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

Aaron Swoboda

December 12, 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).

[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.

		Metric				
		LOOCV	GCV	SCV	AIC	
	GGG	72	0	8	0	3/3 Correct
Model Selected	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,

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

where i denotes the observation, N is the sample size, and $m \in \{0, 1, 2\}$ specifies the model coefficient in question.

The final three columns show which model minimized the Root Mean Square Error for each of the three coefficients in our model. These metrics are only available because of the nature of the experiment - we know the true values of the coefficients in the underlying true model and so can compare the estimated coefficients to their

Table 2: Here is my Title

Model Selected	$\widehat{B_0}$	$\widehat{B_1}$	$\widehat{B_2}$
	Coefficient RMSE		
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

This table shows...

true values. The results are startling. The true model, 'GGG' yields that most accurate estimates of the coefficient in question less than 10 percent of the time. Further inspection shows an interesting pattern. For each of our three coefficients, the model with the most accurate estimates is the model that is correct about the global nature of the coefficient in question, but uses a local model for the other two coefficients. For instance, the most accurate estimates of β_0 occur almost one-third of the time when using the 'GLL' model. In other words, the model that is most likely to have the most accurate estimates for a given coefficient is consistently NOT the correct model, although it tends to correctly identify the spatial (non-)stationarity of the coefficient in question.

As a reminder, four of these are available to a researcher with actual data, while the $\text{RMSE}_{\hat{\beta}_0}$, $\text{RMSE}_{\hat{\beta}_1}$, and $\text{RMSE}_{\hat{\beta}_2}$ are only available to us with the simulated data because we generated the data and have known βs.

We have seven different ways to pick the "best" model (the AIC, GCV, SCV, LOOCV, and RMSEs for the three different coefficients). Here we produce a table showing the relative frequency (in percentage) that each model number was selected by optimizing a given metric. Note that the columns in the following tables may not sum exactly to 100 due to rounding. In this instance the true model was one of complete spatial stationarity (no coefficients varied over space). The true model is model 1.

Let's visualize these results in Figure 1. Some patterns are evident:

- 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.

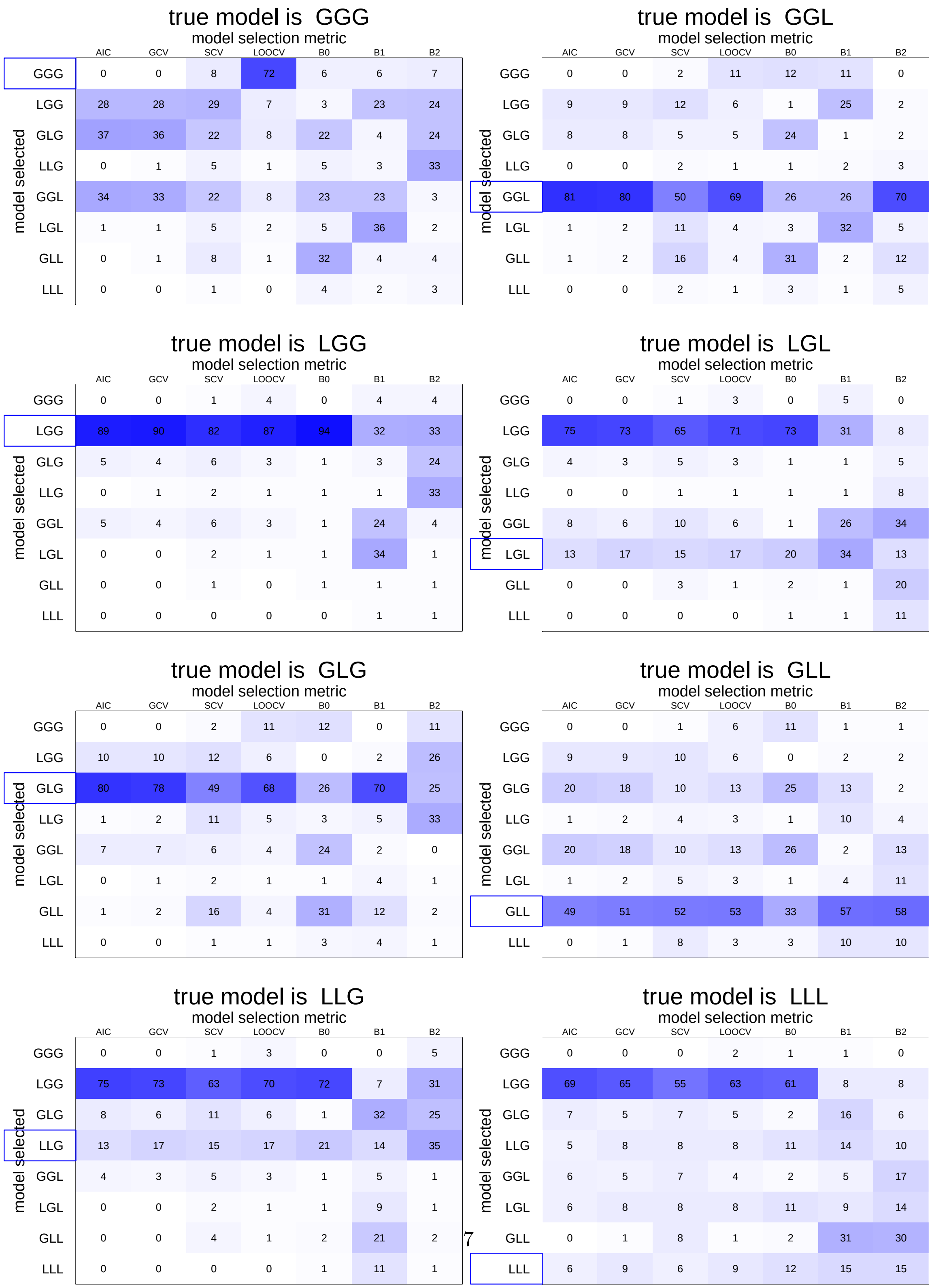

Figure 1: This figure shows

- 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.

2.2 Bandwidth Size

The results of the previous tables must be taken with a grain of salt, as there could times when a "true" model may include variation in a coefficient, but the degree of non-stationarity in the coefficient may be small. In such cases, choosing an incorrect model (such as one that keeps such a coefficient constant) may not be such a big problem. Alternatively, the model with the smallest metric may incorrectly have some local coefficients, but the bandwidth chosen might be large and therefore allows for very little variation in the coefficients.

The idea is that the model selected might look 'overly local' in that it selects models with too many local coefficients. (OK, so need a new variable that is the number of coefficients that are local in the selected model, number in the correct model, difference, and whether or not - even if the number is the same, say one or two, but are they the correct one or two...)

Then, can we look at the bandwidth sizes for different groups of models? For instance, if two local coefficients are selected but only supposed to be one, how do those bandwidths compare to those that correctly identify as one or two?

Let's look at the simplest case first - when the true model was GGG.

Let's construct a table of the bandwidth size selected by model...

Figure 2 shows

Figure 3 shows

metric used AIC

model selected	1	2	3	4	5	6	7
GGG	0	0	0	0	0	0	0
LGG	0	0	1	1.6	1.9	4.4	19
GLG	0.1	0.3	0.9	1.8	2.9	5.6	25.1
LLG	0	0	0	0.1	0	0.2	0
GGL	0.1	0.3	0.8	2	3	4.2	23.2
LGL	0	0	0	0.1	0.1	0.3	0.1
GLL	0	0	0	0.2	0.2	0.1	0
LLL	0	0	0	0	0	0	0

metric used GCV

model selected	1	2	3	4	5	6	7
GGG	0	0	0	0	0	0	0
LGG	0.6	0.7	1.7	2	2	3.9	17.1
GLG	0.4	0.8	1.7	2.3	2.8	5.2	23
LLG	0	0	0.1	0.2	0.1	0.2	0
GGL	0.4	0.8	1.1	2.4	3.2	4.1	21.5
LGL	0	0.1	0	0.2	0.3	0.3	0.1
GLL	0	0	0	0.2	0.3	0.2	0
LLL	0	0	0	0	0	0	0

metric used SCV

model selected	1	2	3	4	5	6	7
GGG	0	0	0	0	0	0	7.6
LGG	0	0.2	2	11.5	9.6	4.8	1
GLG	0	0.2	1.8	6.2	7.3	4.8	1.7
LLG	0	0	0	0.3	1.7	2.3	1
GGL	0	0	1.3	7.2	7.8	4.4	1.4
LGL	0	0	0	0.3	1.6	2.2	1.1
GLL	0	0	0	0.9	3.3	2.7	1
LLL	0	0	0	0	0.1	0.5	0.3

metric used LOOCV

model selected	1	2	3	4	5	6	7
GGG	0	0	0	0	0	0	72.4
LGG	0.6	0.4	1.2	1.6	1	1.5	0.4
GLG	0.3	0.7	1	1.6	1.6	2.3	0.6
LLG	0	0	0	0.2	0.6	0.5	0.1
GGL	0.3	0.8	1.2	1.6	1.6	2	0.6
LGL	0	0	0.1	0.4	0.4	0.7	0.2
GLL	0	0	0.1	0.4	0.5	0.4	0.1
LLL	0	0	0	0	0.1	0.1	0

metric used B0RMSE

model selected	1	2	3	4	5	6	7
GGG	0	0	0	0	0	0	5.8
LGG	0	0	0	0.1	0.1	0.7	2.6
GLG	9.4	3.2	2.3	2	1.9	2	1.3
LLG	0	0	0	0	0	0.2	4.4
GGL	9.4	3.4	2.9	2.6	1.8	1.4	1.4
LGL	0	0	0	0	0	0.3	4.4
GLL	14	5.1	3.7	3.6	2.6	2	1.4
LLL	0	0	0	0	0	0	4.1

metric used B1RMSE

model selected	1	2	3	4	5	6	7
GGG	0	0	0	0	0	0	6.2
LGG	8.4	4.2	3.2	2.7	2	1.6	1.3
GLG	0	0	0	0	0	0.4	3.3
LLG	0	0	0	0	0	0	2.6
GGL	9.9	3	2.5	2.4	1.8	1.8	1.2
LGL	14.5	4.9	4.4	3.6	3.2	2.4	2.7
GLL	0	0	0	0	0	0.1	3.6
LLL	0	0	0	0	0	0.1	1.8

metric used B2RMSE

model selected	1	2	3	4	5	6	7
GGG	0	0	0	0	0	0	7.1
LGG	9.4	3.8	3	2	1.8	1.8	1.7
GLG	9.5	4	2.4	1.3	2.2	2.2	2
LLG	12.1	5.1	4.3	4.3	2.8	2.7	1.8
GGL	0	0	0	0	0	0.4	2.6
LGL	0	0	0	0	0	0.1	2.4
GLL	0	0	0	0	0	0.2	3.7
LLL	0	0	0	0	0	0.1	3

Figure 2: This figure shows

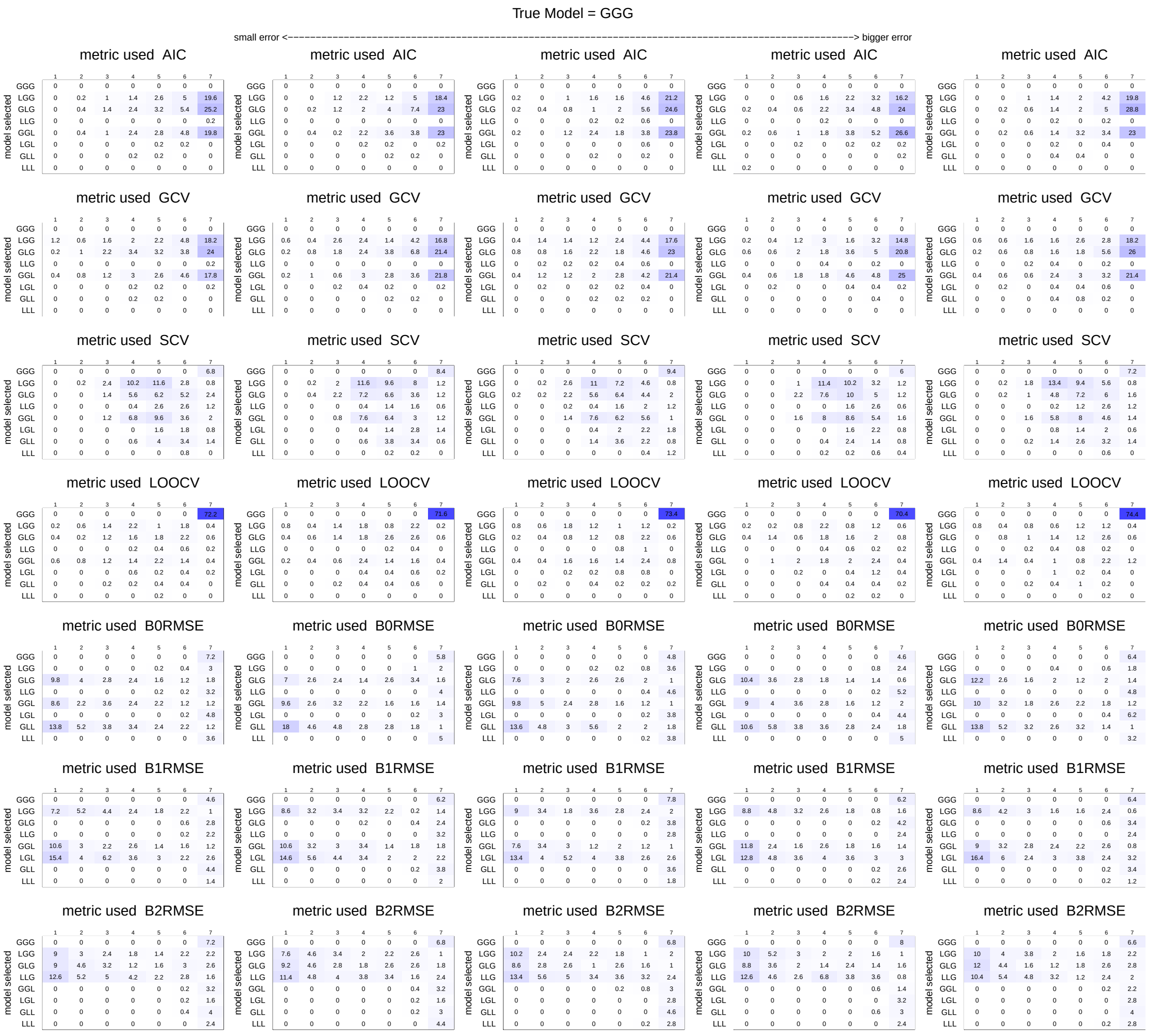

Figure 3: This figure shows the GLG model...

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.