
# ESTIMATING NONLINEARITIES IN SPATIAL AUTOREGRESSIVE MODELS

NICOLAS DEBARSY & VINCENZO VERARDI


UNIVERSITY OF NAMUR

WP 1016

DEPARTMENT OF ECONOMICS
WORKING PAPERS SERIES

# Estimating Nonlinearities in Spatial Autoregressive Models*

Nicolas Debarsy† and Vincenzo Verardi‡

**Abstract**

In spatial autoregressive models, the functional form of autocorrelation is assumed to be linear. In this paper, we propose a simple semiparametric procedure, based on Robinson's (1988) double residual methodology, that relaxes this restriction. Simple simulations show that this model outperforms traditional SAR estimation when nonlinearities are present. We then apply the methodology on real data to test for the spatial pattern of voting for independent candidates in US presidential elections. We find that in some States, votes for "third candidates" in some counties are non-linearly related to votes for "third candidates" in neighboring counties, which encourages for strategic behavior.

**Keywords:** Spatial econometrics, semiparametric estimations

**JEL Classification: C14, C21**

*We would like to thank all our colleagues at CRED, ECARES, and CERPE as well as Emmanuel Flachaire, Marjorie Gassner, Whitney Newey, Christopher Parmeter, Darwin Ugarte, Sébastien Van Bellegem, the editor and anonymous referees for useful comments.

†Corresponding author, CERPE, Facultés Universitaires Notre Dame de la Paix de Namur. E-mail: ndebarsy@fundp.ac.be. Nicolas Debarsy is a Doctoral Researcher of the FNRS and gratefully acknowledges their financial support.

‡CRED, Facultés Universitaires Notre Dame de la Paix de Namur; ECARES, CKE, Université Libre de Bruxelles. E-mail: vverardi@fundp.ac.be. Vincenzo Verardi is an Associated Researcher of the FNRS and gratefully acknowledges their financial support.

# 1 Introduction

In spatial econometrics, autoregressive models (or SAR) have been developed to estimate how changes in a given variable spread over the neighborhood (see Anselin, 1988 and LeSage and Pace, 2009, for further details). This effect is generally assumed to be linear, which is obviously restrictive. The objective of this paper is to propose a simple estimation method, based on Robinson's (1988) double residual estimator, that allows for non-linear spatial interdependence.

The structure of the paper is the following: after this short introduction, in section 2 we propose a procedure to estimate a semiparametric spatial autoregressive model. Besides, we recommend a test that compares parametric adjustment with nonparametric fits aiming at understanding whether assuming linearity in the autoregressive component is legitimate. In section 3 we present some simple simulations to assess the performance of the procedure while section 4 sets out an empirical example based on US presidential elections. Section 5 concludes.

# 2 Estimation method

## 2.1 Nonlinear spatial autoregressive model

The general form of a linear first order spatial autoregressive model is

$$y_i = \mathbf{x}_i\theta + \rho(Wy)_i + \varepsilon_i, \quad i = 1, ..., N \tag{1}$$

where $y_i$ is the value taken by the dependent variable and $\mathbf{x}_i$ is the row vector of characteristics of individual $i$. $Wy_i$ measures the value of $y$ in the neighborhood $N_{(i)}$ of individual $i$ and is defined as $Wy_i = \sum_{j \in N_{(i)}} w_{ij}y_j$, where $w_{ij}$ models spatial interactions between $i$ and $j$. Column vector $\theta$ and the autoregressive spatial parameter $\rho$ are the coefficients to be estimated and $\varepsilon$ is assumed $iid$ with zero mean and constant variance $\sigma^2$. Given the linearity of (1), a unit change in $Wy$ is associated with a $\rho$ units change

in the conditional expectation of $y$, whatever the value of $Wy$. This assumption could be relaxed by considering a more general model of the following type

$$y_i = \mathbf{x}_i\theta + f[(Wy)_i] + \varepsilon_i, \quad i = 1, ..., N \tag{2}$$

where $f$ is not constrained to any specific form. This model can be estimated extending Robinson's (1988) double residual methodology to spatial autoregressive models. A characteristic of these models is the endogeneity of the spatial lag of the dependent variable, $\mathbf{Wy}$. To deal with this, we suggest using the control function approach (CFA hereafter) developed in Newey et al. (1999). This two-step procedure consists in first regressing the troublesome covariate ($\mathbf{Wy}$ in this case) on a set of instruments and fitting the residuals. In the second step, one reintroduces these residuals in the initial model (2) to explicitly account for the endogeneity. As variable $\mathbf{Wy}$ enters equation (2) nonparametrically, all the instruments traditionally considered in the spatial econometrics literature can not be used (see Kelejian and Prucha (1998, 1999) for further details). However, from (2) it is evident that a linear relation exists between $\mathbf{Wy}$ and $\mathbf{Wx}$. The latter ($\mathbf{Wx}$), that can be interpreted as the neighboring values of the explanatory variables, can so be used to instrument the former ($\mathbf{Wy}$). This first step equation can therefore be written as:

$$Wy_i = \mathbf{x}_i\pi + \mathbf{Wx}_i\gamma + \eta_i \tag{3}$$

To implement the CFA, it is necessary to assume that endogeneity in the $\mathbf{Wy}$ term in (2) is accounted for linearly:

$$E[\varepsilon_i|Wy_i, \eta_i] = \delta\eta_i$$

This implies that the error term in (2) can be written as $\varepsilon_i = \delta\eta_i + \omega_i$, where $\omega_i$ is exogenous to regressors. This assumption is a simplified version of the one presented in Newey et al (1999, p.566). Equation (2) can thus be rewritten as

$$y_i = f(Wy_i) + \mathbf{x}_i\theta + \delta\eta_i + \omega_i \tag{4}$$

where endogeneity is now explicitly accounted for through the additional term $\delta\eta$ ($\eta$ is estimated by $\hat{\eta}$, the residual fitted in equation (3)).

To estimate (4), we use Robinson's (1988) semiparametric estimator which consists in computing the conditional expectations of $\mathbf{y}$ and $\mathbf{x}$ with respect to $\mathbf{Wy}$. To do so, we regress nonparametrically $\mathbf{y}$ and each of the regressors (including $\hat{\eta}$) on $\mathbf{Wy}$ and compute the residuals. Formally, we estimate:

$$\begin{aligned} y_i &= f_y(Wy_i) + \epsilon_1 \\ x_{i,k} &= f_{x,k}[(Wy)_i] + \epsilon_{2,k}, \quad k = 1, \ldots, K \\ \hat{\eta}_i &= f_\eta(Wy_i) + \epsilon_3 \end{aligned} \tag{5}$$

where $K$ is the number of covariates entering the parametric part. We then estimate $\theta$ and $\delta$ in

$$\hat{\epsilon}_1 = \hat{\epsilon}_2\theta + \delta\hat{\epsilon}_3 + \nu \tag{6}$$

where $\hat{\epsilon}_1 = \mathbf{y} - \widehat{E(\mathbf{y}|\mathbf{Wy})}$, $\hat{\epsilon}_{2,k} = \mathbf{x}_k - \widehat{E(\mathbf{x}_k|\mathbf{Wy})}$ and $\hat{\epsilon}_3 = \hat{\eta} - \widehat{E(\hat{\eta}|\mathbf{Wy})}$. Residuals $\hat{\epsilon}_1$, $\hat{\epsilon}_2$ and $\hat{\epsilon}_3$ are $iid$.

At least, to assess the relation between $y_i$ and $Wy_i$, we run a nonparametric regression of the $y_i$s (filtered of the parametric part) on $Wy_i$. The nonparametric estimator we consider at all stages is a gaussian kernel weighted local polynomial fit of degree 3 (which is a kernel of order 2).[1] The "optimal" bandwidth used minimizes the conditional weighted mean integrated squared error.

To test for the appropriateness of the linearity (or of any other polynomial adjustment) assumption of the relation between $y_i$ and $Wy_i$, we use a test developed by Hardle and Mammen (1993) which compares the nonparametric and the parametric regression fits using square deviations between them. The test statistic is:

$$T_n = N\sqrt{(h)}\sum_{i=1}^{N}\left(\hat{f}(Wy_i) - \hat{f}(Wy_i, \theta)\right)^2 \pi(Wy_i) \tag{7}$$

where $\hat{f}(Wy_i)$ is the nonparametric function estimated in (4), $\hat{f}(Wy_i, \theta)$ is an estimated parametric function and $h$ is the bandwidth used. $\pi()$ is a weight function set to $1/N$

[1] Higher order kernels could be used.

in this paper. A more complex function could be used if needed (as, for example, to cope with heteroskedasticity).

To obtain critical values for the test, Hardle and Mammen (1993) suggest relying on simulated values calculated by wild bootstrap. Obviously, an absence of rejection of the null (i.e. "accepting" the parametric model) means that the polynomial adjustment is at least of the degree that has been tested.

To assess the performance of the proposed methodology, we present some simple simulations in the following section.

# 3 Simulations

The four following data generating processes (DGP) are considered:

a) $y_i = \mathbf{x}_i\theta + \varepsilon_i$

b) $y_i = 0.75Wy_i + \mathbf{x}_i\theta + \varepsilon_i$

c) $y_i = 0.75Wy_i - 0.4\left(Wy_i\right)^2 + \mathbf{x}_i\theta + \varepsilon_i$

d) $y_i = \left(\frac{1}{1+\exp(-2Wy_i)} - 0.5\right) + \mathbf{x}_i\theta + \varepsilon_i$

where $\mathbf{x}_i$ is a $1 \times 3$ vector whose elements are drawn from three independent $N(0,1)$, $\theta$ is a $3 \times 1$ vector of ones and $\varepsilon_i \sim N(0,1)$. The simulated sample size is 300. The x-coordinates are generated from a $U[0,20]$ and the y-coordinates from a $U[0,50]$. Spatial weights are

$$w_{ij} = \begin{cases} \dfrac{1/b_{ij}}{\sum\limits_j 1/b_{ij}} & \text{if } b_{ij} < \bar{b} \\ 0 & \text{otherwise} \end{cases}$$

where $b_{ij}$ are all pairwise distances. Parameter $\bar{b}$ (the threshold value above which the interaction between $i$ and $j$ is assumed to be negligible) is set to 5. By convention, $w_{ii} = 0$.

To illustrate the fitting performance of the proposed estimation procedure, we generate four samples according to the DGPs discussed above and present the scatter plots, the non-parametric fit (thick plain line) and the true DGP (thin dashed line) in Figure 1. As expected, the results are unambiguous.

[INSERT FIGURE 1 HERE]

In the case of no spatial autocorrelation (panel a), no clear pattern emerges and the non-parametric curve lies close to the horizontal line (the true DGP). In the three other cases (panels b, c and d), the nonparametric estimation of the autocorrelation matches the true functional form quite well. The last two panels (c and d) shed doubt on the appropriateness of a linear approximation for the spatial component.

As mentioned in the previous section, the $T_n$ statistic assesses the adequacy of a polynomial adjustment compared to a nonparametric fit. Table 1 presents the performance of the test for the three first DGPs described above. The rows indicate the order of the polynomial generated in the DGP while the columns designate the order of polynomial tested. Thus, the diagonal elements indicate the size of the test while elements below the main diagonal assess some measure of power. The terms above the main diagonal illustrate the behavior of the test when the order tested is greater than the true polynomial adjustment. The results of Table 1 are obtained by replicating the three DGPs above 1000 times. Each time, a new error term is randomly drawn and a new dependent variable is generated while the design space is kept unchanged. The bootstrap inference for the test is based on 200 simulations. We observe that the test has good rejection rates when the order of the polynomial adjustment tested is lower than the true one. Furthermore, the size of the test (whose theoretical value is set at 5%), presented in diagonal elements, is not far from its nominal value even though the test seems conservative.

[INSERT TABLE 1 HERE]

# 4 Application

In this section, we present an illustrative example of the nonlinear SAR model. The objective of the analysis is to study the voting behavior for independent candidates (focusing on the US presidential elections of 2000) in a given county as a function of the numbers of votes cast for this candidate in neighboring counties (which are assumed to be well anticipated by electors). The hypothesis tested is that electors will not vote for the third candidate if they believe that it might help the candidate they dislike the most to win the elections (as occurred, for instance, in the first round of the French presidential elections of 2002 when Jean-Marie Le Pen, the extreme right wing candidate, obtained 16.86% of the vote and qualified for the second round at the expense of the socialist candidate).

Hence, if interested electors anticipate that the third candidate will collect a limited number of votes, they will vote for him to declare their dissatisfaction with the political establishment. Furthermore, the resulting share of votes is expected to increase jointly with the share of votes in the neighborhood as the message sent will be stronger. However, if electors perceive that the amount of votes obtained by the third candidate will jeopardize the political scenario they will stop voting for him. We therefore expect to observe a concave-shaped spatial autoregressive component in the vote for the third candidate. To test for this, we estimate the relation

$$y_i = \mathbf{x}_i\theta + f[(Wy)_i] + \varepsilon_i, \quad i = 1, ..., N \tag{8}$$

where variable $\mathbf{y}$ is the log of the vote share cast for outsider candidates and $i$ indexes counties. The control variables ($\mathbf{x}$) are those generally considered in this type of regression i.e. (i) the log of the vote shares of independents in the previous elections, (ii) the log of the vote share of Republicans, (iii) the average per capita income, (iv)

the log of the ratio between Democrats and Republicans votes in the previous elections and (v) the ethnic composition (i.e. the proportion of blacks, whites, while proportions of all the others ethnics are grouped in a third variable considered as the reference category).[2] The weighting matrix is defined as follows: counties located in different States do not interact.[3] Within States, we assign a spatial weight proportional to the inverse of the distance between counties' centroids which implicitly assumes that individuals are better informed on closer counties. Data come from Polidata, a national demographic and political data consulting firm in the US.

We focus on New York, North Dakota, Pennsylvania and Tennessee where the concaved-shaped relation is clear.[4]

[INSERT FIGURE 2 HERE]

Figure 2 shows the nonlinear relation between the vote share (in logs) for independent candidates in a county and its neighborhood. For the State of North Dakota, we clearly observe that the sincere voting behavior occurs as long as the vote share for independents in neighboring counties is not too high. Indeed, in this situation, an increase in the vote share for independents in the neighborhood induces an increase in the vote share for independents in the considered county. This is probably due to the fact that electors believe that their vote will strengthen the message conveyed by the neighbors on the dissatisfaction with the political establishment. However, when the votes for outsiders become more numerous in the neighborhood, the "votes for change" start decreasing. This could be explained by the fact that voters realize that they should vote strategically to prevent the candidate they dislike the most from winning the elections. Estimating a traditional (linear) spatial autoregressive model would have led to the conclusion of absence of link between vote shares (in logs) in a county and

[2]Including the log of the vote share for the Democrats instead does not affect results.

[3]This assumption is based on the majoritarian system in place.

[4]The graphs for all the other States are available from the authors upon request.

its neighborhood. This theory seems to hold well in New York, Pennsylvania and Tennessee, where a concave-shaped relation appears.

Table 2 presents the results of the statistic that compares the parametric and non-parametric modeling of the relation between the vote share (in logs) for independents in a county and its neighborhood. Columns 2 to 5 report the p-values of the test when different degrees for the parametric part are assumed. Hence, column 2 assumes the absence of any relation, column 3, a linear one, column 4 supposes a quadratic link while the fifth column assumes a cubic relation between the vote share (in logs) for independent candidates in a county and in its neighborhood. The last column reports the p-value associated to the spatial autoregressive parameter if a standard linear spatial autoregressive model is estimated (as in equation (1)). The inference for the test is based on 200 simulations using wild bootstrap.

[INSERT TABLE 2 HERE]

For the States of New York and Pennsylvania, we cannot reject the absence of relation between $y$ and its spatial lag. This fact is corroborated by the non significance of the $\rho$ parameter (p-values of 79.3% and 80.4% respectively). For Tennessee, according to the test, a quadratic form best models the relation between vote shares (in logs) for independent candidates in a county and its neighborhood. Indeed, column 4 does not reject the null of a quadratic form. However, if we had performed the traditional (linear) spatial autoregressive model, we would have concluded to the absence of spatial effects (p-value of $\rho$ is 14.5%). This can be explained by the concavity of the quadratic relation. For the State of North Dakota, p-values of the comparison test indicate that the polynomial adjustment needed to capture the form of the link between $y$ and $Wy$ should be at least of order 3. Again, estimating only the linear spatial autoregressive model would lead to conclude to the absence of any relation (p-value of $\rho$ is 56.5%).

Let us finally note that although plotting the relation is useful, it should not blindly guide the applied researcher because there might be an optical illusion. Performing the test-statistic for different parametric polynomial degrees is a more valuable tool to

make a decision.

# 5 Conclusion

In spatial econometrics, the spatial autoregressive model is one of the most commonly used. In this paper, we propose a simple generalization for the case of a non-linear spatial autoregressive component. We present some simulations and a simple empirical application to show the usefulness of the procedure.

# References


[1] Anselin L (1988) *Spatial Econometrics, Methods and Models*. Kluwer Academic Publishers, Dordrecht

[2] Hardle W, Mammen E (1993) Comparing nonparametric versus parametric regression fits. *The Annals of Statistics* 21: 1926-1947

[3] Kelejian HH, Prucha IR (1998) A generalized spatial two stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics* 17: 99-121

[4] Kelejian HH, Prucha IR (1999) A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review* 40: 509-533

[5] LeSage J, Pace RK (2009) *Introduction to spatial econometrics*. CRC Press/Taylor and Francis Group, London

[6] Newey WK, Powell JL, Vella F (1999) Nonparametric estimation of triangular simultaneous equation models. *Econometrica* 67: 565-603

[7] Robinson PM (1988) Root-N-consistent semiparametric regression. *Econometrica* 56: 931-954

Table 1: Performance of the comparison test $T_n$

| | Order 0 | Order 1 | Order 2 |
|---|---|---|---|
| Order 0 | 0.067 | 0.018 | 0.031 |
| Order 1 | 1 | 0.02 | 0.04 |
| Order 2 | 1 | 0.94 | 0.045 |

Figures correspond to rejection rates.

Table 2: Results of the comparison test $T_n$

| State | Order 0 | Order 1 | Order 2 | Order 3 | $\rho$ (p-value) |
|---|---|---|---|---|---|
| New York | 0.095 | 0.220 | 0.550 | 0.885 | 0.793 |
| North Dakota | 0.040 | 0.040 | 0.045 | 0.130 | 0.565 |
| Pennsylvania | 0.160 | 0.130 | 0.320 | 0.450 | 0.804 |
| Tennessee | 0.035 | 0.045 | 0.210 | 0.875 | 0.145 |

Figures correspond to the p-values. Order corresponds to the order of the parametric relation assumed (0 for a constant, 1 for linear, 2 for quadratic and 3 for cubic).

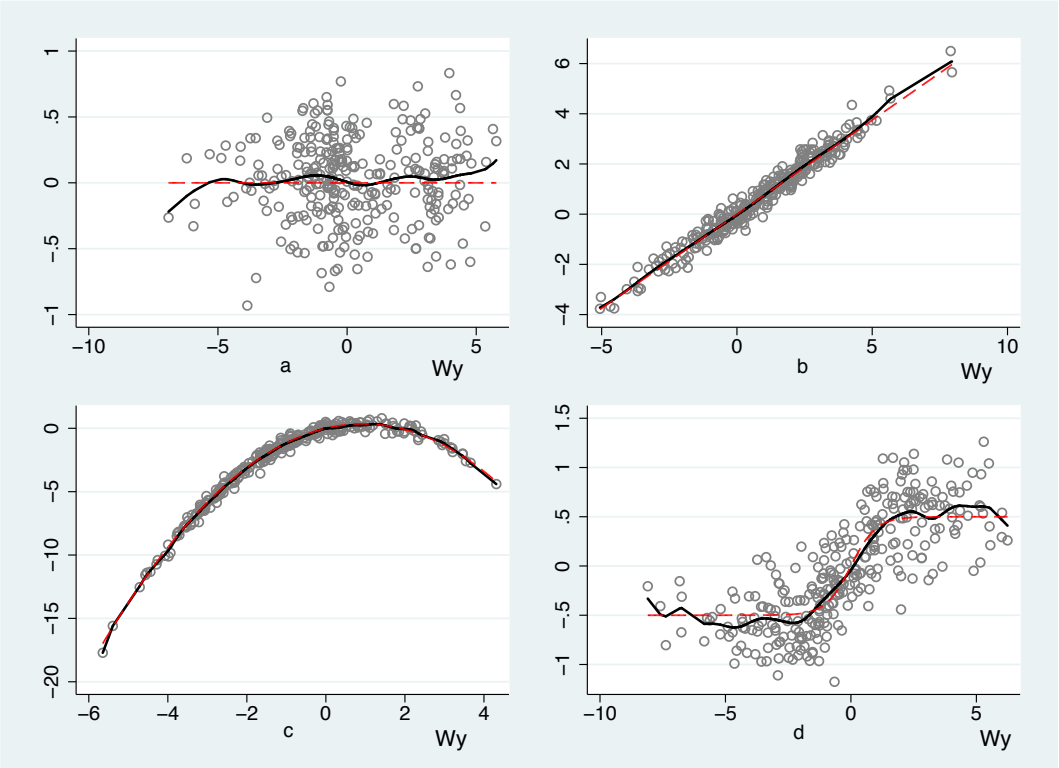


Figure 1: Non-parametric fit of spatial autocorrelation

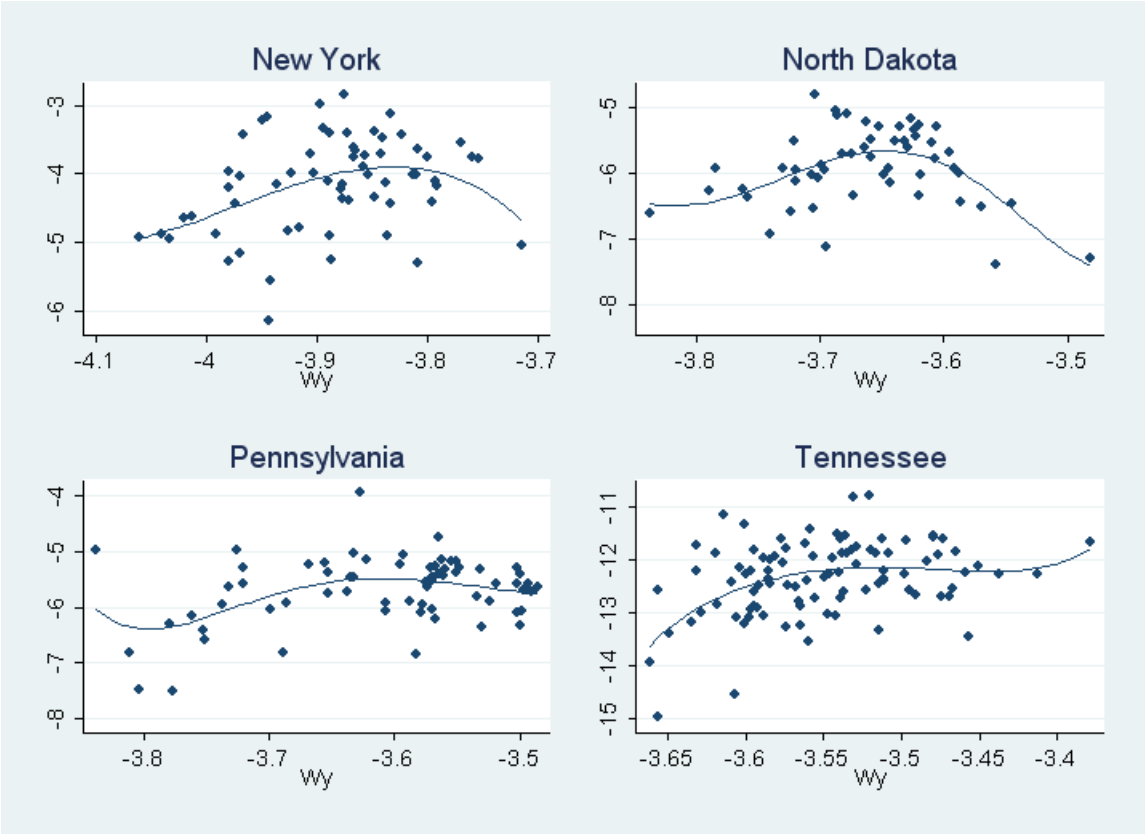


Figure 2: Nonlinear SAR by State