

# Can Conventional Measures Identify Geographically Varying Mixed Regression Relationships? A Simulation-based Analysis of Locally Weighted Regression

Aaron Swoboda

March 15, 2015

## Contents

<b>1</b>	<b>Background</b>	<b>3</b>
1.1	Weighted Regression to the Rescue? . . . . .	3
1.2	Mixed Geographic Models . . . . .	4
<b>2</b>	<b>Experimental Design</b>	<b>5</b>
2.1	Data Generation Process . . . . .	5
2.2	Model Selection . . . . .	7
2.2.1	Leave One Out Cross Validation . . . . .	7
2.2.2	Generalized Cross Validation . . . . .	8
2.2.3	Standardized Cross Validation . . . . .	8
2.2.4	Akaike Information Criterion . . . . .	9
2.3	Examples of citations using different methods... . . . .	9
2.3.1	Uses the AIC . . . . .	9
2.3.2	Uses SCV . . . . .	9
2.3.3	Uses LOOCV . . . . .	9
2.4	Mixed Models? . . . . .	9
<b>3</b>	<b>Simulation Results</b>	<b>10</b>
3.1	Starting Simple: All Coefficients are Spatially Stationary . . . . .	10
3.1.1	Model and Bandwidth . . . . .	12
3.2	All Local Coefficients . . . . .	12
3.3	One Local Coefficient . . . . .	15
3.4	Two Local Coefficients . . . . .	15
<b>4</b>	<b>Discussion</b>	<b>15</b>

## List of Figures

1	A Visual Representation of the Spatial Variation for the Three Coefficients as Described in Equations (5) - (7) . . . . .	6
2	Model and Bandwidth Selected by Metric when the True Model = 'GGG' (All coefficients are stationary) Each subfigure shows the percentage of simulations a given model and bandwidth combination was selected among the 50 possible combinations (7 bandwidths for each of the seven mixed models plus the GGG model) for a given metric. The sum of all values in a given subfigure sum to 100 subject to rounding error. For convenience, cell values less than 0.5 percent are omitted. The color saturation of the cells helps denote the magnitude of the values. . . . .	13
3	Model and Bandwidth Selected by Metric when the True Model = 'LLL' (All coefficients are non-stationary) Each subfigure shows the percentage of simulations a given model and bandwidth combination was selected among the 50 possible combinations (7 bandwidths for each of the seven mixed models plus the GGG model) for a given metric. The sum of all values in a given subfigure sum to 100 subject to rounding error. For convenience, cell values less than 0.5 percent are omitted. The color saturation of the cells helps denote the magnitude of the values. . . . .	14
4	Model and Bandwidth Selected by Metric for 'None', 'Some', and 'More' Spatial Variation in the Three Regression Coefficients Each subfigure shows the percentage of simulations a given model and bandwidth combination was selected among the 50 possible combinations (7 bandwidths for each of the seven mixed models plus the GGG model) for a given metric. The sum of all values in a given subfigure sum to 100 subject to rounding error. For convenience, cell values less than 0.5 percent are omitted. The color saturation of the cells helps denote the magnitude of the values. . . . .	16
5	This figure shows . . . . .	17
6	This figure shows . . . . .	18
7	This figure shows . . . . .	19

## List of Tables

1	Distribution of Model Selected by Metric when True Model = GGG (All Coefficients are Non-Stationary). Cell values denote the percentage of simulations in which each model yielded the best metric value. For instance, the GGG model had the smallest LOOCV value for 72 percent of our simulations. Each column sums to 100 subject to rounding error. Cell shading denotes the number of coefficients that are correctly identified as stationary or not. . . . .	10
---	--	----

2	Distribution of Model Selected by RMSE when True Model = GGG (All Coefficients are Non-Stationary). Cell values denote the percentage of simulations in which each model yielded the lowest RMSE value for each coefficient. For instance, the GGG model had the smallest RMSE $\widehat{B}_0$ value for 6 percent of our simulations. Each column sums to 100 subject to rounding error. Shaded cells denote model for which the spatial non-stationarity of the column coefficient is correct. . . . .	11
---	--	----

# 1 Background

Imagine a simple linear model,

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \epsilon. \tag{1}$$

In addition to the three variables listed above ( $Y$ ,  $X_1$ , and  $X_2$ ), assume we know the geographical location for each of our  $N$  observations. Thus, our data consists of an  $N \times 5$  matrix, where  $Y$  may be house prices,  $X_1$  and  $X_2$  could be the living space and lot size associated with each house, and the final two columns determine the location of the observations (for instance, latitude and longitude, or distances north and east from a prescribed point).

The simple model in (1) exemplifies spatial stationarity in the parameters: the  $\beta$  coefficients are constant over space. Alternatively, the coefficients could exhibit spatial non-stationarity, in which case one, two, or all three of the  $\beta$  coefficients vary with location. This has a natural interpretation in the current real estate example: location matters. However, location can matter in different ways. For instance, if the value of land varies over space, then we would expect the coefficient on lot size to vary over space, while it is also possible that the intercept varies over space to reflect variation in prices of similar houses in different locations.

It is possible to parameterize the variation in coefficients, for instance by including a variable measuring the distance from an observation to an important amenity such as the Central Business District and then this distance variable could be interacted with variables whose value are predicted to vary over space. However, it is plausible that the variation in coefficients might not be easily parameterized (for instance, if land values are a non-monotonic function of distance). Researchers may instead interact variables with fixed effects for cities or census tracts. However, such strategies require the analyst to make assumptions that severely limit the type and degree of variation in the parameters. For instance, interaction terms with geographic boundaries assume discrete differences in the value of parameters across the boundaries, while instead the parameters may instead be a continuous function of location. Additionally, numerous interaction terms may unduly reduce the degrees of freedom.

## 1.1 Weighted Regression to the Rescue?

Locally Weighted Regression (LWR) techniques are one set of potential solutions to the challenge described. It is a weighted least squares methodology in which

regression coefficients are estimated as a function of the local data. Cleveland and Devlin (1988) provides one of the first descriptions of the method, while more recently Brunson et al. (1998b) and Fotheringham et al. (2002) have explicitly incorporated geography/location into the methodology. Specifically, the coefficients are estimated,

$$\hat{\beta}(\text{location}_i) = (X'W(\text{location}_i)X)^{-1}X'W(\text{location}_i)Y, \quad (2)$$

where  $X$  is a  $N \times m$  matrix of independent variables,  $W_i$  is the  $N \times N$  weights matrix, and  $Y$  is the  $N \times 1$  vector of dependent variable values. The weights matrix,  $W_i$  is a diagonal matrix where element  $w_{jj}$  denotes the weight that the  $j^{th}$  data point will receive in the regression coefficients estimated at location  $i$  in the dataset.<sup>1</sup> We employ a bi-square weights function and a k-nearest neighbor bandwidth approach as described in equation (3),

$$w_{jj} = \left[ 1 - \left( \frac{d_{ij}}{d_k} \right)^2 \right]^2 \text{ if } d_{ij} < d_k, \text{ otherwise } = 0, \quad (3)$$

where  $d_{ij}$  denotes the distance between observations  $i$  and  $j$ , and  $d_k$  is the distance from observation  $i$  to the  $k^{th}$  nearest observation. This function assigns weights close to 1 for data points near observation  $i$ , weights positive but closer to zero for observations farther away, and zero for all  $n - k$  observations farther away than the  $k^{th}$  nearest observation.

A key decision in estimating LWR models is choosing the number of observations to include in the bandwidth. Bandwidths that are too large in the presence of spatial non-stationarity create bias in the regression estimates (the large bandwidth creates weights matrices that are similar over space and therefore the regression coefficients are forced to be similar when they should vary over space). Bandwidths that are too small add unnecessary error in our estimates by excluding informative observations. Often, researchers choose a bandwidth by minimizing a cross validation metric.

## 1.2 Mixed Geographic Models

Thus far two models have been discussed - the standard OLS model in which all of the coefficients are stationary and a local model in which all coefficients are estimated locally using some bandwidth. The two cases are identical in the extreme case of an infinite bandwidth. Other “mixed” models are also possible in which some coefficients are assumed to be stationary while others are estimated locally. In the example described in Equation (1) with three coefficients, there are eight ( $2^3$ ) total models, ranging from the LWR model in which all three coefficients are assumed to be “Local”, three models in which two coefficients are “Local” and one is assumed to be “Global”, three models in which one coefficient is “Local” and the remaining two coefficients are “Global” and the OLS model in which all three coefficients are “Global”.

In the presence of mixed models, the researcher is tasked with simultaneously choosing which variables to treat as spatially non-stationary and the bandwidth

---

<sup>1</sup>If all diagonal elements of  $W$  are one, then (2) simplifies to the standard OLS estimation.

at which to perform the analysis. Little is known about model performance when models are selected across multiple mixed models and among multiple different potential bandwidth sizes. This paper uses simulated data generated under multiple conditions to begin to answer some of the outstanding questions in the area of geographically mixed models.

## 2 Experimental Design

### 2.1 Data Generation Process

For a given sample size,  $n$ , we generate an  $n \times 4$  matrix of uniformly independently distributed values between 0 and 1.<sup>2</sup> Let the first two of these columns be called *East* and *North* and denote the location of each of our observations within a two-dimensional unit grid (East, North), while the remaining two columns be  $X_1$  and  $X_2$  and serve as the independent variables in our regression equation.

Our goal is to generate data according to the following process:

$$Y_i = \beta_0(East_i, North_i) + \beta_1(East_i, North_i) * X_{1i} + \beta_2(East_i, North_i) * X_{2i} + \epsilon_i. \quad (4)$$

where  $\epsilon_i$  is normally distributed with mean zero and variance  $\sigma^2$ . Equation 4 states that the value of the dependent variable for a given observation depends on  $X_1$ ,  $X_2$ , a normally distributed error term, and three  $\beta$  coefficients that are a function of geographic location.

For our simulations, the relationship between  $\beta_1$  and location is given by,

$$\beta_1(East, North, \eta_1) = 2 + \eta_1 \times (East - .5), \quad (5)$$

where  $\eta_1 \in \{0, 2, 4\}$  represents a “degree of spatial variation” parameter. When  $\eta_1 = 0$ , equation (5) simplifies to  $\beta_1 = 2$  and there is no spatial variation in the coefficient while  $\beta_1$  exhibits spatial non-stationarity when  $\eta_1 \in \{2, 4\}$ . Due to the functional form of equation (5) and the uniform distribution of the *East* variable between zero and one, the expected value of  $\beta_1$  is two for all  $\eta_1$ . Figure 1 shows a visual representation of the possible values for  $\beta_1$  as a function of location and  $\eta_1$ .

Similar to  $\beta_1$ , the relationship between location and  $\beta_2$  is given by,

$$\beta_2(East, North, \eta_2) = 2 + \eta_2 \times (North - .5) \quad (6)$$

and is displayed in the second row of Figure 1.

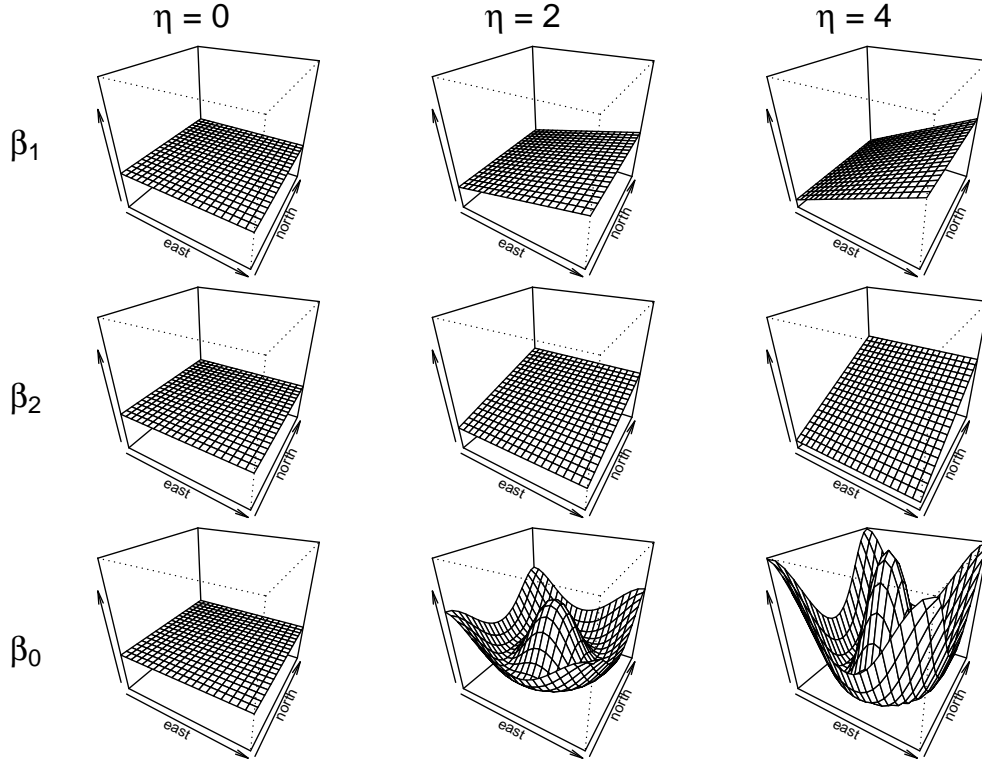
The main difference between equations (5) and (6) is that  $\beta_2$  is a function of *North*, while  $\beta_1$  is a function of *East*, thus the  $\beta_1$  and  $\beta_2$  coefficients are orthogonal to one another. In order to introduce spatial non-stationarity in the  $\beta_0$  coefficient without also creating high and/or perfect collinearity among the coefficients, we generated the  $\beta_0$  coefficient by the following,

$$\beta_0(East, North, \eta_0) = 2 + \eta_0 \times \left[ C + \cos \left( 2\pi \times \sqrt{\frac{(East - .5)^2 + (North - .5)^2}{.5}} \right) \right], \quad (7)$$

---

<sup>2</sup>For our simulation we set  $n \in \{50, 100, 200, 800\}$ .

### Coefficient Spatial Variation



**Figure 1:** A Visual Representation of the Spatial Variation for the Three Coefficients as Described in Equations (5) - (7)

where  $C$  is a constant selected to yield  $E(\beta_0) = 2$ . Equation (7) is presented visually in the third row of Figure 1.

Figure 1 shows the three different formulations possible for the three model coefficients. The leftmost column shows each coefficient constant over space, the middle column shows “some” spatial variation and the rightmost column shows more spatial variation. In other words, the left column shows the stationary, or “Global”, specification for each coefficient, while the middle and right columns show the non-stationary, or “Local”, specifications for the coefficients.

In each instance of our simulation we select a degree of spatial variation (none, some, more) for each of the three coefficients. These 27 different combinations can be categorized into one of eight possible models depending on whether each of the three coefficients is “Global” or “Local”<sup>3</sup>

In addition to changing the degree of spatial variation in the three regression coefficients, we also implement different error term variances,  $\sigma^2$ , to explore the

<sup>3</sup>Specifically, the eight models are GGG, GGL, GLG, LGG, GLL, LGL, LLG, and LLL.

impact of “noisier” data. All told, each simulation can be described as a quintuple choosing from the following values:

$$\begin{aligned}\eta_0 &\in \{0, 2, 4\} \\ \eta_1 &\in \{0, 2, 4\} \\ \eta_2 &\in \{0, 2, 4\} \\ n &\in \{50, 100, 200, 400, 800\} \\ \sigma^2 &\in \{0.25, .5, 1, 2, 3\}\end{aligned}\tag{8}$$

yielding a total of 675 unique combinations. We replicate each combination 100 times for a total of 67,500 simulation iterations.

## 2.2 Model Selection

Gotta say something about Leung et al. (2000). What does it do and what are some potential weaknesses?

They talk about two things... first develop a test of stationarity for the model as a whole, then a step-wise model for adding/subtracting spatially varying variables. Weaknesses? Some quote about “not being able to undo” the addition or subtraction, which may be a problem in the presence of correlated coefficients.

The Fotheringham et al. (2002) text suggests a system of Monte Carlo experiments.

We compare four important cross-validation/information criteria: Leave One Out Cross Validation (LOOCV), Generalized Cross Validation (GCV), Standardized Cross Validation (SCV), and the Akaike Information Criterion (AIC). How frequently can researchers utilizing these metrics identify the correct model among the various possible combinations? Are certain metrics more/less prone to false positive/negatives? Do they suggest no spatial variation when in fact it exists? Do they suggest spatial variation when in fact there is not?

### 2.2.1 Leave One Out Cross Validation

Perhaps the most common cross validation metric used in the literature is the Leave One Out Cross Validation score (LOOCV), which is calculated as follows,

$$LOOCV = \frac{1}{N} \sqrt{\sum_{i=1}^N (y - \hat{y}_{\neq i})^2},\tag{9}$$

where  $\hat{y}_{\neq i}$  represents the dependent variable estimate for observation  $i$  while excluding observation  $i$  from the regression. This prevents the observation from having undue influence in the regression with small bandwidths and overfitting the model. Such a model, while intuitively appealing, can be computationally expensive, as regressions must be estimated first while excluding individual observations to calculate the LOOCV and then again while including the observation to obtain the regression coefficients.

### 2.2.2 Generalized Cross Validation

An alternative cross validation metric is known as the Generalized Cross Validation (GCV) score, which only requires calculating the regressions once per location and explicitly calculates the leverage each observation has over the regression coefficients. The GCV score calculation is detailed in equation (10),

$$n * \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{(n - v_1)^2}, \quad (10)$$

where  $\hat{y}_i$  is the predicted dependent variable value for observation  $i$ , and  $v_1$  can be interpreted as the “effective number of model parameters,” and calculated as  $v_1 = \text{tr}(\mathbf{S})$ , where the matrix  $\mathbf{S}$  is the “hat matrix” which maps  $y$  onto  $\hat{y}$ ,

$$\hat{y} = \mathbf{S}y, \quad (11)$$

and each row of  $\mathbf{S}$ ,  $r_i$  is given by:

$$r_i = X_i(X'W_iX)^{-1}X'W_i. \quad (12)$$

The GCV score is a convenient model selection metric that rewards models that provide a good fit to the data, while penalizing models with a greater number of model parameters (Loader, 1999; McMillen and Redfearn, 2010). (Paez et al., 2011; McMillen and Redfearn, 2010; McMillen, 2012).

### 2.2.3 Standardized Cross Validation

The Standardized Cross Validation Score was suggested by (Farber and Páez, 2007) and elaborated on in (Paez et al., 2011) as an alternative to conventional metrics. This metric is designed to limit the influence of outliers which may disproportionately impact the choice of bandwidth. The Standardized Cross Validation score for a given observation  $i$  and bandwidth  $k$  is,

$$SCV_i(k) = \frac{\sum (y_i - \hat{y}_{-i}(k))^2}{\sum_k (y_i - \hat{y}_{-i})^2}, \quad (13)$$

and the total score for bandwidth  $k$  is then,

$$SCV(k) = \sum_i SCV_i(k). \quad (14)$$

Equation (13) calculates a partial score for each observation as a proportion of the total squared deviance at that observation across the different bandwidths, while (14) then calculates the sum across all observations for a given bandwidth. Note that, contrary to the other metrics described here, the SCV score has to be calculated after all possible bandwidths have been implemented.



### 2.2.4 Akaike Information Criterion

As noted in (Fotheringham et al., 2002), the well-known Akaike Information Criterion is calculated in the geographically weighted regression framework as follows,

$$2 * n * \ln(\hat{\sigma}) + n * \ln(2 * \pi) + n * \frac{n + v_1}{n - 2 - v_1} \quad (15)$$

where  $\hat{\sigma}$  is the estimated standard error of the regression,  $n$  is the sample size, and  $v_1$  remains the “effective number of parameters” estimated by the model as described above.

## 2.3 Examples of citations using different methods...

### 2.3.1 Uses the AIC

Borst and Mccluskey (2007) Axhausen and Löchl (2010) Cahill and Mulligan (2007) Eckey et al. (2007) Foody (2003) Hanink et al. (2010) Haynes (2011) Helbich and Leitner (2009) Kestens et al. (2005) Partridge et al. (2008) Partridge and Rickman (2007) Pineda Jaimes et al. (2010) Sá et al. (2010) Tu and Xia (2008) Yu (2006) Yu (2007) Yu et al. (2007)

### 2.3.2 Uses SCV

No one in our sample, but do a google scholar search and look for articles that cite the original articles

[https://scholar.google.com/scholar?as\\_ylo=2011&hl=en&as\\_sdt=5,24&scioldt=0,24&cites=965449565](https://scholar.google.com/scholar?as_ylo=2011&hl=en&as_sdt=5,24&scioldt=0,24&cites=965449565)

### 2.3.3 Uses LOOCV

Brunsdon et al. (1998a) Cho et al. (2006) Cho et al. (2009) Huang and Leung (2002) Lloyd and Shuttleworth (2005) Lloyd and Shuttleworth (2005) Lloyd and Shuttleworth (2005) Lloyd and Shuttleworth (2005)

## 2.4 Mixed Models?

Cahill and Mulligan (2007) determine that a mixed model might be best because of Monte Carlo simulations.

Huang et al. (2010) find evidence consistent with a mixed model (some coefficients appear to stationary and others non-stationary), but do not estimate one.

Yu (2006) implement a mixed model after using tests of spatial stationarity

Chow et al. (2000) uses the AIC to select which variables are nonstationary.

Mei et al. (2004) seems like an important article to talk about. It proposes a methodology for selecting a mixed model. Google Scholar has info on who cites this article.

[https://scholar.google.com/scholar?cites=11649932945261630952&as\\_sdt=5,24&scioldt=0,24&hl=en](https://scholar.google.com/scholar?cites=11649932945261630952&as_sdt=5,24&scioldt=0,24&hl=en)

### 3 Simulation Results

#### 3.1 Starting Simple: All Coefficients are Spatially Stationary

We begin by examining the simulation results for the spatially stationary data generation process. With no spatial variation for any of the coefficients, these data are consistent with standard OLS regression. We label this model ‘GGG’ to denote that all three coefficients are ‘Global’ rather than ‘Local’.<sup>4</sup> Table 1 displays the percentage of simulation iterations that each of the eight different mixed GWR models was ‘selected’ by each of the four different metrics: LOOCV, GCV, SCV, and AIC. Correspondingly, each column sums to 100 (subject to rounding error).

		Metric				
		LOOCV	GCV	SCV	AIC	
Model Selected	GGG	72	0	8	0	3/3 Correct
	LGG	7	28	29	28	2/3 Correct
	GLG	8	36	22	37	
	GGL	8	33	22	34	
	LLG	1	1	5	0	1/3 Correct
	LGL	2	1	5	1	
	GLL	1	1	8	0	
	LLL	0	0	1	0	0/3 Correct
		100	100	100	100	

**Table 1:** Distribution of Model Selected by Metric when True Model = GGG (All Coefficients are Non-Stationary).

Cell values denote the percentage of simulations in which each model yielded the best metric value. For instance, the GGG model had the smallest LOOCV value for 72 percent of our simulations. Each column sums to 100 subject to rounding error. Cell shading denotes the number of coefficients that are correctly identified as stationary or not.

Table 1 shows a distinct difference between LOOCV and the three other metrics. Almost three-fourths of the time the LOOCV was minimized using the model that was “correct” across all three coefficients. Conversely, the SCV metric selected the correct (‘GGG’) model in less than 10 percent of the simulations and both the GCV and AIC metrics selected the correct model less than 1 percent of the time. Interestingly, while the GCV, SCV, and AIC metrics did not choose the correct model nearly as frequently as the LOOCV metric, they tend to make a correctly identify the spatial (non-)stationarity for two out of three coefficients. The GCV and AIC metric almost exclusively selected one of the ‘LGG’, ‘GLG’, and ‘GGL’ models. Almost 20 percent of the time the SCV metric selected a model that was incorrect about two (‘LLG’, ‘LGL’, ‘GLL’) or all (‘LLL’) of the coefficients stationarity.

At first glance, the high frequency of type one error displayed in Table 1 is frustrating. However, the goal of regression analyses tends to be the efficient, consistent,

<sup>4</sup>The eight GWR models representing the unique mixture/combinations of (non-)stationarity across the three coefficients are labeled: GGG, LGG, GLG, GGL, LLG, LGL, GLL, and LLL.

and unbiased estimation of a particular variable coefficient rather than identifying the exact model specification. We therefore also calculated the Root Mean Square Error for each regression coefficient across all of our model specifications as,

$$\text{RMSE } \hat{\beta}_m = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{\beta}_{mi} - \beta_{mi})^2}, \quad (16)$$

where  $i$  denotes the observation,  $N$  is the sample size, and  $m \in \{0, 1, 2\}$  specifies the model coefficient in question. Unlike the metrics in Table 1, these model performance metrics are not available to researchers with observational data. These measures of estimated coefficient accuracy can only be calculated because we know the true underlying data generating process. Table 2 shows the distribution of models selected by having the smallest RMSE for each coefficient.

	Coefficient RMSE		
	$\widehat{B}_0$	$\widehat{B}_1$	$\widehat{B}_2$
Model Selected	GGG	6	7
	LGG	3	24
	GLG	22	4
	GGL	23	3
	LLG	5	33
	LGL	5	2
	GLL	32	4
	LLL	4	3
	100	100	100

**Table 2:** Distribution of Model Selected by RMSE when True Model = GGG (All Coefficients are Non-Stationary).

Cell values denote the percentage of simulations in which each model yielded the lowest RMSE value for each coefficient. For instance, the GGG model had the smallest RMSE  $\widehat{B}_0$  value for 6 percent of our simulations. Each column sums to 100 subject to rounding error. Shaded cells denote model for which the spatial non-stationarity of the column coefficient is correct.

An interesting pattern emerges upon examination of Table 2. The largest value in each column represents approximately one-third of the simulations, but is not the perfect model. The ‘GGG’ row of Table 2 shows that the correct model yielded the smallest RMSE for a given  $\widehat{B}$  in only 6 to 7 percent of simulations. Instead, the model most frequently minimizing the RMSE for  $\widehat{B}$  correctly identifies the spatial stationarity of the coefficient in question, but incorrectly treats both of the other coefficients as non-stationary. For instance, in 32 percent of these simulations with a true ‘GGG’ model, the smallest RMSE for  $\widehat{B}_0$  was obtained using the ‘GLL’ model. In each column the four respective models that correctly identify the spatial non-stationarity of the respective coefficient yield the most accurate estimates of the coefficient in question for approximately 85 percent of our simulations. It is relatively rare for the most accurate stationary regression coefficient estimates to

be obtained from a model that incorrectly identifies it as spatially non-stationary. Approximately half of our simulations yielded minimum RMSEs using models that were correct about the coefficient in question, but were incorrect about one of the two remaining coefficients.

### 3.1.1 Model and Bandwidth

The previous section explored the model selected by different metrics in the presence of an underlying globally stationary data generation process. The results showed that the correct model, ‘GGG’, was only selected relatively frequently by the LOOCV metric. However, we have also seen that some of the most accurate estimates of the individual regression coefficients come from incorrect models. We have not yet quantified how wrong these incorrect model are. For instance, a model may incorrectly identify a coefficient as spatially non-stationary, but might estimate a very small degree of variation in the coefficient by using a large bandwidth relative to our sample size. That is, it is potentially very different to select the ‘GGL’ model instead of the ‘GGG’ with a large vs. small bandwidth. A small bandwidth can yield more variation in coefficient estimates across our sample, while a large bandwidth will restrict the non-stationary coefficients to be more similar.

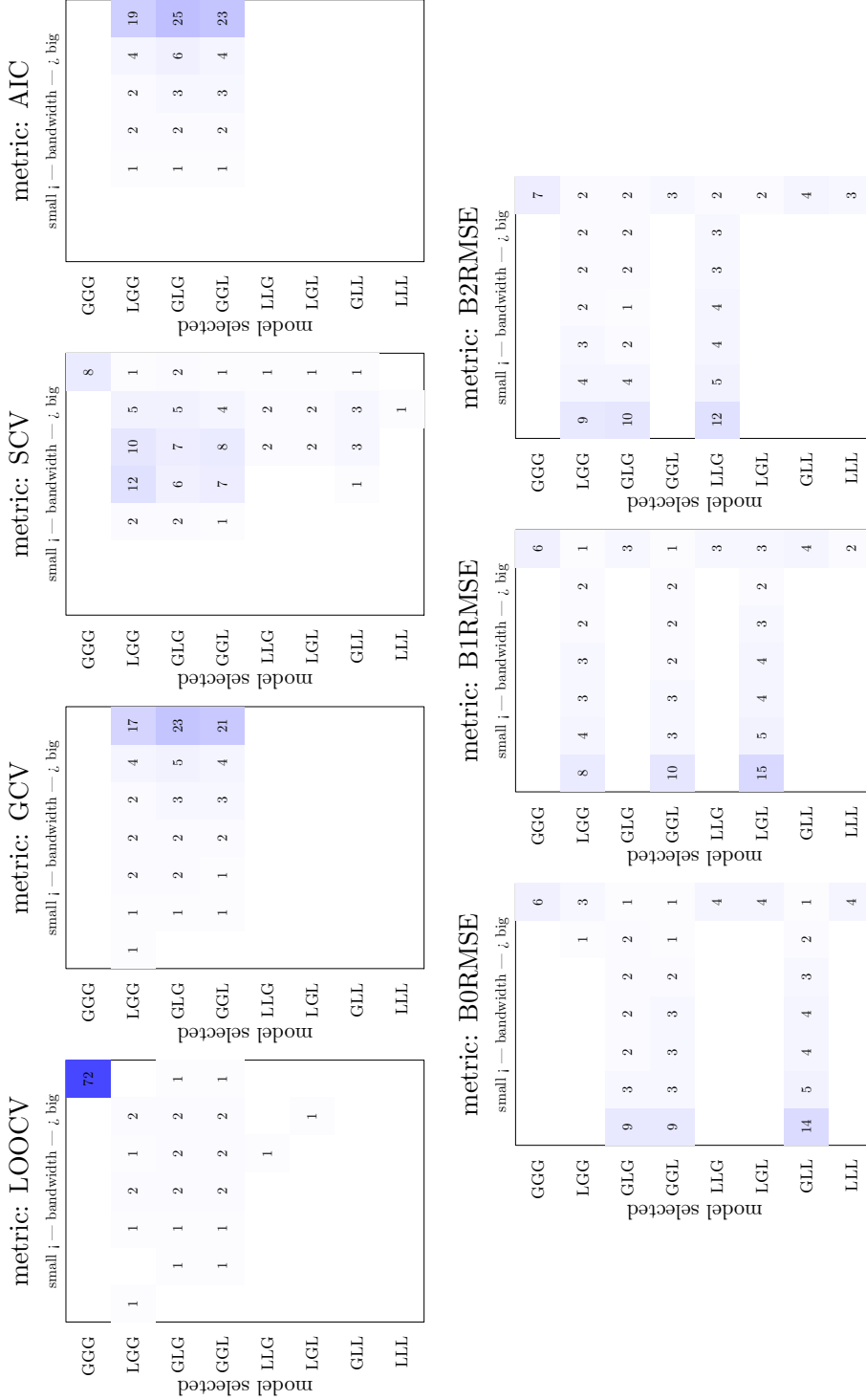
In each of our simulations we estimated the seven models allowing non-stationarity in at least coefficient for seven different bandwidths, ranging from using all of the data to just under 10 percent of the observations in the mixed models with the smallest bandwidth. Figure 2 shows the distribution of model selected and bandwidth size for each of the seven metrics we’ve discussed.

## 3.2 All Local Coefficients

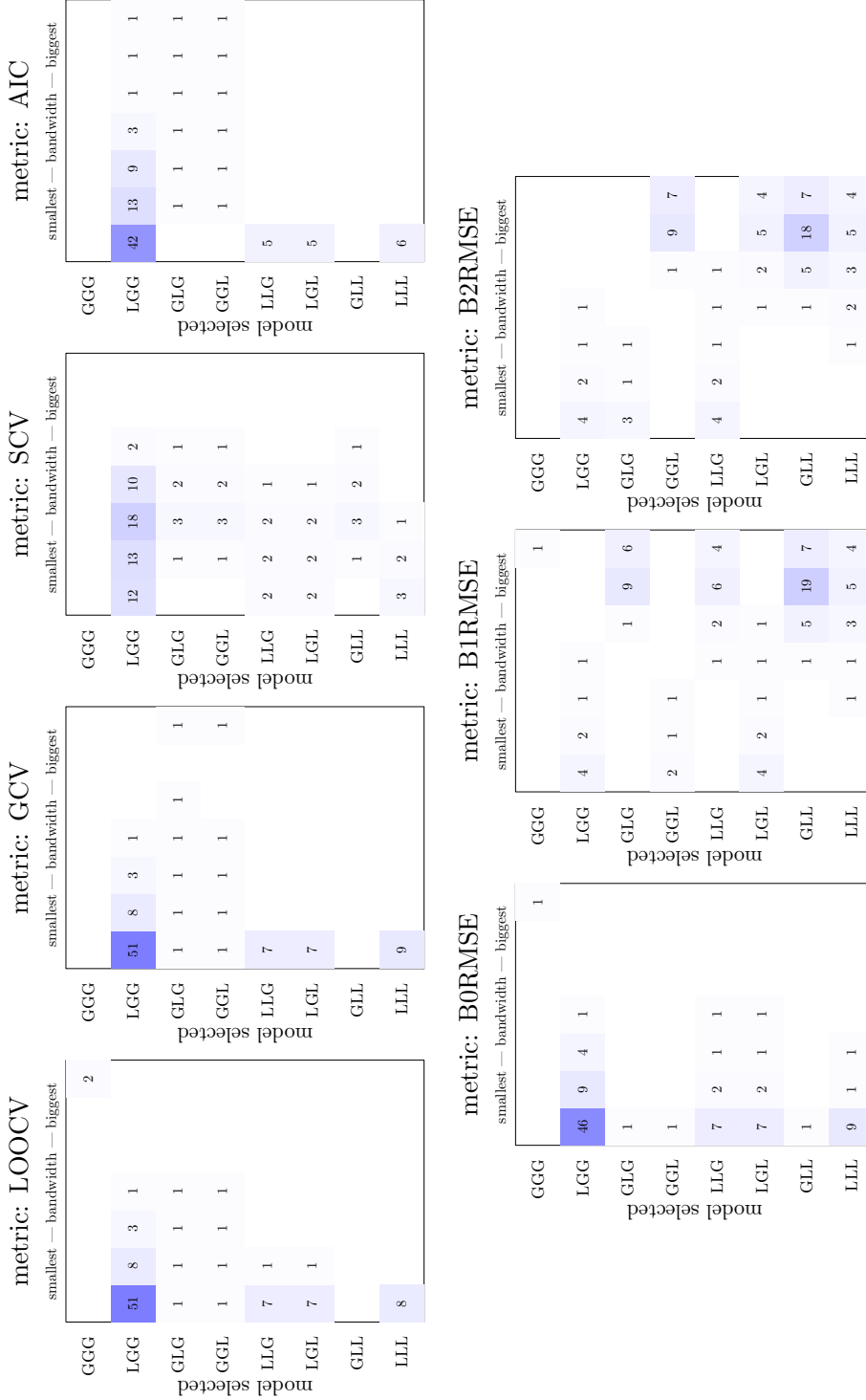
The previous section investigated the simulation results when the true model was stationary across all three coefficients. In this section we examine the opposite extreme: all non-stationary coefficients. To begin, we construct a figure similar to Figure 2 but with a true model of ‘LLL.’

Figure 3 reveals a stark pattern. Although the true underlying model is non-stationary, none of the four metrics choose the ‘LLL’ model in more than 10 percent of our simulations. Instead, the ‘LGG’ model is selected over 50 percent of the time by each metric. The LOOCV, GCV, and AIC metrics all select the ‘LGG’ model with the smallest bandwidth much more frequently than any other model/bandwidth combination. For instance, LOOCV and GCV select the ‘LGG’ model with the smallest bandwidth over 50 percent of the time.

The second row of results in Figure 3 displays the model and bandwidth combinations that yield the smallest RMSE for the coefficients in our model. The pattern for the smallest RMSE associated with the intercept term,  $\beta_0$  closely resembles the pattern of models and bandwidths for LOOCV, GCC, and AIC. Roughly 10 percent of simulations have the smallest RMSE using the ‘LLL’ model, while roughly half of the simulations selected the ‘LGG’ model with the smallest bandwidth. However, we see a different pattern for  $\beta_1$  and  $\beta_2$ . In both instances, no single model and bandwidth combination yielded the smallest RMSE more than 20 percent of the time (the ‘GGL’ model with the second largest bandwidth). That is, the intercept



**Figure 2:** Model and Bandwidth Selected by Metric when the True Model = ‘GGG’ (All coefficients are stationary)  
Each subfigure shows the percentage of simulations a given model and bandwidth combination was selected among the 50 possible combinations (7 bandwidths for each of the seven mixed models plus the GGG model) for a given metric. The sum of all values in a given subfigure sum to 100 subject to rounding error. For convenience, cell values less than 0.5 percent are omitted. The color saturation of the cells helps denote the magnitude of the values.



**Figure 3:** Model and Bandwidth Selected by Metric when the True Model = 'LLL' (All coefficients are non-stationary)  
Each subfigure shows the percentage of simulations a given model and bandwidth combination was selected among the 50 possible combinations (7 bandwidths for each of the seven mixed models plus the GGG model) for a given metric. The sum of all values in a given subfigure sum to 100 subject to rounding error. For convenience, cell values less than 0.5 percent are omitted. The color saturation of the cells helps denote the magnitude of the values.

was fixed while the other two coefficients were treated as non-stationary and the bandwidth was comprised of approximately two-thirds of the observations.

The simulations contained in Figure 3 actually contain a lot of variation in the amount of coefficient non-stationarity. Recalling that we implemented two amounts of spatial variation in each coefficient (think of it as ‘some’ and ‘more’) there are actually eight combinations of ‘LLL’ models ranging from “some, some, some,” to “more, more, more.” We therefore present a comparison of the following three models, the ‘GGG’ model we’ve already discussed, the ‘LLL’ model containing the smallest amount of overall spatial non-stationarity in all coefficients (the “some, some, some” model), and the largest amount of overall spatial non-stationarity (the “more, more, more” model).

### 3.3 One Local Coefficient

What have we learned in the previous sections?

1. LOOCV does the best identifying the non-stationary models
2. most accurate  $\hat{\beta}$  estimates do not necessarily come from the correct model
3. SCV sucks
4. something else.

Now we explore the mixed model cases in which some of the coefficients are stationary and some are non-stationary. We begin with the case of a single non-stationary coefficient.

Figure 5 shows that most metrics can correctly identify the proper mixed model when only one coefficient is non-stationary.

### 3.4 Two Local Coefficients

Now we explore the mixed model cases in which two of the coefficients are non-stationary.

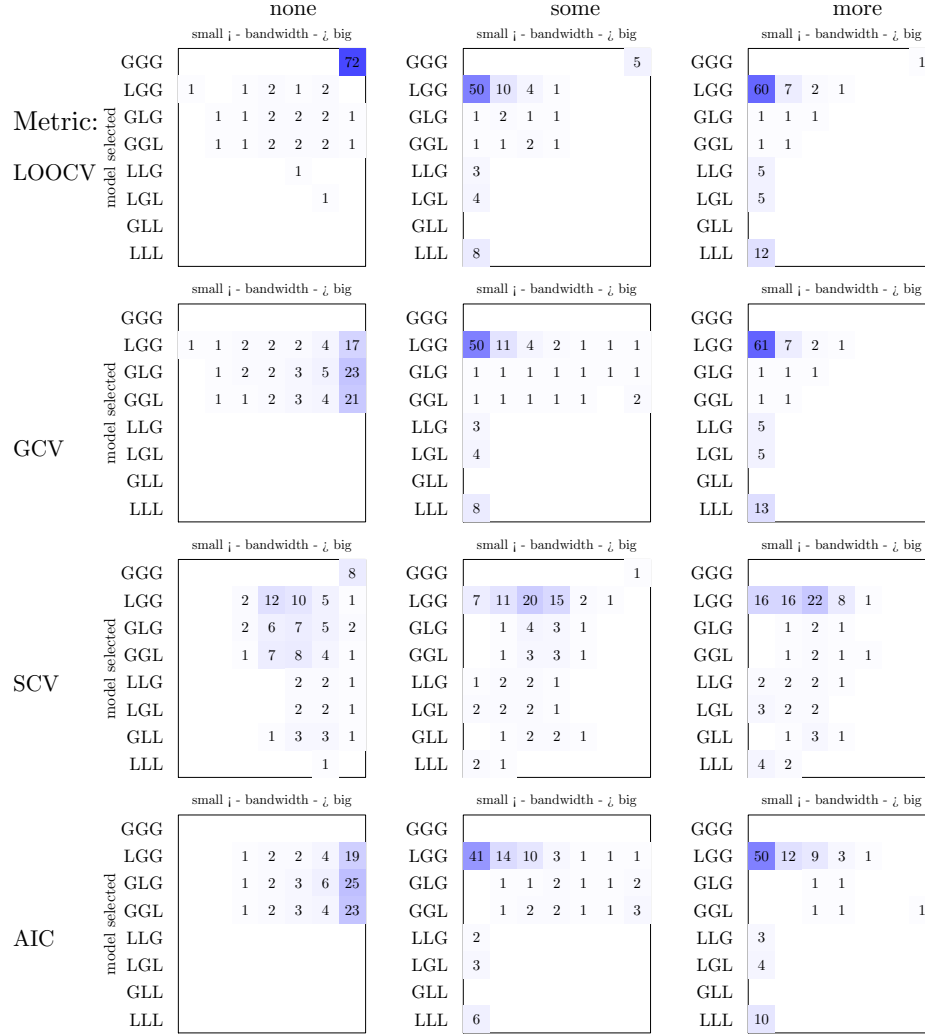
Figure 6 shows that most metrics can.

## 4 Discussion

Let’s show the model selection by metric across our 8 different models.

The left (right) column in Figure 7 contains the four models in which the intercept coefficient is (non-)stationary over the study area. In the top row of the figure both  $\beta_1$  and  $\beta_2$  are stationary, the middle two rows contain one stationary and one non-stationary coefficient, and in the bottom row both  $\beta_1$  and  $\beta_2$  are non-stationary. An interesting pattern is visible. The models on the left tend to be more accurately identified by the metrics than the models on the right. That is, if the intercept term is stationary, our information metrics do a pretty good job identifying whether or not the other coefficients are stationary or not. However, when the intercept term is stationary, our metrics commonly choose the ‘LGG’ model even when one or both of the other coefficients is non-stationary.

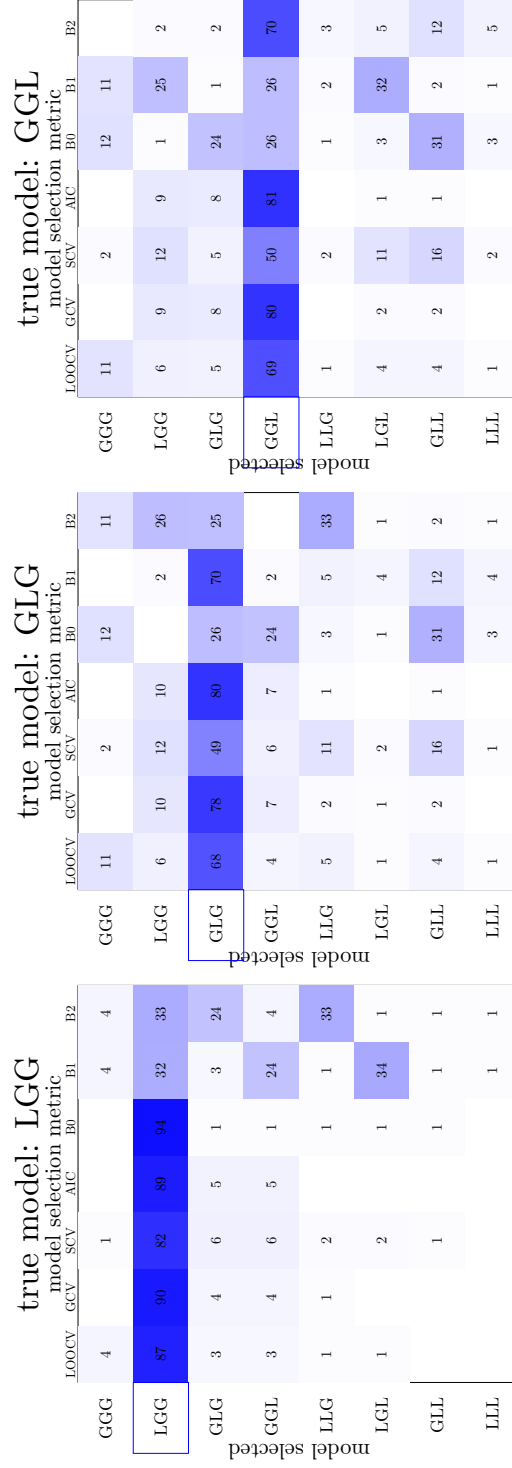
### Spatial Variation in All Three Coefficients:



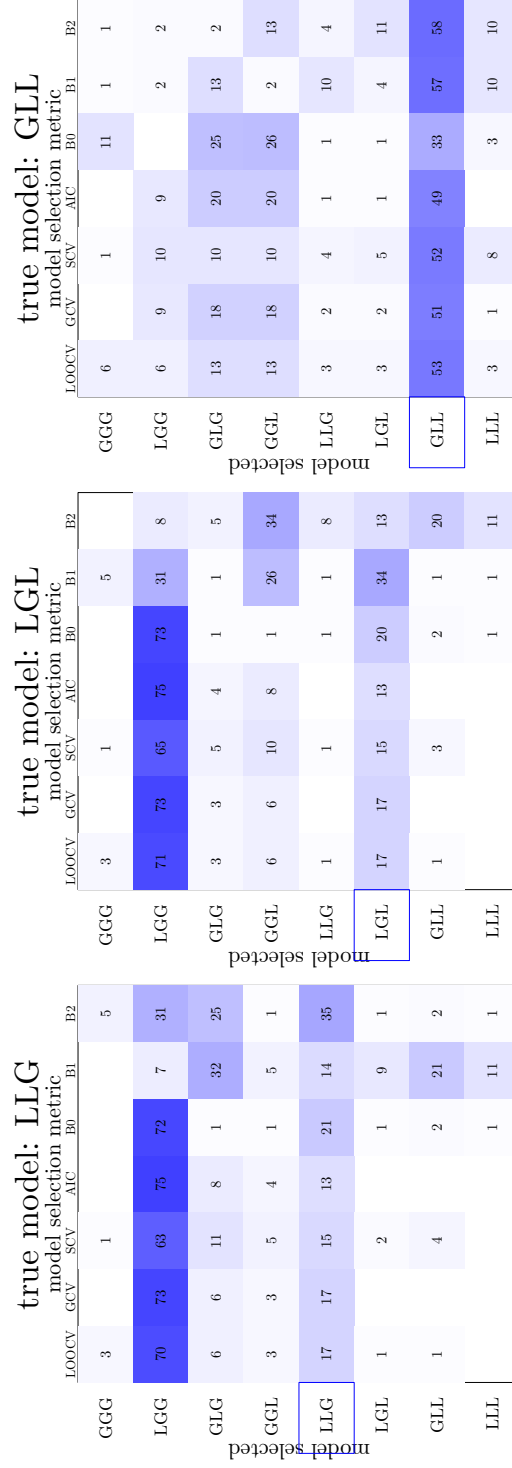
**Figure 4:** Model and Bandwidth Selected by Metric for ‘None’, ‘Some’, and ‘More’ Spatial Variation in the Three Regression Coefficients

Each subfigure shows the percentage of simulations a given model and bandwidth combination was selected among the 50 possible combinations (7 bandwidths for each of the seven mixed models plus the GGG model) for a given metric. The sum of all values in a given subfigure sum to 100 subject to rounding error. For convenience, cell values less than 0.5 percent are omitted. The color saturation of the cells helps denote the magnitude of the values.





**Figure 5:** This figure shows



**Figure 6:** This figure shows

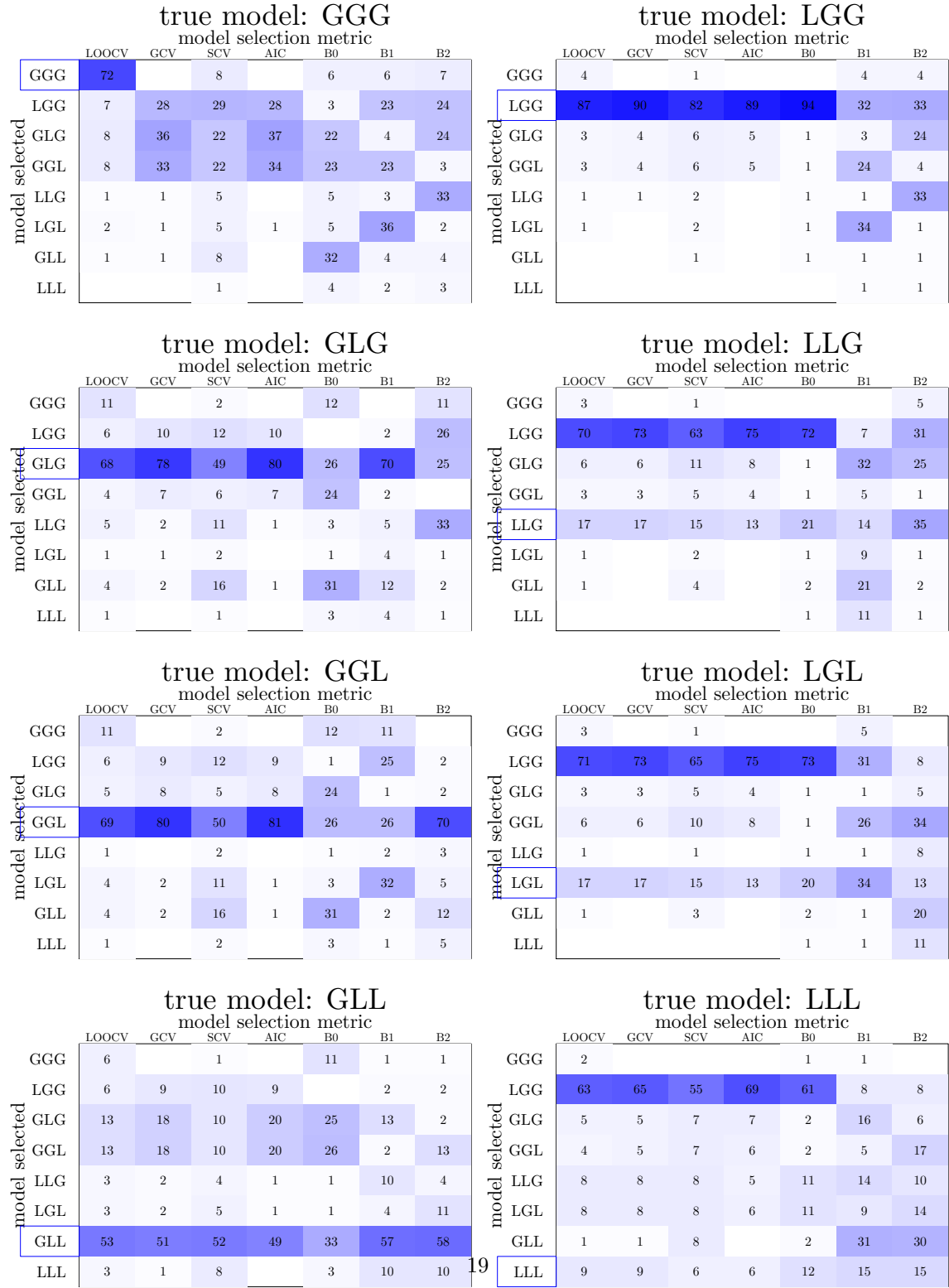


Figure 7: This figure shows

## References

- Kay W Axhausen and Michael Löchl. Modeling hedonic residential rents for land use and transport simulation while considering spatial effects. *Journal of Transport and Land Use*, 3(2):39–63, 2010. doi: 10.1598/jtlu.v3i2.117.
- B Y Richard A Borst and William J Mccluskey. Using Geographically Weighted Regression to Detect Housing Submarkets: Modeling Large-Scale Spatial Variations in Value. *Journal of Property Tax Assessment Administration*, 5(1):21–54, 2007. ISSN 15483614. URL <http://www.redi-bw.de/db/ebsco.php/search.ebscohost.com/login.aspx?direct=true&db=buh&AN=31846286&site=ehost-live>.
- Chris Brunsdon, Stewart Fotheringham, and Martin Charlton. Geographically weighted regression-modelling spatial non-stationarity. *Journal of the Royal Statistical Society Series D The Statistician*, 47(3):431–443, 1998a. ISSN 00390526. doi: 10.1111/1467-9884.00145. URL <http://www.jstor.org/stable/2988625>.
- Chris Brunsdon, Stewart Fotheringham, and Martin Charlton. Geographically weighted regression-modelling spatial non-stationarity. *Journal of the Royal Statistical Society Series D The Statistician*, 47(3):431–443, 1998b. ISSN 00390526. doi: 10.1111/1467-9884.00145. URL <http://www.jstor.org/stable/2988625>.
- M Cahill and G Mulligan. Using Geographically Weighted Regression to Explore Local Crime Patterns. *Social Science Computer Review*, 25(2):174–193, 2007. ISSN 08944393. doi: 10.1177/0894439307298925. URL <http://ssc.sagepub.com/cgi/doi/10.1177/0894439307298925>.
- Seong-hoon Cho, J. M Bowker, and William M. Park. Measuring the Contribution of Water and Green Space Amenities to Housing Values: An Application and Comparison of Spatially Weighted Hedonic Models. *Journal of Agricultural and Resource Economics*, 31(3):485–507, 2006.
- Seong-hoon Cho, D Christopher, M William, and Seung Gyu Kim. Spatial and Temporal Variation in the Housing Market Values of Lot Size and Open Space. *Land Economics*, 85(1):51–73, 2009.
- Lee-fang Chow, Fang Zhao, Xuemei Liu, Min-tang Li, and Ike Ubaka. Transit Ridership Model Based on Geographically Weighted Regression. *Transportation Research Record*, 1972(1972):105–114, 2000. ISSN 03611981. doi: 10.3141/1972-15. URL <http://dx.doi.org/10.3141/1972-15>.
- William S Cleveland and Susan J Devlin. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403):596–610, June 1988. doi: 10.1080/01621459.1959.10501996.
- Hans-Friedrich Eckey, Reinhold Kosfeld, and Matthias Türck. Regional Convergence in Germany: a Geographically Weighted Regression Approach. *Spatial Economic Analysis*, 2(1):45–64, 2007. ISSN 17421772. doi: 10.1080/

17421770701251905. URL <http://www.tandfonline.com/doi/abs/10.1080/17421770701251905>.

S Farber and A Páez. A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. *Journal of Geographical Systems*, 9(4):371–396, 2007.

G.M M Foody. Geographical weighting as a further refinement to regression modelling: An example focused on the NDVIrainfall relationship. *Remote Sensing of Environment*, 88(3):283–293, December 2003. ISSN 00344257. doi: 10.1016/j.rse.2003.08.004. URL <http://eprints.soton.ac.uk/14528/http://linkinghub.elsevier.com/retrieve/pii/S0034425703001949>.

A. Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically Weighted Regression: the analysis of spatially varying relationships*. John Wiley & Sons, West Sussex, England, 2002.

Dean M Hanink, Robert G Cromley, and Avraham Y Ebenstein. Spatial Variation in the Determinants of House Prices and Apartment Rents in China. *The Journal of Real Estate Finance and Economics*, 2010. ISSN 08955638. doi: 10.1007/s11146-010-9262-3. URL <http://www.springerlink.com/index/10.1007/s11146-010-9262-3>.

K E Haynes. The Employment Effects of New Business Formation: A Regional Perspective. *Economic Development Quarterly*, 25(3):282–292, 2011. ISSN 08912424. doi: 10.1177/0891242411407310. URL <http://edq.sagepub.com/cgi/doi/10.1177/0891242411407310>.

Marco Helbich and Michael Leitner. Spatial Analysis of the Urban-to-Rural Migration Determinants in the Viennese Metropolitan Area. A Transition from Suburbia to Postsuburbia? *Applied Spatial Analysis and Policy*, 2(3):237–260, 2009. ISSN 1874463X. doi: 10.1007/s12061-009-9026-8. URL <http://www.springerlink.com/content/pkq04616t6577843>.

Bo Huang, Bo Wu, and Michael Barry. Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science*, 24(3):383–401, 2010. ISSN 13658816. doi: 10.1080/13658810802672469. URL <http://www.informaworld.com/openurl?genre=article&doi=10.1080/13658810802672469&magic=crossref>.

Yefang Huang and Yee Leung. Analysing regional industrialisation in Jiangsu province using geographically weighted regression. *Journal of Geographical Systems*, 4(2):233–249, 2002. ISSN 14355930. doi: 10.1007/s101090200081. URL <http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s101090200081>.

Yan Kestens, Marius Thériault, and François Des Rosiers. Heterogeneity in hedonic modelling of house prices: looking at buyers household profiles.

*Journal of Geographical Systems*, 8(1):61–96, 2005. ISSN 14355930. doi: 10.1007/s10109-005-0011-8. URL <http://www.springerlink.com/index/10.1007/s10109-005-0011-8>.

Yee Leung, Chang-Lin Mei, and Wen-Xiu Zhang. Testing for spatial autocorrelation among the residuals of the geographically weighted regression. *Environment and Planning - Part A*, 32(5):871–890, 2000. ISSN 0308518X. doi: 10.1068/a32117. URL <http://www.envplan.com/abstract.cgi?id=a32117>.

Chris Lloyd and Ian Shuttleworth. Analysing commuting using local regression techniques: scale, sensitivity, and geographical patterning. *Environment and Planning - Part A*, 37(1):81–103, 2005. ISSN 0308518X. doi: 10.1068/a36116. URL <http://www.envplan.com/abstract.cgi?id=a36116>.

Clive Loader. *Local Regression and Likelihood*. Springer-Verlag, New York, NY, 1999.

Daniel P. McMillen. Perspectives on Spatial Econometrics: Linear Smoothing With Structured Models. *Journal of Regional Science*, 52(2):192–209, May 2012. ISSN 00224146. doi: 10.1111/j.1467-9787.2011.00746.x. URL <http://doi.wiley.com/10.1111/j.1467-9787.2011.00746.x>.

Daniel P. McMillen and Christian L. Redfearn. Estimation and Hypothesis Testing for Nonparametric Hedonic House Price Functions. *Journal of Regional Science*, 50(3):712–733, April 2010. ISSN 00224146. doi: 10.1111/j.1467-9787.2010.00664.x. URL <http://doi.wiley.com/10.1111/j.1467-9787.2010.00664.x>.

Chang-Lin Mei, Shu-Yuan He, and Kai-Tai Fang. A NOTE ON THE MIXED GEOGRAPHICALLY WEIGHTED REGRESSION MODEL. *Journal of Regional Science*, 44(1):143–157, 2004. ISSN 00224146. doi: 10.1111/j.1085-9489.2004.00331.x. URL <http://doi.wiley.com/10.1111/j.1085-9489.2004.00331.x>.

Antonio Paez, Steven Farber, and David Wheeler. A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. *Environment and Planning A*, 43(12):2992–3010, 2011.

M D Partridge, D S Rickman, K Ali, and M R Olfert. The Geographic Diversity of U . S . Nonmetropolitan Growth Dynamics : A Geographically Weighted Regression Approach. *Land Economics*, 84(2):241–266, 2008. ISSN 00237639. URL <http://www.scopus.com/inward/record.url?eid=2-s2.0-44649124821&partnerID=40&md5=1385582b8ae5c778c531d9481977e14c>.

Mark D Partridge and Dan S Rickman. Persistent pockets of extreme American poverty and job growth: Is there a place-based policy role? *Journal of Agricultural and Resource Economics*, 32(1):201–224, 2007. ISSN 01621912. URL <https://libserv7.princeton.edu:82/pul/nph-pul2.cgi/000000A/http/www.jstor.org/stable/info/40987359=3f&Search=3dyes&searchText=3dpersistent&searchUri=3d=252Faction=252FdoBasicSearch=253FQuery=253Dpersistent=252Brural=252Bpoverty=2526acc=253Don=2526wc=253Don>.

- Noel Bonfilio Pineda Jaimes, Joaquín Bosque Sendra, Montserrat Gómez Delgado, and Roberto Franco Plata. Exploring the driving forces behind deforestation in the state of Mexico (Mexico) using geographically weighted regression. *Applied Geography*, 30(4):576–591, 2010. ISSN 01436228. doi: 10.1016/j.apgeog.2010.05.004. URL <http://linkinghub.elsevier.com/retrieve/pii/S0143622810000561>.
- Ana C L Sá, José M C Pereira, Martin E Charlton, Bernardo Mota, Paulo M Barbosa, and A Stewart Fotheringham. The pyrogeography of sub-Saharan Africa: a study of the spatial non-stationarity of fire-environment relationships using GWR. *Journal of Geographical Systems*, 13(3):227–248, 2010. ISSN 14355930. doi: 10.1007/s10109-010-0123-7. URL <http://www.springerlink.com/index/10.1007/s10109-010-0123-7>.
- Jun Tu and Zong-Guo Xia. Examining spatially varying relationships between land use and water quality using geographically weighted regression I: Model design and evaluation. *Science of the Total Environment*, 407(1):358–378, 2008. URL <http://www.ncbi.nlm.nih.gov/pubmed/18976797>.
- Dan-Lin Yu. Spatially varying development mechanisms in the Greater Beijing Area: a geographically weighted regression investigation. *The Annals of Regional Science*, 40(1):173–190, January 2006. ISSN 0570-1864. doi: 10.1007/s00168-005-0038-2. URL <http://www.springerlink.com/index/10.1007/s00168-005-0038-2>.
- Danlin Yu. Modeling owner-occupied single-family house values in the city of milwaukee: A geographically weighted regression approach. *GIScience Remote Sensing*, 44(3):267–282, 2007. ISSN 15481603. doi: 10.2747/1548-1603.44.3.267. URL <http://bellwether.metapress.com/openurl.asp?genre=article&id=doi:10.2747/1548-1603.44.3.267>.
- Danlin Yu, Yehua Dennis Wei, and Changshan Wu. Modeling spatial dimensions of housing prices in Milwaukee, WI. *Environment and Planning B: Planning Design*, 34(6):1085–1102, 2007. ISSN 02658135. doi: 10.1068/b32119. URL <http://www.envplan.com/abstract.cgi?id=b32119>.