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

Aaron Swoboda

December 17, 2014

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 β coefficients are constant over space. Alternatively, the coefficients could exhibit spatial non-stationarity, in which case one, two, or all three of the β coefficients are a function of 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 not implausible to believe 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 Geographically Weighted Regression to the Rescue?

Locally Weighted Regression (also referred to as Geographically Weighted Regression) is one possible solution to the challenge presented by spatially non-stationary regression coeffiecients. Locally Weighted Regression (LWR) techniques (also known as Geographically Weighted Regression) are described in detail by Cleveland and Devlin (1988), Brunsdon et al. (1998), Fotheringham et al. (2002), and others. It is a weighted least squares methodology in which regression coefficients are estimated over space as a function of the local data as described in Equation (2),

$$\hat{\beta}(location_i) = (X'W(location_i)X)^{-1}X'W(location_i)Y, \tag{2}$$

where X is a $N \times 2$ 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. We employ a bi-square weights function and a k-nearest neighbor bandwidth approach as described in equation (3),

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

where d_{ij} denotes the distance between observations i and j, and d_{ik} 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 unneccessary error in our estimates by excluding informative observations. Often, researchers choose a bandwidth my minimizing a cross validation metric.

This choice is further complicated in the context of mixed models where only some coefficents exhibit spatial stationarity (in contrast to standard models in which all coefficients are treated as spatially stationary or LWR models in which no coefficients are treated as stationary). Little is known about model performance when models are selected across multiple mixed models and among multiple different potenial 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. 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?

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,

$$LOOCV = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(y - \hat{y}_{\neq i})^2}, \tag{4}$$

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.

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 (5),

$$n * \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{(n - v_1)^2}, \tag{5}$$

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 $v_1 = \mathrm{tr}(\mathbf{S})$, where the matrix $\mathbf{S}$ is the "hat matrix" which maps y onto $\hat{y}$,

$$\hat{y} = \mathbf{S}y, \tag{6}$$

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

$$r_i = X_i(X'W_iX)^{-1}X'W_i. \tag{7}$$

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).

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,

$$SCV_i(k) = \frac{\sum(y_i - \hat{y}_{-i}(k))^2}{\sum_k (y_i - \hat{y}_{-i})^2}, \tag{8}$$

and the total score for bandwidth k is then,

$$SCV(k) = \sum_i SCV_i(k). \tag{9}$$

Equation (8) calculates a partial score for each observation as a proportion of the total squared deviance at that observation across the different bandwidths, while (9) 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.

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

$$2 * n * ln(\hat{\sigma}) + n * ln(2 * \pi) + n * \frac{n + v_1}{n - 2 - v_1} \tag{10}$$

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.

1.2 Experimental Design

We generate data in the following format:

$$Y = \beta_0(location) + \beta_1(location) * X_1 + \beta_2(location) * X_2 + \epsilon, \tag{11}$$

where sometimes the coefficient is in fact stationary, $\beta_m(location) = \beta_m$, and other times it is non-stationary, $\beta_m(location_p) \neq \beta_m(location_q)$. With three coefficients, $m = \{0, 1, 2\}$, each having the possibility of being stationary or not, there are eight different possible combinations of the three parameters, ranging from (stationary, stationary, stationary) to (non-stationary, non-stationary, non-stationary). We refer to any parameter combination containing both stationary and non-stationary coefficients as "mixed."

We generate data using all eight different combinations and then estimate all eight possible LWR models across seven different bandwidth sizes. We then calculate different Cross-Validation metrics and compare their values across models and bandwidths.

We have three different values for each coefficient in our DGP, no variation, some variation, and more variation. We also change the sample size of our data as well as the variance of the model error term.

2 Simulation Results

2.1 Starting Simple: All Coefficients are Spatially Stationary

We begin by examining the simulation resuts for the spatially stationary data generation process. With no spatial variation for any of the coeffients, these data are consistent with standard OLS regression. We label this model 'GGG' to denote

that all three coefficients are 'Global' rather than 'Local'.[1] 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	
	GLG	8	36	22	37	2/3 Correct
	GGL	8	33	22	34	
	LLG	1	1	5	0	
	LGL	2	1	5	1	1/3 Correct
	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, 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,

$$\text{RMSE } \widehat{\beta}_m = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{\beta}_{mi} - \beta_{mi})^2}, \tag{12}$$

[1]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.

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	6	7
	LGG	3	23	24
	GLG	22	4	24
	GGL	23	23	3
	LLG	5	3	33
	LGL	5	36	2
	GLL	32	4	4
	LLL	4	2	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 estiamtes 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.

2.1.1 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 dentify 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 1 shows the distribution of model selected and bandwidth size for each of the seven metrics we've discussed.

2.2 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 1 but with a true model of 'LLL.'

Figure 2 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 2 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, β_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 β_1 and β_2. In both instances, no single model and bandwidth combination yielded the smallest RMSE more than 20 percent of the time (the 'GLL' model with the second largest bandwidth). That is, the intercept 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 2 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

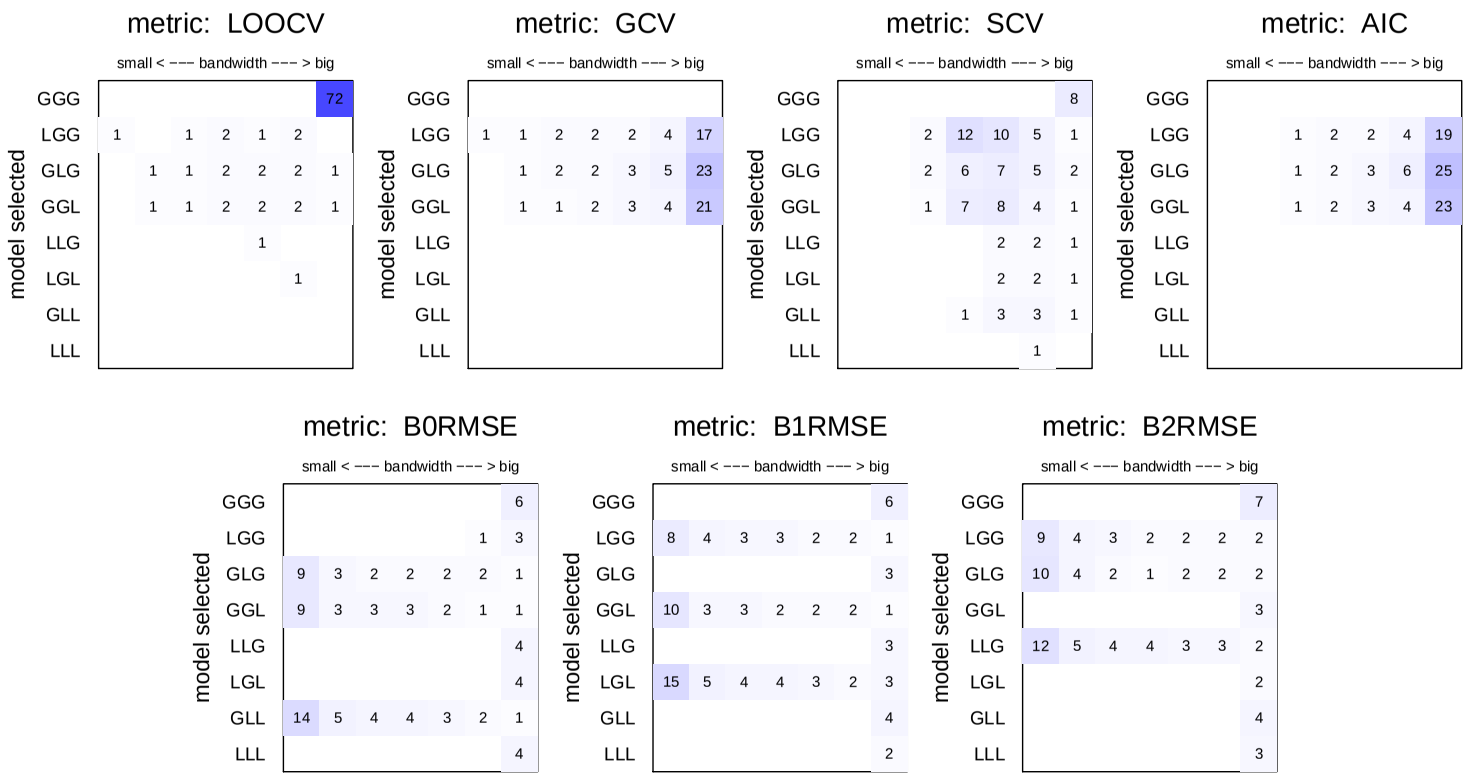

Figure 1: Model and Bandwidth Selected by Metric when the True Model = 'GGG' (All coefficients are stationary)
Each subfigure shows the percentage of simulations a given model and bandwidth combination was selected among the 50 possible combinations (7 bandwidths for each of the seven mixed models plus 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.

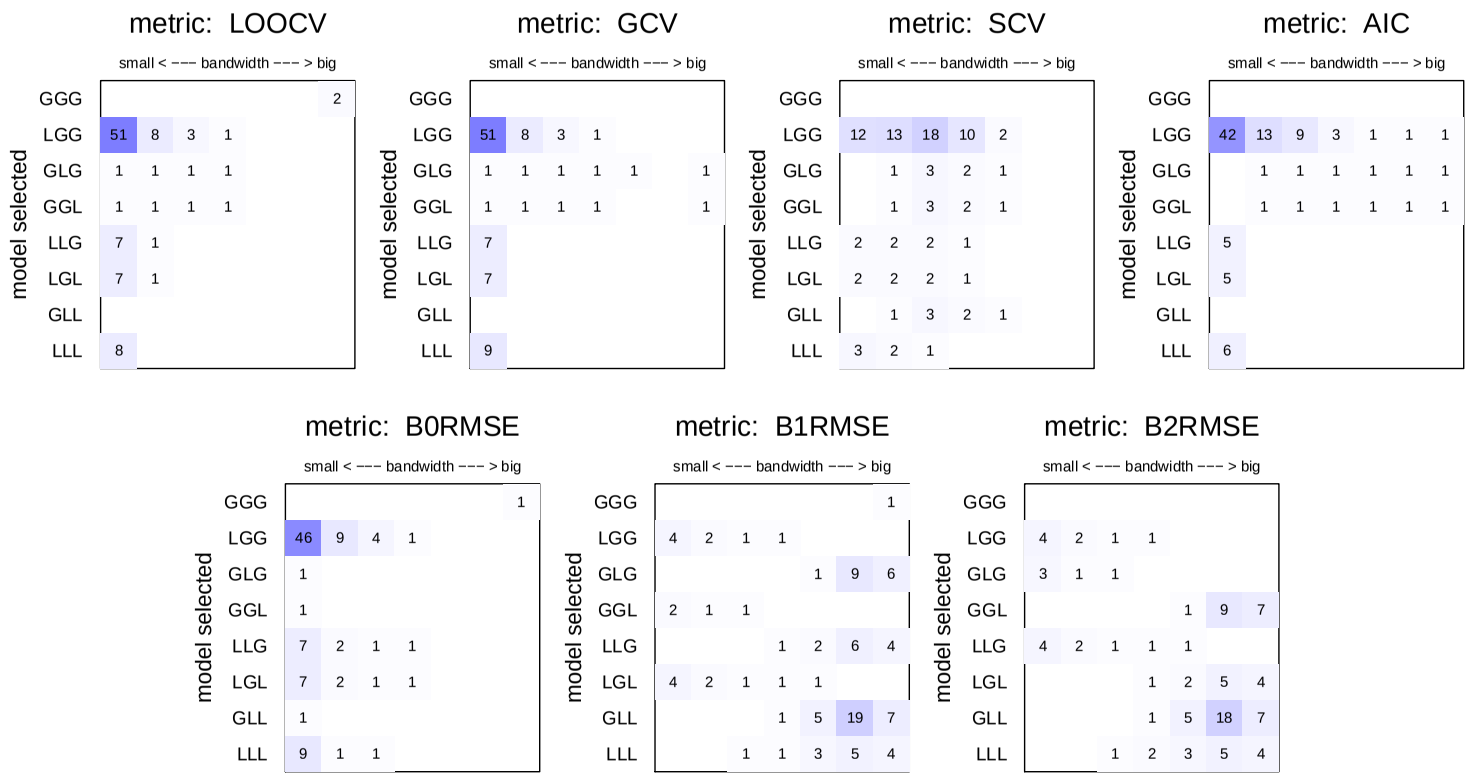

Figure 2: Model and Bandwidth Selected by Metric when the True Model = 'LLL' (All coefficients are non-stationary)
Each subfigure shows the percentage of simulations a given model and bandwidth combination was selected among the 50 possible combinations (7 bandwidths for each of the seven mixed models 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.

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).

- The AIC and GCV metrics almost *never* select the 'GGG' model, even when it is the actual model.
- When exactly one variable is non-stationary, AIC, GCV, and LOOCV do a very good job identifying the true model(over two-thirds of the time), while SCV does slightly less well (only 50 percent of the time if the non-stationary coefficient isn't the intercept term).
- Frequently, when there are two or more non-stationary variables, AIC, GCV, SCV, and LOOCV over selected the 'LGG' model.
- There are several occasions where the model/bandwidth combination with the smallest RMSE is not the "correct" model.
- It is frequently the case that the models with the smallest RMSE for a given coefficient have the (non-) stationarity of the individuals coefficient correctly identified but incorrectly identify the (non-)stationarity of the other coefficients.

Figure 4 shows

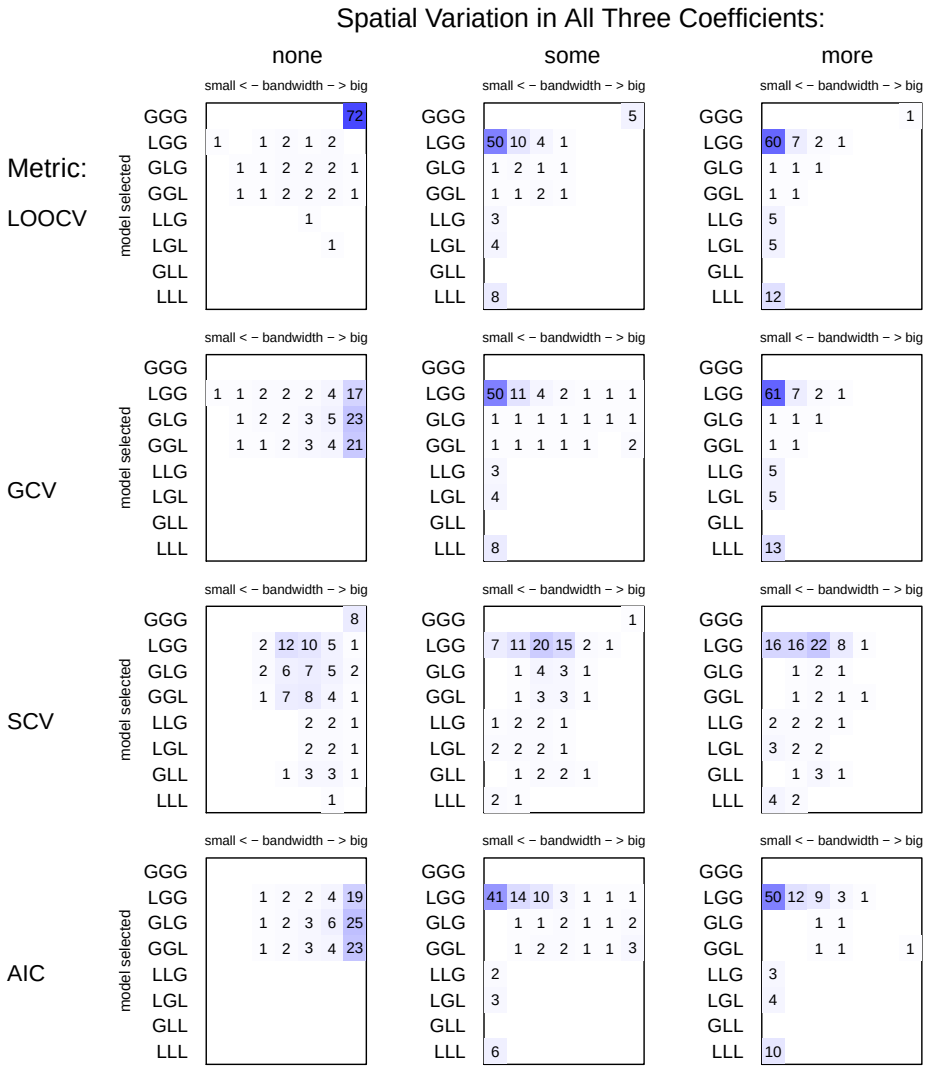

Figure 3: Model and Bandwidth Selected by Metric for 'None', 'Some', and 'More' Spatial Variation in the Three Regression Coefficients
Each subfigure shows the percentage of simulations a given model and bandwidth combination was selected among the 50 possible combinations (7 bandwidths for each of the 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.

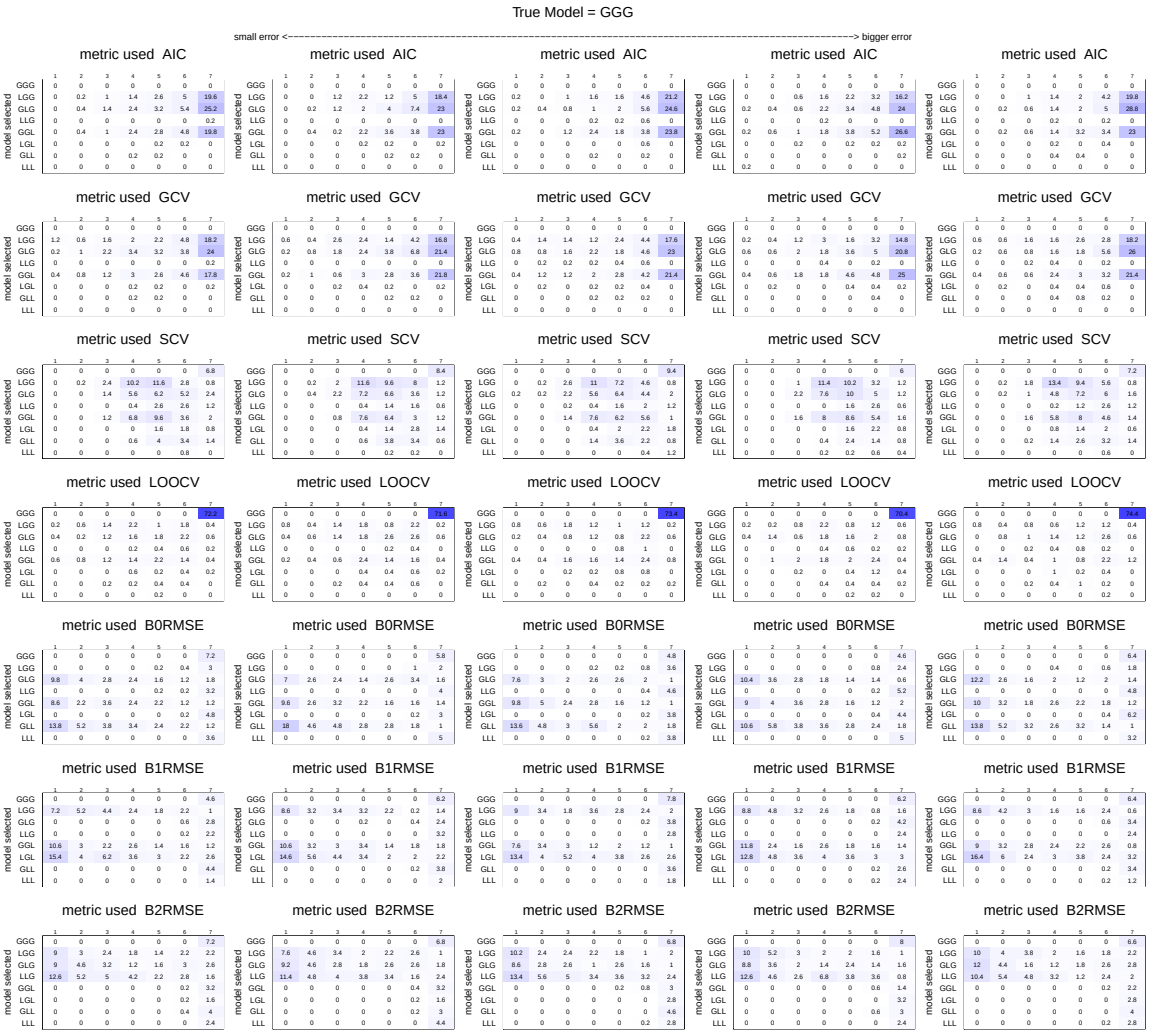

Figure 4: This figure shows the GLG model...

References

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, 1998. ISSN 00390526. doi: 10.1111/1467-9884.00145. URL `http://www.jstor.org/stable/2988625`.

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.

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.

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

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`.

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.