Policy Research Working Paper 8757

Pollution and Expenditures in a Penalized Vector Spatial Autoregressive Time Series Model with Data-Driven Networks

Bo Pieter Johannes Andrée
Phoebe Spencer
Sardar Feredun Azari
Andres Chamorro
Dieter Wang
Harun Dogo

WORLD BANK GROUP
Environment and Natural Resources Global Practice
February 2019

Policy Research Working Paper 8757

Abstract

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. It discusses the model and estimation, focusing on 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.

This paper is a product of the Environment and Natural Resources Global Practice . It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/research. The authors may be contacted at bandree@worldbank.org.

Produced by the Research Support Team

Pollution and Expenditures in a Penalized Vector Spatial Autoregressive Time Series Model with Data-Driven Networks

Bo Pieter Johannes Andrée[1a,1b,2,3,*], Phoebe Spencer[1a], Andres Chamorro[1a], Dieter Wang[1a,2,4,5], Sardar Feredun Azari[1b], and Harun Dogo[1a]

[1a] *Worldbank Group, Environment and Natural Resources Global Practice*
[1b] *Worldbank Group, Geo-Operations Support Team*
[1c] *Worldbank Group, Information and Technology Solutions*
[2] *Tinbergen Institute*
[3] *Department of Spatial Economics/SPINlab, VU University Amsterdam, Netherlands*
[4] *Department of Finance, VU University Amsterdam*
[5] *Department of Econometrics, VU University Amsterdam*
[*] *Corresponding Author: email b.p.j.andree@vu.nl or bandree@worldbank.org*

Keywords: Poverty, Pollution, Penalized Inference, Spatial Models, Impulse Response.

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.

Contemporaneous regression models alone often fail to provide conclusive evidence. This is mainly due to their inability to decide on the direction of causality. To pin down the direction, these models require the support of economic theory. Furthermore, standard approaches often confound direct effects with feedback effects. This leads many researches to adopt an instrumental approach, which often proves to be a non-trivial task. 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).[1] 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. Testing for Granger-causality in this framework results in repeatedly comparing different models estimated on different data sets. This led some to argue that the notion of Granger-causality is non-operable, as one cannot simply remove or add lags of a variable from a model and test if the effect is significant, see Hendry (2017) for discussion. In this paper 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.[2] Results that are selected based on information optimality therefore benefit from being accompanied by goodness of fit diagnostics.

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 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.

[1] 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.

[2] As measured by a divergence metric w.r.t. the correct hypothesis.

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, if 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. Notable results include that growth in household expenditures precedes pollution reduction, particularly when the expenditures of poorer households increase; that increasing pollution is followed by reduced growth in expenditures, particularly those of poorer households; and that there are significant spillovers from bottom-up growth in expenditures. We did not find evidence for top-down growth spillovers, that is, we do not find evidence that growth in average household expenditures consistently precedes subsequent growth in bottom income household expenditures.

It is important to note that the Granger-causality concept involves contrasting the probabilistic forecasting performance of a univariate and bivariate specification. This differs from the deterministic causal framework of Pearl (2000). In fact, Pearl (2000) explicitly mentions that Granger-causality is not causality, and that the concepts of "strong exogeneity" Engle et al. (1983) and Granger-causality are only statistical concepts, see also the review by Neuberg et al. (2003). In two essays critical to modern econometrics, Haavelmo (1943, 1944) argued that the very nature of economic behavior itself implies that economic theories should be stochastically formulated to make these simplifications of reality elastic enough for application. This led to the now standard approach in regression analysis to include error terms in otherwise exact relationships. In a more critical note, Kalman (1983) goes further to argue that the classical/deterministic model of reality developed in physics is inapplicable to the problems of economics. In a sequence of more recent publications, it has been shown that simple questions about important concepts in economics, such as choice and uncertainty, can even in very simplistic settings not be answered within Pearl's framework. White and Chalak (2009); White et al. (2014) extend Pearl's causal model to include optimization, equilibrium, learning concepts, and choice that are integral parts of economics, game theory, and social systems in which agents act and react under uncertainty. Under the extended Pearl-causal framework, White and Lu (2010) forge the previously missing link between Granger-causality and structural causality by showing that, given a corresponding conditional form of exogeneity, Granger-causality holds *if and only if* a corresponding form of structural causality holds. Eichler and Didelez (2010) provide conditions under which Granger non-causality implies that an intervention has had no effect. White et al. (2011) show that tests for Granger-causality can be used to test for structural causality in sequential systems, and Lu et al. (2017) produce tests for cross-section and panel data valid in a general case that does not assume linearity, monotonicity in (un)observables, or separability between observed and unobserved variables in the structural relations. White and Pettenuzzo (2014) show that instead of relying on exogeneity (weak, strong or super) conditional on the model or DGP, structurally causal effects can also be consistently estimated by relying on correct specification which is the basis for the arguments in this work. While recent literature thus suggests that structural causality can be shown to hold using statistical approaches valid in this paper's setting, it must be reminded that Granger-causality only implies that a *corresponding* form of structural causality must hold. Given the level of granularity of our data, it is not possible to infer directly what that corresponding from is. Particularly, changes in budgets of households are events that by themselves are not mechanical, hence they cannot lead mechanically to a change in pollution. The interpretation of our Granger-causal result is thus that other events that occur alongside the changes in expenditures, for example changes methods of production, must have structural effects on pollution in a manner that corresponds to our estimated results.

The remaining part of this paper is as follows. Section 2 introduces the model. Specifically, we detail the process equations, and our approach to build connectivity up from the data using kernels. Section 3 discusses the properties of the model, specifically stability, invertibility, non-uniqueness, and the IRF. Section 4 provides the tools needed for estimation and develops diagnostics to assess model fit. 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 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 concludes.

2 Spatial Vector Autoregressive Moving Average Model

This section details VARMA approaches for multiple panels that exhibit spatial feedback. Figure 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.

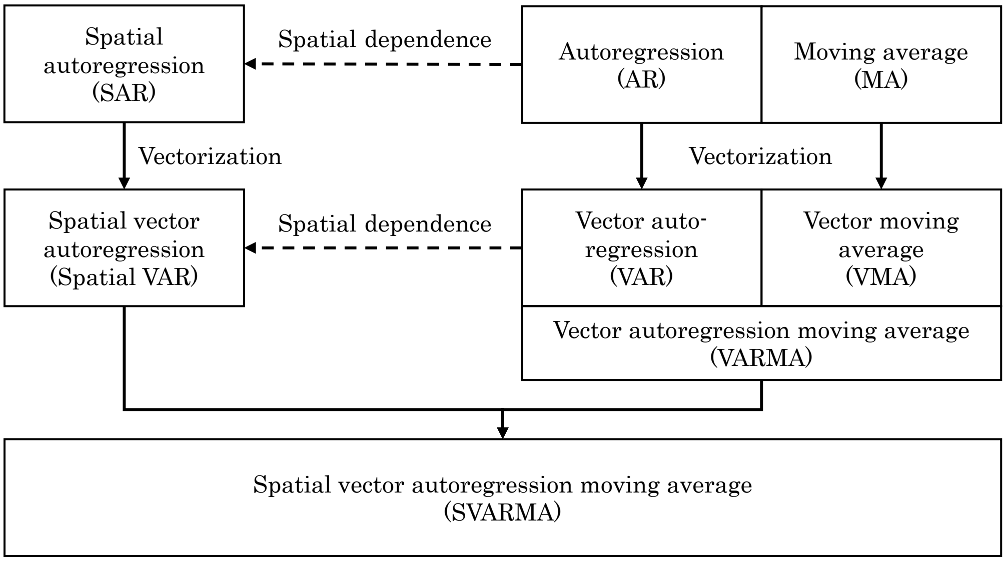

Spatial autoregression (SAR)
Spatial dependence
Autoregression (AR)
Moving average (MA)
Vectorization
Vectorization
Spatial vector autoregression (Spatial VAR)
Spatial dependence
Vector auto-regression (VAR)
Vector moving average (VMA)
Vector autoregression moving average (VARMA)
Spatial vector autoregression moving average (SVARMA)

Figure 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 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.

[3] Neutralizing serial dependence is required to satisfy the martingale property of the score needed to apply a standard CLT.

[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.

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 $\underline{c}$onnectivity matrix.

2.1 Vector Autoregressive Moving Average (VARMA)

In the multiple univariate sequence case, $w := (x, y)$, $\varepsilon := (\varepsilon^x, \varepsilon^y)$, a VARMA is a process

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

with parameter matrices structured as

$$A := \begin{bmatrix} a^{xx} & a^{xy} \\ a^{yx} & a^{yy} \end{bmatrix}, M := \begin{bmatrix} m^{xx} & m^{xy} \\ m^{yx} & m^{yy} \end{bmatrix}. \tag{2}$$

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

$$\mathbf{A}_{0:p} := \begin{bmatrix} A^{xx}_{0:p} & A^{xy}_{0:p} \\ A^{yx}_{0:p} & A^{yy}_{0:p} \end{bmatrix}, \mathbf{M}_{0:p} := \begin{bmatrix} M^{xx}_{0:p} & M^{xy}_{0:p} \\ M^{yx}_{0:p} & M^{yy}_{0:p} \end{bmatrix}, \mathbf{I} := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}, \tag{3}$$

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

$$\mathbf{A}_0 \mathbf{w}_t + \mathbf{A}_1 \mathbf{w}_{t-1} + ... + \mathbf{A}_p \mathbf{w}_{t-p} = \mathbf{M}_0 \boldsymbol{\varepsilon}_t + \mathbf{M}_1 \boldsymbol{\varepsilon}_{t-1} + ... + \mathbf{M}_q \boldsymbol{\varepsilon}_{t-q} \; \forall \; t \in \mathbb{Z}, \tag{4}$$

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. (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. (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.

2.2 Spatial Vector Autoregression (Spatial VAR)

Consider first a simple bivariate VAR with spatial dependencies within cross-sections (SVAR) in scalar notation,

$$\begin{aligned} \mathbf{x}_t &= \rho^x C_{n_x}\mathbf{x}_t + \phi^{xy}\mathbf{y}_{t-1} + \phi^{xx}\mathbf{x}_{t-1} + m_0^x\boldsymbol{\varepsilon}_t \\ \mathbf{y}_t &= \rho^y C_{n_y}\mathbf{y}_t + \phi^{yy}\mathbf{y}_{t-1} + \phi^{yx}\mathbf{x}_{t-1} + m_0^y\boldsymbol{\varepsilon}_t \end{aligned}, \tag{5}$$

with usually but not necessarily $m_0^x = m_0^y = 1$. By inverting the contemporaneous feedback we can write

$$\begin{aligned} \mathbf{x}_t &= (I_{n_x} - \rho^x C_{n_x})^{-1}(\phi^{xy}\mathbf{y}_{t-1} + \phi^{xx}\mathbf{x}_{t-1} + m_0^x\boldsymbol{\varepsilon}_t) \\ \mathbf{y}_t &= (I_{n_y} - \rho^y C_{n_y})^{-1}(\phi^{yy}\mathbf{y}_{t-1} + \phi^{yx}\mathbf{x}_{t-1} + m_0^y\boldsymbol{\varepsilon}_t) \end{aligned}, \tag{6}$$

which we can also write as

$$\begin{aligned} \mathbf{x}_t &= S^x\phi^{xy}\mathbf{y}_{t-1} + S^x\phi^{xx}\mathbf{x}_{t-1} + S^x m_0^x\boldsymbol{\varepsilon}_t \\ \mathbf{y}_t &= S^y\phi^{yy}\mathbf{y}_{t-1} + S^y\phi^{yx}\mathbf{x}_{t-1} + S^y m_0^y\boldsymbol{\varepsilon}_t \end{aligned}, \tag{7}$$

by introducing $S^x = (I_{n_x} - \rho^x C_{n_x})^{-1}$ and $S^y = (I_{n_x} - \rho^y C_{n_y})^{-1}$ as spatial multipliers. In this $S\phi$ are $n \times n$ matrices that arise from the combination of autoregressive and spatial forces, similarly Sm are $n \times n$ matrices producing spatial autoregressive moving averages. The spatial VAR with scalar time effects defined at the cross-sectional level thus has a VAR representation with parameter matrices defining the heterogeneous dependence structure at the observational level.

2.3 Spatial Vector Autoregressive Moving Average (SVARMA)

We can write SVARMA using $\mathbf{M} = \mathbf{I}$ by defining $\mathbf{A}_0$ in eq. (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,

$$(\mathbf{I} + \mathbf{A}_\mathbf{C})\mathbf{w}_t + \mathbf{A}_1\mathbf{w}_{t-1} + ... + \mathbf{A}_p\mathbf{w}_{t-p} = \boldsymbol{\varepsilon}_t + \mathbf{M}_1\boldsymbol{\varepsilon}_{t-1} + ... + \mathbf{M}_q\boldsymbol{\varepsilon}_{t-q} \;\forall\; t \in \mathbb{Z}. \tag{8}$$

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:

$$\mathbf{w}_t + \mathbf{S}\mathbf{A}_1\mathbf{w}_{t-1} + ... + \mathbf{S}\mathbf{A}_p\mathbf{w}_{t-p} = \mathbf{S}\boldsymbol{\varepsilon}_t + \mathbf{S}\mathbf{M}_1\boldsymbol{\varepsilon}_{t-1} + ... + \mathbf{S}\mathbf{M}_q\boldsymbol{\varepsilon}_{t-q} \;\forall\; t \in \mathbb{Z}, \tag{9}$$

with $\mathbf{S} = (\mathbf{I} + \mathbf{A}_\mathbf{C})^{-1}$. We refer to eq. (9) 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

$$\mathbf{w}_t + \mathbf{H}_1\mathbf{w}_{t-1} + ... + \mathbf{H}_p\mathbf{w}_{t-p} = \boldsymbol{\epsilon}_t + \mathbf{M}_1\boldsymbol{\epsilon}_{t-1} + ... + \mathbf{M}_q\boldsymbol{\epsilon}_{t-q} \;\forall\; t \in \mathbb{Z}. \tag{10}$$

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.

2.4 Kernel-driven spatial weight matrices

Key to the contemporaneous effects is specifying a network structure that defines the spillover channels between cross-sectional observations. In spatial literature, the weights matrix is based on geographical distances, but it is equally possible to define networks based on economic distances. Below, we propose a flexible approach based on Gaussian kernels that can produce weights matrices based on distances within a set of variables $\mathbf{v}$. Specifically, connectivity matrices C can be constructed by computing a Gaussian kernel

$$G = k(\mathbf{v}_i, \mathbf{v}_j; b) = \exp\left(\frac{-\|\mathbf{v}_i - \mathbf{v}_j\|^2}{b}\right), \tag{11}$$

with $\|\mathbf{v}_i - \mathbf{v}_j\|$ being the Euclidean distance, and b being a bandwidth parameter that determines the network smoothness.[5] For $b > 0$, the kernel k can be understood as a measure of similarity, which is seen by applying a Cauchy-Schwarz inequality

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

revealing that if $\mathbf{v}_i$ and $\mathbf{v}_j$ are similar, then $k(\mathbf{v}_i, \mathbf{v}_j; b)_{b>0}$ will be close to 1, and close to 0 when $\mathbf{v}_i$ and $\mathbf{v}_j$ are dissimilar. For positive b, few but strong network links arise for small bandwidths. For large $|b|$, strong links result. Negative bandwidths produce networks based on dissimilarities. If $\mathbf{v}_i$ and $\mathbf{v}_j$ are similar, then $k(\mathbf{v}_i, \mathbf{v}_j; b)_{b<0}$ will be close to 1, but larger than 1 when $\mathbf{v}_i$ and $\mathbf{v}_j$ are dissimilar

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

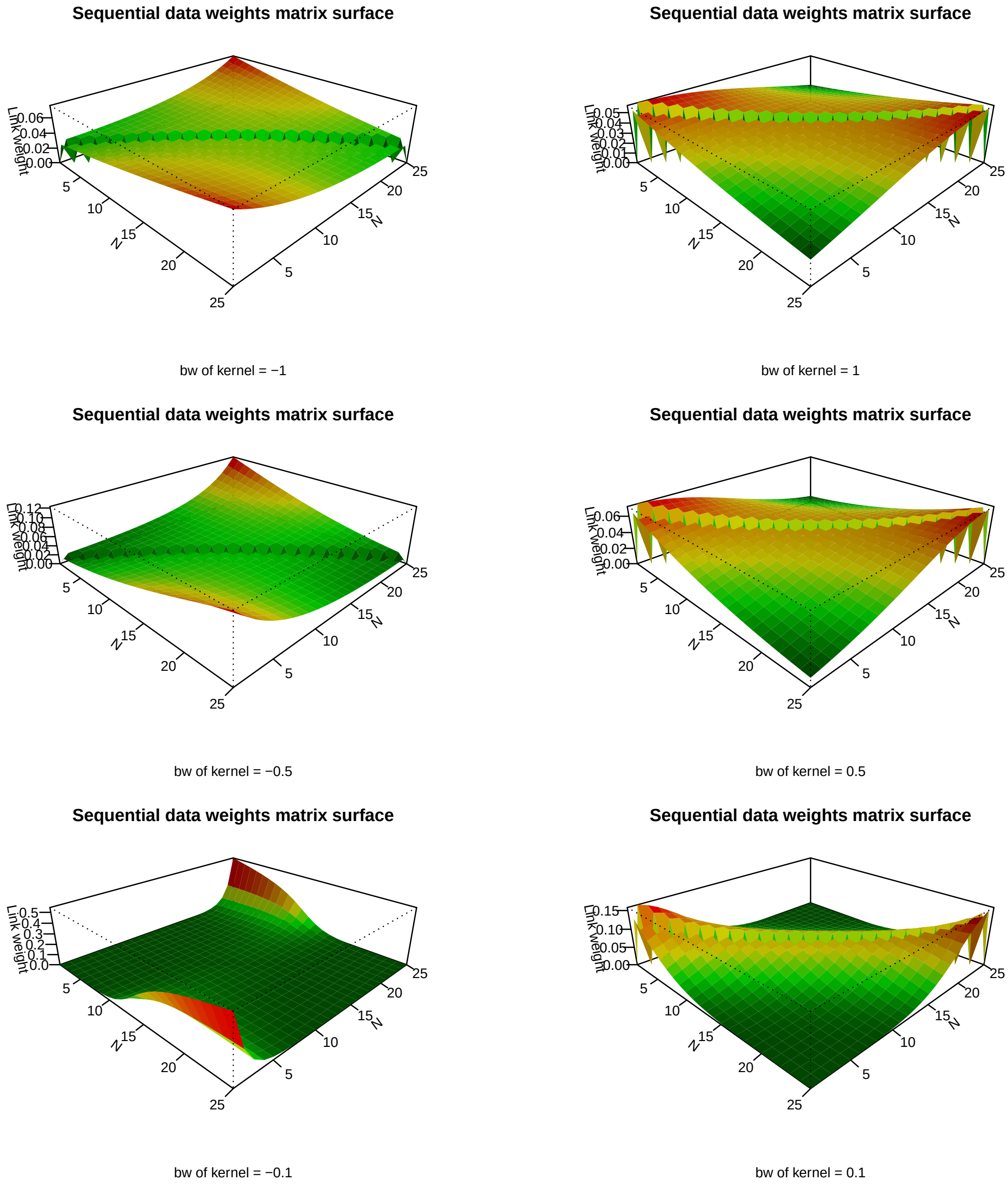
Both have empirical relevance. If the kernel is drawn around the levels of a cross-sectional time-series, the resulting contraction between dissimilar observations mimics the error-correction effect of a VECM process between local observations. For positive bandwidths, the similarity view of the kernel approach caries the interpretation of Tobler's law that underlies the intuition

[5] Geographic weight matrices can equally be constructed if $\mathbf{v}$ describes the physical locations of observations. Non-Gaussian kernels may also be used.

of the SAR. For any b, the diagonals of R are unit. We build a matrix D:

$$D = G - I = k(\mathbf{v}_i, \mathbf{v}_j; b) = \exp\left(\frac{-\|\mathbf{v}_i - \mathbf{v}_j\|^2}{b}\right) - I \qquad (12)$$

and build a spatial weight matrix C by row-normalizing D.



bw of kernel = −1

bw of kernel = 1

bw of kernel = −0.5

bw of kernel = 0.5

bw of kernel = −0.1

bw of kernel = 0.1

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

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

$$\mathbf{H}(L)\mathbf{w}_t = \mathbf{M}(L)\boldsymbol{\epsilon}_t \;\forall\; t \in \mathbb{Z}, \tag{13}$$

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 (13) is convenient notation for the SVARMA because it allows us to condition theory directly on components similar to the standard case of eq. (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.[6] 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, these properties are directly related to the properties of the model itself and the range of parameter values considered.[7] 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] See for example Blasques et al. (2018) for results on the relation between filters and *DGP*s.

[7] 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. (2017) show that multivariate ARMA structures can generate geometrically Ergodic data even when a nonlinear observation-driven spatial dependence process is considered.

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

$$\mathbf{w}_t = \mathbf{\Phi}\mathbf{w}_{t-1} + \boldsymbol{\epsilon}_t \ \forall \ t \in \mathbb{Z}, \tag{14}$$

with $\mathbf{\Phi}z = -\mathbf{H}_1 Lz - ... - \mathbf{H}_p L^p z$. Assuming some form of fading memory, eq. (14) may be expanded by a process of infinite back-substitution, giving rise to an infinite-order multivariate spatial autoregressive moving average:

$$\mathbf{w}_t = \{\boldsymbol{\epsilon}_t + \mathbf{\Phi}\boldsymbol{\epsilon}_{t-1} + \mathbf{\Phi}^2\boldsymbol{\epsilon}_{t-2} + ... + \mathbf{\Phi}^{\infty}\boldsymbol{\epsilon}_{t-\infty}\} \ \forall \ t \in \mathbb{Z}. \tag{15}$$

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 3.3. Stationarity and invertibility conditions that apply to eq. (13) are naturally an extension of this first order autoregressive case, which is itself a generalization of the scalar ARMA case. 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

$$\mathbf{w}_t = \mathbf{\Psi}(L)\boldsymbol{\epsilon}_t = \{\mathbf{\Psi}_0\boldsymbol{\epsilon}_t + \mathbf{\Psi}_1\boldsymbol{\epsilon}_{t-1} + \mathbf{\Psi}_2\boldsymbol{\epsilon}_{t-2} + ... + \mathbf{\Psi}_{\infty}\boldsymbol{\epsilon}_{t-\infty}\} \ \forall \ t \in \mathbb{Z}. \tag{16}$$

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

$$\mathbf{H}(z)\mathbf{\Psi}(z) = \mathbf{M}(z). \tag{17}$$

The conditions

$$\mathbf{H}_0 := \begin{bmatrix} I_{n_x} \ O_{n_x} \\ O_{n_y} \ I_{n_y} \end{bmatrix}, \ \mathbf{M}_0 := \begin{bmatrix} I_{n_x} \ O_{n_x} \\ O_{n_y} \ I_{n_y} \end{bmatrix}, \tag{18}$$

imply that

$$\mathbf{\Psi}_0 := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}. \tag{19}$$

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

$$\boldsymbol{\epsilon}_t = \mathbf{\Pi}(L)\mathbf{w}_t = \{\mathbf{\Pi}_0\mathbf{w}_{t-1} + \mathbf{\Pi}_1\mathbf{w}_{t-1} + \mathbf{\Pi}_2\mathbf{w}_{t-2} + ... + \mathbf{\Pi}_\infty\mathbf{w}_{t-\infty}\} \; \forall \; t \in \mathbb{Z}. \tag{20}$$

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

$$\mathbf{M}(z)\mathbf{\Pi}(z) = \mathbf{H}(z). \tag{21}$$

The conditions eq. (18) imply that

$$\mathbf{\Pi}_0 := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}. \tag{22}$$

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:

$$\mathbf{w}_t = \mathbf{H}^{-1}(L)\{\mathbf{M}(L)\boldsymbol{\epsilon}_t\} = \mathbf{M}(L)\mathbf{\Xi}_t \; \forall \; t \in \mathbb{Z}, \tag{23}$$

where $\mathbf{\Xi}_t = \mathbf{H}^{-1}(L)\boldsymbol{\epsilon}_t$.

Equation (23) 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. (23) to derive the companion matrix, and see that stability follows if the eigenvalues of $\mathbf{\Phi}$ lie inside the unit circle. Additional details are provided in the Appendix, appendix B.

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. (13). 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.

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

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

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

$$\mathbf{w}_t = \mathbf{\Psi}(L)\{\mathbf{p}(t)\mathbf{e}\}\ \forall\ t \in \mathbb{Z}. \tag{24}$$

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. (24) 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.

4 Penalized Estimation

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 (Eucledian 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 LASSO 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 ridge 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 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:

$$\hat{\boldsymbol{\theta}}_T := \arg\min_{\boldsymbol{\theta}\in\Theta} Q_T(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta}) + \lambda\gamma(\boldsymbol{\theta}), \tag{25}$$

with the ML criterion defined as

$$\begin{gathered} Q_T := \ell_T(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta}) = \sum_t^T \ell_t(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta}), \\ \ell_t(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta}) = \ln p_\varepsilon(\mathbf{w}_t - f(\mathbf{v}, \mathbf{w}_t; \boldsymbol{\theta}), \boldsymbol{\Sigma}; \boldsymbol{\nu}), \end{gathered} \tag{26}$$

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

$$\lambda\gamma(\boldsymbol{\theta}) = 1/\sqrt{NT} \sum |\mathbf{M}|^2. \tag{27}$$

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

$$\ell_t(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta}) = D(\boldsymbol{\theta}^{\mathbf{S}}, \mathbf{v}) + K(\boldsymbol{\theta}) + E(\boldsymbol{\theta}, \mathbf{v}, \mathbf{w}_t), \tag{28}$$

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

$$D(\boldsymbol{\theta}^{\mathbf{S}}, \mathbf{v}) := \ln\det \mathbf{S}\Big(\boldsymbol{\theta}^{\rho}, \mathbf{C}(\mathbf{v}; \mathbf{b})\Big), \tag{29}$$

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

$$\mathbf{S}(\boldsymbol{\theta}^{\boldsymbol{\rho}}, \mathbf{C}(\mathbf{v};\mathbf{b})) = \Big(\mathbf{I} - \boldsymbol{\rho} \circ \mathbf{C}(\mathbf{v};\mathbf{b})\Big)^{-1}, \tag{30}$$

with $\mathbf{C}(\mathbf{v};\mathbf{b})$ constructed as detailed in section 2.4. Importantly, the log determinant equals the sum of the log determinants of its diagonal blocks, as the off-diagonal blocks are zero

$$D(\boldsymbol{\theta}^{\mathbf{S}}, \mathbf{v}) = \ln\det \mathbf{S}\Big(\boldsymbol{\theta}^{\boldsymbol{\rho}}, \mathbf{C}(\mathbf{v};\mathbf{b})\Big) = \ln\det S^x\Big(\rho^x, C_{n_x}(\mathbf{v};b^x)\Big) + \ln\det S^y\Big(\rho^y, C_{n_y}(\mathbf{v};b^y)\Big), \tag{31}$$

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

$$\boldsymbol{\rho} \circ \mathbf{C}(\mathbf{v};\mathbf{b}) = \begin{bmatrix} \rho^x C_{n_x}(\mathbf{v};b^x) & O_{n_x} \\ O_{n_y} & \rho^y C_{n_y}(\mathbf{v};b^y) \end{bmatrix}. \tag{32}$$

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

$$K(\boldsymbol{\theta}) := \ln\Gamma\Big((\nu+N)/2\Big)\Big[\det\Sigma^{\frac{1}{2}}(\nu\pi)^{\frac{N}{2}}\Gamma\left(\nu/2\right)\Big]^{-1}, \tag{33}$$

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

$$\begin{aligned} E(\boldsymbol{\theta}, \mathbf{v}, \mathbf{w}_t) :=& \\ &-\tfrac{1}{2}(\nu^{\mathbf{x}}+N)\ln\Big(1 + \nu^{\mathbf{x}-1}(\mathbf{x}_t - f^{\mathbf{x}}(\mathbf{v}, \mathbf{w}_t;\boldsymbol{\theta}^{\mathbf{x}}))'\Sigma^{\mathbf{x}-1}(\mathbf{x}_t - f^{\mathbf{x}}(\mathbf{v}, \mathbf{w}_t;\boldsymbol{\theta}^{\mathbf{x}}))\Big) \\ &-\tfrac{1}{2}(\nu^{\mathbf{y}}+N)\ln\Big(1 + \nu^{\mathbf{y}-1}(\mathbf{y}_t - f^{\mathbf{y}}(\mathbf{v}, \mathbf{w}_t;\boldsymbol{\theta}^{\mathbf{y}}))'\Sigma^{\mathbf{y}-1}(\mathbf{y}_t - f^{\mathbf{y}}(\mathbf{v}, \mathbf{w}_t;\boldsymbol{\theta}^{\mathbf{y}}))\Big). \end{aligned} \tag{34}$$

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. Parallelization is key to feasible numerical optimization of the model under current computer systems, specifically when more than two variables are considered as in our application.[8] Since $f(\mathbf{v}, \mathbf{w}_T;\boldsymbol{\theta})$ depends on the moving

[8] We found that it is helpful to perform an initial estimation of a restricted SVARMA, after which numerical results can be passed on as the starting point for numerical optimization of the Likelihood for the unrestricted

averages that in turn result as difference combinations of $\mathbf{w}_t - f(\mathbf{v}, \mathbf{w}_T; \boldsymbol{\theta})$, the components of eq. (34) 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. When working with restrictions on the moving averages the likelihoods of each channel may be evaluated independently. Initializing a numerical search at the (P)MLE of the restricted model can further help reduce the search length of the sum of eq. (28) significantly. It should furthermore be noted that the penalty, and each of the three components of eq. (28), can be evaluated independently of each other. Furthermore eq. (29), or equivalently eq. (31), and eq. (33) do not depend on time if $\mathbf{v}$ is time-invariant, hence their values can be computed once and recycled T times. Parallelizing eq. (31) is much faster than calculating eq. (29) directly and can be performed independently from evaluating eq. (34).

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 $\boldsymbol{\theta}^{\boldsymbol{\rho}}$, which in turn may bias the autoregressive and moving average parameters by shifting cross-sectional dependence to a time channel. 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. Due to the dependence on moving averages that are not available as difference combinations for the first q periods are unavailable, the 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 Blasques et al. (2018). For small T however, the impact remains. We focus our simulations on the small T case to investigate this.

model. These first two steps are computationally demanding since the parameter vector is large and the search may cover a large range within Θ. We had good results using a $c++$ implementation of the Likelihood using an MKL compiled BLAS/LAPACK, and numerically solving for a maximum using HOPSPACK combined with an MPI compiler to spread computation across a cluster. Once an initial result is available, further dimensionality reduction is a less intensive task. Once a solution has been found for the full parameter vector, we had reasonable computation time with the optMaxlik package available at https://github.com/BPJandree/optMaxlik linked against an MKL optimized BLAS/LAPACK using a BFGS algorithm.

4.1 Goodness of fit

As always, goodness of fit is a matter of neutralizing residual correlations and explaining a significant share of the variance in the data. We propose two straightforward diagnostics.

4.1.1 Residual correlation

To assess how well the estimated structure fits the data, we propose estimating a cross-sectional AR model on the residuals on an equation-by-equation basis. Specifically, we suggest running r models, with lags $1, ..., r$, on the residuals with the first q periods dropped. Under the null, we estimate models on random data and we will expect 1 out of 10 lags to be significant at .10 purely out of chance. We suggest computing r individual LR ratios 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 should be used to report on the significance of residual correlations for each channel. For small T, the test statistic is still affected by the initialization. A second difficulty for small T is that the residual models can only fit on $N \times (T - q - r)$ observations, the residual time-dimension shrinks fast. We suggest $r = \max(q, p)$ can be used to diagnose whether major variables are omitted, and preferably $r > \max(q, p)$ to diagnose whether sufficient variables and lags have been included in the model. In our empirical case we use $r = \max(q, p) + 1$ and focus on the Student's-t case with inferred degrees of freedom.

4.1.2 Pseudo-R^2

To decide between competing models, (penalized) likelihood values can be compared. As the kernel bandwidth is unidentified when the spatial dependence is zero, and the kernel function is unidentified at $b = 0$, we suggest to follow an Information Criteria approach. Likelihood or penalized likelihood without a reference provide by itself little insight into model fit. We suggest 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),

$$\hat{R}^2 = 1 - \frac{\sum_1^T |\mathbf{w}_T - f(\mathbf{w}_T; \hat{\boldsymbol{\theta}})|^2}{\sum_1^T |\mathbf{w}_T - f(\mathbf{w}_T; \boldsymbol{\theta}_{\boldsymbol{\theta}=0})|^2}, \tag{35}$$

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. For small T this is not equivalent to the *true* R^2, as eq. (35) is influenced by the initialization of the residual vectors. However, as $T \to \infty$, the initialization effect fades and eq. (35) converges to the *true* R^2. In our empirical application we work with small T, and we suggest that eq. (35) provides a crude approximation that is still useful to inform the researcher on the degree of explained in-sample variance.

5 Application to Subnational Pollution and Household Income 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., 2018). 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.

5.1 Data

Our analysis relies on two longitudinal data sets. As a proxy for air pollution, we use the global estimates of fine particulate matter developed by van van Donkelaar et al. (2016). We use annual averages of monthly household expenditures for the average household and for the bottom quintile household. The expenditure data are taken from the Indonesia Database for Policy and Economic Research (INDO-DAPOER, World Bank Group).[9]

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.

INDO-DAPOER contains key economic, social and demographic indicators at the district-level, primarily sourced from various surveys and the Indonesia Central Bureau of Statistics (BPS). We use the annual means of monthly per capita household expenditures (in IDR) and the same indicator for the average across the 20 percent poorest households, which are available from 1999 to 2014. We are primarily interested in distinguishing between average economic developments and changes in poor household income. The data set includes poverty rates and local estimates of GDP, but the coverage is poor.

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 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.

Since we are particularly interested in pollution, we drop any areas that at one or more

[9] 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.

points in time have a concentration below 6 mcg/m^3. We 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 outliers, we are left with a final number of 113 areas that consist of polluted urban clusters. Apart from the 113 areas that we defined as polluted urban areas, we perform an additional estimation focusing on 60 heavily polluted areas that exceed the WHO air quality guidelines in all years.

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, appendix B.2.

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.

5.3 Results

Tables tables 2 and 4 in the appendix present the estimation results for the SVARMA(AICc). VARMA(AICc) results are contained in tables 1 and 3. 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 $\mathrm{PM}_{2.5} > 6$ and by 277.103 points at $\mathrm{PM}_{2.5} > 10$, indicating that the SVARMA improves the conditional density implied by the model significantly over the VARMA. Our $\hat{R}^2$ estimates suggest that we explain 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 $\mathrm{PM}_{2.5} > 6$, 0.732 versus 0.722 at $\mathrm{PM}_{2.5} > 10$). In both cases the SVARMA, however, uses less ARMA parameters (29 versus 34 at $\mathrm{PM}_{2.5} > 6$, 27 versus 31 at $\mathrm{PM}_{2.5} > 10$.) implying that the improvements are significant. Our residual correlation tests also favor the SVARMA representation (the VARMA at $\mathrm{PM}_{2.5} > 6$ retains significant residual correlations). The rejections of residual correlations, and reasonable $\hat{R}^2$, lead us to conclude that no major components are missing in either of the SVARMA results, hence we interpret the parameters and standard errors in their usual context.

5.3.1 Clustering effects

The spatial filtering improves the model fit considerably, we can also see that the bandwidths that control the network smoothness are different in each channel of the model. Figure 13 in the appendix plots the network surfaces, we have ordered the link weights from high to low. This reveals that the bandwidths at $\mathrm{PM}_{2.5} > 6$ 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 within the network. Average expenditures have a higher bandwidth value than bottom expenditures, hence the results indicate that bottom expenditures spillover in smaller but stronger clusters, than the average expenditures. In the $\mathrm{PM}_{2.5} > 10$ model that has smaller cross-sections, we see that the network structures are relatively more similar. However, from the bandwidth values we can see that also in this smaller subset of heavily polluted urban areas, pollution is shared across fewer and stronger links within the entire network, bottom expenditures are in turn shared across more but weaker links, while average expenditures have the smoothest network through which indirect feedback effects are more easily transmitted to far-away observations.

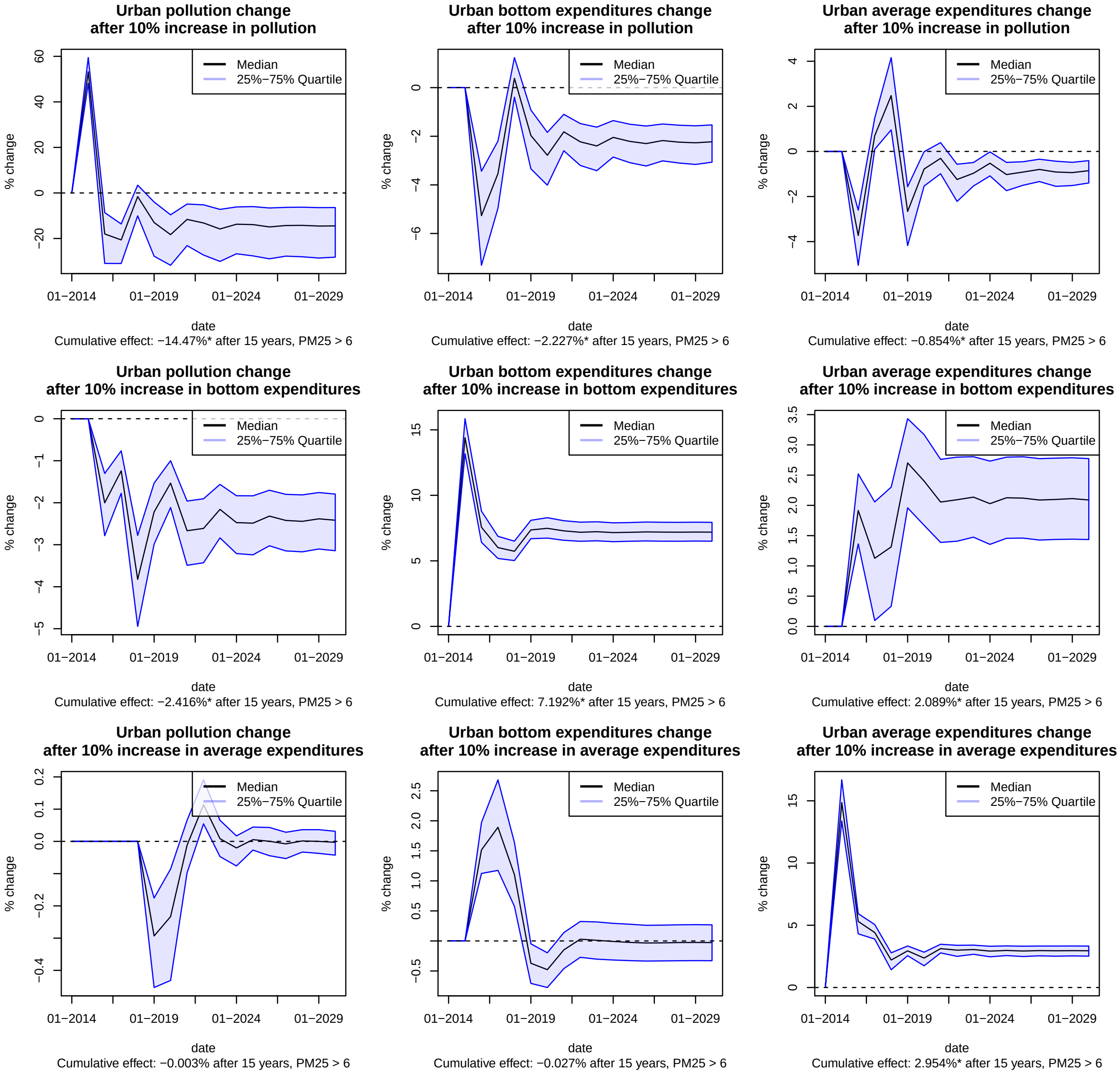
5.3.2 Impulse Response analysis

To explore the dynamics implied by the estimated results, we use the parameters to simulate IRFs. 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. Figure 3 shows the results for the model estimated at $\text{PM}_{2.5} > 6$, and fig. 4 shows the results from the model estimated at $\text{PM}_{2.5} > 10$. The figures are produced by $10,000$ random draws and show the cumulative effects resulting from compounding the percentage changes including spatial feedback effects.

We find that across all districts with $\text{PM}_{2.5} > 6$, 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 strengthens (-4.171%). Growth in average expenditures in these highly polluted areas is also found to reduce pollution, albeit with smaller impact (-.235%). Exogenous pollution in both models has 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 long-run effects and 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. This suggests that a region remains polluted as long as exogenous effects continue to enter the system, and pollution levels decline when these structural contributions stabilize and continued income growth takes over as a predominant driver of pollution decline.

Another result is that at both $\text{PM}_{2.5} > 6$ and $\text{PM}_{2.5} > 10$, average growth is non-inclusive. At $\text{PM}_{2.5} > 6$, an increase in average household expenditures does not significantly spill over to bottom households in the long-run, and at $\text{PM}_{2.5} > 10$ the long-run impact is -0.634%. Growth in bottom expenditures, on the other hand, boosts the average (7.192% at $\text{PM}_{2.5} > 6$

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 $PM_{2.5} > 10$).



Urban pollution change after 10% increase in pollution

Cumulative effect: −14.47%* after 15 years, PM25 > 6

Urban bottom expenditures change after 10% increase in pollution

Cumulative effect: −2.227%* after 15 years, PM25 > 6

Urban average expenditures change after 10% increase in pollution

Cumulative effect: −0.854%* after 15 years, PM25 > 6

Urban pollution change after 10% increase in bottom expenditures

Cumulative effect: −2.416%* after 15 years, PM25 > 6

Urban bottom expenditures change after 10% increase in bottom expenditures

Cumulative effect: 7.192%* after 15 years, PM25 > 6

Urban average expenditures change after 10% increase in bottom expenditures

Cumulative effect: 2.089%* after 15 years, PM25 > 6

Urban pollution change after 10% increase in average expenditures

Cumulative effect: −0.003% after 15 years, PM25 > 6

Urban bottom expenditures change after 10% increase in average expenditures

Cumulative effect: −0.027% after 15 years, PM25 > 6

Urban average expenditures change after 10% increase in average expenditures

Cumulative effect: 2.954%* after 15 years, PM25 > 6

Figure 3: 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 *.

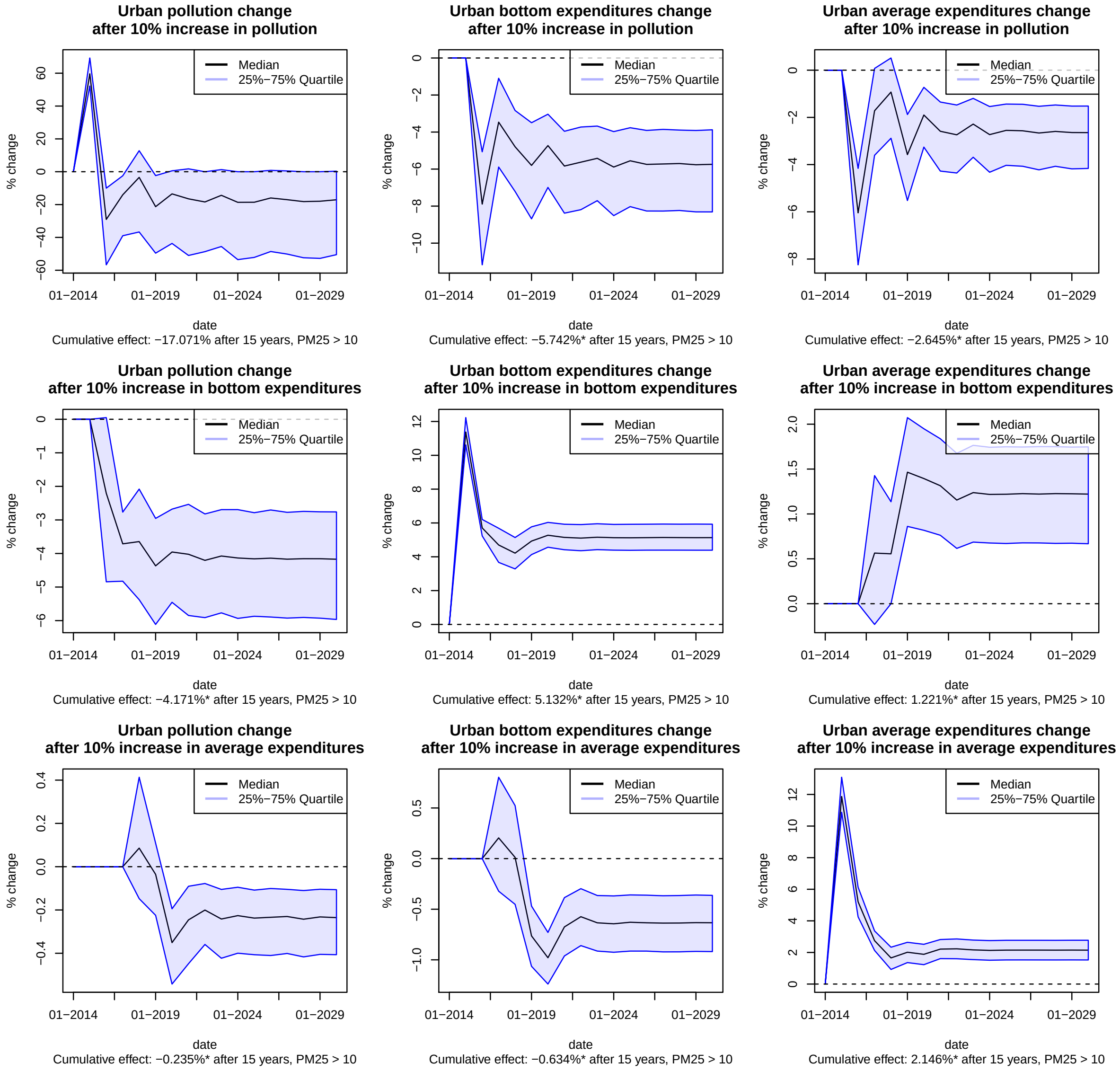


Figure 4: IRF plots for exogenous shocks in pollution, bottom household expenditures and average household expenditures $PM_{2.5}$ >10. Effects that exclude zero in the final year, are marked by *.

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.

5.3.3 Economic significance

Pollution damages account for considerable loss. Using the converged IRF impact, and using the 2017 dollar conversion rate, we can draft the following crude impacts of 10% country wide pollution increase based on the average expenditure levels per distinguished household group. 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 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 between 6 and 10 $PM_{2.5}$.[10] Weighing the effect on expenditures on pollution by population, the total damages across all households of a 10% pollution increase would be over 3 billion dollars. Of that, approximately half a billion dollars are lost expenditures of poor households. Various factors can further add to this number in the future, including continued growth in urban population, income and pollution levels. 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.

[10] As a reference, the United Nations put the total Indonesian population at 261,115,456 in 2016.

6 Conclusion

In this paper we discussed and estimated a fat-tailed Spatial Vector Autoregressive Moving Average model that enables analyzing high-dimensional interactions between multiple cross-sectional time series. This type of model is particularly useful to study Granger-causal interactions. The model requires specifying multiple matrices that define cross-sectional spillover channels. Networks may be part geographic, but also relate to economic distances. We introduced a framework to dynamically construct networks from the data. The smoothness of the network is controlled by a bandwidth parameter that can be integrated into the Likelihood framework.

We estimated the model parameters using remotely sensed pollution statistics of urban areas in Indonesia, together with subnational household expenditure data from surveys. We estimated two models with respectively 113 urban and 60 highly polluted urban areas over the period 1999-2014. Our networks are based on the similarities in sample moments and quantiles of local returns in the data, and the smoothness has been estimated within the likelihood framework. We contrasted the spatial model with a non-spatial counterpart and found that the spatial framework improves considerably in terms of various diagnostics, while using fewer ARMA parameters. Our approach to network modeling is not only favored by the data, it also provides additional insights. We find that cross-sectional dependencies in pollution are centered in smaller, but stronger, clusters than the economic variables. Expenditures of poor households, spill over more locally than those of average households.

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. Earlier research has indicated that economic impacts of air pollution stem mainly from health effects that decrease length and quality of life, increase health expenditures, and reduce labor supply and productivity. 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). For those facing negative health impacts, personal wealth can suffer immensely, between $240 billion and $630 billion are spent each year on health care costs related to pollution (Preker et al., 2016). Quality of life is affected too, Levinson (2012) found individual average willingness to pay up to $42 per day for a one-standard-deviation improvement in air quality in the United States. Reduced air pollution has also been shown to increase short-run labor supply due to a healthier population, in one case by 3.5 percent, translating to increased income for individual workers (Hanna and Oliva, 2015). Worker productivity in certain sectors also relies on health from clean air, with a 10 ppb change in average ozone exposure, for instance, resulting in a 5.5 percent change in a study of agricultural worker productivity (Zivin and Neidell, 2012). In that case, reducing ozone pollution could result in annual cost savings in labor expenditure—totaling approximately $700 million in the United States if a 10 ppb reduction was implemented.

While many of the results point toward an economic failure, we also see 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.[11] In these cases, lowering household 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.

Importantly, we see that the economic impacts of pollution growth are higher in 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 leads in either case 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 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).

References

Andree, B. P. J., Blasques, F., and Koomen, E. (2017). Smooth Transition Spatial Autoregressive Models. *Tinbergen Institute Discussion Papers.*

Andree, B. P. J., Spencer, P., Chamorro, A., and Dogo, H. (2018). Penalized Kernel Learning of Deforestation, Pollution and Carbon. *World Bank Policy Research Working Papers, forthcoming.*

Bao, Y. and Ullah, A. (2007). Finite sample properties of maximum likelihood estimator in spatial models. *Journal of Econometrics*, 137(2):396–413.

Beenstock, M. and Felsenstein, D. (2007). Spatial Economic Analysis Spatial Vector Autoregressions Spatial Vector Autoregressions. *Spatial Economic Analysis*, 2(2):167–196.

[11] 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).

Blasques, F., Gorgi, P., Koopman, S. J., and Wintenberger, O. (2018). Feasible invertibility conditions and maximum likelihood estimation for observation-driven models. *Electronic Journal of Statistics*, 12(1):1019–1052.

Boudjellaba, H., Dufour, J.-M., and Roy, R. (1992). Testing Causality Between Two Vectors in Multivariate Autoregressive Moving Average Models. *Journal of the American Statistical Association*, 87(420):1082.

Brockwell, P. J. and Davis, R. A. (2002). *Time series : theory and methods.*

Covey, T. and Bessler, D. A. (1992). Testing for Granger's Full Causality. *The Review of Economics and Statistics*, 74(1):146.

Deschenes, O., Greenstone, M., and Shapiro, J. S. (2012). Defensive Investments and the Demand for Air Quality: Evidence from the NOx Budget Program. *NBER Working Paper No. 18267*.

Dufour, J.-M., Pelletier, D., and Renault, É. (2006). Short run and long run causality in time series: inference. *Journal of Econometrics*, 132(2):337–362.

Dufour, J.-M. and Renault, E. (1998). Short Run and Long Run Causality in Time Series: Theory. *Econometrica*, 66(5):1099.

Dufour, J.-M. and Taamouti, A. (2010). Short and long run causality measures: Theory and inference. *Journal of Econometrics*, 154(1):42–58.

Eichler, M. and Didelez, V. (2010). On Granger causality and the effect of interventions in time series. *Lifetime Data Analysis*, 16(1):3–32.

Engle, R. F., Hendry, D. F., and Richard, J.-F. (1983). Exogeneity. *Econometrica*, 51(2):277.

Granger, C., King, M. L., and White, H. (1995). Comments on testing economic theories and the use of model selection criteria. *Journal of Econometrics*, 67(1):173–187.

Granger, C. W. (1980). Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424.

Grossman, G. M. and Krueger, A. B. (1995). Economic Growth and the Environment. *The Quarterly Journal of Economics*, 110(2):353–377.

Haavelmo, T. (1943). The Statistical Implications of a System of Simultaneous Equations. *Econometrica*, 11:1–12.

Haavelmo, T. (1944). The Probability Approach in Econometrics. *Econometrica*, 12((Suppl.)):1–115.

Hanna, R. and Oliva, P. (2015). The effect of pollution on labor supply: Evidence from a natural experiment in Mexico City. *Journal of Public Economics*, 122:68–79.

Hendry, D. F. (2017). *European journal of pure and applied mathematics.*, volume 10.

Kalman, R. (1983). Identifiability and Modeling in Econometrics. *Developments in Statistics*, 4:97–136.

Lee, L. (2012). Environmental poverty, a decomposed environmental Kuznets curve, and alternatives: Sustainability lessons from China. *Ecological Economics*, 73(1):86–92.

Levinson, A. (2012). Valuing public goods using happiness data: The case of air quality. *Journal of Public Economics*, 96(9-10):869–880.

Lu, X., Su, L., and White, H. (2017). Granger Causality and Structural Causality in Cross-section and Panel Data. *Econometric Theory*, 33(02):263–291.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis.* Springer Berlin Heidelberg, Berlin, Heidelberg.

Neuberg, L. G., Neuberg, and Gerson, L. (2003). Causality: Models, Reasoning, and Inference, by Judea Pearl, Cambridge University Press, 2000. *Econometric Theory*, 19(04):675–685.

Nsiri, S. and Roy, R. (1993). On the Invirtibility on Multivariate Linear Processes. *Journal of Time Series Analysis*, 14(3):305–316.

Panayatou, T. (1997). Demystifying the environmental Kuznets curve: turning a black box into a policy tool. *Environment and Development Economics*, 2(4):S1355770X97000259.

Pearl, J. (2000). *Causality : models, reasoning, and inference.* Cambridge University Press.

Pötscher, B. M. and Prucha, I. R. (1997). *Dynamic Nonlinear Econometric Models.* Springer Berlin Heidelberg, Berlin, Heidelberg.

Preker, A. S., Adeyi, O. O., Lapetra, M. G., Simon, D. C., and Keuffel, E. (2016). Health Care Expenditures Associated With Pollution: Exploratory Methods and Findings. *Annals of Global Health*, 82(5):711–721.

Quiggin, J. (2009). Six Refuted Doctrines. *Economic Papers*, 28(3):239–248.

Ranieri, R. and Almeida Ramos, R. (2013). Inclusive growth: Building up a concept.

Reinsel, G. C. (2003). *Elements of multivariate time series analysis.* Springer.

Roy, A., McElroy, T. S., and Linton, P. (2014). Estimation of Causal Invertible VARMA Models.

Shapiro, J. S. and Walker, R. (2016). Why is Pollution from U.S. Manufacturing Declining? The Roles of Environmental Regulation, Productivity, and Trade. *Cowles Foundation Discussion Paper No. 1982R.*

Sims, C. A. (1972). Money, Income, and Causality. *The American Economic Review*, 62(4):540–552.

Sin, C.-Y. and White, H. (1996). Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics*, 71(1):207–225.

Soumyananda, D. (2004). Environmental Kuznets Curve Hypothesis: A Survey. *Ecological Economics*, 49(4):431–455.

Stagl, S. (1999). Delinking Economic Growth from Environmental Degradation? A Literature Survey on the Environmental Kuznets Curve Hypothesis. *Wirtschafts Universitat Wien Working Paper No. 6.*

Stelzer, R. (2008). Multivariate Markov-switching ARMA processes with regularly varying noise. *Journal of Multivariate Analysis*, 99(6):1177–1190.

Stern, D. I. (1998). Progress on the Environmental Kuznets Curve? *Environment and Development Economics*, 3(2):173–196.

Stern, D. I. (2004). The Rise and Fall of the Environmental Kuznets Curve. *World Development*, 32(8):1419–1439.

Stern, D. I., Common, M. S., and Barbier, E. B. (1996). Economic growth and environmental degradation: The environmental Kuznets curve and sustainable development. *World Development*, 24(7):1151–1160.

van Donkelaar, A., Martin, R. V., Brauer, M., Hsu, N. C., Kahn, R. A., Levy, R. C., Lyapustin, A., Sayer, A. M., and Winker, D. M. (2016). Global Estimates of Fine Particulate Matter using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors. *Environmental Science & Technology*, 50(7):3762–3772.

White, H. and Chalak, K. (2009). Settable Systems: An Extension of Pearl's Causal Model with Optimization, Equilibrium, and Learning. *Journal of Machine Learning Research*, 10(Aug):1759–1799.

White, H., Chalak, K., and Lu, X. (2011). Causality in Time Series Linking Granger Causality and the Pearl Causal Model with Settable Systems. *JMRL: Workshop and Conference Proceedings 12*, pages 1–29.

White, H. and Lu, X. (2010). Granger Causality and Dynamic Structural Systems. *Journal of Financial Econometrics*, 8(2):193–243.

White, H. and Pettenuzzo, D. (2014). Granger causality, exogeneity, cointegration, and economic policy analysis. *Journal of Econometrics*, 178:316–330.

White, H., Xu, H., and Chalak, K. (2014). Causal discourse in a game of incomplete information. *Journal of Econometrics*, 182(1):45–58.

World Bank (1992). World Development Report 1992. Development and the Environment. Technical report.

World Bank and Institute for Health Metrics and Evaluation (2016). The cost of air pollution : strengthening the economic case for action.

Zheng, T., Xiao, H., and Chen, R. (2015). Generalized ARMA models with martingale difference errors. *Journal of Econometrics*, 189(2):492–506.

Zivin, J. G. and Neidell, M. (2012). The impact of pollution on worker productivity. *American Economic Review*, 102(7):3652–3673.

Supplementary Appendix

Pollution and Expenditures in a Penalized Spatial Vector Autoregressive Moving Average with Data-Driven Networks

Bo Pieter Johannes Andrée, Phoebe Spencer, Andres Chamorro,

Dieter Wang, Sardar Feredun Azari, Harun Dogo.

A 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

$$\begin{bmatrix} \mathbf{x}_t + H_1^{xx}\mathbf{x}_{t-1} + H_1^{xy}\mathbf{y}_{t-1} + ... + H_p^{xx}\mathbf{x}_{t-p} + H_p^{xy}\mathbf{y}_{t-p} \\ \mathbf{y}_t + H_1^{yx}\mathbf{y}_{t-1} + H_1^{yy}\mathbf{x}_{t-1} + ... + H_p^{yx}\mathbf{y}_{t-p} + H_p^{yy}\mathbf{x}_{t-p} \end{bmatrix} =$$

$$\begin{bmatrix} \boldsymbol{\epsilon}_t^x + M_1^{xx}\boldsymbol{\epsilon}_{t-1}^x + ... + M_q^{xx}\boldsymbol{\epsilon}_{t-q}^x \\ \boldsymbol{\epsilon}_t^y + M_1^{yy}\boldsymbol{\epsilon}_{t-1}^y + ... + M_q^{yy}\boldsymbol{\epsilon}_{t-1}^y \end{bmatrix} \forall\ t \in \mathbb{Z}. \tag{36}$$

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

$$\mathbf{H}_0 := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}, \mathbf{H}_{1:p} := \begin{bmatrix} H_{1:p}^{xx} & H_{1:p}^{xy} \\ H_{1:p}^{yx} & H_{1:p}^{yy} \end{bmatrix}, \mathbf{M}_0 := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix},\ \mathbf{M}_{1:p} := \begin{bmatrix} M_{1:p}^{xx} & O_{n_x} \\ O_{n_y} & M_{1:p}^{yy} \end{bmatrix}. \tag{37}$$

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. (36) is of the form:

$$\begin{bmatrix} \boldsymbol{\epsilon}_t^x + M_1^{xx}\boldsymbol{\epsilon}_{t-1}^x + M_1^{xy}\boldsymbol{\epsilon}_{t-1}^x + ... + M_q^{xx}\boldsymbol{\epsilon}_{t-q}^x + M_q^{xy}\boldsymbol{\epsilon}_{t-q}^x \\ \boldsymbol{\epsilon}_t^y + M_1^{yx}\boldsymbol{\epsilon}_{t-1}^y + M_1^{yy}\boldsymbol{\epsilon}_{t-1}^y + ... + M_q^{yx}\boldsymbol{\epsilon}_{t-q}^y + M_q^{yy}\boldsymbol{\epsilon}_{t-q}^y \end{bmatrix}. \tag{38}$$

The matrix representation results from

$$\mathbf{H}_0 := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}, \mathbf{H}_{1:p} := \begin{bmatrix} H^{xx}_{1:p} & H^{xy}_{1:p} \\ H^{yx}_{1:p} & H^{yy}_{1:p} \end{bmatrix}, \mathbf{M}_0 := \begin{bmatrix} I_{n_x} & O_{n_x} \\ O_{n_y} & I_{n_y} \end{bmatrix}, \ \mathbf{M}_{1:p} := \begin{bmatrix} M^{xx}_{1:p} & M^{xy}_{1:p} \\ M^{yx}_{1:p} & M^{yy}_{1:p} \end{bmatrix}. \tag{39}$$

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.[12] From a practical aspect it is useful to first consider models of the type eq. (37) first, and use the results to feed numerical algorithms to estimate models of the eq. (39) type.

B 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,

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

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

$$\mathbf{H}_{0:p} := \begin{bmatrix} H^{xx}_{0:p} & H^{xy}_{0:p} \\ H^{yx}_{0:p} & H^{yy}_{0:p} \end{bmatrix}, \ \mathbf{M}_{0:p} := \begin{bmatrix} M^{xx}_{0:p} & M^{xy}_{0:p} \\ M^{yx}_{0:p} & M^{yy}_{0:p} \end{bmatrix},$$

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.

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:

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

Using the state vector $\boldsymbol{\Xi}_t = \left[\boldsymbol{\Xi}_{1t}, \boldsymbol{\Xi}_{2t}, ..., \boldsymbol{\Xi}_{p_t}\right]'$ we can now write the system after defining $\mathbf{O} = 0 \circ \mathbf{I}$:

$$\begin{bmatrix} \boldsymbol{\Xi}_1(t) \\ \boldsymbol{\Xi}_2(t) \\ \vdots \\ \boldsymbol{\Xi}_p(t) \end{bmatrix} = \begin{bmatrix} -\mathbf{H}_1 & \dots & -\mathbf{H}_{p-1} & -\mathbf{H}_p \\ \mathbf{I} & \dots & -\mathbf{O} & \mathbf{O} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{O} & \dots & \mathbf{I} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Xi}_1(t-1) \\ \boldsymbol{\Xi}_2(t-1) \\ \vdots \\ \boldsymbol{\Xi}_p(t-1) \end{bmatrix} + \begin{bmatrix} \mathbf{I} \\ \mathbf{O} \\ \vdots \\ \mathbf{O} \end{bmatrix} \boldsymbol{\epsilon}(t),$$

with measurement equation

$$\mathbf{w}(t) = \mathbf{M}_0\boldsymbol{\Xi}_1(t) + ... + \mathbf{M}_{p-1}\boldsymbol{\Xi}_p(t) \; \forall \; t \in \mathbb{Z}.$$

Stability can now be expressed in terms of the companion matrix $\boldsymbol{\Phi}$. 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 us

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

that 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., 2017) are of help. While the stability conditions of SVARMA are 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.

B.1 Small sample distribution of the 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 $\lambda = 1/\sqrt{NT}$ in situations where T is small. We set the sample size and 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 spatial structures. We explore whether the spatial structure improves the ARMA estimates, and explore robustness to over-fitting under the null of an ARMA process. The *DGP* is

$$\mathbf{y}_t = 0.6C(\mathbf{x}; b)\mathbf{y}_t - 0.35\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t + 0.25\boldsymbol{\varepsilon}_{t-1}, \tag{40}$$

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. We focus on $t - p = 12$, which is identical to, and $N = (10, 25, 75, 125)$ which covers, our empirical cases.

As we can see in fig. 6 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, causing a bias towards zero. In fact, by increasing N and fixing T, the impacts increase further as the ratio of distorted information N/T grows. Nonetheless, the ARMA parameters are jointly well behaved, even when both N and T are small. We conclude that inference on the joint parameters is therefore valid, while statements that involve differentiation between short- and long-term effects should be made with caution in small T panels. Figure 7 shows the results for the MLE. It is clear that the penalization improves the empirical distribution of the bandwidth parameter substantially.

Figure 8 and fig. 9 document results for $b = 2$. Again the unpenalized distribution of b is not well-behaved. The penalized distribution of b improved substantially and is also better than the distribution under $b = .15$. In both results, the ARMA parameters remain jointly well-behaved confirming that the model is useful for inference about the time dynamics.

Figure 10 shows results for ARMA estimation on the identical *DGP*s. This reveals that when the cross-sectional process exhibits both ARMA and spatial effects, and the spatial effects are not modeled, the ARMA parameters become severely biased. Combined, all the simulation results not only confirm that the model performs well in empirically relevant situations, but also that not specifying the spatial effects results in biased results.

Finally, to investigate the behavior of the kernel procedure under the alternative when the ARMA is in fact correct, and no spatial structure is needed, we present Figure 5 below. The bandwidth density is centered around 0, Note that the kernel structure is not identified at this value. Note also that if spatial dependence is zero, the bandwidth could take on any value, which might allow the structure to eventually find some (dis)similarities that produce significant cross-sectional dependencies. The results show that the penalization of the bandwidth prevents the values to wander off into the extreme, while bringing some minor improvements to the spatial dependence estimate. Either case, the spatial dependence parameter is well-behaved, allowing the research to decide between SARMA and ARMA mechanics while estimating the weights structure, simply by using a Wald test around the spatial dependence parameter.

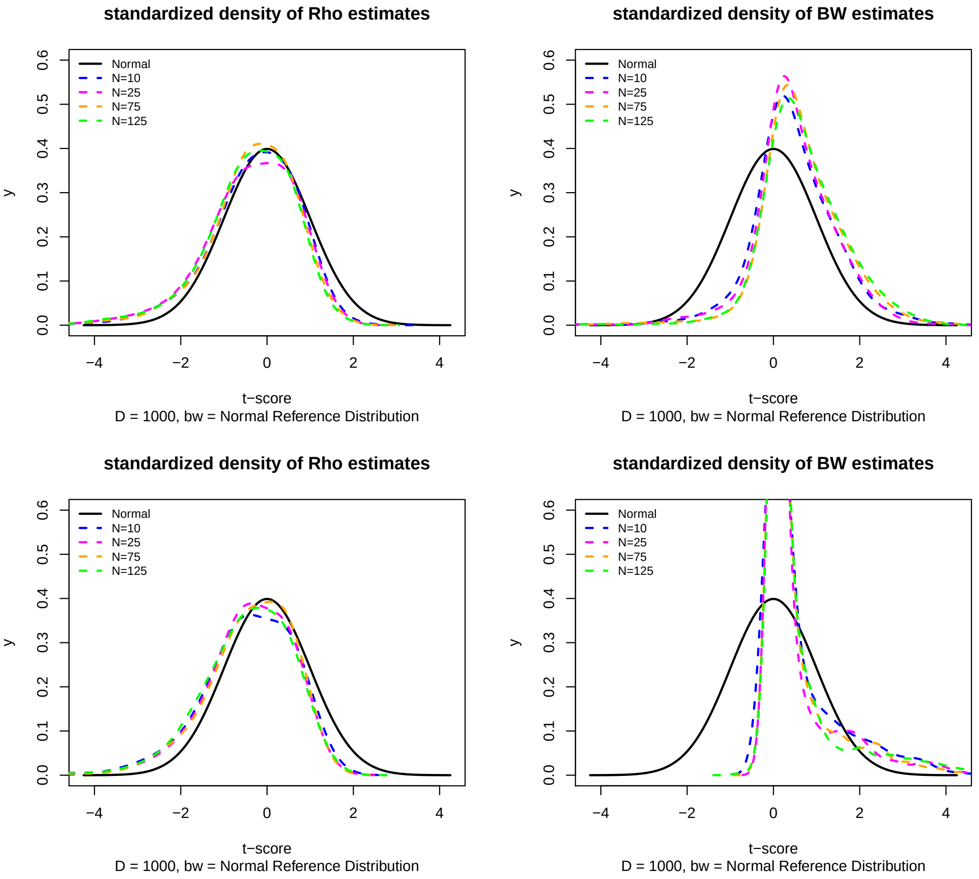


Figure 5: Penalized (upper) small sample distributions of bandwidth and spatial parameter in the SARMA, when the *true* process is a cross-sectional ARMA with zero spatial effects. The bandwidth density is centered around 0, note that the kernel structure is not identified at this value.

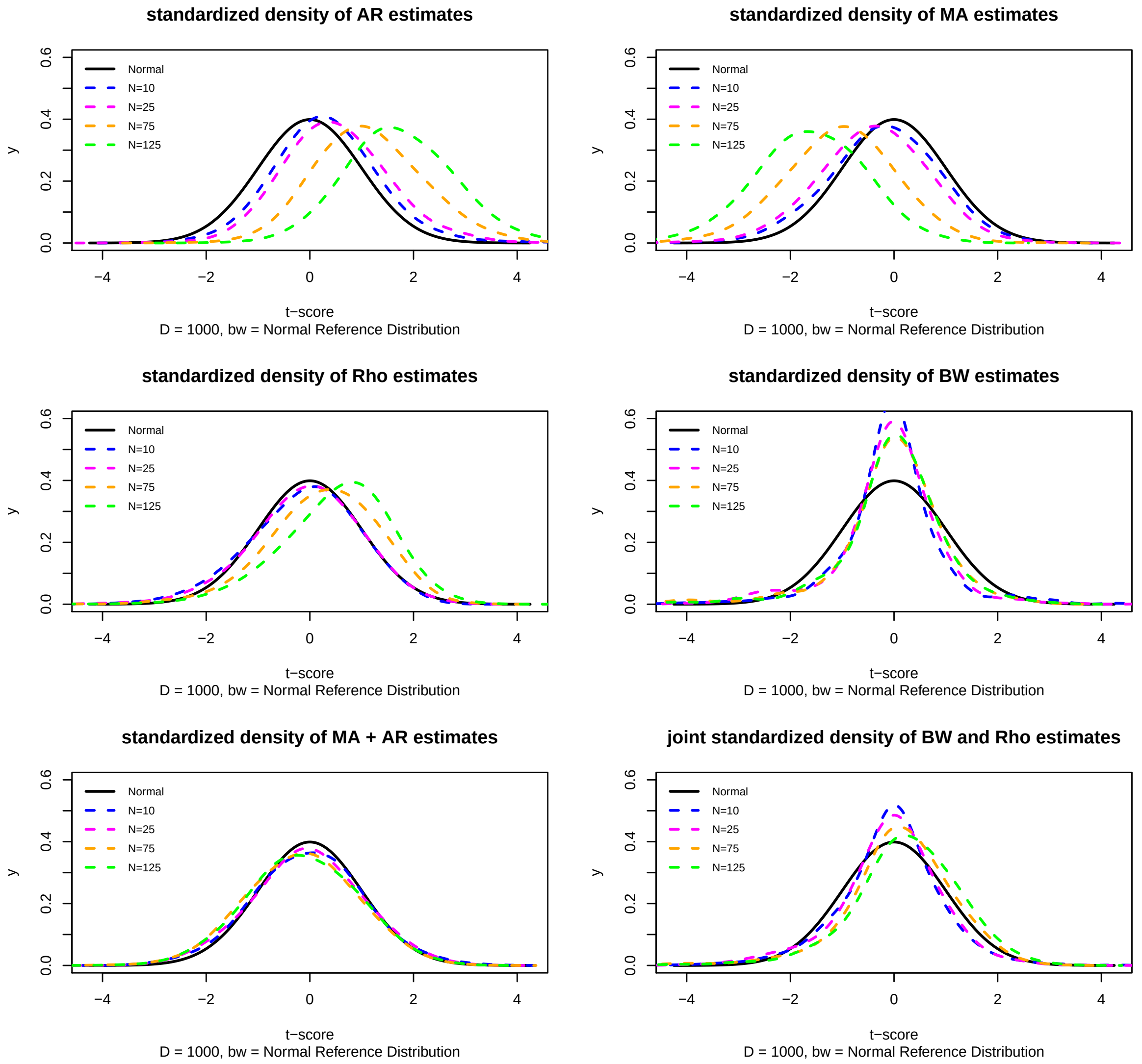


Figure 6: Penalized small sample distributions of the correctly specified SARMA, bandwidth of the spatial kernel matrix in the DGP set to .15

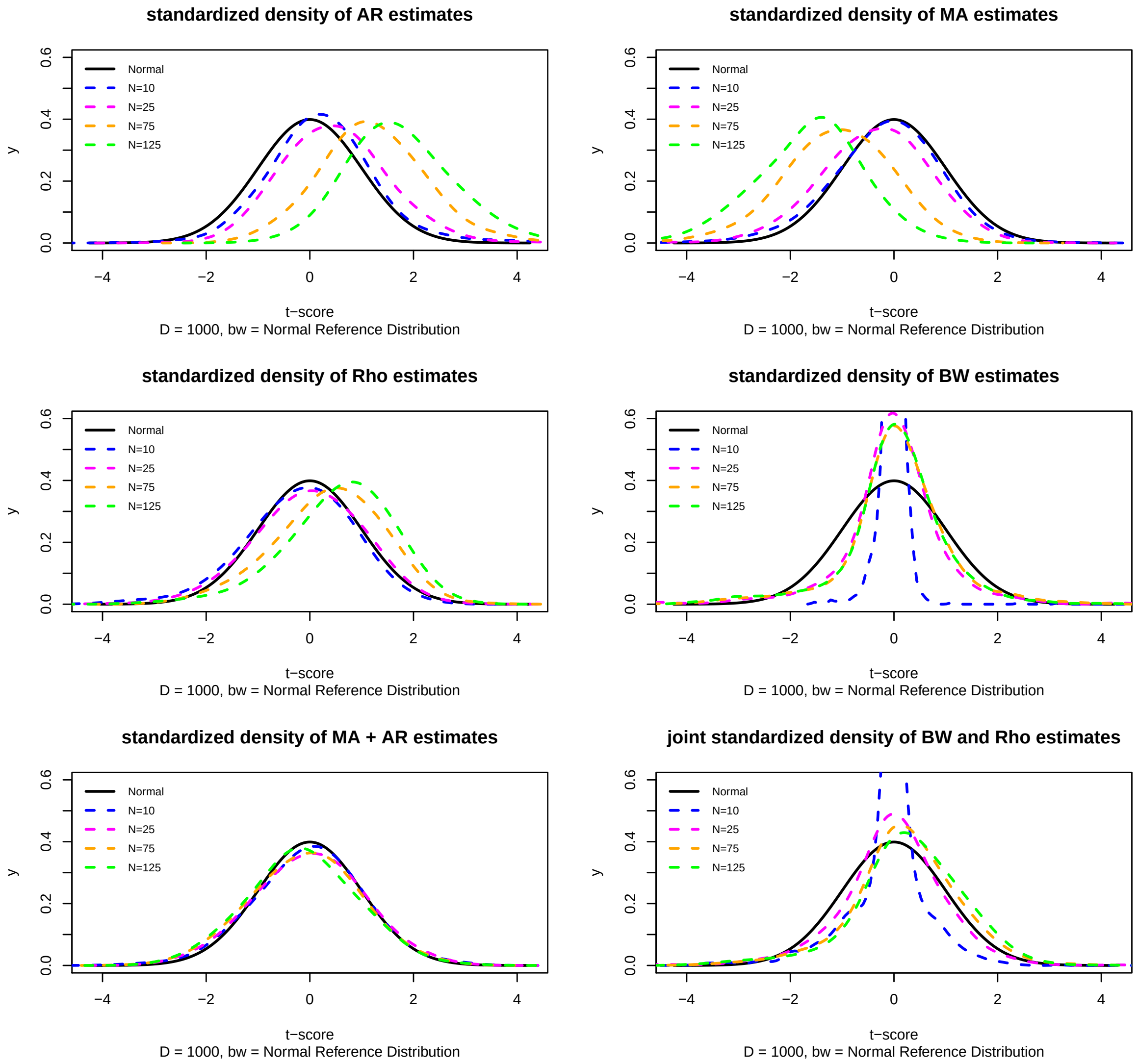
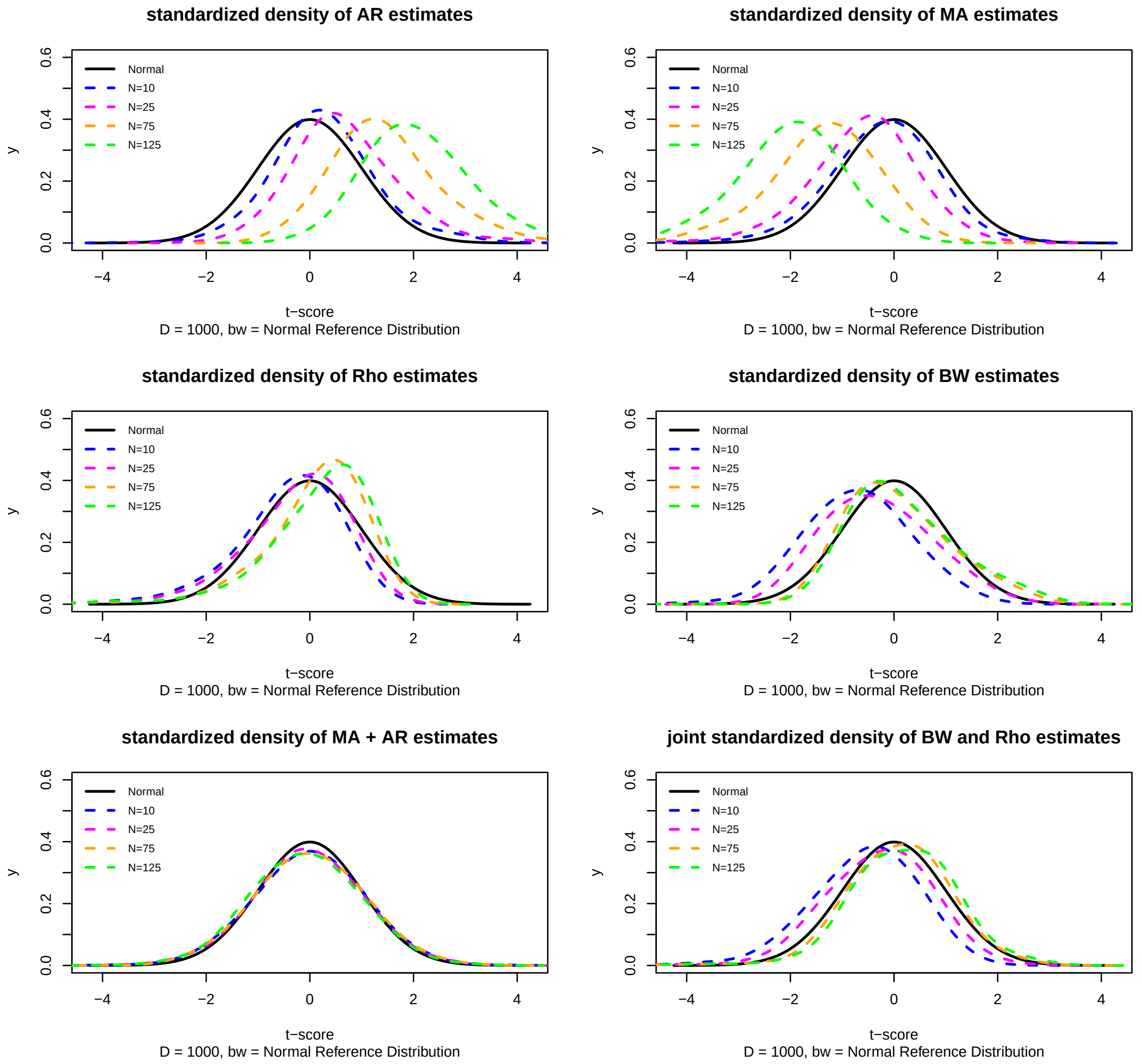


Figure 7: Unpenalized small sample distributions of the correctly specified SARMA, bandwidth of the spatial kernel matrix in the DGP set to .15.


standardized density of AR estimates

standardized density of MA estimates

standardized density of Rho estimates

standardized density of BW estimates

standardized density of MA + AR estimates

joint standardized density of BW and Rho estimates

Figure 8: Penalized small sample distributions of the correctly specified SARMA, bandwidth of the spatial kernel matrix in the DGP set to 2.

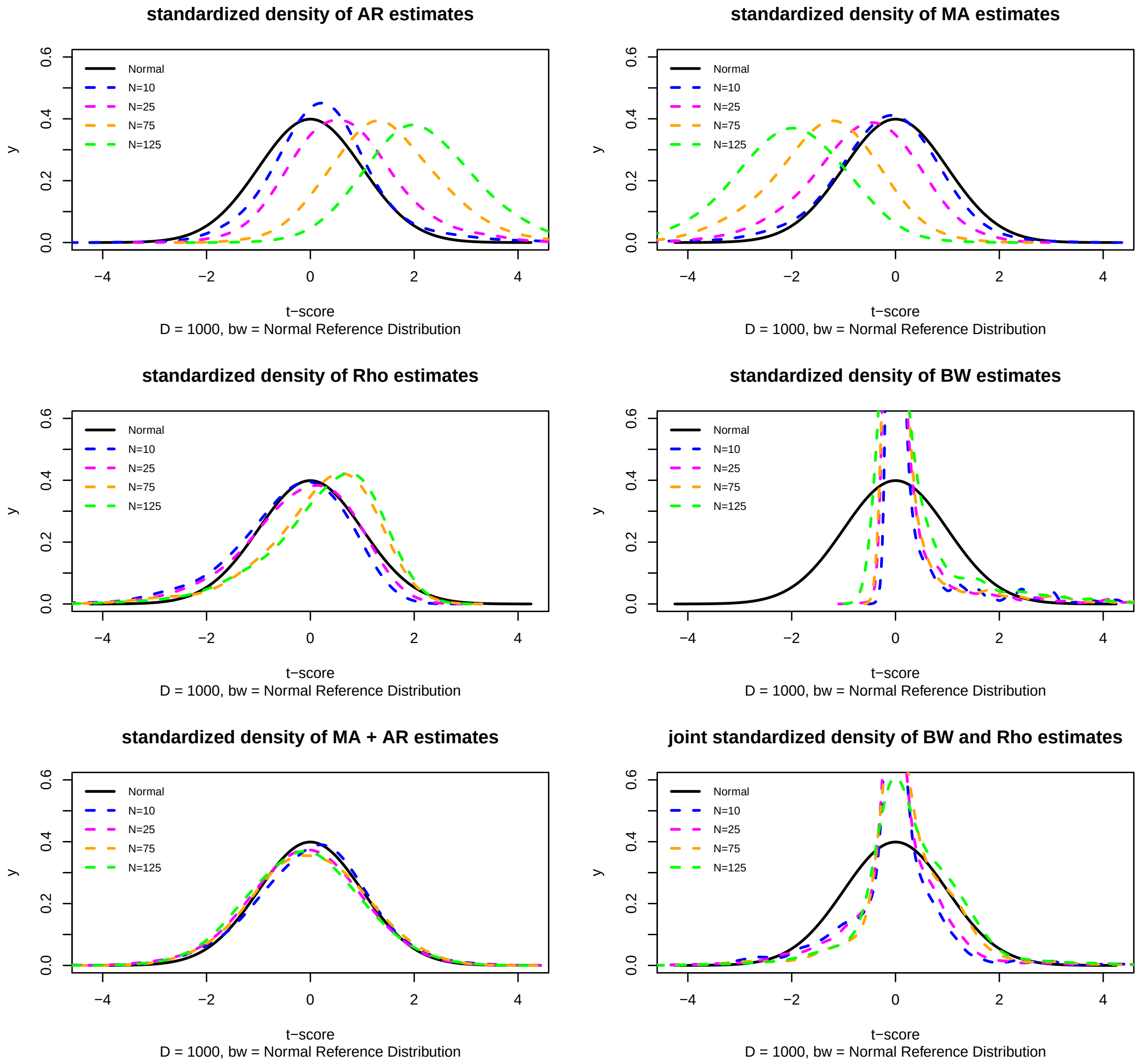
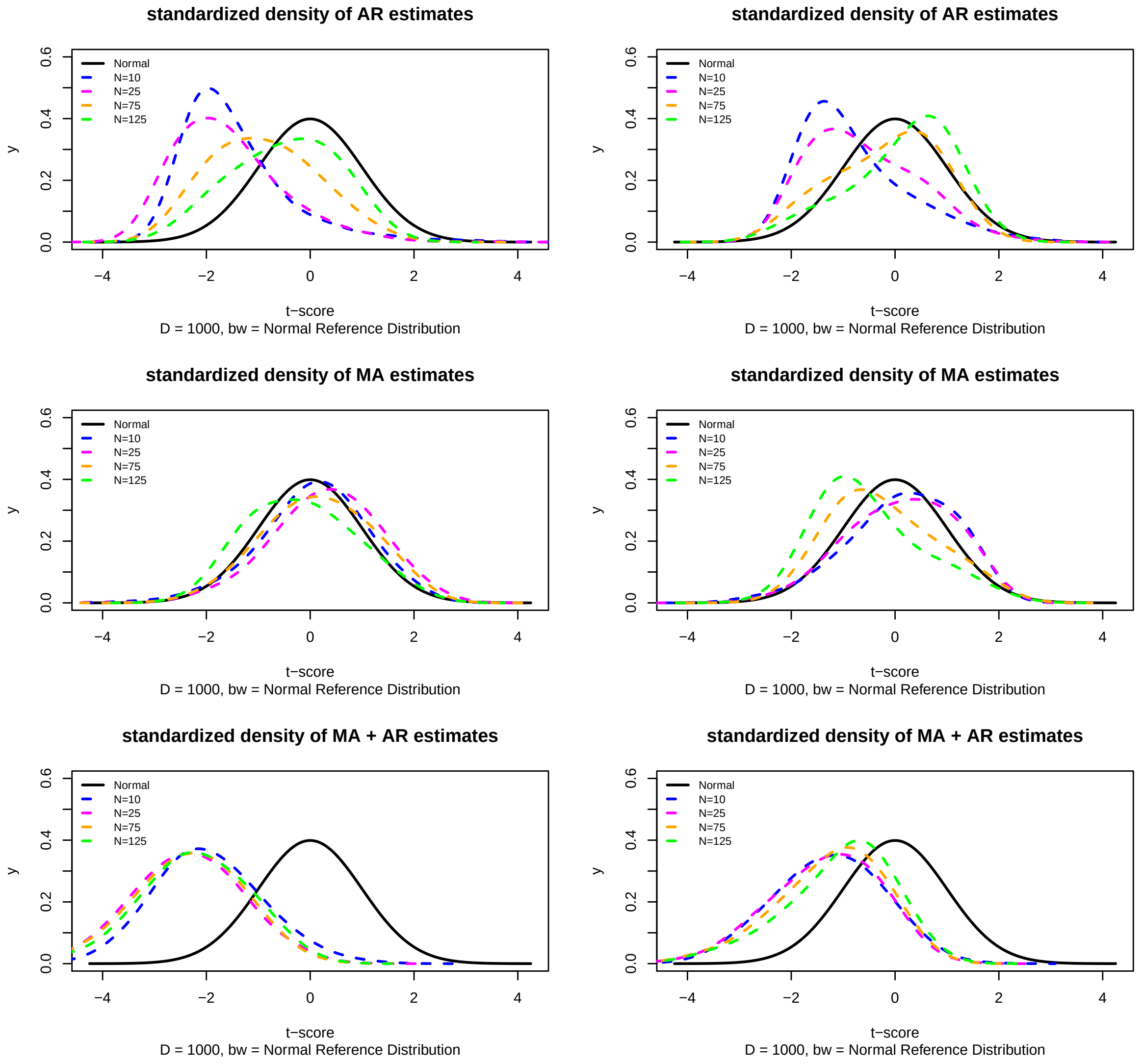


Figure 9: Unpenalized small sample distributions of the correctly specified SARMA, bandwidth of the spatial kernel matrix in the DGP set to 2.



standardized density of AR estimates

standardized density of AR estimates

standardized density of MA estimates

standardized density of MA estimates

standardized density of MA + AR estimates

standardized density of MA + AR estimates

Figure 10: 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).

B.2 Data

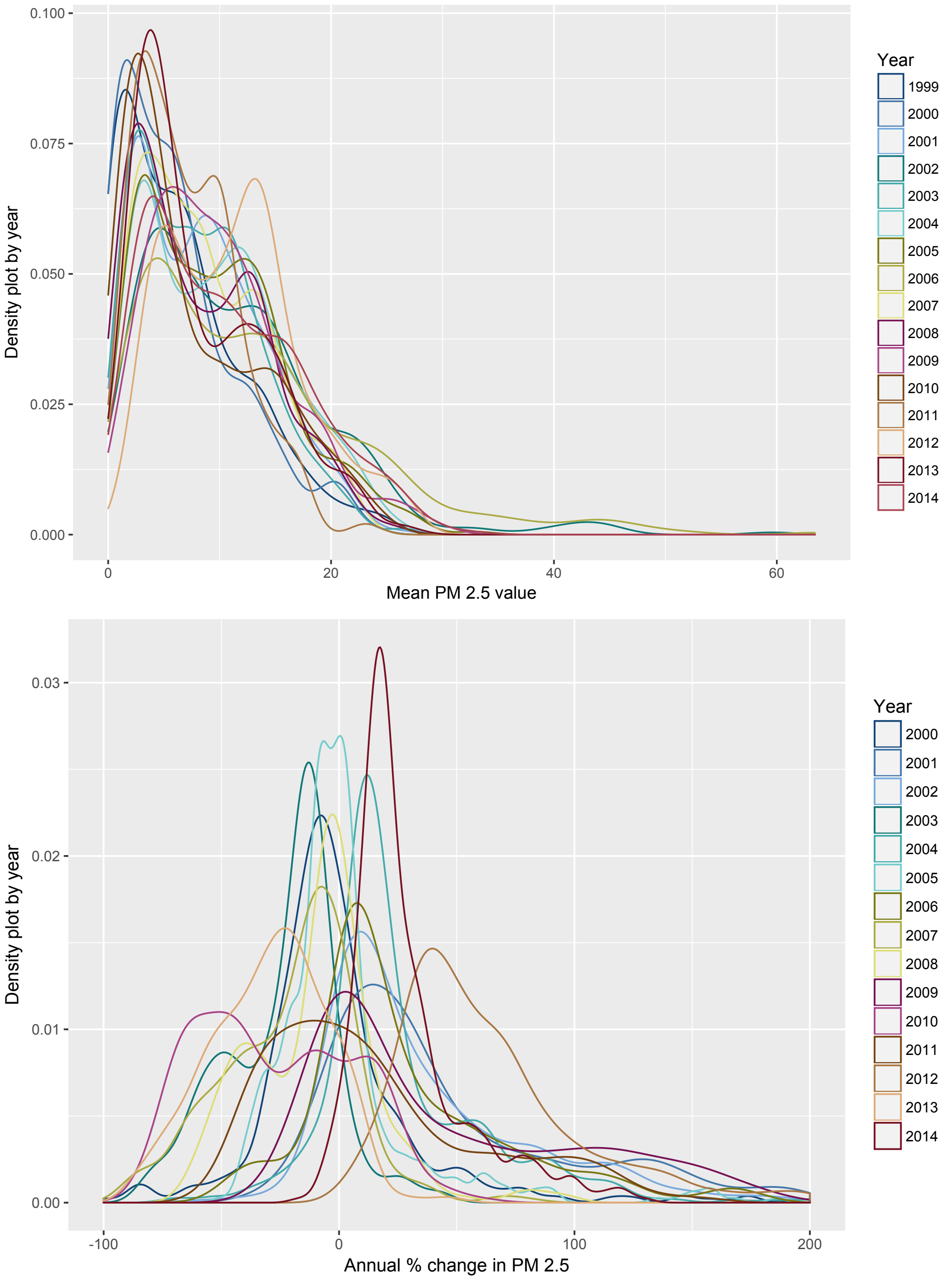


Figure 11: Densities of pollution levels (left) and changes in pollution (right) for 219 areas with an urban patch of over 5,000 people and densities of 300 per square kilometer or higher.

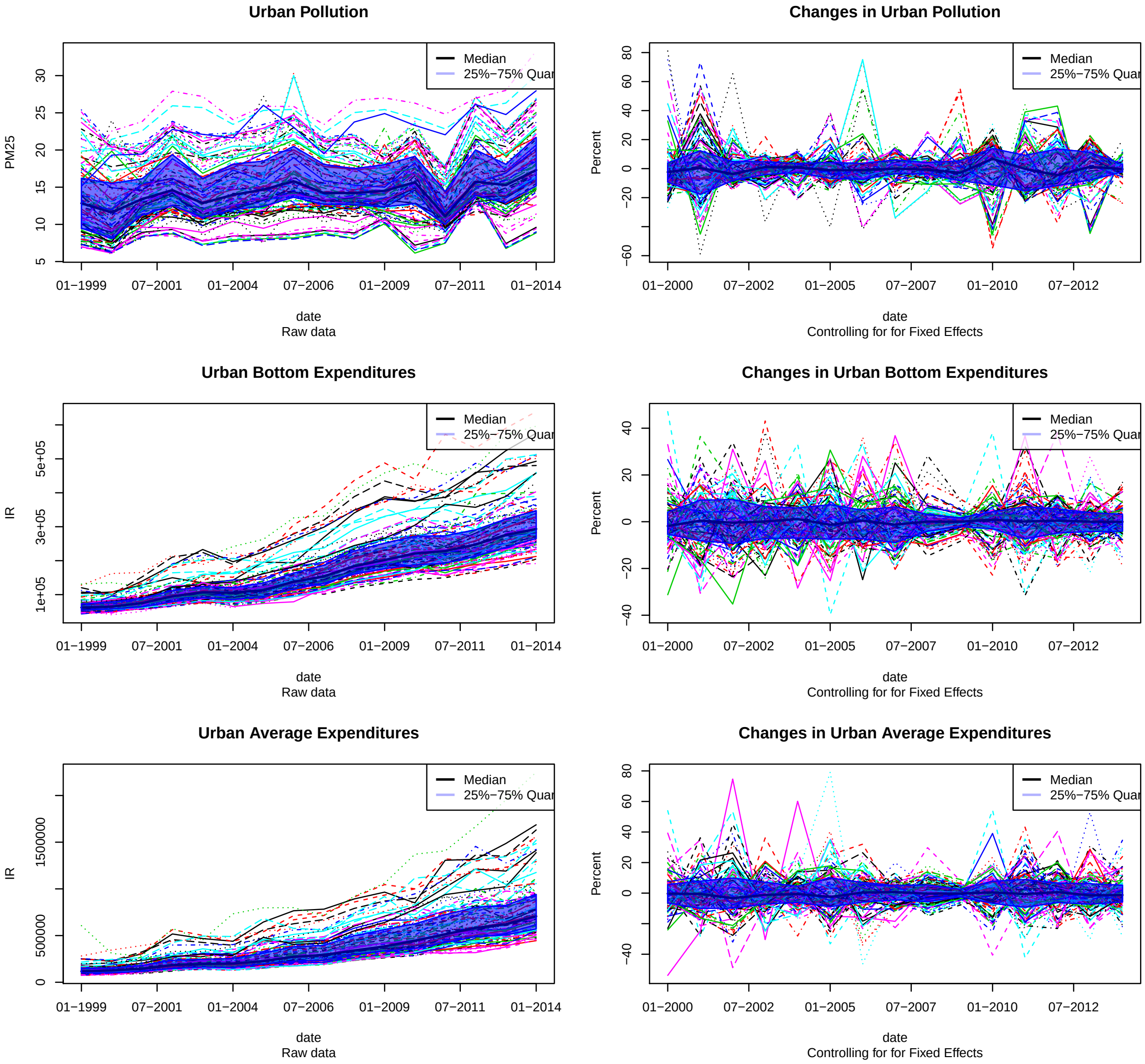
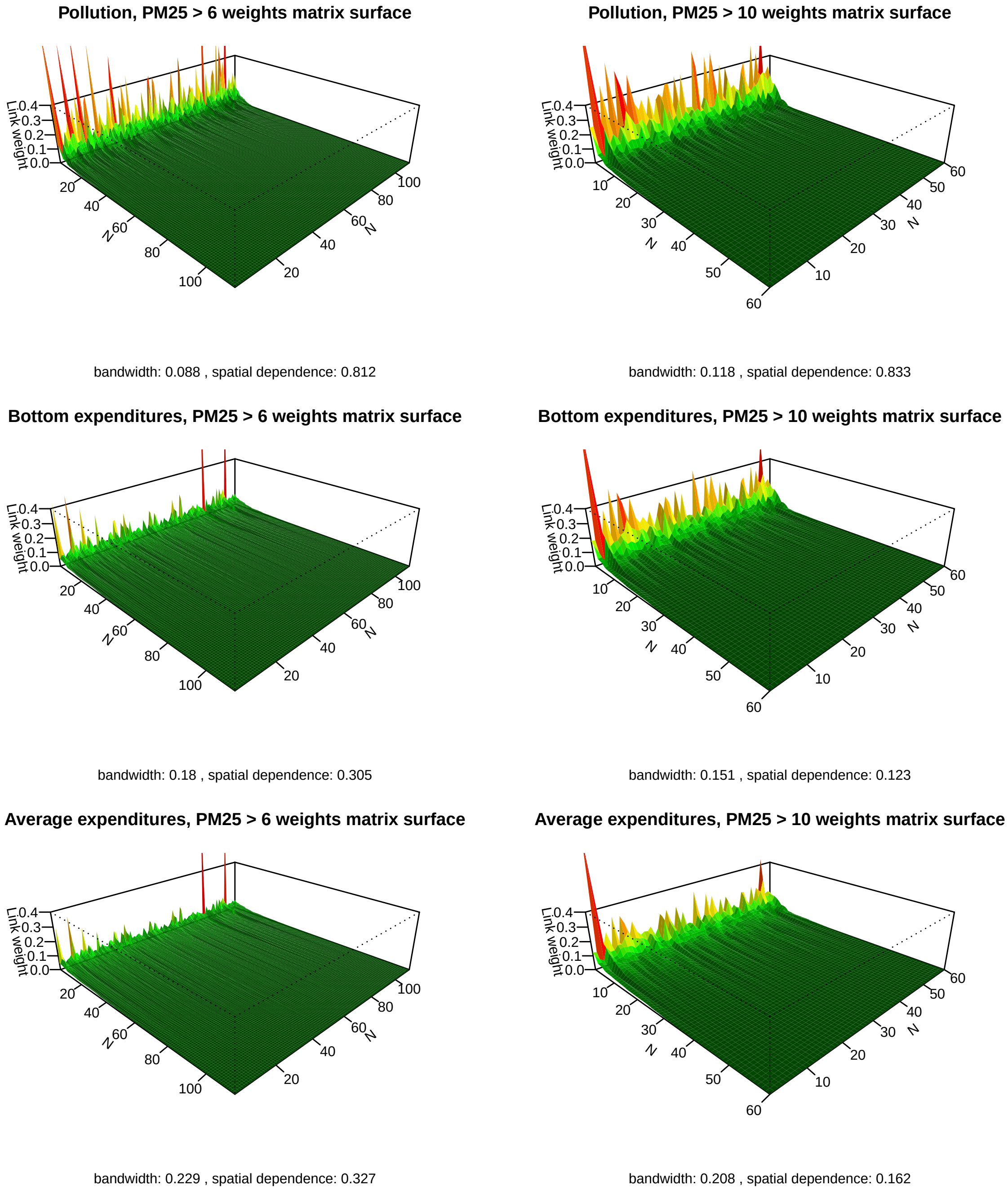


Figure 12: Cross-sectional time series plots of 113 urban areas with the quantiles and medians shaded. Left levels, right percentage changes.

B.3 Regression Results


Pollution, PM25 > 6 weights matrix surface

bandwidth: 0.088 , spatial dependence: 0.812

Pollution, PM25 > 10 weights matrix surface

bandwidth: 0.118 , spatial dependence: 0.833

Bottom expenditures, PM25 > 6 weights matrix surface

bandwidth: 0.18 , spatial dependence: 0.305

Bottom expenditures, PM25 > 10 weights matrix surface

bandwidth: 0.151 , spatial dependence: 0.123

Average expenditures, PM25 > 6 weights matrix surface

bandwidth: 0.229 , spatial dependence: 0.327

Average expenditures, PM25 > 10 weights matrix surface

bandwidth: 0.208 , spatial dependence: 0.162

Figure 13: 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 1: 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
	(pol)	(bot)	(exp)
$\phi\ pol_{t-1}$	-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 2: 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
	(pol)	(bot)	(exp)
$\phi\ pol_{t-1}$	-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 3: VARMA(AICc) results at $\mathrm{PM}_{2.5} > 10$, $\hat{R}^2 = 0.722$, 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
	(pol)	(bot)	(exp)
$\phi\ pol_{t-1}$	-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 4: 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
	(pol)	(bot)	(exp)
ϕ pol_{t-1}	-0.097**	-0.150***	-0.085**
	(-2.503)	(-2.803)	(-2.367)
ϕ pol_{t-2}		-0.073**	-0.049
		(-2.048)	(-1.482)
ϕ pol_{t-3}	0.041*	-0.045	
	(1.773)	(-1.446)	
ϕ bot_{t-1}	0.092**		
	(2.446)		
ϕ bot_{t-2}		-0.271***	-0.200***
		(-4.573)	(-3.518)
ϕ bot_{t-3}			-0.093**
			(-2.326)
ϕ exp_{t-1}			-0.324***
			(-5.421)
ϕ exp_{t-2}		0.133***	
		(2.995)	
ϕ 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.

Table 5: Cumulative effects after 15 years following an initial 10% increase per variable.

	$PM_{2.5} > 6$			$PM_{2.5} > 10$		
	25%	50%	75%	25%	50%	75%
Impulse: Pollution						
Pollution	−28.203	−14.470	−6.403	−50.507	−17.071	0.334
Bottom	−3.066	−2.227	−1.534	−8.312	−5.742	−3.876
Average	−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	6.504	7.192	7.940	4.389	5.132	5.928
Average	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	−0.329	−0.027	0.268	−0.919	−0.634	−0.363
Average	2.522	2.954	3.330	1.525	2.146	2.772

Table 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 expenditures 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