A}_1 + ... + \mathbf{I} + \mathbf{A}_p z^p) \neq 0 \; \forall \; |z| \leq 1.$$

Finally, this only differs from the standard condition cited in VARMA literature that

$$\det(I - \Phi(z)) = \det(A(z)) = \det(A_0 + A_1 + ... + I + A_p z^p) \neq 0 \; \forall \; |z| \leq 1,$$

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 $\boldsymbol{\rho}$ 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

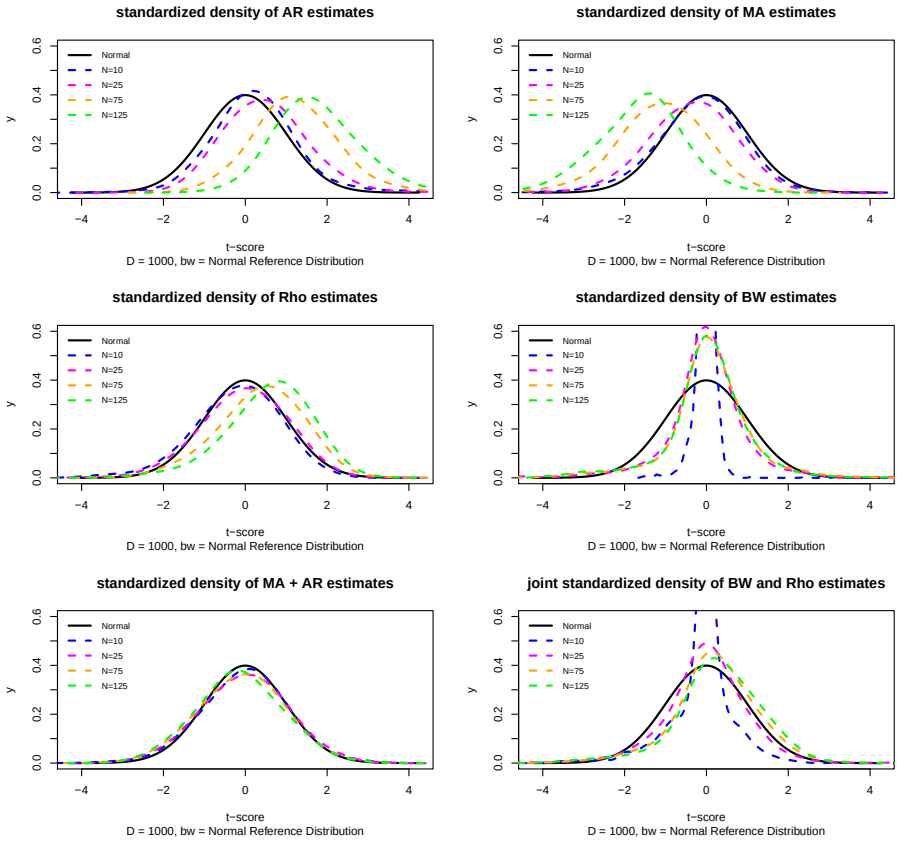

Figure 6.5: Unpenalized small sample distributions of the correctly specified SARMA, bandwidth of the spatial kernel matrix in the DGP set to .15.

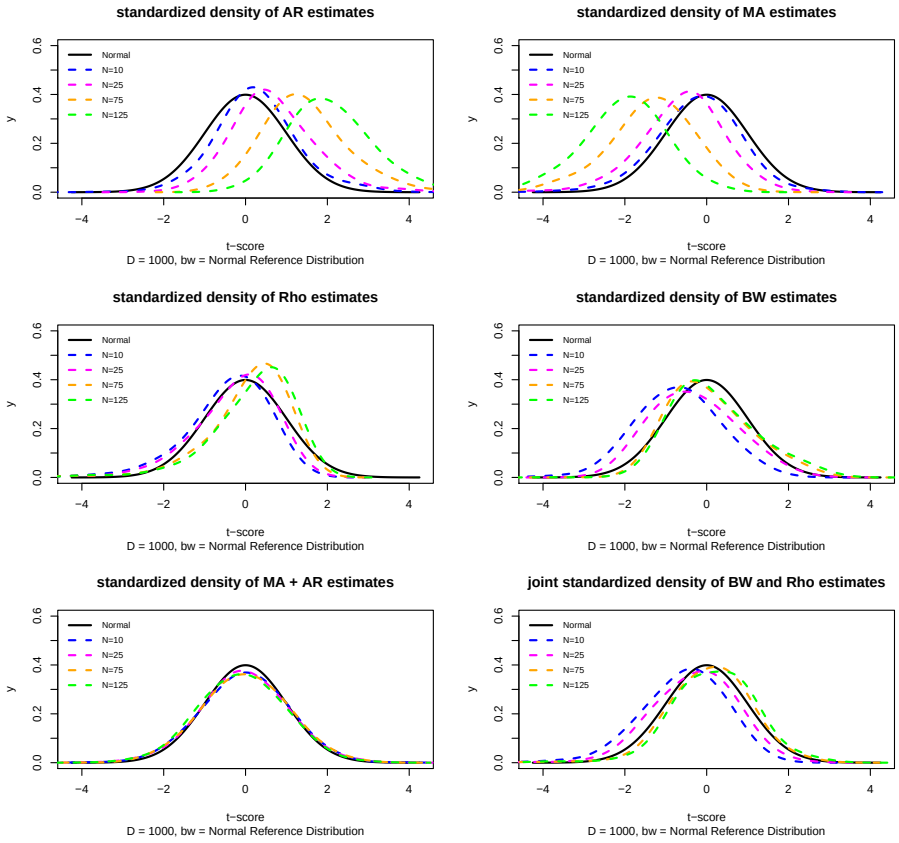

Figure 6.6: Penalized small sample distributions of the correctly specified SARMA, bandwidth of the spatial kernel matrix in the DGP set to 2.

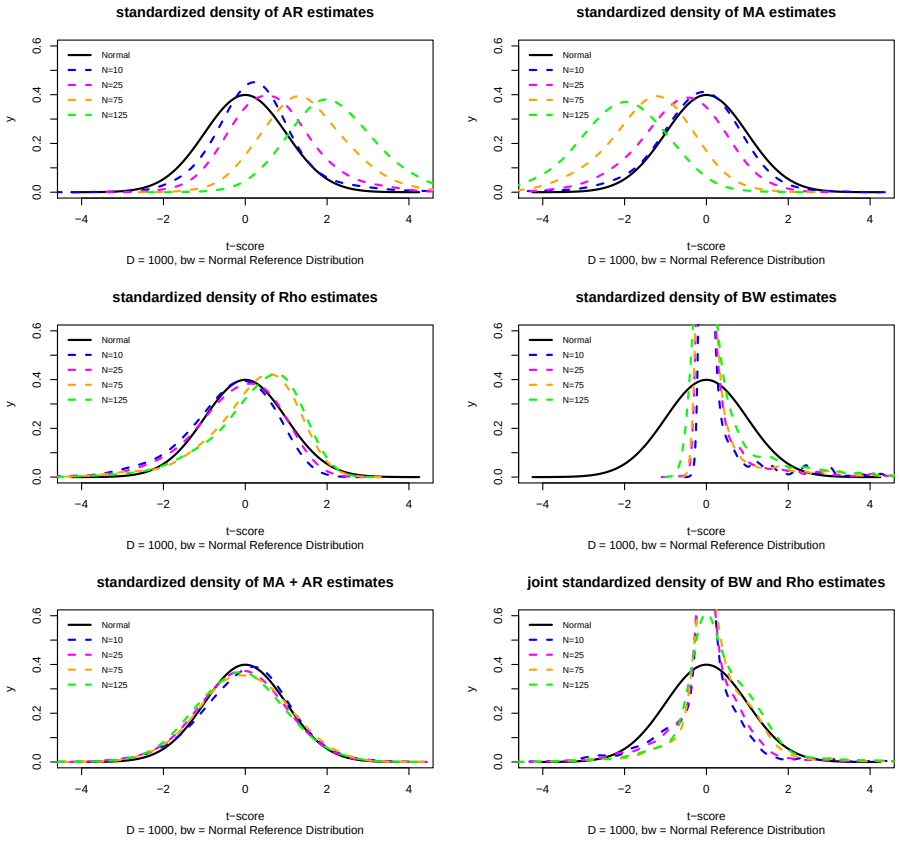

Figure 6.7: Unpenalized small sample distributions of the correctly specified SARMA, bandwidth of the spatial kernel matrix in the DGP set to 2.

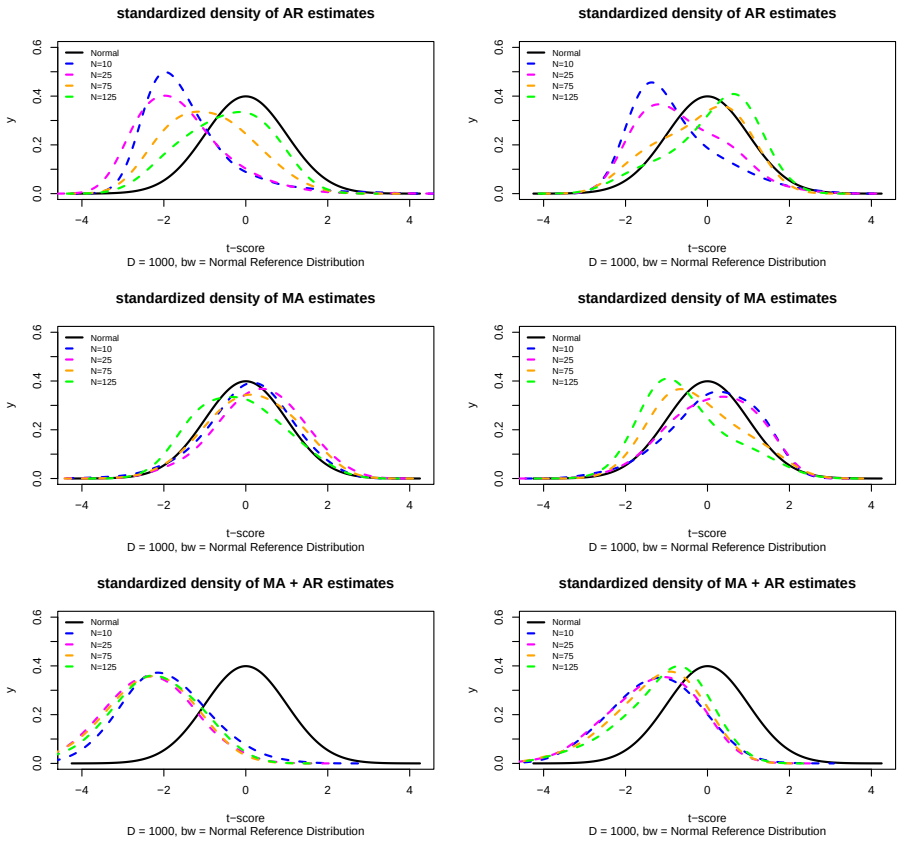

Figure 6.8: Unpenalized small sample distributions of the miss-specified ARMA, when the *true* process is an SARMA with bandwidth of the spatial kernel matrix set to .15 (left) and 2 (right).

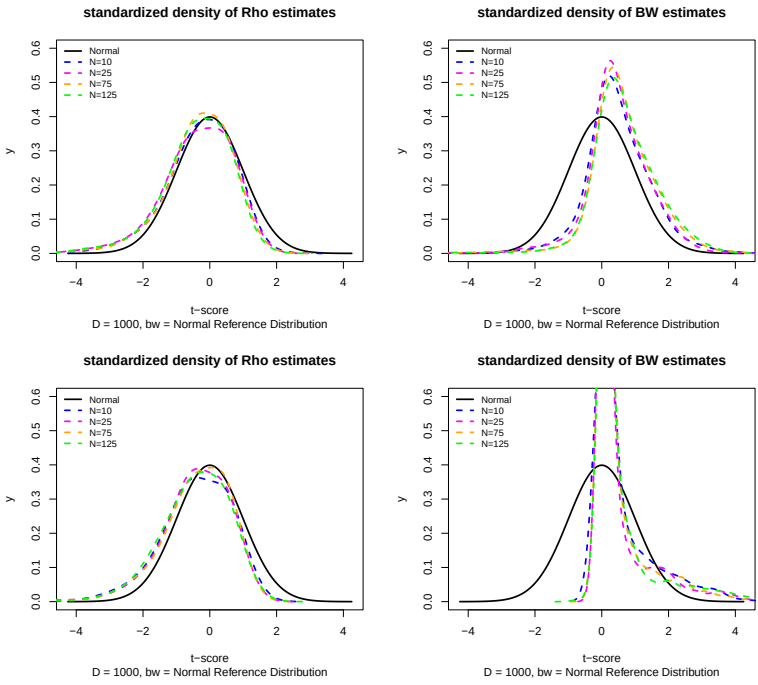

Figure 6.9: Penalized (upper) small sample distributions of bandwidth and spatial parameter in the SARMA, when the *true* process is a cross-sectional ARMA with zero spatial effects. The bandwidth density is centered around 0, note that the kernel structure is not identified at this value.

6.7.4 Pollution data

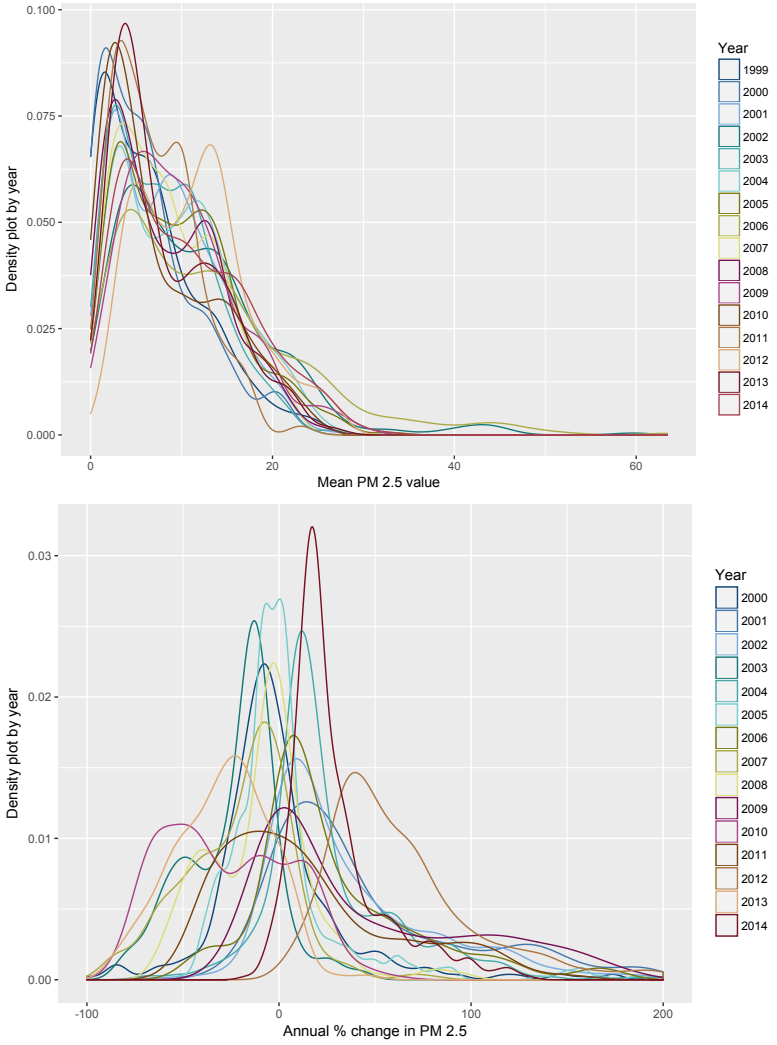

Figure 6.10: Densities of pollution levels (left) and changes in pollution (right) for 219 areas with an urban patch of over 5,000 people and densities of 300 per square kilometer or higher.

6.7.5 Additional regression results

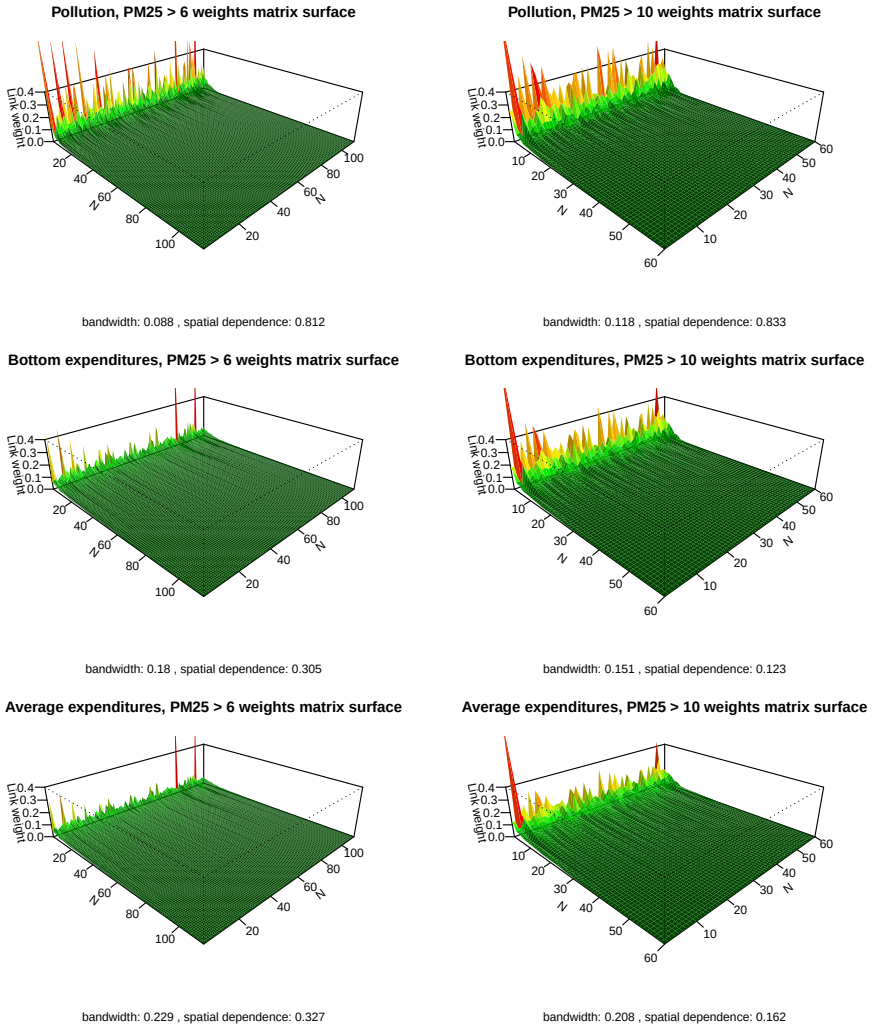

Figure 6.11: Surfaces of estimated spatial weights, ordered by link strengths (observations in no particular order), revealing the different links and links strengths across the different channels of the SVARMA structure.

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***	-0.155***	0.124*
	(-6.407)	(-2.745)	(1.862)
$\phi\ pol_{t-2}$	-0.201***	-0.101***	
	(-6.171)	(-3.215)	
$\phi\ pol_{t-3}$			0.078***
			(3.081)
$\phi\ bot_{t-1}$	0.101**	-0.119**	
	(2.019)	(-2.02)	
$\phi\ bot_{t-2}$		-0.138***	-0.123***
		(-4.645)	(-2.655)
$\phi\ bot_{t-3}$	-0.056**	-0.094***	-0.156***
	(-2.362)	(-3.333)	(-3.825)
$\phi\ exp_{t-1}$		0.069***	-0.277***
		(2.884)	(-4.884)
$\phi\ exp_{t-2}$		0.159***	
		(4.557)	
$\phi\ exp_{t-3}$		0.072***	
		(3.099)	
$M\ pol_{t-1}$	-0.142*	0.145**	-0.196***
	(-1.853)	(2.444)	(-2.7)
$M\ pol_{t-2}$	-0.100**		0.129***
	(-2.236)		(2.788)
$M\ pol_{t-3}$	0.068*	-0.085***	-0.088**
	(1.806)	(-2.877)	(-2.225)
$M\ bot_{t-1}$	-0.148***	-0.235***	0.095**
	(-2.585)	(-3.855)	(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***	-0.356***
		(-3.658)	(-8.198)
$M\ exp_{t-3}$			-0.137***
			(-3.75)
σ	0.109	0.087	0.100
ν	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***	-0.542***	-0.413***
	(-17.043)	(-4.875)	(-2.811)
$\phi\ pol_{t-2}$	-0.234***	-0.320***	-0.204**
	(-4.764)	(-4.866)	(-2.459)
$\phi\ pol_{t-3}$	0.064*		
	(1.883)		
$\phi\ bot_{t-1}$	0.138*	-0.440***	-0.248**
	(2.327)	(-11.281)	(-2.558)
$\phi\ bot_{t-2}$	0.083**		
	(2.374)		
$\phi\ bot_{t-3}$		-0.061*	
		(-1.85)	
$\phi\ exp_{t-1}$		0.151***	-0.172**
		(3.049)	(-2.654)
$\phi\ exp_{t-2}$		0.056*	-0.219***
		(1.837)	(-5.579)
$\phi\ exp_{t-3}$			
$Mpol_{t-1}$		0.569***	0.342**
		(4.903)	(2.244)
$Mpol_{t-2}$	-0.184***		
	(-3.11)		
$Mpol_{t-3}$	-0.134***	-0.294***	-0.208***
	(-2.594)	(-4.529)	(-2.751)
$Mbot_{t-1}$	-0.188***		0.348***
	(-2.644)		(3.247)
$Mbot_{t-2}$		-0.344***	-0.154**
		(-6.453)	(-2.138)
$Mbot_{t-3}$			
$Mexp_{t-1}$		-0.106**	-0.303***
		(-2.021)	(-4.569)
$Mexp_{t-2}$	-0.079***		
	(-2.579)		
$Mexp_{t-3}$	0.054*		-0.286***
	(1.789)		(-6.83)
ρ			
b			
σ	0.094	0.079	0.101
ν	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**	-0.150***	-0.085**
	(-2.503)	(-2.803)	(-2.367)
$\phi\ pol_{t-2}$		-0.073**	-0.049
		(-2.048)	(-1.482)
$\phi\ pol_{t-3}$	0.041*	-0.045	
	(1.773)	(-1.446)	
$\phi\ bot_{t-1}$	0.092**		
	(2.446)		
$\phi\ bot_{t-2}$		-0.271***	-0.200***
		(-4.573)	(-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***	0.198***	
	(-6.212)	(3.19)	
$M\ pol_{t-2}$	-0.105**		
	(-2.33)		
$M\ pol_{t-3}$	0.120***		
	(2.96)		
$M\ bot_{t-1}$	-0.146***	-0.447***	
	(-3.43)	(-12.057)	
$M\ bot_{t-2}$		0.194***	0.260***
		(2.864)	(3.825)
$M\ bot_{t-3}$		-0.216***	
		(-4.225)	
$M\ exp_{t-1}$			-0.106
			(-1.496)
$M\ exp_{t-2}$		-0.139***	-0.400***
		(-2.911)	(-8.092)
$M\ exp_{t-3}$			-0.159***
			(-3.934)
ρ	0.833***	0.123	0.162
	(24.272)	(1.374)	(1.46)
b	0.118	0.151	0.208
σ	0.906	0.080	0.101
ν	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

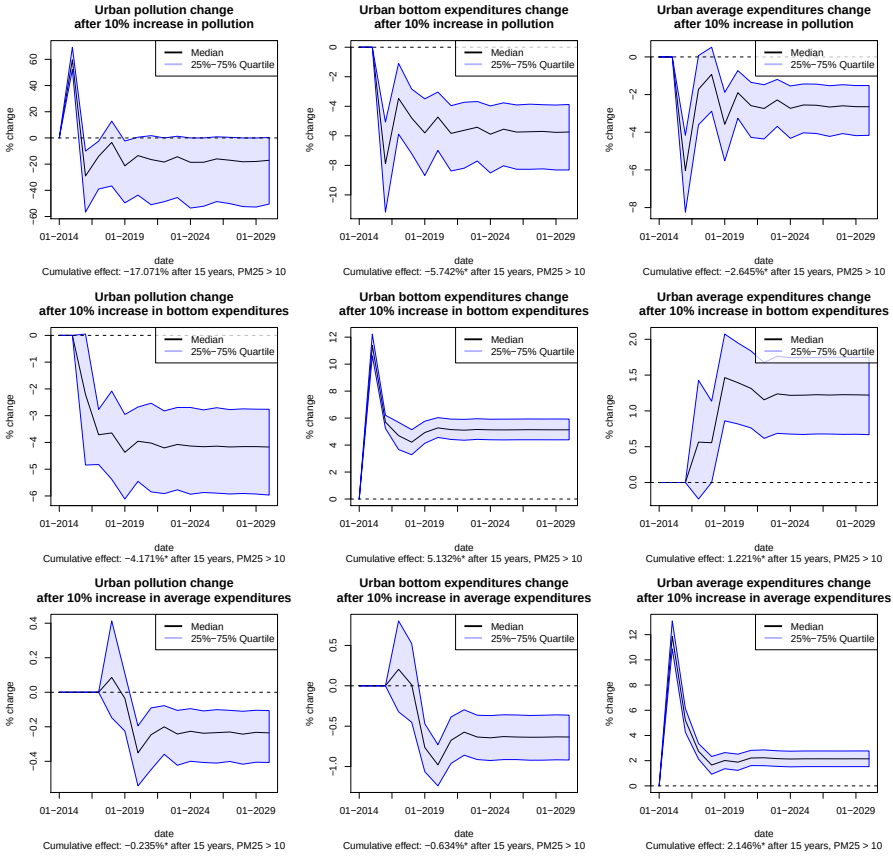

Figure 6.12: IRF plots for exogenous shocks in pollution, bottom household expenditures and average household expenditures $PM_{2.5}$ >10. Effects that exclude zero in the final year, are marked by *.

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