# The Local Voter: A Geographically Weighted Approach to Ecological Inference

**Ernesto Calvo** University of Houston
**Marcelo Escolar** Universidad de Buenos Aires

*Drawing inferences about individual behavior from aggregate ecological data has been a persistent problem in electoral and behavioral studies, in spite of important methodological advances. In a recent article Anselin and Tam Cho (2002) provided Monte Carlo evidence that King's Ecological Inference (EI) solution will produce biased estimates in the presence of extreme spatial heterogeneity. In this article we provide further empirical evidence that supports their findings and shows that in the presence of spatial effects the residuals of Goodman's naïve model exhibit the same spatial structure that King's local* $B_i^B$ *estimates. Solving for extreme spatial heterogeneity, it is argued here, requires controlling the omitted variable bias expressed in the spatial structure of much ecological data. In this article we propose a Geographically Weighted Regression approach (GWR) for solving problems of spatial aggregation bias and spatial autocorrelation that affect all known methods of ecological inference. The estimation process is theoretically intuitive and computationally simple, showing that a well-specified GWR approach to Goodman and King's Ecological Inference methods may result in unbiased and consistent local estimates of ecological data that exhibit extreme spatial heterogeneity.*

In the last few years significant advances have been made both in the statistical exploration of local politics (quantitative geography) and in the estimation of individual-level behavior from aggregate data (ecological inference). Nevertheless, little research has been conducted combining the different strengths of both approaches.[1] As a result, two important problems—*extreme spatial heterogeneity*[2] and *spatial autocorrelation*[3]—have not been properly addressed by most of this literature, and a good deal of information has not been used because of this lack of attention to the importance of location in local politics.

We show here that the two most popular methods of ecological inference, Goodman's regression and King's Ecological Inference (EI), are far from immune to the estimation problems that arise from data exhibiting extreme spatial heterogeneity. The results are particularly troublesome in the case of King's EI: running an uncorrected model produced estimates that were significantly biased. This bias results from the covariance parameter $\rho$ of the truncated bivariate normal distribution used by EI modeling the spatial structure of the data rather than the presumed covariance between the explanatory variables $X_i$ and 1-$X_i$. In effect, the higher versatility of King's method to describe local variation also makes EI more susceptible to spatial aggregation problems.

We develop a geographically weighted autoregressive[4] approach (GW-AR) to correct for extreme spatial heterogeneity with applications to Goodman's

We thank Jon Agnew, Isabella Alcañiz, Charles Brunsdon, Tulia Falleti, Chuck Finocchiaro, Edward Gibson, Mark Jones, Gary King, Ricardo Martini, Vicky Murillo, Ben Schneider, Wendy Tam Cho, and two anonymous reviewers for their comments and suggestions. A previous version of this article was presented at the 2002 Conference on Ecological Inference, Harvard University. Stata files (do. files) to estimate GWR models or replicate this study available at ecalvo@uh.edu or peel@utdt.edu.

[1] See Anselin and Tam Cho (2002) for a notable exception.

[2] "The phenomenon were a model (i.e., parameters, functional specification, error specification, etc.) is not constant across spatial observations" (Anselin and Tam Cho 2002). As a result, the model specification changes to accommodate different spatial regimes (Anselin, 1988) or contextual effects (Fotheringham, Brundson, and Charlton 2000; Graaff, Florax, and Nijkamp 1998).

[3] The estimation problems that arise when $y_i$ is related to contiguous observations $y_j$.

[4] More precisely, we use an autoregressive spatial error model estimated by a semi-parametric geographically weighted regression. The approach is similar to a nonlinear estimation of the error term by a general additive model (Brunsdon, Charlton, and Fotheringham 2000; Hastie and Tibshirani 1996; Ormoneit and Hastie 2000). This will be further explained in the third section.

***American Journal of Political Science,*** Vol. 47, No. 1, January 2003, Pp. 189–204

ISSN 0092-5853

regression and King's EI. This procedure, however, can be easily adapted to correct similar problems in other ecological inference models.

The organization of this article is as follows: we first provide an overview of the spatial estimation problems found in ecological data and discuss the relationship between geographic context and ecological inference. We then analyze center-right voter turnout in the City of Buenos Aires to evaluate the performance of Goodman's regression and King's EI and compare the results with baseline models that display no aggregation bias. Geographical and numerical estimates of the error term are presented for both Goodman and EI showing the existence of extreme spatial heterogeneity and spatial autocorrelation in the data. The spatial structure $B_{shi}$ of the voter's turnout is estimated by a Geographically Weighted Regression of the *predicted* turnout on the error term and new corrected models are computed.[5] Finally, we comment on the relative performance of the corrected models and describe the properties of the geographically weighted parameters and quantities of interest.

## Local Politics and Ecological Inference

Recent work on ecological inference has come a long way toward reexamining the theoretical and technical boundaries of a problem that has a long tradition in the social sciences: how to estimate individual level behavior from aggregate data or, more precisely, geographically located aggregate data. King and Agnew's restatement of the claim that *all politics is local* has gained significant ground in our disciplines, and new statistical methods have shifted the focus from summarizing global parameters to modeling local social heterogeneity (Fotheringham 1997; Fotheringham, Brunsdon, and Charlton 2001). Interestingly, however, the conceptual transition toward local politics among political scientists has been blind to the geographic location of such politics. Grouped data has been generally treated as randomly clustered rather than location-specific grouped data, and the dismissal of contextual analysis as a valid methodological strategy has led scholars and policymakers to disregard the omitted variable bias that was recognized by most contextual approaches.[6]

The importance of controlling for spatial dependence when using ecological data was downplayed in King's 1997 seminal book *A Solution to Ecological Inference*. When describing the theoretical assumptions behind EI, King declares that a "third and final assumption of the basic [EI] model is that, conditional on X, Ti and Tj are (spatially) independent for all $i \neq j$. Fortunately, because spatial dependence only affects the variance $\phi$, the consequences of violating this last assumption are usually not very serious" (King 1997, 164).[7] Such a statement contrasts markedly with the conventional wisdom among geographers and regional science scholars, who have shown that serial correlation in multi-dimensional space affects the properties of most estimators to a much larger extent than in time-series analyses (Anselin 1988; Graaff et al. 1998). Moreover, spatial dependence is often a marker for deeper aggregation problems in the spatial structure of the data. These aggregation problems result in *regions* of the data being explained by different spatial regimes or contextual variables.

A recent article by Anselin and Tam Cho (2002) revisits this methodological gap between local politics and location politics that authors such as Agnew (1987, 1996a, 1996b) and King (1996, 1997) debated five years ago. Anselin and Tam Cho compare the relative performance of Goodman's naive ecological inference model and King's EI, finding that EI not only provides poorer individual estimates of data exhibiting extreme spatial heterogeneity but that these estimates also display significant bias. Anselin and Tam Cho, however, do not describe the specific mechanism that induces bias in EI's method in the presence of spatial effects.

[5] For notation purposes, observations with the subscript i should be read as $i_{(east,north)}$, indicating that every observation is associated with a specific geographic location. The spatial parameter $B_{shi}$ is indexed sh for spatial heterogeneity and $i_{(east,north)}$ for every observation in the dataset.

[6] Agnew 1996a; Agnew 1996b; King 1996; King 1997; Flint 1996; Brustein 1996.

[7] We follow King's notation unless otherwise noted. Goodman's accounting identity for ecological inference is $T_i = X_i + (1\text{-}X_i)$, where the dependent variable $T_i$ (turnout) is explained as a function of the variable $X_i$ (black) and $1\text{-}X_i$ (non-black) without a constant. Whether $T_i$ and $T_j$ are spatially independent conditional on $X_i$ adapts the classic time-series autocorrelation interpretation to spatial data. Two differences between time-series autocorrelation and spatial autocorrelation are nevertheless worth noticing. (1) There are no exogenous lags in spatial data to explain the contextual averages found across the data. (2) There are generally no *multiplier effects* in the data that can explain the lagged effect as an adjustment process from an equilibrium point A to a new equilibrium B except when some proper diffusion mechanism is specified, i.e., disease transmission, land productivity, etc. In consequence, it is difficult to explain spatial heterogeneity without considering explicitly other exogenous explanatory variables.

Let us describe the problem with an example. In Argentina, public employees receive relatively equal pay for equal work across the country. However, average wages across all employee groups (both public and private) change dramatically across regions. Average wages in the poor provinces are close to one-third those in the city of Buenos Aires. As a result, public employees receive nearly half the wage of private employees in Buenos Aires but close to twice the average provincial wage in Santiago del Estero. Highly qualified workers are attracted to the public sector in the poor provinces but move to the private sector in the rich ones. Ecological inference methods designed to estimate the pro-Peronist vote of public employees in Argentina should control for this difference in the relative value of public employment across regions to produce valid estimates. Such differences are important for explaining both the spatial variation exhibited by the dependent variable as well as changes in the global relationship between public-sector employees and the Peronist vote (Gibson and Calvo 2000). Similar problems have to be addressed when exploring the social composition of the Nazi vote (O'Loughlin 2000: Germany), differences in racial turnout (Kohfeld and Sprague 2002: St. Louis), racial polarization (King 1997: New Jersey; Miller and Voss 2001: Kentucky) or split-voting (Burden and Kimball 1998: United States; Johnston and Patti 2001: New Zealand; Benoit, Giannetti, and Laver 2000: Italy; Agnew and Shin 2002 Italy).

As stated earlier, within King's framework the preceding example would describe an aggregation bias problem rather than one of *spatial dependence*.[8] Entering a variable Z that controls for the value of public employee wages by provincial average wages would solve the contextual problem and provide consistent estimates of our parameters of interest, $B_i^B$.[9]

A different type of systematic spatial variation may occur when there is an endogenous covariance between the dependent variable $T_i$ and the explanatory variable $X_i$ with distinctive spatial patterns. For example, in multiparty systems a regional increase in the vote for a particular third party can induce other voters to shift their preferences toward that same party for strategic reasons. Party vote $X_i$ may be endogenously related to $T_i$ and nevertheless exhibit distinctive geographic patterns as strategic voting shifts from one successful candidate in one region to different successful candidates and parties in other regions. Take for example the last presidential race in Mexico, where a large number of voters from the leftist PRD in Mexico voted for the liberal PAN to defeat the ruling PRI. Similar strategic voting occurred at the subnational level in Mexico, where PAN voters moved to support the PRD and PRD voters moved to support the PAN according to which one was the likely winner in a particular province. In those cases the likelihood of strategic voting should be read as an endogeneity problem that systematically affects the covariance between $T_i$ and $X_i$. The likelihood of electoral success would induce higher voting averages for different parties across states. In such cases, how do we distinguish the effect of the interaction between $B_i^B$ and $B_i^W$ with regard to the spatial aggregation bias induced by the systematic spatial pattern described by our data?

In both these cases, the methodological problem is not just that the dependent variable T is spatially heterogeneous, adding noise to our estimated errors, but that an unobserved variable Z intervenes systematically across different geographic contexts. A random effects model such as EI cannot give information about the relative impact of $Z_i$—different social compositions of public employees—to produce consistent estimates of $B_i^B$. Rather, the solution is to model the unknown systematic error directly by some form of spatial autoregressive method.

In the next section we compare Goodman's naive ecological regression model and King's EI model under extreme spatial heterogeneity using electoral turnout data from the City of Buenos Aires. We find that the error structure of the estimates in both approaches exhibits extreme spatial heterogeneity, and, as was shown by Anselin and Tam Cho (2002), significant bias is evident in King's EI posterior $B_i^B$.

## The Spatial Structure of Voter Turnout in the City of Buenos Aires

The City of Buenos Aires constitutes the core of a larger metropolitan area (Metropolitan Area of Buenos Aires, MABA) that contains almost a third of the Argentine population. Located at the center of the MABA, the City of Buenos Aires has almost 10 percent of the national population—3 million people—and is considerably wealthier and more politically diverse than the rest

[8] Note that, in contrast to King, the term spatial dependence is used in the spatial econometrics literature as a description of both the less problematic lack of independence between the error terms xi, xj and for the more substantive dependence that results from omitted variable bias.

[9] In King's EI, instead of a global parameter $B^B$, different parameters $B_i^B$ are estimated for every location in the dataset.

of the country. Candidates from 11 different parties and alliances compose the state legislature, far more than those observed in any of the other state legislatures in the country. This variety of parties in the city of Buenos Aires provides an extremely suitable laboratory for comparing ecological inference models, as we have important parties whose core constituencies (Gibson 1996) and voters are distributed according to different *spatial regimes* (Anselin 1988).

Since the beginning of the twentieth century, the City of Buenos Aires has been politically dominated by a middle-class, moderate, catch-all party, the Union Cívica Radical (UCR). The strongest political party in the country for the last 60 years, the Peronist (PJ), has traditionally performed poorly in the city.[10] The center-right parties have traditionally been stronger in the north shore of the city, obtaining some significant results with the conservative U.Ce.De. in the early 1990s and with Action for the Republic in the late 1990s—the party of the former minister of Economy Domingo Cavallo. The single district proportional representation formula used for electing representatives for both the national and local legislatures is responsible to a large extent for the high level of party competition.

Figure 1 shows the geographic distribution of parties in the City of Buenos Aires. Center-right parties, as we already mentioned, have traditionally been stronger in the northern neighborhoods of the city which is depicted in the first map of Figure 1. The Peronist party (PJ) has been stronger in the south and southwest side of the city. On the other hand, the UCR-Alliance, the dominant party in the City of Buenos Aires for the last twenty years, has a spatially homogenous distribution with a slight advantage in the geographical center of the city (see Figure 1). The Peronist (PJ) and Action for the Republic (APR) have more spatially heterogenous votes than that of the UCR-Alliance.

In contrast to the United States, voting in Argentina is mandatory and registration requires using the Citizen's ID card (DNI). After voting, the DNI is stamped to certify that the person participated. Legally, the lack of this stamp entails a small fine and a judge must intervene to normalize the document in order to avoid further administrative problems. Consequently, voter turnout is considerably higher than that of the U.S., as are various forms of "*protest votes*" (i.e., blank votes and null votes). Not voting is legally justified when the citizen is 500+ kilometers away from his designated voting place,[11] is 70 years of age or older (voluntary vote), physically incapacitated, etc.

As Figure 2 shows, voter turnout in the City of Buenos Aires exhibits a clear pattern of extreme spatial heterogeneity. The east and northeast areas of the city exhibit considerably lower turnout levels than the west and southwest areas of the city.

The reasons for such differences are open to speculation but a number of variables that may be relevant include the existence of commercial offices in downtown where many business owners from the provinces have their legal (and voting) address,[12] the lower value of fines and legal hassles to induce voting among the wealthier city dwellers of the northeast shore, and so on.

The social composition of the areas where turnout is the lowest provides an interesting contrast with the conventional wisdom in political science that holds that participation tends to be higher among the wealthy and more highly educated. Still, we have little reason to believe that such interpretation would be anything but spurious and the result of other omitted variables at work.

Because we have no individual-level information on the relationship between party vote and turnout but we do have ballot-box data, we decided to approach the problem of cross-level inference by considering the ballot box[13] relationship between the center-right vote (APR-*Acción por la República*) and turnout as the "individuals" whose behavior needed to be estimated by the precinct-level information. Six thousand ballot boxes provide our baseline observations to be estimated by aggregating this data in 209 precincts. Knowing that individual level "least squares is merely an algebraic convenience here, and is equivalent to doing a cross-tabulation and reading off the answer," we used "least squares applied to this individual-level

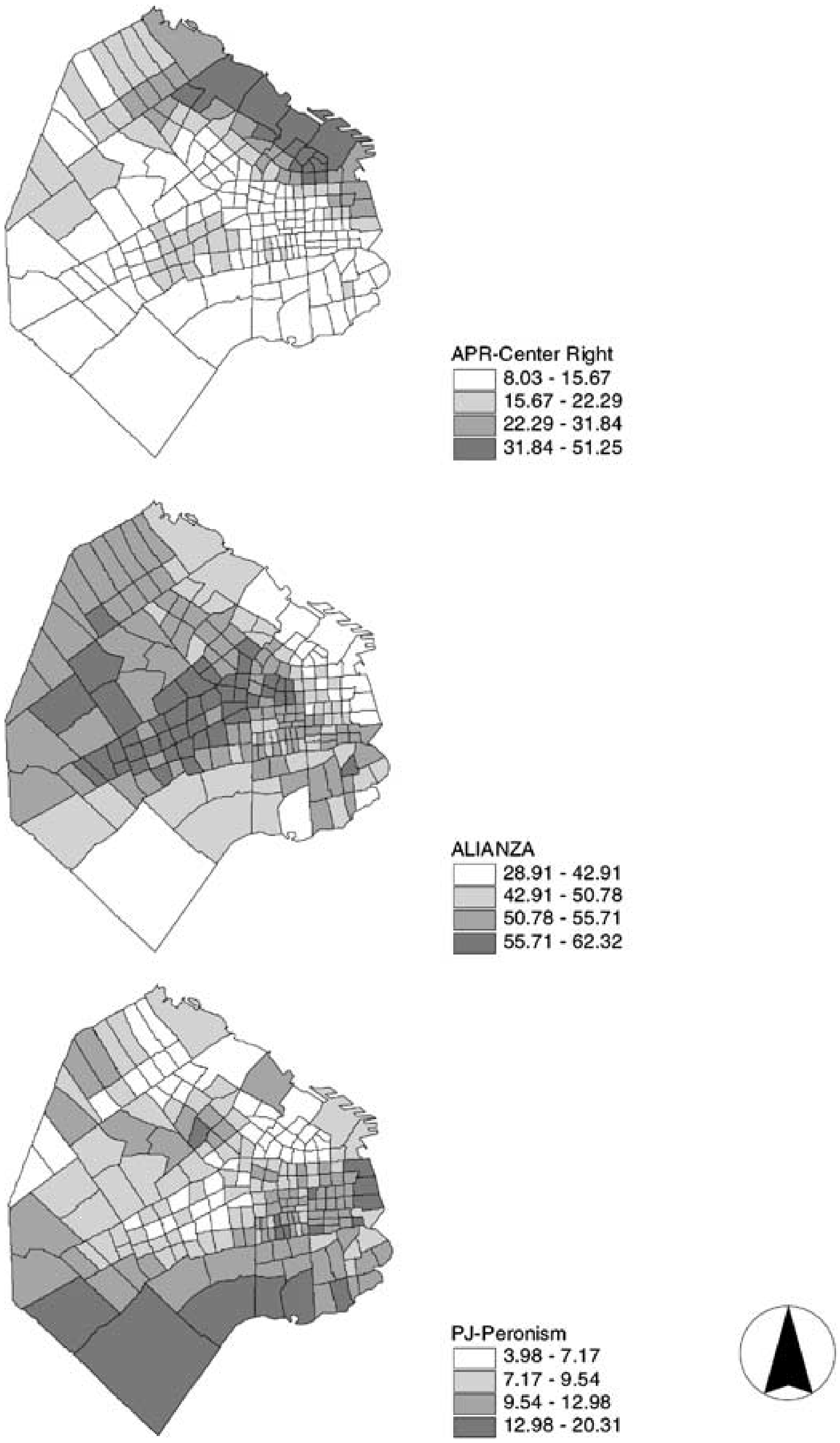
[10] Peronist (PJ) only recently won a legislative race (in 1993), shortly thereafter plunging 20 to 30 points below its national average of 40 percent.

[11] Some folkloric consequences of this rule include the recent creation of a "social movement" called the *Quinientos uno* (the *five hundred and one movement*) which is composed mostly of middle-class young voters who spend the day of the election more than five hundred kilometers away from the City of Buenos Aires to protest.

[12] The 500+ kilometer rule applies in that case.

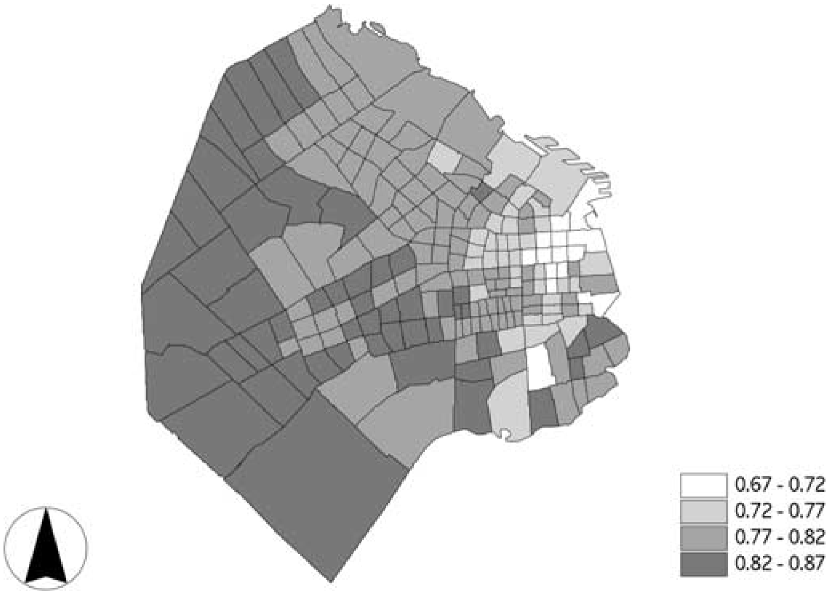
[13] We are actually using the polling station for our baseline model but in Argentina every polling station has only one ballot box. Voters have preassigned a specific polling station, which provides official turnout records and votes. For comparative purposes, every *mesa* (booth box) has on average 386 voters with a standard deviation of 13 voters. The smallest *mesa* has 184 voters and the largest one, 400. By contrast, there are 209 precincts which have on average 12,119 voters, with a standard deviation of 12,230 voters. The smallest precinct has 770 voters and the largest one has 123,658.

FIGURE 1 **The City of Buenos Aires and Three Distributions of the Party Vote**



equation [to obtain] exactly B, the parameter of interest" (King 1997, 47). Our baseline estimation is not, therefore, the individual's vote but the vote of the *mesa* (ballot box). This gives us the possibility of testing the estimation power of the different models against a *true* parameter. To be fair to the different estimation strategies of Goodman's regression and King's EI, we also provide a baseline EI estimation which, as it was

FIGURE 2 Voter Turnout in the 209 Precincts of the City of Buenos Aires.



expected, did not differ substantively from the Goodman baseline.

## Spatial Heterogeneity in the Uncorrected Cross-Level Inference Procedure

The preliminary estimation results of turnout by APR voters are described in Table 1. The baseline models show the global parameter estimates of turnout for APR voters to be around 75 percent, 7 points below the average participation rate of 82 percent for the other political parties. Goodman and King's EI produce comparable baseline results. The estimations obtained from the precinct-level datasets, however, do not compare as well.

Goodman's naïve model slightly underestimates APR voter turnout while King's EI posterior $B^{APR}$ is almost 7 percent off the mark. Moreover, EI's turnout estimates for non-APR voters are, surprisingly, lower than for APR voters. The correlation parameter $\rho$ that measures the

TABLE 1 Estimated Turnout by APR and Non-APR voters; Baseline ("*true*") Estimates, Naïve Goodman Estimates and King's EI Estimates

| | Goodman Baseline Turnout Model$^{c/}$ | EI Baseline Turnout Model$^{a/}$ | Goodman's Naive Model$^{c/}$ | King's EI Model$^{b/}$ |
|---|---|---|---|---|
| $B^{APR}$ | .75 | .766 | .73*** | .834*** |
| Turnout | (.008) | (.022) | (.024) | (.0185) |
| $B^{NONAPR}$ | .829 | .812 | .828*** | .80*** |
| Turnout | (.002) | (.004) | (.006) | (.004) |
| $\rho$ (Cov $B^{APR}$ $B^{N\text{-}APR}$) | — | −.283 | — | .632 |
| N | 6545 | 6545 | 209 | 209 |

*Note*: Standard errors in parenthesis. The full distribution of $\beta_i$ is in the scatterplot of Figure 5. King's EI usually reports the variance of the estimates rather than the SE, to facilitate interpretation we used the SE in the table. Estimated variances for $^{a/}\sigma^B = .148$ and $\sigma^w = .06$. Estimated variances for $^{b/}\sigma^B = .148$ and $\sigma^w = .06$. $^{c/}$Computed with population weights. Dependent Variable *Turnout* relationship between $B^{APR}$ and $B^{NONAPR}$ is also backwards, showing that a higher turnout for non-APR voters drove APR voters to turn out in larger numbers.[14]

Although in this particular case Goodman's naïve estimates are not far from the true baseline, the extent of the bias in OLS will depend on the functional form of the omitted spatial structure.[15] In this case, correct inferences have been drawn from a spatial structure that may as well lead to biased estimates.[16]

As expected, Goodman standard errors in the naïve model are larger than the baseline estimates, reflecting the usual heteroskedastic pattern found in heavily trended ecological data. EI standard errors, on the other hand, are closer to the baseline model because such heteroskedastic patterns are partially explained by the covariance parameter $\rho$ and fit within the local bounds.

Figure 3a maps the Goodman residuals model. Similarly, Figure 3b maps EI's posterior density $B_i^{APR}$, which provides local summaries of the data equivalent to fitted values in OLS. The geographic distribution of EI's $B_i^{APR}$ is suggestive, showing similar spatial patterns as those of the dependent variable turnout in Figure 2. Moreover, the spatial distribution of the $B_i^{APR}$ estimates is strongly correlated with the Goodman residuals in Figure 4, reflecting how the spatial structure present in the data is absorbed by EI's $B_i^{APR}$.

Such a relationship should in principle be considered a good feature of King's EI. After all, EI results are informative of the local-level turnout that Goodman's regression says nothing about. The model, however, is providing biased estimates of the parameters of interest and overfitting the residuals within the sample bounds.

The results should be of concern for those using EI to evaluate policy making in cases where extreme spatial heterogeneity is expected. The estimated turnout levels for our two groups have been reversed and the local $B_i^{APR}$ appears to be describing the spatial structure of the data rather than the presumed covariance between the dependent variable $T_i$ and the explanatory variables.

In this example, there is no theoretical basis for expecting that APR turnout should increase in response to higher turnout by other parties as shown by the model. But if we are analyzing black and white voter turnout, we may well describe the spatial structure of the data reflected in EI's $B_i^W$ as a case of racial polarization where increased turnout by blacks leads whites to vote in larger numbers. The consequences are politically and theoretically problematic, entering the unknown contextual effects into the $B_i^B$ quantities of interest and forcing unwarranted conclusions that may be inaccurate.

## Controlling for Spatial Heterogeneity Using GWR

While context should not matter it clearly affects many ecological inference analyses and we should control for it. Whether we presume an omitted variable bias in space lags, a space-structured inflation factor in the grouped data (Palmquist 1993), or extreme spatial heterogeneity within distinctive geographical regimes (Anselin and Tam Cho 2002), context should be accounted for rather than presumed not to count. In this section we provide a geographically weighted correction procedure to control for spatial effects in ecological inference.

Controlling for spatial effects means modeling the assumption that values in adjacent geographic locations are likely to be linked to each other by some underlying spatial structure. This spatial structure may be itself the result of other omitted local variables or some diffusion mechanism that force T to be spatially dependent on contiguous values.[17] One way to account for such spatial structure would be to use an extra explanatory variable describing the mean value of the dependent variable for neighboring observations. Such a procedure would be equivalent to including a time lag in time-series analysis. In ecological data a spatial matrix-lag of mean $y_i$ values can also be entered into the equation. However, different from time series, the matrix lag is multidimensional, and the lags modeled into the equation cannot be considered exogenous.[18]

[14] As it will become clear later, in cases of extreme spatial heterogeneity $\rho$ provides a linear approximation to the nonlinear spatial structure $B_{shi}$ present in the dependent variable.

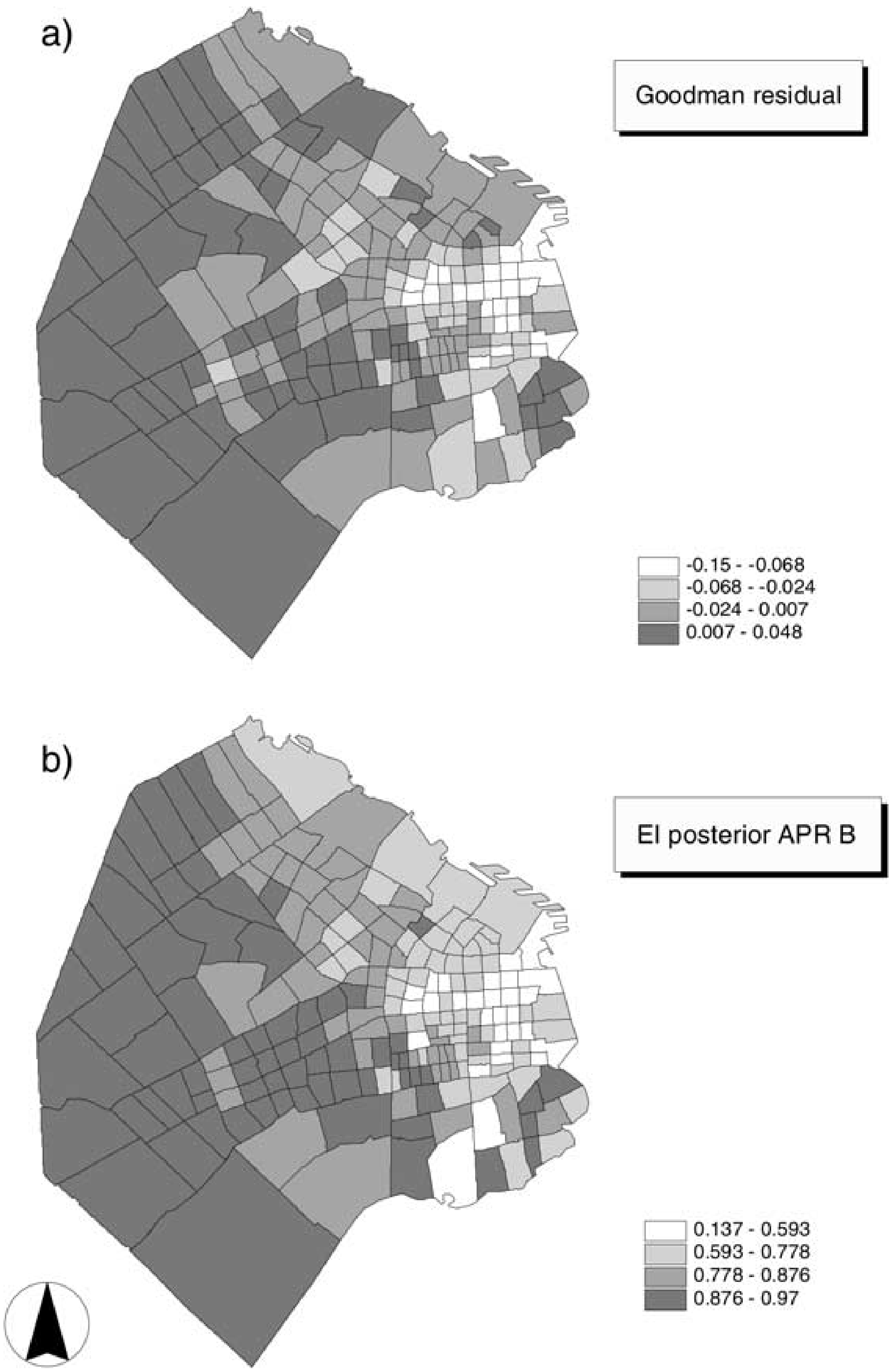
[15] The implemention of SARMA (Spatial autoregressive and moving average) models is significantly more complex in cases of spatial rather than temporal autocorrelation.

[16] In Calvo and Escolar (2002) we use Monte Carlo simulations to evaluate the performance of a semi-parametric GWR method to recover different spatial structures. OLS estimates were often both biased and inefficient.

[17] Note that spatial structure on the dependent variable $T_i$ always implies autocorrelation. However, spatial structure may or may not result in aggregation bias—extreme spatial heterogeneity. Recovering the underlying spatial structure present in a particular dataset should both improve the efficiency of the estimates in cases of autocorrelation and control for ommited spatial effects when spatial dependence leads to aggregation bias—extreme spatial heterogeneity.

[18] We use the notation of Fotheringham, Brunsdon, and Charlton (2000) to describe the spatial autoregressive model.

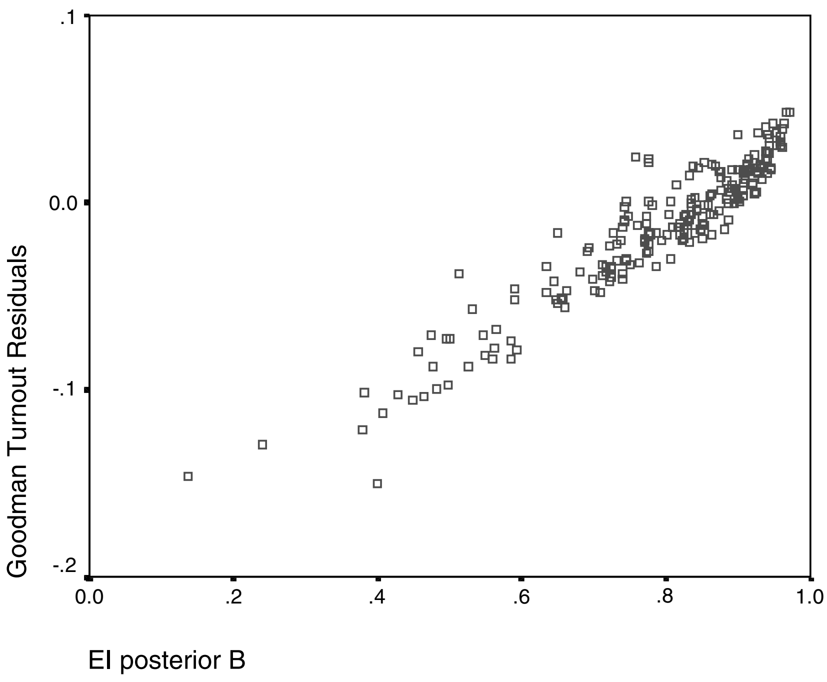
FIGURE 3 Naïve Goodman Residuals and King's Uncorrected Posterior $B^{APR}$



The matrix of the lag-dependent variables can be written as **Wy**, where $w_{ij}$ describes an observation in location j as adjacent to point i if $w_{ij} = 1$ or not adjacent if $w_{ij} = 0$. Notice that, if $w_{ij} = 1$ then $y_i$ and $y_j$ are geographically located next to each other. Therefore, $y_j$ will be entered as a lagged value of $y_i$, and $y_i$ will also be entered as a lagged value of $y_j$. The extended model can be written as:

$$y = \mathbf{XB} + \rho\mathbf{Wy} + \varepsilon \tag{1}$$

**FIGURE 4 Scatterplot of Goodman Residuals and King's Posterior B in the Uncorrected Model**



where ρ is the coefficient for the adjacent mean variable.[19] As it occurs with standard time-series autoregressive models, the number of autoregressive lags can vary—i.e., a first-order spatial lag would include observations that are contiguous to $w_{ij}$, second-order spatial lags would be contiguous to the first-order lag of $w_{ij}$, etc. Different from time-series, however, observations that are distant may still be related to $w_{ij}$. Therefore, it is important to model the entire spatial structure of the data into **Wy**. Such an alternative is possible through kriging, the expansion method (Casseti) or, alternatively, through a geographically weighted regression of the residuals.

A common variation for the model just described is the autoregressive error model, which assumes that the error term is spatially dependent as described in the following equation:[20]

$$y = \mathbf{XB} + (\mathrm{I} - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon} \qquad (2)$$

As it is the case in standard time-series analyses, Equation 2 can be estimated by decomposing the spatially dependent error term **ε** into a grid that describes the spatial trend **ρWu**, and the usual stochastic error term $u_{i,}$ i.e., recovering the spatial structure in the error term. This error term **ε** has, using a more classic notation, a systematic spatial component $v_i$ and the usual stochastic error term $u_i$. In the example of the previous section, the systematic component $v_i$ could be easily identified in the spatial distribution of Figure 4.

Five years ago Brunsdon, Fotheringham, and Charlton created a Geographically Weighted Regression (GWR) method for exploring what they define as spatial nonstationarity: the condition by which "a simple 'global' model cannot explain the relationship between sets of variables" (1996, 1). In what can be defined as a statistical

[19] Taking Equation 1, subtracting ρ**Wy** from both sides, and factoring, we have:

$$(\mathrm{I} - \rho\mathbf{Wy})y = \mathbf{XB} + \varepsilon.$$

After transforming the X matrix (Brundsdon, Fotheringham, and Charlton 2000), we obtain a spatial autoregressive model

$$y = (\mathrm{I} - \rho\mathbf{W})^{-1}\mathbf{XB} + (\mathrm{I} - \rho\mathbf{W})^{-1}\varepsilon, \text{ where}$$
$$\text{the variance} - \text{covariance matrix}$$
$$\mathrm{Cov}(y) = \sigma^2[(\mathrm{I} - \rho\mathbf{W})^{-1}]'(\mathrm{I} - \rho\mathbf{W})^{-1}.$$

We can see that the last two equations are equivalent to those of the standard OLS, but with an error term that is a linear transformation of the original spatially dependent vector ε. We can observe that the main problem is, therefore, finding an acceptable value of ρ to substitute into the ecological inference model to control for the spatial structure present in the ecological data.

[20] Graaff et al. (2001).

renaissance of contextual analysis, GWR provided a tool for modeling social variables in spatially heterogenous settings.

Similar to King's EI (1996), GWR estimates local parameters for every observation i in a dataset but, different from EI, it uses distance weights to reestimate the changing relationship among variables within different spatial regimes. Such weights give declining salience to cases that are further away geographically, measuring the distance from each observation to all others in the dataset. The distance is usually computed from the geographical center of each observation—centroid—entered in the estimation process by their east-north coordinates. Examples of different geographical centroids are the east-north center of a precinct, a *circuito*, a state, and so on.

The problem of how to find an optimum weight to describe the spatial structure of the data is facilitated by the use of different geographical bandwidths but a description of its implementation is beyond the scope of this article (Brunsdon, Fotheringham, and Charlton 1999).

If used to explore the relationship between variables, as intended, GWR might be criticized as a tool that transforms meaningful social parameters into unaccounted geographical contexts. Nevertheless, such a property provides an excellent foundation to estimate $B_{\mathrm{shi}}$ parameters—a vector of parameters describing the nonstationary nature of the error term—that captures the nonlinear spatial structure present in most ecological inference data.

As it is described in Appendix A, the GWR procedure generates a variable $B_{\mathrm{shi}}$ fitting a different regression line in every centroid (observation) of the map.[21] Such different lines are OLS estimates that weight more heavily the observations that are closer to a centroid for every point in a map. For example, we might have a substantively different turnout average for black and white voters in St. Louis (Kohfield and Sprague 2000). If there is extreme spatial heterogeneity, and we compute two different standard regression lines between race and turnout, one for the black precincts and one for the white precincts, the estimated global parameters $B^B$ and $B^W$ should be flat, centered on the population average for blacks and whites, respectively. The GWR parameter $B_{\mathrm{shi}}$ captures spatial nonstationarity, describing spatial context much in the same way that dummy variables would control for the different average turnout between blacks and whites. This method is, therefore, an alternative to the more common spatial weight matrix approaches used for exploring extreme spatial heterogeneity.

As Anselin and Tam Cho (2002) argue, spatial effects can be captured by the use of spatial lags that explain the systematic variation in the error term $\nu_{\mathrm{i}}$ at the local level.[22] However, the use of spatial matrix weighted autoregressive lags can be more problematic than distance weighting schemes like GWR for at least two different reasons. First, lagged values of the dependent variable tend to carry a heavy load of explanatory power with little, if any, substantive interpretation. As a result, we may easily recover predicted values for the full equation but it will be hard to recover intuitive $B_{\mathrm{i}}^B$ values that express the individual behavior we seek to explain. GWR has mean 0 and a GW variance of 1 that allows the original parameter to be read just as in the naïve Goodman model or as the posterior $B_{\mathrm{i}}^B$ in EI.

Second, variations in the spatial structure of a model beyond contiguous values—AR(2+) spatial lags—would be hard to process by most weighted matrix approaches. Matrix weights are good for exploring "neighborhood effects," but they do not capture meaningful local observations that extend beyond nearby observations that are first-order contiguous. Second- and third-order contiguity, when entered into the equation, would further complicate the substantive interpretation of the results and of the model's total explanatory power, rendering our parameters of interest $B_{\mathrm{i}}^B$ statistically meaningless.

By contrast, a GWR approach to modeling the spatial structure $B_{\mathrm{shi}}$ will include all observations in an area that is calibrated by different bandwidths. Moreover, it will always have mean 0 and a GW variance of 1. Technically, a GWR approach to the error $\nu_{\mathrm{i}}$ allows $\nu_{\mathrm{i}} \sim (0,_{\mathrm{GW}}1)$ which, when entered in the classic Goodman equation, maintains the interpretation of $B_{\mathrm{i}}^B$ just as it was in the naïve Goodman model or King's EI.[23]

## The Procedure

The estimation procedure for a GWR Goodman or King model requires four relatively simple steps.

1. First, we compute the naïve Goodman's regression model of APR's vote ($X_{\mathrm{i}}$) and other parties' vote ($1\text{-}X_{\mathrm{i}}$) on the precinct turnout ($T_{\mathrm{i}}$) and save the predicted values and the residuals. Population weights may also be entered in this stage if necessary.

[21] See Appendix A for a technical description of GWR.

[22] In their own words, "imposing spatial structure" (Anselin and Tam Cho 2002, 3).

[23] Notice that $\nu_{\mathrm{i}} \sim \mathbf{N}(0,_{\mathrm{GW}}1)$—normality—is a choice, not a requirement. Non-Gaussian family schemes can also be implemented.

2. Second, we map (ArcView or equivalent) the spatial structure of the residuals and conduct tests of spatial autocorrelation between our *residuals* and the *predicted* turnout, i.e., Moran's I, GWR Monte Carlo testing. A scatterplot of the residuals against the east and north coordinates of the data can also provide a simple visual test for spatial aggregation bias.
3. In the presence of spatial autocorrelation, we compute a GWR of the ***predicted turnout*** on the first stage ***residuals*** and save the local parameter $B_{shi}$—technically equivalent to estimating an instrument for the spatial distribution of the error described in Equation 2. We can save the $B_{shi}$ parameter given that GWR fits a regression line in every observation of our dataset. Because we are regressing the predicted dependent variable of the original Goodman model on the residuals, $B_{shi}$ will have mean 0 and a GW variance 1, describing the spatial structure of the error term in the first stage.
4. Finally, for a GWR Goodman,
   (a) we regress the new model as

$$T_i = B^{APR}X_i + B^{NONAPR}(1 - X_i) + B_3B_{shi} + u_i \qquad (6)$$

where $B^{APR}$ describes the turnout level of APR voters, $B^{NONAPR}$ describes the turnout level of non-APR voters, $B_3$ describes the direct effect of the spatial parameters $B_{shi}$ on the ecological inference estimate and $u_i$ describes the stochastic error. As in the first stage, Equation 2 has no constant. It is important to note that by using the $B_{shi}$ parameter to predict the spatial structure of APR turnout, we can both obtain local estimates and aggregate quantities of interest as in King's EI. A GWR Goodman will provide, therefore, local estimates that will be much closer to King's EI. If the nonlinear spatial parameter $B_{shi}$ explains no variation in the dependent variable T, the results will be similar to the standard Goodman's regression. For iterating the procedure, predict a new dependent variable T from the previous model and repeat steps 1, 3, and 4. As in most semi-parametric smoothing techniques, there are small efficiency gains by iterating the procedure. However, more important than iterating the procedure is properly choosing bandwidths and kernel functions.[24]

   For a GWR EI,
   (b) run the GWR EI model by entering the $B_{shi}$ parameter estimated in (3) as a covariate $Z^B$. No second stage is required.

In Appendix A, we describe the results of step 3, the GWR estimation of the $B_{shi}$ spatial-structure variable. Here we move forward describing the final models from step 4.

Table 2 describes the result of applying this GWR correction to the Goodman naïve model and King's EI. The GWR Goodman results are good, showing a closer and more efficient fit. The new error term from the GWR Goodman model shows no sign of spatial autocorrelation, which can be observed in the checkerboard distribution of the SE in Figure 5a. Also, the fitted line is closer to the true turnout value of 75 percent for APR voters, and the residuals show lower heteroskedasticity. Further testing for spatial autocorrelation shows no new adjustments are required.

The GWR Goodman standard errors are much smaller than those of the uncorrected model described in Table 1, as $B_{shi}$ is now explaining much of the spatial structure that induces heteroskedasticity in the data. In subsequent testing we have found no evidence that the model is overfitting the sample data. In fact, if no spatial effects are observed in the dependent variable T, entering the spatial parameter $B_{shi}$ should have no sensible effect on the original standard errors. However, if spatial effect was detected in the data, the corrected model should provide adequate standard errors.[25]

The GWR EI results are also good, showing that the bias has been properly corrected. The posterior $B_i^B$ shows a good fit with the *true* baseline and displays no sign of spatial dependence (Figure 5b). A new scatterplot of the GWR Goodman residuals and EI posterior $B_i^B$ shows no correlation,[26] and the covariance parameter ρ in the corrected GWR EI model is now negative, just as it was in the baseline model.[27]

[24] It is worth noticing that this procedure describes a semi-parametric autoregressive error model. Therefore, the $B_{shi}$ parameter is not entered as a weighting function of the original equation $T_i = B_i^B + B_i^W(1\text{-}X_i)$ but as an instrument for the underlying spatial structure in in the dependent variable T. For further details see "Semi-Parametric Smoothing Approaches" in Brunsdon, Charlton, and Fotheringham (2001, Chapter 7) and Hastie and Tibshirani (1996).

[25] Monte Carlo testing has shown that the amount of change explained by $B_{shi}$ in the Goodman equation is approximately equal to the amount of variation $B_{shi}$ explains from the original spatial variable. On the other hand, if there is no spatial dependence in the original data, the GWR recovered parameter has no effect on the original equation. That is also a good feature given that incorrectly controlling for spatial dependence has no consequences.

[26] Tomography plots show an adequate distribution centered around the *true* baseline parameter in the corrected model.

[27] Differences in the magnitude of ρ are difficult to assess because we lack some measure of precision for this parameter.

TABLE 2 **Estimated Turnout by APR and Non-APR voters: Baseline ("*true*") Estimates, GWR Goodman and GWR EI Estimates**

| | Baseline Turnout Estimates (Booth) | Baseline Turnout Estimate (Booth Level EI)[b/] | GWR Goodman[a/] | GWR EI[b/] |
|---|---|---|---|---|
| $B^{APR}$ | .75*** | .766 | .755*** | .74*** |
| Turnout | (.008) | (.022) | (.01) | (.019) |
| $B^{NONAPR}$ | .829*** | .812 | .825*** | .824*** |
| Turnout | (.002) | (.004) | (.002) | (.0041) |
| ρ | — | −.283 | — | −.66 |
| $B_{shi}$ | — | — | .873*** | $Z^{Bsh}$ |
| Spatial Heterogeneity | | | (.029) | |
| N | 6545 | 209 | 209 | 209 |

*Note:* Standard errors in parenthesis. [a/]Computed with population weights. [b/]Convergence bandwidth used for the GWR. Dependent Variable *Turnout*

It is worth noticing that the recovered $B_i^B$ quantities of interest in EI are slightly different from GWR Goodman, as they retain the properties of the random effects model and do not incorporate the covariate $Z_i$ explicitly. Therefore, to obtain the *contextual* local quantities of interest, which incorporates the spatial effect, it is necessary to add the local $B_i^B$ and $Z_i$. A comparison of the different quantities of interest recovered by the two models is described in Table 3.

While the results of the GWR Goodman and the GWR EI are relatively similar, there are different advantages that should be taken into consideration when deciding which one to use.[28] In EI, once we control for spatial dependence, the parameter ρ will correctly estimate the covariance between X and (1-X) which might be of interest for researchers that study cases such as racial polarization or strategic response models. Of course, researchers have to accept many of the unverifiable assumptions that come with the use of King's truncated bivariate random-effects model. On the other hand, the GWR Goodman results basically require no other assumptions than those standard for any OLS analysis, but the local estimates for $X_i$ and 1-$X_i$, while different on average, will only change as a function of $Z_i$. That should not be a concern when 1-$X_i$ represents a residual category in the analysis, and the true methodological challenge is finding a more adequate solutions to the estimation of tables bigger than 2 × 2, rather than the covariance between $X_i$ and its complement.

Monte Carlo testing has shown the proposed GWR solution to significantly improve results when spatial heterogeneity affects the dependent variable $T_i$ (Calvo and Escolar 2002). This is an important step as spatial dependence on $T_i$ generally leads to biased estimates while spatial dependence in the explanatory variables will generally affect the efficiency of the estimates. The GWR spatial autoregressive error model also performed relatively well when spatial dependence affected the independent variable $X_i$ and 1-$X_i$ was not entered into the equation. However, the quadratic form of the spatial effect with $X_i$ and 1-$X_i$ in the equation may result in poor estimates of the $B_{shi}$ parameter.

## Conclusion

In this article we show that, as Anselin and Tam Cho (2002) pointed out recently, extreme spatial heterogeneity can add additional problems to the estimation of individual-level behavior using aggregate data. We also detect significant bias in King's EI estimates in data displaying extreme spatial heterogeneity. The results should send a warning sign to those using an incorrectly specified EI approach to ecological inference in cases where extreme spatial heterogeneity or spatial autocorrelation is expected.

This article also shows that a Geographically Weighted Regression approach to Ecological Inference can provide a straightforward solution to the problems normally associated with extreme spatial heterogeneity and autocorrelation. Such a correction factor has mean 0 and a GW variance of 1 that can be easily introduced

[28] Of course we still recommend running both models and comparing their performance.

FIGURE 5 GWR Goodman Residuals and GWR EI Posterior $B^{APR}$

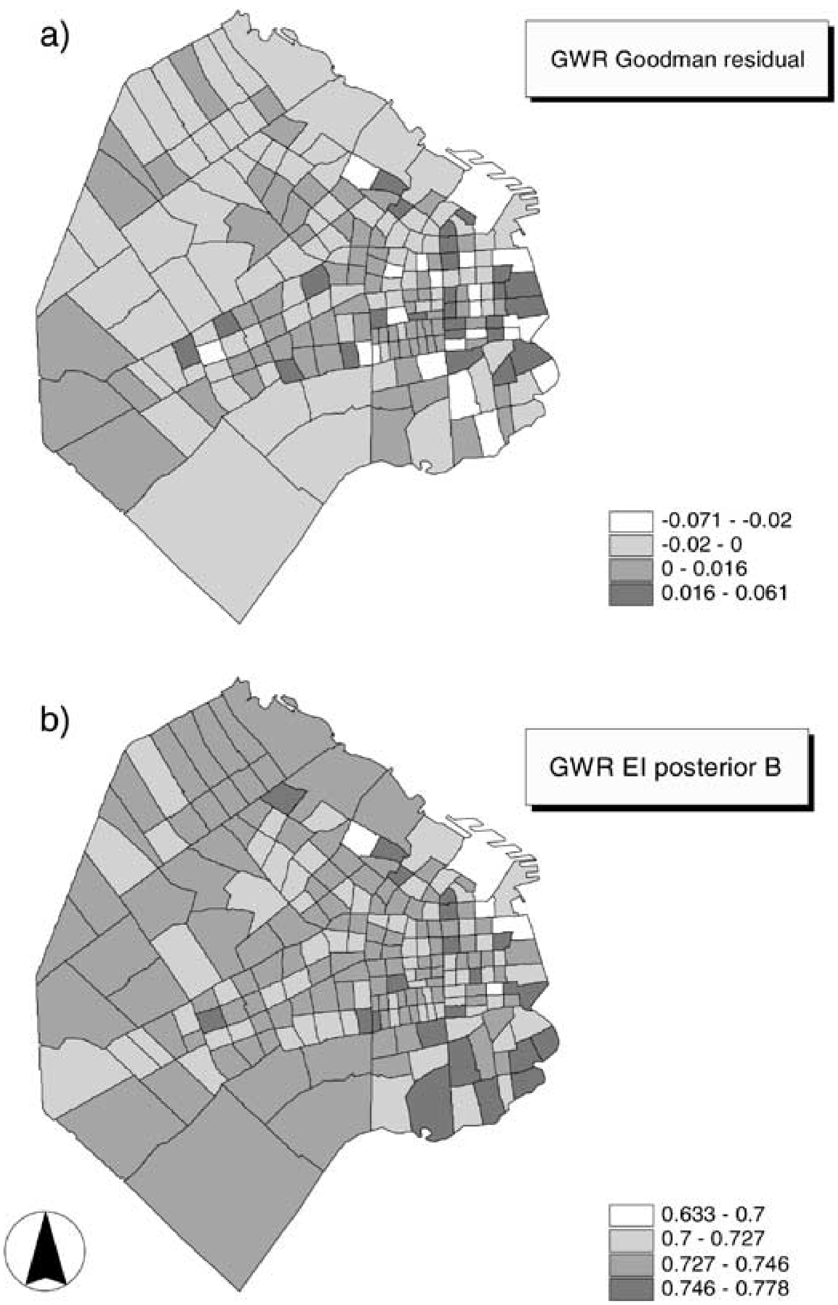


TABLE 3 Local Quantities of Interest of the GWR EI and the GWR Goodman Models

| | Basic Model Quantities of Interest | GWR Local Quantities of Interest |
|---|---|---|
| EI | $T_i = (B_i^B + e_i^B)X_i + (B_i^W + e_i^W)\ (1-X_i)$ | $T_i = (B_i^B + e_i^B)X_i + (B_i^W + e_i^W)\ (1-X_i) + Z_i$ |
| Goodman | $T_i = B^B X_i + B^W (1-X_i)$ | $T_i = B^B X_i + B^W (1-X_i) + Z_i$ |

into Goodman's regression, as it just adds a second ancillary parameter to the original equation. This variable also provides new information to compute local values that express meaningful relationships, allowing researchers to compute aggregate quantities of interest within Goodman's framework that are similar to those produced by King's EI.

A geographically weighted correction factor can also be computed for King's EI obtaining results that were equally adequate. King's EI truncated bivariate normal strategy is a much more versatile method, which effectively captures local variation in ecological inference. However, in data that exhibits extreme spatial heterogeneity, King's EI produces bias results as $\rho$, the parameter used to capture the covariance between different groups of voters in the truncated bivariate normal, and tends to be strongly affected by the systematic spatial patterns found in ecological data. Setting the covariance parameter $\rho$ equal to zero provided results that were similar to those of the original baseline and to the naïve Goodman. However, the researcher needs substantive information that there is no contextual aggregation bias and that $\rho$ is approximately equal to zero. Therefore, there will be no information left to estimate interesting political problems such as racial polarization or strategic voting. GWR restitutes the spatial independence properties of the data allowing EI to obtain proper estimates of $\rho$.

Monte Carlo testing has also shown significantly improved estimates when extreme spatial heterogeneity affects the dependent variable T. More moderate improvements were obtained when spatial autocorrelation affected the independent variable X and 1-X.

Until recently, methodological collaboration between quantitative geographers and political scientists had been extraordinarily limited, in spite of the enormous potential for combining the different analytical insights from both disciplines. We show here that it is possible to take advantage of these different approaches to find new sources of statistical information that improve upon our current ecological inference procedures.

## Appendix: An Estimation of the Spatial Structure Parameter $\mathrm{B}_{\mathrm{shi}}$ by GWR

Geographically weighted regression is a procedure that fits standard OLS lines at each point i according to a weighting function *g*. Such function *g* measures the distance from observation $\mathrm{i}_{(\mathrm{east,north})}$ to every other observation $j_{(\mathrm{east,north})}$ in the dataset. The resulting local parameters $\mathrm{B}_{\mathrm{shi}}$ are estimated from an equation that uses the east-north coordinates of each point *i* to build a distance weight function to estimate Equation A.1.

$$\boldsymbol{\varepsilon}_{\mathrm{i}} = \mathrm{B}_{\mathrm{shi}}(\boldsymbol{g})\check{\mathbf{T}}_{i} + u_{i} \qquad \text{(A.1)}$$

Where $\check{\mathbf{T}}_{\mathrm{i}}$ is the predicted turnout of the first-stage equation and $\boldsymbol{\varepsilon}_{\mathrm{i}}$ is the residual of the first-stage equation. It is worth noticing that in our particular case Equation A.1 has no constant. As presented by Fotheringham, Brunsdon, and Charlton, "the weights are chosen such that those observations near the point in space where the parameter estimates are desired have more influence on the result than observations further away"(2001, 122). The function used for the Gaussian scheme, shown in Equation A.2, uses the h parameter to calibrate different bandwidth values to describe spatial variation (nonstationarity).

$$\mathrm{w}_{\mathrm{i}}(\boldsymbol{g}) = e^{-(\mathrm{d}/\mathrm{h})^{2}}$$

As the value of h decreases, the local effect captured by the data is enhanced. The choice of extremely low h values, however, has to be carefully considered as the predictive power of the GWR may decline significantly and the $\mathrm{B}_{\mathrm{shi}}$ may become flatter. The standard procedures for choosing a bandwidth is cross-validation (Brunsdon, Fotheringham, and Charlton 1999).

We used the GWR procedure in STATA 6.0, created by Mark S. Pearce, to generate a parameter $\mathrm{B}_{\mathrm{shi}}$ describing the spatial structure of the residuals in Goodman's naïve model. For that purpose, we ran a GWR of the predicted turnout on the Goodman residuals of the first stage, with no constant.

The GWR test of nonstationarity gave a statistically significant value of 17.70, with $p < .05$, showing that, as it was observed in the Map 4a, the data displays significant nonstationarity. The local $\mathrm{B}_{\mathrm{shi}}$ are shown in the kernel density graph of Figure A.1 and mapped in Figure A.2. *We strongly recommend to set* ***reps(0)*** *in STATA 6.0 once the first test has been conducted. Computing time will be reduced from hours to less than one minute.*

As it is possible to observe in Figure A.2, the full map of the parameters $\mathrm{B}_{\mathrm{shi}}$ is far superior to first- and second-order contiguity matrices, displaying systematically the full spatial structure of the error term observed in our data. Just as it was observed in the kernel density scatterplot, the full spatial distribution of $\mathrm{B}_{\mathrm{shi}}$ is bimodal, with a median observation below the expected mean zero. These parameters $\mathrm{B}_{\mathrm{shi}}$ were entered in step 4 to control for the spatial heterogeneity problem detected in the first stage of our ecological inference models.

**FIGURE A.1 Kernel Density Scatterplot of the Spatial Structure Parameter $B_{shi}$**

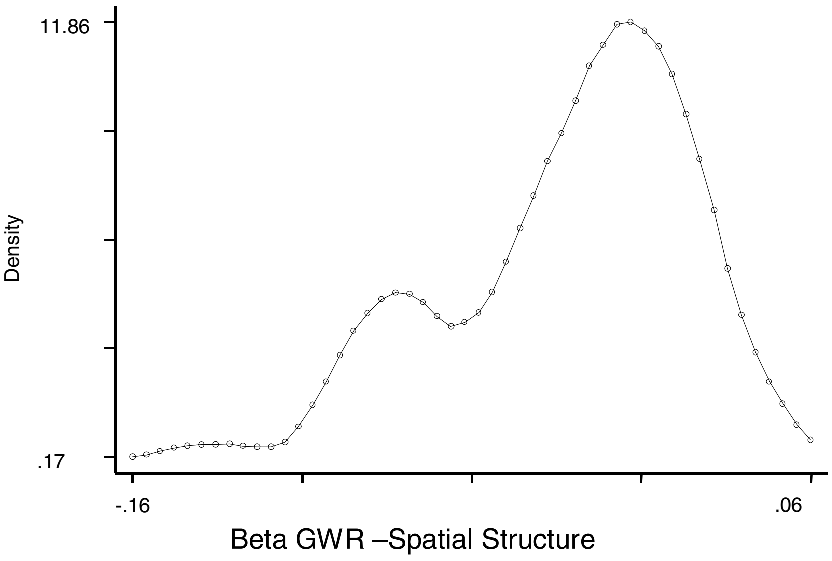
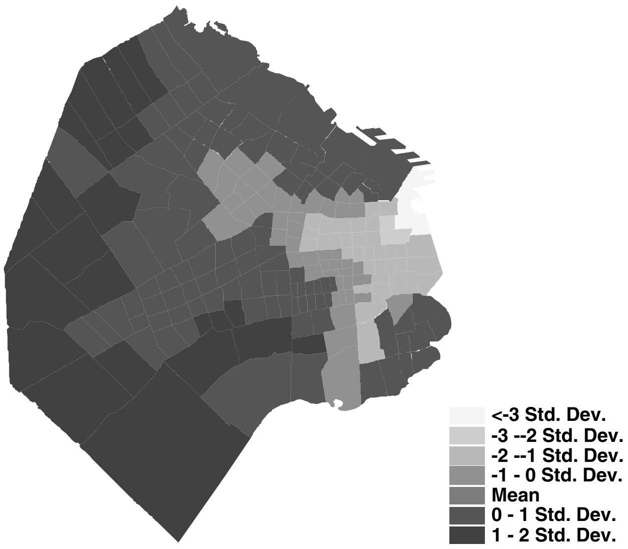


**FIGURE A.2 Distribution of the GWR $B_{shi}$ Parameters**



# References

Achen, C.A., and W.P. Shively. 1995. *Cross-Level Inference*. Chicago: University of Chicago Press.

Agnew, J. 1996a. "Mapping Politics How Context Counts in Electoral Geography." *Political Geography* 15(2):129–46.

Agnew, J. 1996b. "Maps and Models in Political Studies: A Reply to Comments." *Political Geography* 15(2):165–67.

Agnew, J. 1987. *Place and Politics: The Geographical Mediation of Stat and Society*. London: Allen and Unwin.

Agnew, Jon, and Michael Shin. 2002. "The Geography of Party Replacement In Italy, 1987–1996." *Political Geography* 21(2):221–42.

Anselin, Luc. 1988. *Spatial Econometrics, Methods and Models*. Boston: Kluwer Academic.

Anselin, Luc, and Wendy Tam Cho. 2002. "Spatial Effects and Ecological Inference." *Political Analysis* 10(3):276–97.

Benoit, Kenneth, Daniela Giannetti, and Michael Laver. 2000. "Strategic Voting in Mixed-Member Electoral Systems: The Italian Case." Presented at the 2000 Annual Meeting of the American Political Science Association.

Brunsdon, C. 1999. “Some Notes on Parametric Significance Tests for Geographically Weighted Regression.” *Journal of Regional Science* 39(3):497–524.

Brunsdon, C., A. Stewart Fotheringham, and M. Charlton. 1996. “Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity.” *Geographical Analysis* 28(4):281–98.

Brunsdon, Chris, A. Stewart Fotheringham, and M. Charlton. 1998. “Spatial Nonstationarity and Autoregressive Models.” *Environment and Planning* 30:957–73.

Brunsdon, C., A. Stewart Fotheringham, and M. Charlton. 2001. *Quantitative Geography: Perspectives on Spatial Data Analysis*. London: Sage.

Brustein, William. 1996. “Mapping Politics: How Mode of Production Counts in Electoral Geography. *Political Geography* 15(2):153–58.

Burden, Barry C., and David C. Kimball. 1998. “A New Approach to the Study of Ticket Splitting.” *American Political Science Review* 92(3):533–44.

Calvo, Ernesto, and Marcelo Escolar. 2002. “Between Places and Relationships: Uncovering Context through a Geographically Weighted Autoregressive Model.” Working Paper. PEEL. Universidad Di Tella.

Flint, Collin. 1996. “Whither the Individual, Whither the Context.” *Political Geography* 15(2):147–51.

Fotheringham, A. Stewart. 1997. “Trends in Quantitative Methods I: Stressing the Local.” *Progress in Human Geography* 21(1):88–96.

Fotheringham, A.S., C. Brunsdon, and M. Charlton. 2000. *Quantitative Geography: Perspectives on Spatial Data*. London: Sage.

Gibson, Edward. 1996. “Class and Conservative Parties: Argentina in Comparative Perspective.” Baltimore: Johns Hopkins University Press.

Gibson, Edward, and Ernesto F. Calvo. 2000. “Federalism and Low-Maintenance Constituencies: Territorial Dimensions of Economic Reform in Argentina.” *Studies in Comparative International Development* 35(5):32–55.

Graaff, Thomas, Raymond J.G.M. Florax, Peter Nijkamp, Aura Reggiani. 1998. *Diagnostic Tools for Nonlinearity in Spatial Models*. Tinbergen Institute Discussion Papers Number 98-072/3. Amsterdam: Tinbergen Institute.

Hastie, Trevor, and Robert Tibshirani. 1996. Generalized Additive Models. In *Encyclopedia of Statistical Science*, ed. Samuel Kotz, Campbell Read, and Norman Johnson. John Wiley & Sons.

Hastie, Trevor, and Dirk Ormoneit. 2000. Optimal kernel shapes for local linear regression. In *Advances in Neural Information Processing Systems 12*, ed. S.A. Solla, T.K. Leen, and K.-R. Müeller. Boston: MIT Press, 540–46.

King, G. 1996. “Why Context Should Not Count.” *Political Geography* 15(2):159–64.

King, G. 1997. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton: Princeton University Press.

Kohfeld, Carol W., and John Sprague. 2002. “Race, Space and Turnout.” *Political Geography* 21(2):175–93.

Miller, Penny, and Stephen Voss. 2001. “Following a False Trail: The Hunt for White Backlash in Kentucky’s 1996 Desegregation Vote.” *State Politics and Policy Quarterly* 1(March):141–82.

O’Loughlin, John. 2000. “Can King’s Ecological Inference Method Answer a Social Scientific Puzzle: Who Voted for the Nazi Party in Weimar Germany?” Boulder: University of Colorado Press.

Palmquist, Bradley L. 1993. “Ecological Inference, Aggregate Data Analysis of US Elections, and the Socialist Party of America.” Ph.D. Dissertation, University of California, Berkeley.