

# Places and Relationships in Ecological Inference: Uncovering Contextual Effects through a Geographically Weighted Autoregressive Model <sup>1</sup>

Ernesto Calvo  
University of Houston  
[ecalvo@uh.edu](mailto:ecalvo@uh.edu)

Marcelo Escolar  
Universidad de Buenos Aires  
[marceloescolar@fibertel.com.ar](mailto:marceloescolar@fibertel.com.ar)

June/2003

## 1 Introduction

One of the most salient but less studied features of ecological inference is the presence of spatial structure inducing aggregation bias in the observed data. This lack of attention is due to the fact that in most ecological inference models aggregation bias<sup>2</sup> and spatial aggregation bias<sup>3</sup> have been confounded into one and the same thing. However, provided that we know the location of the observable ecological units, there exists considerable more information about spatial aggregation bias than about most other non-spatial sources of bias.

In this article we take advantage of a geographically weighted auto-regressive approach (GW-AR) to ecological inference that incorporates information about the underlying sources of spatial aggregation bias in ecological data. This spatial information can be then incorporated into most ecological inference methods although we will focus on spatial auto-regressive controls for the Goodman regression and King's EI. In doing so, we will also shed light on the different performance of the standard Goodman and EI models in the presence of spatial effects (Anselin and Tam Cho, 2002; King, 2002; Calvo and Escolar, 2003) and their different local estimates (Herron and Shotts, 2001; Adolph and King, 2002; Herron and Shotts, 2002).

There are a number of different procedures that can be used to explore spatial aggregation bias in ecological data. Geographically Weighted Regression (GWR) provides a theoretically sound and computationally simple alternative within the classical framework. We also provide a distance weighted MCMC alternative in Appendix A, in the spirit of that presented by Haneuse and Wakefield in Chapter 13 of this volume.

---

<sup>1</sup>We thank Charles Brunsdon, Noah Kaplan, Gary King, Sebastien Haneuse, Keith Poole, Jon Wakefield, and an anonymous reader for their comments and suggestions. In particular, we want to thank Sebastien Haneuse and Jon Wakefield for their advice in programming the distance weighted WinBugs alternative provided in Appendix A.

<sup>2</sup>The “grouping-induced correlation between  $X_i$  and  $e_i$ -error term” (King, 1997; pg. 55).

<sup>3</sup>The correlation between  $X_i$  and a spatially non-stationary error term, the result of the data being explained by different spatial regimes. See also “extreme spatial heterogeneity” in Anselin, 1988.

The order of presentation of this article is as follows: first, in sections 2 and 3, we introduce a common statistical perspective to discuss local contexts and global relationships. We then describe the Geographically Weighted Auto-Regressive (GW-AR) approach to control for spatial effects in ecological inference. In section 6 we provide Monte Carlo evidence on the performance of the GW-Goodman and GW-EI models. The results converge with previous literature showing that in the presence of spatial effects EI may provide estimates that are both biased and closer to the true  $\beta_i^b$  than Goodman's  $\beta_i^b$  (Voss, Chapter 3 in this volume; Anselin and Tam Cho, 2002; Herron and Shotts, 2001). Finally, in section 7, we exemplify the method with an analysis of the relationship between the Peronist vote and turnout in Argentina.

## 2 Contextual Effects and Global Relationships

Maps can be read in many ways. They provide information about the shortest route to our destination but they also provide information about social structures and processes. Poverty maps are meaningful because wealth is not randomly distributed in cities; southern and northern democrats have geographically distinctive political agendas; and city areas like Chinatown, the magnificent mile, or Cabrini Green in Chicago all express different social structures and relationships that construct these locations as meaningful *places*. Yet, is it only recently that we have started to explore the statistical implications of this diversity in political science rather than just searching for a solution to their *problematic* effects (Ward, 2000; Ward, 2002; Sprague, 2002; Anselin and Tam Cho, 2002).

In a strictly statistical sense location matters, just as places matter in a much more substantive way (Ward, 2002). People with similar incomes choose different neighborhoods to live for reasons that shape their school choices, and select schools for reasons that affect their vote, and decide their vote for reasons that are not unrelated to their housing and neighbor preferences. Similarly, party machines can register voters more successfully in some counties, close races for local candidates can drive voters to turn out in larger numbers in one state but not in others, and even differences in the average age of citizens' across different Florida counties may have a significant impact on the State level of political participation or their vote.

In most ecological inference methods these peculiarities are construed as noise, even though this noise often displays a fairly systematic spatial structure.<sup>4</sup> It usually expresses contextual relationships that shape our variables of interest, affecting the estimation of *global* parameters by the continuous intervention of a geographically located world. And, while these spatial effects are usually the result of local omitted variables (King, 1996; Agnew, 1996), it is generally impossible to account for all the contextual variables that shape social phenomena over space. As it often observed in public opinion, preferences among different groups of voters tend to display trends over time in response

---

<sup>4</sup>See Haneuse (Chapter X) and Voss (Chapter X) for exceptions.

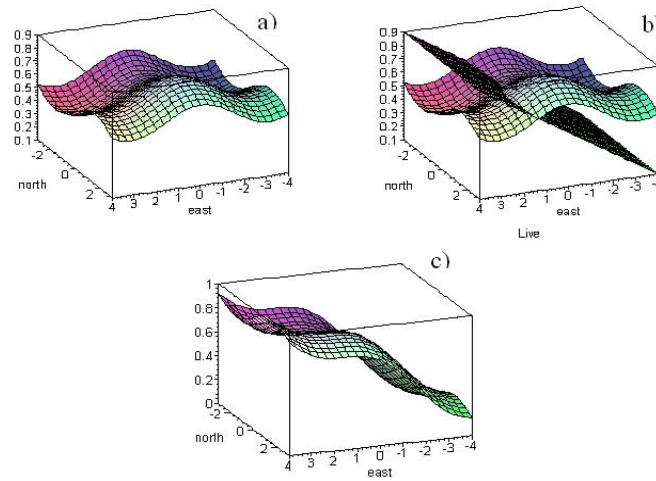
to significant political events. For example, in panel data it is often observed that the support for a candidate shifts up or down for all voters in a sample in response to a political scandal even though different group preferences continue to be affected by other intervening variables like wealth or education. Similarly, a candidate’s scandal in a local community can move the average vote for Party  $i$  down holding other meaningful variables constant. As a result, spatially structured data –i.e. ecological data- often displays large margin errors, heteroscedastacity, and highly uninformative scatter plot distributions (Anselin, 1988).

The geographic extent to which these *neighborhood effects* (Johnston, 1986a; Johnston, 1986b) depress voting, however, is much more than a factor to be corrected. It provides information as to how communities are linked, what populations fall within the influence area of a particular territorial politics and how these contextual *topographies* are linked to other socially relevant phenomena (Fotheringham and O’Kelly, 1989; Anselin, 1988; Sui, 2002). In other words, how much a *place* –contextual variables- explains social phenomena and how global relationships really are.

### 3 Linear Relationships and non-linear Spaces

Imagine a three dimensional Euclidean representation of Party A vote in the unrealistically squared city depicted in Figure 1.

Figure 1: Spatial Dependence in the Vote for Party i



The geography of this city is mapped by its east and north coordinates and  $y_i$  describes the mean vote for Party A in every coordinate  $i$  of its spatial surface. In the absence of *spatial effects*, the average local vote  $y(east, north)$  in any

particular region of the city would be similar to the city’s average (Guillore H.; Levy, J, 1992). However, in the presence of spatial effects, different regions of the city would be characterized by different expected local means.<sup>5</sup> For example, in the area show by the [2,-2] coordinates<sup>6</sup> of Figure 1a the mean expected vote for Party A is around 35%. Meanwhile, in the area [0,1] the mean expected vote for Party A is close to 65%.

These differences could be explained by a number of global variables –i.e. spatial distribution of wealthy voters in different regions of the city—, local variables –i.e. financial scandal of local alderman Smith—, or diffusion effects – degree of integration of the City’s media, transportation, etc—. Mean differences in  $y_i$  could also be explained by the endogenous properties of the covariates if the distribution of the white and black vote has a spatial structure (Haneuse and Wakefield, 2002 in this volume) –i.e. racial polarization leads to higher turnout in particular regions of the data (King, 1997 and Voss 2002 in this volume).

Without introducing further variables to account for spatial effects, the basic model presented in Figure 1a presumes that  $y_i$  is to some degree explained by an underlying spatially heterogeneous structure  $sh_i$ .

$$y_i = f(sh_i) + u_i \text{ (Eq. 1)}$$

Because spatial non-stationarity means that nearby observations are clustered together, an instrument for  $f(sh_i)$  in equation 1 is provided by the more familiar spatial auto-regressive model in which  $Wy_j$ , contiguous observations of  $y_i$ , explain some of the variation in  $y_i$ .

$$y_i = \rho \mathbf{W}y_j + u_i \text{ (Eq. 2)}$$

Where  $\mathbf{W}$  describes a contiguity matrix that takes the value of 1 if  $y_j$  is next to  $y_i$  and 0 otherwise,  $\rho$  is a parameter indicating the magnitude of variation in  $y_i$  as the mean value of contiguous observations change, and  $u_i$  describes the stochastic error term.

Now assume that an exogenous variable  $X_i$  is both linearly related to  $y_j$  and linearly increasing from east to west, as shown in Figure 1b.<sup>7</sup> If we were unaware of the spatial structure in Figure 1b, we could run the basic Goodman regression and obtain OLS estimates of the global parameters of interest by the equation:<sup>8</sup>

$$y_i = \beta^w + (\beta^b - \beta^w)X_i + u_i \text{ (Eq. 3)}$$

This model, however, provides a linear approximation to Figure 1c rather than to the more appropriate data generation process described by Figure 1b. Therefore, the omitted spatial structure of Figure 1a will lead to inefficient and often biased ecological estimates of the parameters of interest.

---

<sup>5</sup>Spatial Non-stationarity, spatial regimes (quantitative geography), Random Fields (statistics, image technology) are growing research areas analyzing the properties of spatial structures.

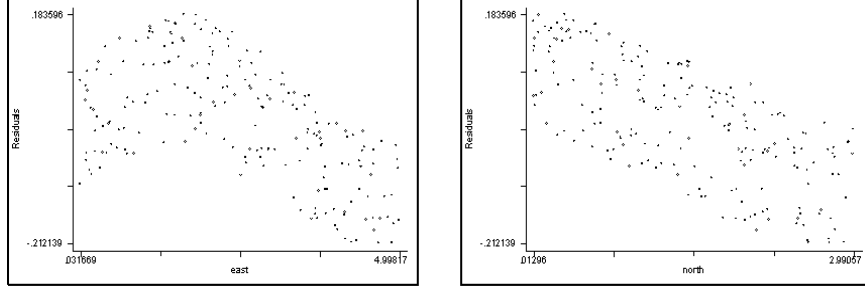
<sup>6</sup>The first number indicates the east coordinate and the second number indicates the north coordinate.

<sup>7</sup>We impose the east-west restriction so that we can represent more intuitively the linear relation between  $X_i$  and  $y_i$  into the City’s surface.

<sup>8</sup>We use  $y_i = \beta^w + (\beta^b - \beta^w)X_i + u_i$  instead of the more familiar  $y_i = \beta^b X_i + \beta^w(1 - X_i) + u_i$  to be consistent with Figure 1. See Grofman and Merrill in this volume (Ch. 5).

There are currently a large number of tests that can be used to detect spatial effects<sup>9</sup> but in many cases a visual exploration of the relationship between the residuals and the east-north coordinates (Figure 2) will clearly show the presence of contextual effects in the data.<sup>10</sup>

Figure 2: Spatial Dependence in the Residuals of the Goodman Model (East, North Coordinates of the Precinct centroids)



Provided that the social process that generated  $Y$  corresponds to that depicted in Figure 1b, the basic Goodman identity should be corrected to allow for the presence of spatial effects in the data.

$$y_i = \beta^w + (\beta^b - \beta^w)X_i + f(sh_i) + u_i \text{ (Eq. 4)}$$

or,

$$y_i = \beta^b X_i + \beta^w(1 - X_i) + f(sh_i) + u_i \text{ (Eq. 5)}$$

To those familiar with General Additive Model (GAM), equations 4 and 5 should ring familiar (Hastie and Tibshirani, 1990). Turnout is here explained as a linear function of black and white's turnout (standard Goodman model) while  $f(sh_i)$  estimates the non-linear spatial structure in the data. The basic problem, however, is that we lack an appropriate instrument to assess the non-linear structure of spatial –contextual- effects.

Luckily, quantitative geographers, regional scientists, and epidemiologists have developed a number of models to deal with issues of spatial heterogeneity and auto-correlation in their data. In the next section we will focus on a distance weighted alternative and introduce a geographically weighted autoregressive control (GW-AR) for ecological inference in the presence of spatial effects. Using Brunsdom, Charlton and Fotheringham (1997) Geographically Weighted Regression, we show that it is possible to recover a spatial vector parameter  $B_{sh_i}$  that provides substantive information about the relative impact of context and its spatial structure.

<sup>9</sup>Some popular alternatives include Moran's I, which estimates the correlation between every observation  $y_i$  and its neighbors  $y_j$ ; and Geti's G, which provides local correlation estimates for every point in the map.

<sup>10</sup>These two plots were obtained from the Monte Carlo simulations that will be presented in the next section.

## 4 Geographically Weighted Regression and its Alternatives

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  $y_i$  to be spatially dependent on contiguous values.<sup>11</sup>

### 4.1 Contiguity

As we already showed in equations 2, 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 multi-dimensional and the lags modeled into the equation cannot be considered exogenous.<sup>12</sup> The matrix of the lag dependent variables can be written as  $\mathbf{W}\mathbf{y}$ , 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:

$$\mathbf{y} = \mathbf{X}\mathbf{B} + \rho\mathbf{W}\mathbf{y} + \varepsilon \text{ (Eq. 6)}$$

where  $\rho$  is the coefficient for the adjacent mean variable.<sup>13</sup>

As it occurs with standard time-series auto-regressive models, the number of auto-regressive 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

---

<sup>11</sup>Note that spatial structure on the dependent variable  $y_i$  always implies auto-correlation. 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 auto-correlation and control for omitted spatial effects when spatial dependence leads to aggregation bias –extreme spatial heterogeneity.

<sup>12</sup>We use the notation of Fotheringham, Brunsdon, and Charlton (2000) to describe the spatial auto-regressive model.

<sup>13</sup>Taking equation 1, subtracting  $\rho \mathbf{W}\mathbf{y}$  from both sides and factoring we have that:

$$(\mathbf{I} - \rho\mathbf{W})\mathbf{y} = \mathbf{X}\mathbf{B} + \varepsilon.$$

After transforming the X matrix (Brunsdon, Fotheringham and Charlton; 2000), we obtain a spatial auto-regressive model

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1} \mathbf{X}\mathbf{B} + (\mathbf{I} - \rho\mathbf{W})^{-1} \varepsilon, \text{ where the variance-covariance matrix}$$

$\text{Cov}(\mathbf{y}) = \sigma^2 [(\mathbf{I} - \rho\mathbf{W})^{-1}]' (\mathbf{I} - \rho\mathbf{W})^{-1}$ . 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  $\varepsilon$ . The main problem is, therefore, finding an acceptable value of  $\rho$  to substitute into the ecological inference model to control for the spatial structure present in the ecological data.

to model the entire spatial structure of the data into  $\mathbf{W}\mathbf{y}$ . Such an alternative is possible through kriging, the expansion method (Casetti) or, alternatively, through a geographically weighted regression of the residuals.

A common variation for the model just described is the auto-regressive error model, which assumes that the error term is spatially dependent as described in the following equation:<sup>14</sup>

$$\mathbf{y} = \mathbf{X}\mathbf{B} + (\mathbf{I} - \rho\mathbf{W})^{-1} \varepsilon \text{ (Eq. 7)}$$

As it is the case in standard time series analyses, equation 2 can be estimated by decomposing the spatially dependent error vector  $\varepsilon$  into a grid that describes the spatial trend  $\rho$ , and the usual stochastic error  $u_i$  -i.e. recovering the spatial structure in the error term.

## 4.2 Distance Weights: GWR

An alternative to contiguity matrices are distance weighting schemes, modeling the assumption that nearby observations  $y_j$  have more influence in the estimation than observations that are further away. Seven years ago Brunson, Fotheringham and Charlton (1996) created Geographically Weighted Regression for exploring what they define as spatial nonstationarity: the condition by which “a simple ‘global’ model cannot explain the relationship between sets of variables” (pg.1).

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 *cirquito*, a state, etc.

Geographically Weighted Regression implements one local regression model for every observation of the dataset according to the equation:

$$y_i = \sum_k \beta_k(e_i, n_i) X_{ik} + \varepsilon_i \text{ (Eq. 8)}$$

Where  $y_i$  describes in the expected local mean,  $\beta_k$  describes the estimated local parameters  $\beta_o$  through  $\beta_k$ ,  $(e_i, n_i)$  describes a distance weighting function of the  $i^{th}$  observation by its east-north coordinates,  $X_{ik}$  describes the explanatory variables which may include a vector of ones,  $X_{io}$ , if there is a constant. The model assumes that data near to point  $i$  have more influence in the estimation of  $\beta_k(e_i, n_i)$  than observations that are further away from  $i$ . In matrix notation, GWR can be written as

$$\mathbf{B}_{(ei,ni)} = \mathbf{X}^T \mathbf{W}_{(ei,ni)} \mathbf{X}^{-1} \mathbf{X}^T \mathbf{W}_{(ei,ni)} \mathbf{y} \text{ (Eq. 9)}$$

Where  $\mathbf{W}$  describes a matrix of weights whose off diagonal elements are zero and its diagonal elements  $w_{i1} \dots w_{in}$  provide a decay functions for points further away from  $i$ . Notice that a number of different weights can be used to estimate

---

<sup>14</sup>De Graff, Florax and Nijkamp (2001).

this local regression. For example, if all  $w_{i1} = 1$  then no decay is represented by this matrix and the local regression model will be identical to the global OLS. On the other hand, if  $w_{i1} = 1$  for 50% of the sample whose observations are closer to  $i$ , all local regressions in  $w_{i1} = 1$  will be identical to the OLS estimated for the full sub-sample.

The problem of how to find an optimum weight to describe the spatial structure of the data requires assumptions about either the proper distance-decay function and/or the proper sub-sample of points. As usual, under- and over-smoothing are some problems that can arise from a poorly calibrated model. The most common choice for distance weights is Gaussian, which gives declining weights to observations as

$$W_{i1} = \exp(-d_{i1}^2/h^2) \text{ (Eq. 10)}$$

Where  $d$  describes the distance from observation  $i$  to observation  $1$  and  $h$  describes a smoothing bandwidth. As  $h$  increases the level of smoothing increases, therefore, the local regression parameter  $\beta_{ik}$  converges to the global parameter  $\beta_k$ .<sup>15</sup> As  $h$  decreases, the local estimates become more spiked and the parameter becomes more distinctively local.

As it can be shown, a geographically weighted Goodman regression can be implemented within this framework by specifying equation 8 as

$$y_i = \beta_{(ei,ni)}^b X_i + \beta_{(ei,ni)}^w (1 - X_i) + \varepsilon_i \text{ (Eq. 11)}$$

In general, however, we lack theoretical reasons to assume that the full model will vary over space. More importantly, there are good reasons to think that changing the spatial scale of support for the model (MAUP) will result in similar spatial aggregation problems over the newly restricted sub-sample when estimating  $\beta_i^b$  and  $\beta_i^w$ . Rather, the alternative we present ahead is to model the spatial structure of the error term directly to build a semi-parametric auto-regressive error model.

### 4.3 A Semi-Parametric Model using the GWR

Now that the GWR has been presented, we can write again equation 5 as

$$y_i = f(e_i, n_i) + \sum_k \beta_k X_{ik} + \varepsilon_i \text{ (Eq. 12)}$$

Where the dependent variable  $y_i$  is explained by a set of linear predictors  $\beta_k$  and a non-parametric spatial structure  $f(e_i, n_i)$  over the east-north coordinates. If we knew the error term,  $\varepsilon_i$ , a spatial smoothing could be applied to estimate the full equation. However, as  $\varepsilon_i$  is unknown, we have to both estimate  $f(e_i, n_i)$  and  $\hat{\beta}$  as a two stage model. First we have to smooth the error term

$$f(e_i, n_i) = \sum_i w_{ei,ni} u_i \text{ (Eq. 13)}$$

to find an instrument for the true spatial structure  $f_{sh(e_i, n_i)}$ . Then, we can use the estimated  $f(e_i, n_i)$  to estimate  $\hat{\beta}$ . As shown by Hastie & Tibshirani (1990), and Fotheringham, Brunsdom, and Charlton (2000); a semi-parametric

---

<sup>15</sup>Therefore, as noted by Beck and Jackman, “least square can be thought as an infinitely smooth scatterplot smoother”. Beck and Jackman, 1998: pg. 606.

model with only one smoother can be analytically derived.<sup>16</sup> Therefore, it is not necessary to iterate between  $f(e_i, n_i)$  and  $\hat{\beta}$  for convergence.<sup>17</sup>

We can then use this semi-parametric GWR procedure to estimate equation 5.

## 5 The Procedure

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

1. First, we compute the naïve Goodman regression model regressing  $X$  and  $1-X$  on  $y$ ; saving the predicted values and the residuals. Population weights may also be entered in this stage if necessary, as shown in the Peronist example of section 4.
2. Second, we map (ArcView or equivalent) the spatial structure of the residuals and conduct tests of spatial auto-correlation 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 auto-correlation we compute a Geographically Weighted Regression of the *predicted*  $\hat{y}_i$  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 in equation 13. 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

$$y_i = \beta^b X_i + \beta^w (1 - X_i) + \beta_3 B_{shi} + u_i \text{ (Eq. 14)}$$

where  $\beta^b$  describes blacks’ turnout,  $\beta^w$  describes whites’ turnout,  $\beta_3$  describes the direct effect of the spatial parameters  $B_{shi}$  on the ecological inference estimate and  $u_i$  describes the stochastic error. It is important to note that by using the  $B_{shi}$  parameter to predict the spatial structure of  $\beta^b$  and  $\beta^w$ , 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 non-linear spatial parameter  $B_{shi}$  explains no variation in the dependent variable  $y_i$ , the results will be similar to the standard

<sup>16</sup>Hastie & Tibshirani (1990), pg.118. Fotheringham, Brunsdom, and Charlton (2000), pg. 180.

<sup>17</sup>However, iteration as it will be described ahead can prevent over smoothing in some applications, particularly in the presence of local outliers.

Goodman regression. For iterating the procedures predict a new dependent variable  $\hat{y}_i$  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 that iterating the procedure is properly choosing bandwidths and kernel functions.<sup>18</sup>

For a GWR EI,

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

## 6 A Simple Monte Carlo Test of the GW-AR Approach to Ecological Inference

How well GW-AR does recover the underlying spatial structure observed in the data? In this next section we provide a preliminary answer to this question. In doing so, we also provide new evidence on the relative performance the Goodman Regression and King’s EI in the presence of spatial effects (Anselin and Tam Cho, 2002; King, 2002).

Because the underlying spatial structure that generated the data is unknown, and the result of omitted local and global variables at work, Monte Carlo experiments are particularly well suited to test the performance of spatially heterogeneous ecological inference models. We can (i) produce a dependent variable  $Y$  that is a function of  $X$ ,  $(1-X)$  and a known spatially heterogeneous structure  $f(sh_i)$ ; (ii) evaluate the performance of ecological inference methods when this spatial structure is not entered into the model; and (iii) evaluate how close is the recovered GW-AR spatial parameters  $B_{shi}$  to the designed spatial structure.

There has been considerable debate as to the proper data generation process that should be implemented to test for spatial heterogeneity and spatial autocorrelation in ecological inference (Anselin and Tam Cho, 2002; King, 2002; Adolph and King, 2002). Therefore, it seems appropriate to briefly describe the Monte Carlo design used in this article.

Following King (2002), we generated the data using an untruncated random effects model design with  $\beta^b \sim N(.4,.02)$ ,  $\beta^w \sim N(.6,.02)$  and  $X \sim N(.6,.04)$ . The *true* spatial structure was created by imposing a *wiggling* functional form to the east-north coordinates of a virtual city, as in Figure 1a. This wiggling spatial structure was  $S_i = ((\sin(\text{east}) + (\cos(\text{north}))/10-.015)$ ; with  $\text{east}, \text{north} \sim U(-5,5)$ . Using a uniform distribution on the east and north coordinates provides an even squared grid while the  $\sin(\text{east})$  and  $\cos(\text{north})$  over the specified -5,5 range generated a wiggling spatial structure similar to Figure 1a.<sup>19</sup> The

<sup>18</sup>It is worth noticing that this procedure describes a semi-parametric auto-regressive error model. Therefore, the  $B_{shi}$  parameter is not entered as a weighting function of the original equation  $y_i = \beta_i^b + \beta_i^w(1 - X)$  but as an instrument for the underlying spatial structure in the dependent variable  $y_i$ . For further details see “Semi-parametric smoothing approaches” in Brunson, Charlton, and Fotheringham (2001) and Hastie and Tibshirani (1990).

<sup>19</sup>Many other distributions are possible by either shifting the scale of the east north coordinates or a different functional form.

range of variation of the spatial structure was reduced from  $[-2,2]$  to  $[-.2,.2]$  and a remainder (.015) was discounted to guarantee that the mean of the spatial structure was 0. Notice that the spatial structure has mean 0 but no variance to reduce the error-in-variables attenuation bias (Adolph and King, 2002). Normal error terms, on the other hand, were introduced into the parameters  $\beta_i^b$  and  $\beta_i^w$ . The basic equation generated was therefore:

$$y_i = \beta_i^b X + \beta_i^w (1 - X_i) + S_i \text{ (Eq. 15)}$$

This model has certain nice features for testing spatial effects including the fact that it will generally stay within bounds but it does not require explicit truncation or the presumption of a truncated bivariate normal distribution. However, avoiding more extreme truncation datasets will also result in a larger number of within bound Goodman estimates. This should be taken into consideration when comparing the relative performances of the Goodman and EI models (Mattos, Chapter 15 in this volume). The distance weighted grid imposed on the data has also many advantages over contiguity matrices. First, complex spatial structures with different functional forms can be intuitively incorporated in this grid to analyze the estimation problems that occur as a result of the observable units being further apart, different in size, unevenly spread, etc. We gave similar population weights to every precinct ( $N=1$  for every  $i$ ) and no aggregation bias different from the noted spatial aggregation bias was imposed onto the data. The descriptive information of the 100 simulations used in this paper is in the Appendix A.

In Figure 3 we show that there is a good fit between true spatial structure  $S_i$  and the  $B_{shi}$  parameter recovered from the residuals in the first Monte Carlo simulation. In fact, in all simulations the  $r^2$  between the true and the recovered spatial structure was above .9. The quality of this fit, however, should in general vary as a function of the level of association between the spatial structure  $S_i$  and  $y_i$  and the noise in the data, as in any GAM models.

In Table 1 we provide comparative information of the global parameters  $\beta^b$  and  $\beta^w$  estimated by EI, Goodman regression, GW-EI and GW-Goodman. The dispersion of the global  $\beta^b$  and  $\beta^w$  parameters around the true (designed) values in the uncorrected models is considerably higher than in corrected models.

More importantly, in only 36% of the naïve Goodman and 67% of EI the  $\beta^b$  estimates were in the .35-.45 interval including the true value of .4. Comparative Kernels of the naïve Goodman and EI models with their GW-AR show the corrected models to provide a more adequate fit to the data. In the corrected models, the percentage of global  $\beta^b$  within the .35-.45 interval was 89% for the GW Goodman and 90% for the GW EI.

Together with Table 1, the kernel graphs in Figure 4 provide some interesting evidence of the impact of spatial effects on the standard Goodman and EI estimates. In Figure 4.1 we see a kernel density graph of EI and GW-EI  $\beta^b$  estimates. As we can see, the uncorrected EI is biased to the right and shows two smaller modes away from the expected mean of 4.

This result is consistent with the attenuation bias in point estimates described by Herron and Shots (2002) which are used by EI to construct the global quantities of interests reported in Table 1. On the other hand, the corrected

Figure 3: Scatterplot of the True Spatial Structure and the GWR Parameter  
(Monte Carlo Simulations 1, 2, 3 and 4)

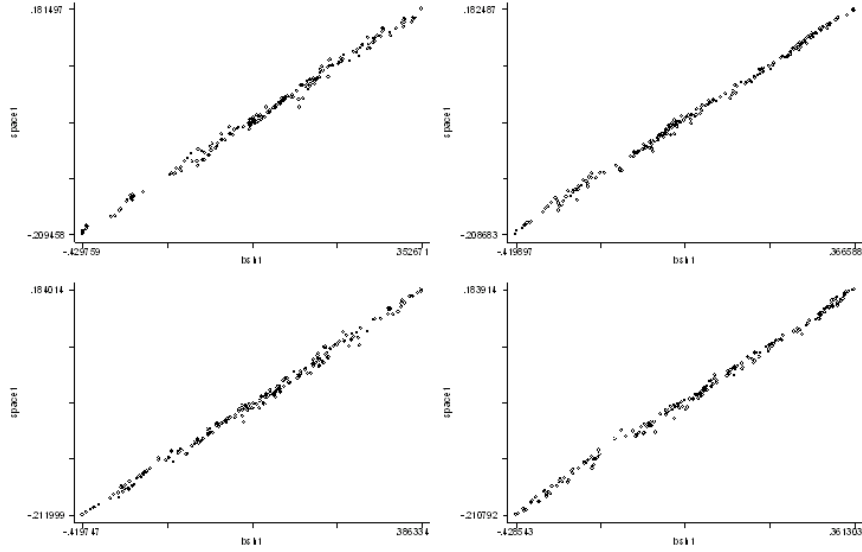


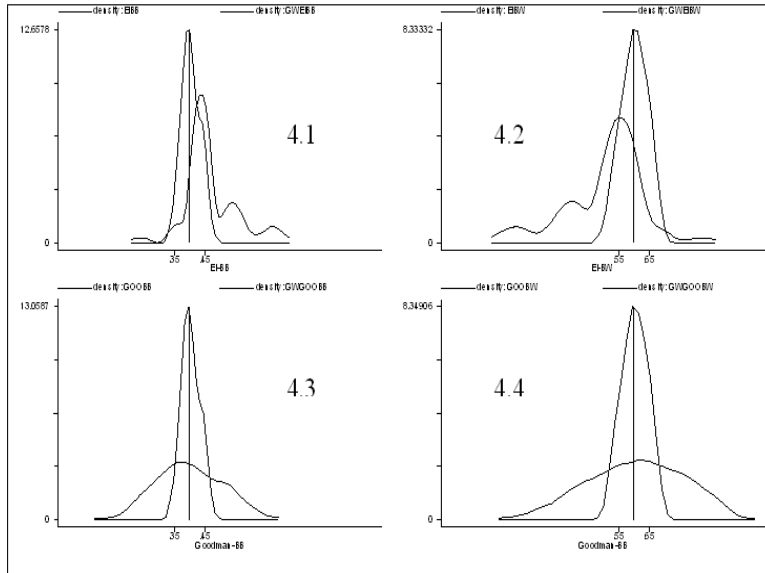
Table 1: Goodman, EI, GW-AR Goodman and GW-AR EI

	$\beta^b$	$\beta^w$
EI	.468	.5061
Min,Max	(.085) .23-.70	(.123) .17-.82
Goodman	.401	.608
Min,Max	(.104) .13-.65	(.155) .21-1.03
GWEI	.405	.6011
Min,Max	(.042) .34-.48	(.03) .50-.69
GW-Goodman	.404	.603
Min,Max	(.029) .35-.48	(.041) .51-.69

GW-EI is centered on the expected value of  $\beta^b = .4$  and displays a narrower variance. Figure 4.2 displays kernel estimates for the EI and GW-EI  $\beta^w$  estimates with results comparable to those of Figure 4.1. We again observe bias to the left which is corrected in the GW-EI model. In Figure 4.3 and 4.4, the uncorrected Goodman models are on average centered on the proper  $\beta^b = .4$  and  $\beta^w = .6$  values. However, the variance is extremely large which would lead us to expect a rather large number of estimated models with unreliable global estimates.

As noted by Voss, chapter X in this volume, extensive debate exists as to the relative performance of EI to recover both local quantities of interest and more adequate global estimates than those of the Goodman model. Extensive applied research displaying sensible results contrast with Monte Carlo evidence showing EI to produce biased estimates in the presence of spatial effects (Anselin and Tam Cho, 2002; Calvo and Escobar, 2003).

Figure 4: Kernel Density Estimates for EI and GW



The Monte Carlo evidence presented in Figure 4.1 to 4.4 may explain some of these conflicting accounts. The figures show that the global EI estimates are closer to the true expected values of  $\beta^b = .4$  and  $\beta^w = .6$  and display smaller variances than the Goodman's estimates. However, the global  $\beta^b$  and  $\beta^w$  estimates are biased.<sup>20</sup> Moreover, the kernel density estimates for the uncorrected EI show

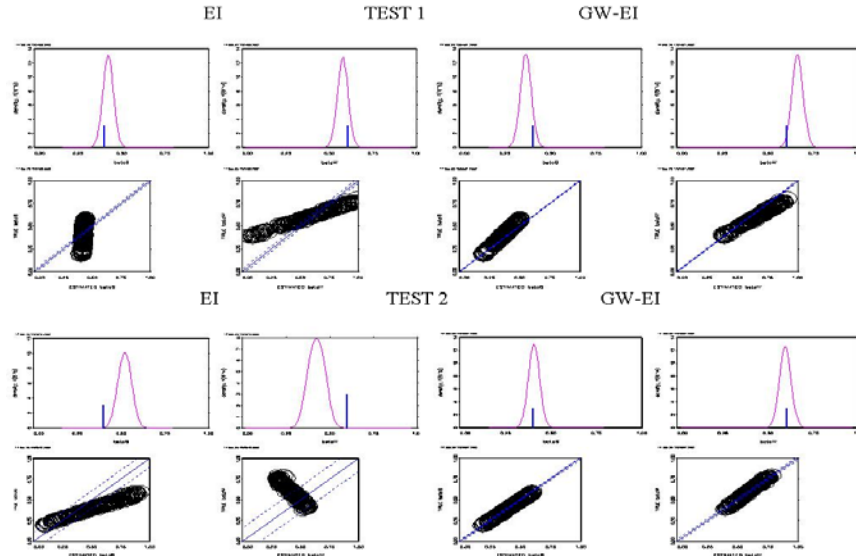
<sup>20</sup>King (2002) suggested that the Monte Carlo evidence provided by Anselin and Tam Cho (2002) may be faulty, the result of a poorly designed spatial experiment. He pointed out that truncation was imposed to the data perhaps leading to some sort of selection bias. We obtained similar results to Anselin and Tam Cho, however, implementing King's suggested Monte Carlo design. There was no truncation and no replacement in our simulated datasets.

that in the presence of spatial effects EI may often flip backwards, finding local maxima away from the true global  $\beta^b$  and  $\beta^w$ .<sup>21</sup>

The standard Goodman regression estimates, on the other hand, show wider variances and display a large number of estimates further away from the true values of .4 and .6. However, the average of these estimates remains unbiased. EI would then appear to produce more sensible results for researchers, although there is little guarantee that those results are in fact centered around the true mean. However, it is worth noticing that the use of a weakly truncated Monte Carlo design overstates the number of *reasonable* estimates produced by the Goodman regression. To conclude, in the GW corrected models, the GW-EI and GW-Goodman global estimates are practically identical.

Analyzing in more detail two of the simulated datasets provides further insights into the estimation of precinct level quantities of interest in the presence of extreme spatial heterogeneity.<sup>22</sup> First it is worth noticing that in the presence of spatial aggregation bias the local quantities of interest can diverge dramatically from the true local values even when the proper global estimates are computed.

Figure 5: Comparing Local Estimates from two Monte Carlo Simulations (EZI 1.5 “True” Graphs



For example, in Figure 5 we provide comparative estimates of local quantities of interest for the first two Monte Carlo simulations produced by our Stata script

<sup>21</sup>Sign reversals were also found by Anselin and Tam Cho (2002), Calvo and Escobar (2003) and Herron and Shots (2002).

<sup>22</sup>In all simulations the  $r^2$  between the true and the recover spatial structure was above .9. The quality of this fit, however, should in general vary as a function of the level of association between the spatial structure and Y as in any GAM models. Both the local estimates of the GWR-Goodman (Figure X.2) and the GWR-EI are also much closer to the true local values.

(Test 1 and Test 2). The plots in Figure 5 show that the global  $\beta^b$  and  $\beta^w$  estimates of Test1 in the uncorrected model (left) are centered near the true design values of .4 and .6 respectively. The local  $\beta_i^b$  and  $\beta_i^w$ , however, provide a poor fit with the true local parameters. In the case of Test 2 we can observe that both the true global  $\beta^b$  and  $\beta^w$  lie outside of EI estimates (kernel plots). The local estimates are also significantly different from their design values.

The fact that confidence intervals of Test 1 are significantly narrower than those of Test 2 does not conform to any substantive information about the performance of EI to find proper local estimates, raising doubts on the exact relationship between the global  $\beta^b$  estimates and the local precinct bounds. The information added by the GW-AR parameter  $B_{shi}$  does provides both EI and Goodman with information to fit different local mean values. In general, the GW-Goodman and GW-EI local parameters are similar. However, given that EI also fits  $\beta_i^b$  and  $\beta_i^w$  within the local bounds, some differences will surface particularly in the presence of local outliers. To obtain similar results, the local Goodman estimates should be adjusted by minimizing the distance from the point estimates to the bounds.<sup>23</sup>

In summary, compared with the standard Goodman regression, the uncorrected EI model was closer to the true but biased estimates of the global  $\beta^b$  and  $\beta^w$  parameters, as noted in previous research (Anselin and Tam Cho, 2002; Calvo and Escolar, 2003). These results were obtained using the random effects model and the Monte Carlo procedure proposed by King (2002) and adding a spatial structure  $sh_i$ . On the other hand, the GW-Goodman and GW-EI models provided adequate, and similar, global and local parameters in our Monte Carlo simulations.

In the next section we use the GW-AR procedure to control for spatial effects in the estimation of the Peronist turnout in the city of Buenos Aires and revisit the problem of obtaining proper local estimates of the quantities of interest.

## 7 An Ecological Inference of the Peronist Turnout in 1999 in the City of Buenos Aires

Analyzing the Peronist vote in Argentina is by itself an important research agenda (Mora y Araujo, 1974; Gibson and Calvo, 2001; Levitsky, 2002). However, for reasons of space and presentation, we will restrict our analysis to the estimation problems as they appear in the data.<sup>24</sup>

In Figure 6 we map the spatial distribution of turnout (Top) and of the Peronist vote (Middle) and precinct size (Bottom) in the City of Buenos Aires. The figure shows significant spatial structure in all three cases. Turnout is

<sup>23</sup>However, we do not have a theory to explain why the minimum distance from the predicted  $\beta_i^b$  (or the posterior  $\beta_i^b$ ) to the local bounds provides an acceptable local estimate for our model (Herron and Shotts, 2002; pg 2). We will revisit this problem in the next section using the example of the Peronist vote in the City of Buenos Aires.

<sup>24</sup>See Calvo and Escolar (2002) for a more extensive description of the electoral geography of the city of Buenos Aires.

considerably lower in the northeast part of the City. The Peronist vote is significantly higher in the south of Buenos Aires where there are larger numbers of registered voters. This would lead us to expect a spatially induced positive correlation between the Peronist vote and the number of registered voters.

Estimating the Peronist turnout in the City of Buenos Aires provides an interesting and challenging case for ecological inference because it combines a number of different problems: (i) there is evidence of aggregation bias across precincts with different population sizes, (ii) there is evidence of spatial aggregation bias across different locations, and (iii) there are significant differences in the spatial structure and distribution of the ecological units. A preliminary review of the first two problems may help to clarify some of the distinctions we have shown in this article.

1. The non-spatial aggregation bias can be readily observed in the increasing proportion of Peronist voters in more populated precincts. These precincts also have higher turnout levels for all voters across the board. A preliminary observation of the naive Goodman residuals against the number of voters shows that differently sized precincts have a distinctive behavior that is also correlated with the Peronist vote (Figure 7, Bottom Left). This aggregation problem results in severely biased Goodman coefficients if the model is run without population weights.
2. The spatial aggregation bias can be observed in the significant relationship between the residuals of the Goodman equation and the east-west dimension (Figure 7, Top Left). The north-south dimension also displays significant heteroscedasticity which should increase the variance around the estimated mean (Figure 7, Top Right).

Since we do not have precinct level true values for Peronist turnout but we do have voting booth values, we decided to approach the estimation process as a Modifiable Areal Unit Problem (MAUP). We used our precinct level data to recover the booth level parameters of Peronist and non-Peronist turnout rather than the individual voters' parameters. In total, 6509 voting booths were used to estimate baseline models of Peronist turnout, which we then compared with our precinct level aggregates.

As shown in Table 2, both baseline models show the Peronist turnout to be above the other parties' turnout. However, we have little reason to believe that this is due to a particularly higher mobilization capacity and expect this difference to be the result of other variables at work (Calvo and Escobar, 2003). Substantive aggregation bias leads to poor estimates for the naïve Goodman model when population weights are not entered in the model. On the other hand, the global estimates of the weighted Goodman model are particularly close to the baseline values although there is a high variance around the mean estimated  $B^{PJ}$ .

The EI estimates are considerably better than those of the unweighted Goodman regression but worst than the weighted Goodman. The  $B^{PJ}$  estimates of EI are 8% below the baseline estimates and below the reported non-peronist

Figure 6: Turnout, Registered Voters, and Peronist Vote in the City of Buenos Aires

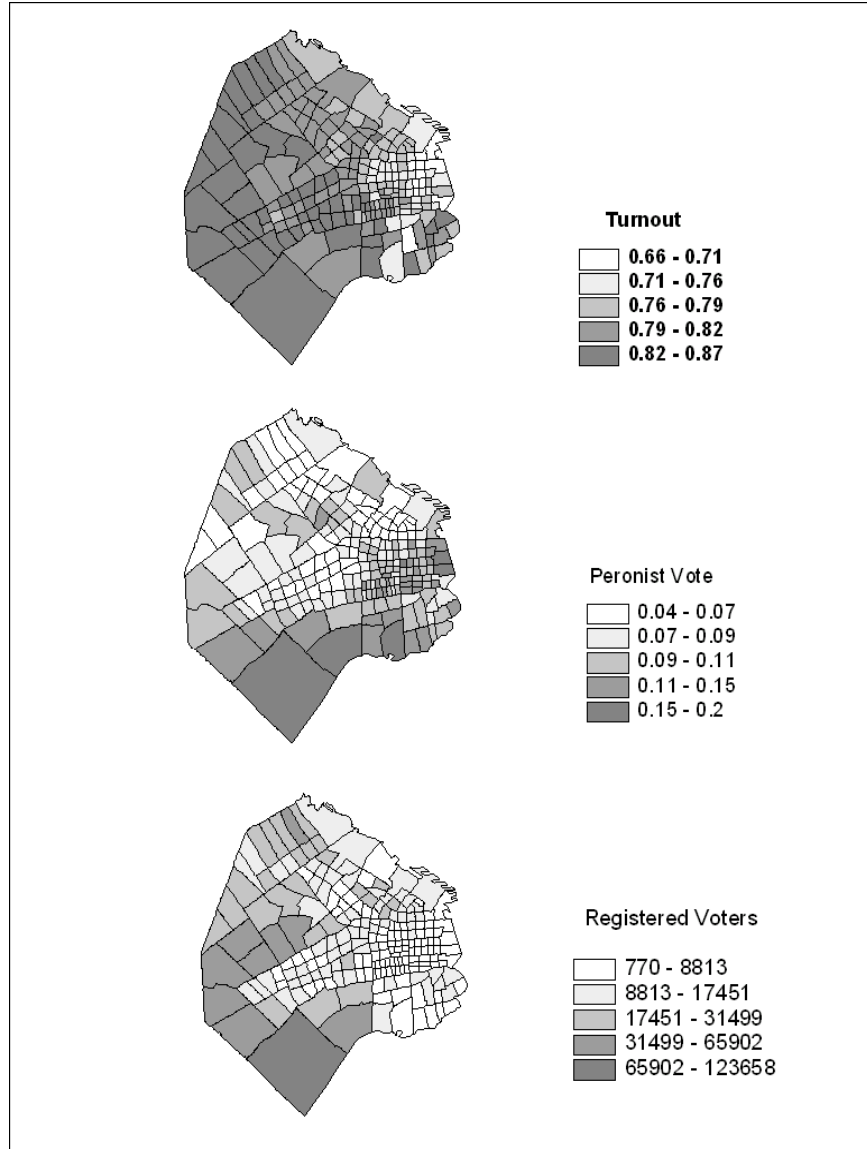


Figure 7: Exploration of the Goodman Residuals against N and the East-West Coordinates

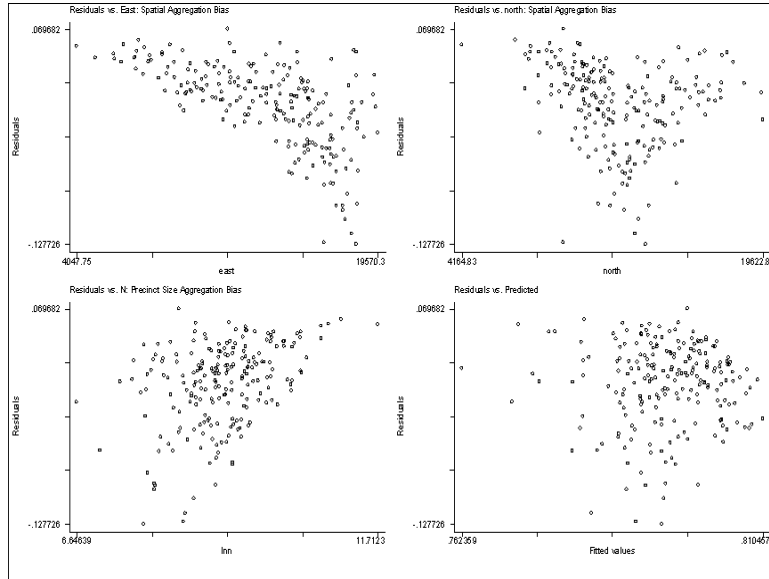
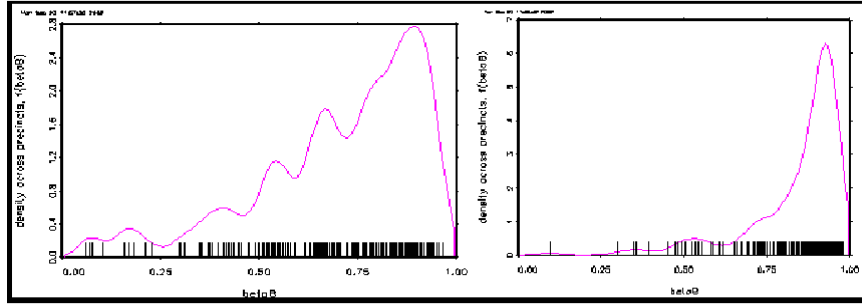


Table 2: Estimates of Peronist and Non-Peronist Turnout in the City of Buenos Aires, All Models

	Baseline Goodman	Baseline EI	Goodman w/o pop weights	Goodman w pop weights	EI	GW-Goodman	GW-EI
$B^{PJ}$	.889 (.011)	.8961 (.04)	.528*** (.09)	.88*** (.066)	.793*** (.015)	.859*** (.02)	.877*** (.012)
$B^{NP}$	.801 (.001)	.80 (.001)	.822*** (.009)	.80*** (.007)	.811*** (.0015)	.805*** (.002)	.802*** (.0012)
Rho	-	.0929	-	-	.720	-	-.1077
$B_{shi}$	-	-	-	-	-	.889***	
N	6509	6509	209	209	209	209	209

turnout  $B^{NOPJ}$ . These results are particularly problematic at the local level, where a rather large number of feasible (within bounds) but unlikely local estimates were computed by EI (Figure 8).

Figure 8: Comparative Kernel Plots of EI(left)and GW-EI(right) Local  $B_i^b$ Estimates;(EZI 1.5 “Results” Graphs



Now we turn our attention to the geographically weighted models. The GW EI estimates are similar to the baseline estimates. Also, the kernel of the local  $B^{PJ}$  estimates displays only one mode compared with three in the uncorrected model (Figure 8). Therefore, as expected, the spatial parameter  $B_{shi}$  provides EI with information to produce narrower precinct estimates. The GW Goodman estimates, however, were 2% below the baseline estimates which is farther from the true values estimated by the weighted Goodman regression. Still, the corrected model displays more adequate standard errors that include the baseline estimates at  $p < .1$  and provides local  $B^{PJ}$  estimates as EI.

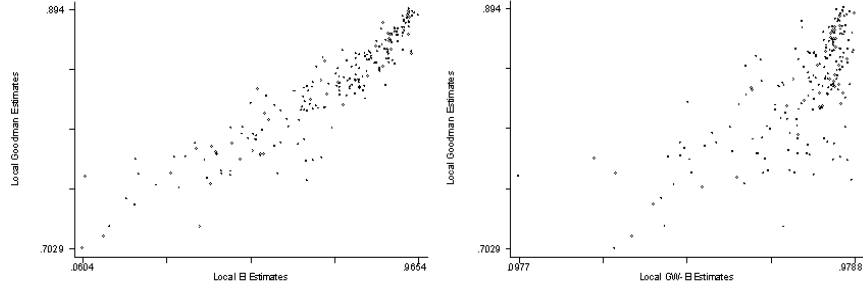
One of the most appealing advantages of EI for many researchers is the possibility of obtaining local estimates that can either provide rich descriptions of the geographic nature of social relationships and new data to conduct further research. There has been, however, little debate about what makes these local estimates good estimates. One of the most interesting problems that emerge from comparing the local  $B_i^{PJ}$  from EI and GW-Goodman (Figure 9) is that while the spatial structure is clearly the same, the scale for the  $B_i^{PJ}$  estimate varies substantively.<sup>25</sup> That is, while the local estimates for EI  $B_i^{PJ}$  go from a minimum turnout of .06 to a maximum of .96, the range of the GW-Goodman goes from .70-894. However, a high correlation between the two sets of estimates show the same contextual effects generating these different local estimates.

When we compare the GW-EI and the GW-Goodman estimates the relationship between the two sets of local estimates fades given that much of the local variation in EI's  $B_i^{PJ}$  is now explained by the covariate  $Z^{PJ}$ . However, the range of variation in  $B_i^{PJ}$  for the corrected GW-EI is still significantly larger than that of the GW-Goodman model. These differences are the result of dif-

<sup>25</sup>It is worth noticing that the local  $B_i^{NOPJ}$  estimates of EI and GW-Goodman were almost identical, as a result of tighter bounds for EI estimates. Less informative bounds, far away from the TBN core, were more problematic.

ferent regions of the data not being explained by the original TBN in EI but still forced within the local bounds.

Figure 9: Comparing Local PJ Turnout Estimates of the Goodman vs. EI (left), and Goodman Vs. GW EI (right) Models



And while EI does provides researchers with tools that described these extreme observations to be characterized by less informative bounds, researchers have usually reported these estimates, and used them for second stage analysis, without acknowledging the different information provided by these local estimates.

The problem is not less dramatic for the GW-Goodman model. Smoothing the spatial surface of the ecological relationship to account for local variations in the mean estimates allow researchers to obtain local values that will not be forced within the observed bounds. However, they generally will fall outside the unit square of the precinct bounds. An alternative modification to the GW-Goodman model would be to minimize the distance between the model's local estimates and the precinct bounds, generating  $B_i^{PJ}$  similar to those of EI (See Merrill, Chapter in this volume). We do not, however, have a theory that provides a rational for such minimization strategy. After all, if the local  $B_i^{PJ}$  in region  $g$  is poorly explained by the overall model, why would the closest point from the model's estimate to the bound be a better predictor of the true local quantity of interest than any other point in the unit square?

## 8 Concluding Remarks

In this paper we described a simple distance-weighted auto-regressive model to control for spatial aggregation bias (extreme spatial heterogeneity) in ecological inference. Using Monte Carlo simulations with a random effects untruncated design (King, 2002) we find that EI produces biased estimates in the presence of spatial effects as previous literature has shown (Anselin and Tam Cho, 2002; Calvo and Escobar, 2003). The Geographically Weighted auto-regressive parameters  $B_{shi}$  was able to restore the spatial independence properties of the ecological data and produce more adequate global estimates, as shown both by the Monte Carlo evidence and the analysis of the Peronist vote. The use of a geographically

weighted control also allowed us to compute local estimates within the classical Goodman framework and compare them to EI's local estimates. The results show EI to produce feasible, but unlikely, local estimates with a wider range of variance than the estimates produced by GW-Goodman. Such results are problematic both when local estimates are used for descriptive purposes and used in second stage inference (Herron and Shotts, 2002). In our view, considerably more research is needed to define statistically acceptable local estimates that fall within the local bounds.

## 9 Bibliography

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

Agnew, J. 1996a. Mapping politics how context counts in electoral geography. *Political Geography* 15 (2):129-146.

———. 1996b. Maps and models in political studies: a reply to comments. *Political Geography* 15 (2):165-167.

———. 1987. *Place and Politics: The geographical mediation of Stat and Society*. London: Allen and Unwin.

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

Anselin, Luc, and Rosina Moreno. 2001. Properties of tests for spatial error components. Urbana, Illinois - Barcelona, Spain: University of Illinois - University of Barcelona.

Anselin, Luc, and Wendy Tam Cho. 2002. Spatial effects and ecological inference. *Political Analysis*, Vol. 10(3).

Benoit, Kenneth; Daniela Giannetti, and Michael Laver. 2000. Strategic voting in mixed-member electoral systems: The Italian case. Prepared for delivery at the 2000 Annual Meeting of the American Political Science Association, Marriott Wardman Park August 31-September 3.

Brundson, C., A. Stewart Fotheringham, and M Charlton. 2000. *Quantitative Geography: Perspectives on Spatial Data Analysis*. Sage Publications. London.

Brundson, C., A. Stewart Fotheringham, and M Charlton. 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis* 28 (4): 281-298.

———. 1999. Some notes on parametric significance tests for geographically weighted regression. *Journal of Regional Science* 39 (3):497-524.

Brunsdon, Chris, A. Stewart Fotheringham, and Martin E. Charlton. 1998. Spatial nonstationarity and autoregressive models. *Environment and Planning* 30:957-973.

Brunstein, William. 1996. Mapping Politics: How mode of production counts in electoral geography. *Political Geography* 15 (2):153-158.

Burden, Barry C. and David C. Kimball. 1998. A new approach to the study of ticket splitting. *The American Political Science Review* 92(3) September:

533-544.

Calvo, Ernesto. 2003. A slow and painful MCMC strategy to model Spatial Dependence . . . and its quick but more limited alternative. Under Review.

Calvo, Ernesto and Marcelo Escolar. 2002. A Geographically Weighted Approach to Ecological Inference. *American Journal of Political Science*. Vol. 47(1);pg. 188-209.

Flint, Collin. 1996. Whither the Individual, Whither the Context. *Political Geography* 15 (2):147-151.

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

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.

Guillourel H.; Levy, J (1992) "Space and electoral system", **Political Geography**, 11 (2): 205-224.

Hastie, Trevor and R.J.Tibshirani. 1990. Generalized Additive Models. Chapman and Hall. London.

Herron, Michael and Shotts, Kenneth. 2001. Using Ecological Inference Point Estimates as Dependent Variables in Second Stage Linear Regression. *Political Analysis*. Vol.11.

Herron, Michael and Shotts, Kenneth. 2001. Cross-Contamination and the Troubled Future of EI-R. Forthcoming in *Political Analysis*.

Johnston, R. J. (1986a) "A Space for Place (or a Place to Space) in a British psychology", *Environment and Planning A*, 19: 599-618.

Johnston, R. J. (1986b) "The neighborhood effect revisited: spatial science or political regionalism", *Environment and Planning D, Society and Space*, 4: 41-55.

Johnston, Ron and Charles Patti. 2000. Ecological Inference and Entropy-Maximizing: An Alternative Estimation Procedure for Split-Ticket Voting. *Political Analysis*. 8(4):333-345.

Jones, K., Jhonston, R.J.; Pattie, C.J. (1992) "Peoples , Places and Regions: Exploring the Use of Multi-level Modelling in the Analisis of Electoral Data", *Brithish Journal of Political Science*, 22 (3): 343-380

King, G. 1996. Why context should not count. *Political Geography* 15 (2):159-164.

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

Kohfeld, Carol W., and John Sprague. 2002. Race, Space and Turnout. *Political Geography*. 21(2): 175-193.

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, Colorado: University of Colorado at Boulder.

Sui, D. Z, Hugill, P.J. (2002) "A GIS-based espatial analysis on neighborhood effects and voter turn-out: a case of study in Colledge Station, Texas." *Political Geography* 21: 159-173.

Ward, Michael D. , and Kristian S. Gleditsh. 2000 - March. Location, location, location: an MCMC approach to modeling spatial context with categorical variables. Paper read at New methodologies for the social sciences: the development and application of spatial analysis for political methodology, at Boulder, Colorado.

## 10 Appendix A: A Distance Weighted WinBugs Model for Ecological Inference in the presence of Spatial Dependence

We provide here a distance weighted alternative to the intuitive model of Haneuse and Wakefield (2003) in this volume. A more extensive description can be found in Calvo (2003). In our example, the distance weighted model also estimates separate spatial structures for whites and blacks. Different from Haneuse and Wakefield, who proposed estimating two separate binomial equations –one for blacks and one for whites—, we estimate a logistic General Linear Model with only one binomial equation which, in our case, facilitated convergence for the two different spatial structures. The model derives directly from King’s random error treatment of the Goodman Identity. Following King (1997:96) we define the local parameters  $\beta_i^b$  and  $\beta_i^w$  as a function of the global mean,  $B^b$  and  $B^w$  and two spatially dependent error terms,  $\varepsilon_i^b$  and  $\varepsilon_i^w$ :

$$\beta_i^b = B^b + \varepsilon_i^b \quad \beta_i^w = B^w + \varepsilon_i^w$$

Substituting these parameters into the Goodman identity equation we have  $= (B^b + \varepsilon_i^b)X_i + (B^w + \varepsilon_i^w)(1 - X_i)$  and replacing the  $\varepsilon_i^b$  and  $\varepsilon_i^w$  by Diggle, Tawn, and Moyeed (1997) spatially correlated matrix of error terms conditional on  $X_i$  and  $1 - X_i$  we obtain  $= B^b X_i + B^w (1 - X_i) + S_i$  where  $S_i = \sigma_b^2 \{1 - \rho(u_i^b)\} X_i + \sigma_w^2 \{1 - \rho(u_i^w)\} (1 - X_i)$  Following Diggle, Tawn, and Moyeed (1998), similar to the GWR approach in our paper, we presume a distribution function for the spatial auto-correlation  $\rho$  as a zero-mean stationary Gaussian process with variance  $\sigma_b^2$  and a correlation function  $\rho(u_i^b) = \exp[-(\alpha d)^k]$  Where  $\alpha > 0$  provides an estimate of the declining correlation with distance  $d$ , and  $0 < k < 2$  describes the level of smoothing over observations. Because the level of smoothing  $k$  and  $\alpha$  are interchangeable, it is common to fix  $k=1$  and to estimate alpha explicitly either by calibrating a semi-variogram or by Bayesian kriging. However, because we are modeling two spatial smoothers simultaneously, the calibration of a semi-variogram for the joint spatial distributions conditional on  $X$  and  $1-X$  may be hard to achieve.

The Winbugs Model

Using the “*spatial.exp*” function in WinBugs, it is possible to estimate the model described. The model is a close relative of universal kriging, with two zero-mean stationary Gaussian spatial smoothers. Hierarchically centering the

spatial structures both follows from the formal model and facilitated convergence. Because we use  $x[i]$  and  $y[i]$  to describe the east and north spatial coordinates, we use  $z[i]$  to describe the percent of black voters usually written as  $X_i$ , and  $t[i]$  for  $T_i$ . Finally, we take advantage of winBugs flexibility to compute and sample from the precinct level *quantities of interest* for blacks (qibbi[i]) and whites (qibwi[i]).

```

model { W[1:N] ~ spatial.exp(mu[], x[], y[], w.tau, w.phi,1)
M[1:N] ~ spatial.exp(mu[], x[], y[], m.tau, m.phi,1)
for (i in 1:N){
t[i] ~ dnorm(g[i], taup[i])
taup[i] <- p[i]/(g[i]*(1-g[i]))
logit(g[i]) <- betab*z[i]+ betaw*(1-z[i])+ space[i]
space[i] <- (M[i]*z[i])+(W[i]*(1-z[i]))
mu[i] <- 0
##quantities of interest
qibbi[i] <- exp(M[i]+betab)/(1+(exp(M[i]+betab)))
qibwi[i] <- exp(W[i]+betaw)/(1+(exp(W[i]+betaw)))
##un-transformed quantities of interest
qib[i] <- M[i]+betab
qiw[i] <- W[i]+betaw
}
betaw ~ dnorm(.001,.001)
betab ~ dnorm(.001,.001)
w.phi ~ dunif(.001,5)
m.phi ~ dunif(.001,5)
w.tau ~ dgamma(.01,.01)
m.tau ~ dgamma(.01,.01)
}

```