Mixed GWR Simulation Write-up

Aaron Swoboda

August 22, 2014

What are our research questions? Basically:

1. Can we find the "true" model among the eight different possibilities with three model parameters?
2. Are there differences in the results based on the metric used?
3. What happens as we change the sample size and amount of error in the model?
4. How much does it really matter if we are concerned with coefficient estimates?
   - How well does our selected model perform as measured by beta RMSE?
   - Does our model perform better when we select the correct model?
   - Can we control for whether we selected a model with the correct spatial variation for a given parameter?
5. What about using bandwidth size as a dependent variable?
6. What happens when we use other decision tools to help with model selection? (Monte Carlo simulations and test statistics)

1 Background

Imagine a simple linear model,

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \epsilon. \quad (1)$$

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 $Nx5$ 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.

While it is possible to parameterize the variation in coefficients, for instance researchers often (CITATION?) include 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.

1.1 Local Regression to the Rescue?

Locally Weighted Regression (also known 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),

$$\hat{\beta}_i = (X'W_iX)^{-1}X'W_iY, \tag{2}$$

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.

2 Title Needed

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 (how many citations?) 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 =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 (CITATION)

AIC

2.1 Experimental Design

We generate data in the following format:

$$Y = \beta_0(location) + \beta_1(location) * X_1 + \beta_2(location) * X_2 + \epsilon, \tag{8}$$

where sometimes the coefficient is in fact stationary, $\beta_m(location) = \beta$, 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, ranging from (stationary, stationary, stationary) to (non-stationary, non-stationary, non-stationary).

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.

3 Simulation Results

We have seven different ways to pick the "best" model (the AIC, GCV, SCV, LOOCV, and RMSEs for the three different coefficients). Here are tables 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.

```
for (i in 1:8) {
  temp2 = which(mcOutput[,"True Model" ,] == i, arr.ind = TRUE)
  temp3 = mcOutput[8:14, "Model Number", unique(temp2[, 2])]
  temp4 = factor(temp3)
```

```
  newdata = data.frame(ModelNum = temp4,
                       Metric = factor(rownames(temp3), levels = rownames(temp3)))
  cat(paste("\n true model =", i, "\n"))
  cat("spatial variation...\n")
  print(models[i, ])
  print(round(table(newdata)*100*7/sum(table(newdata)), 0))
}
```

```
##
##  true model = 1
## spatial variation...
##   beta0 beta1 beta2
## 1    no    no    no
##         Metric
## ModelNum AIC GCV SCV LOOCV B0RMSE B1RMSE B2RMSE
##        1   0   0   8    72      6      6      7
##        2  28  28  29     7      3     23     24
##        3  37  36  22     8     22      4     24
##        4   0   1   5     1      5      3     33
##        5  34  33  22     8     23     23      3
##        6   1   1   5     2      5     36      2
##        7   0   1   8     1     32      4      4
##        8   0   0   1     0      4      2      3
##
##  true model = 2
## spatial variation...
##   beta0 beta1 beta2
## 2   yes    no    no
##         Metric
## ModelNum AIC GCV SCV LOOCV B0RMSE B1RMSE B2RMSE
##        1   0   0   1     4      0      4      4
##        2  89  90  82    87     94     32     33
##        3   5   4   6     3      1      3     24
##        4   0   1   2     1      1      1     33
##        5   5   4   6     3      1     24      4
##        6   0   0   2     1      1     34      1
##        7   0   0   1     0      1      1      1
##        8   0   0   0     0      0      1      1
##
##  true model = 3
## spatial variation...
##   beta0 beta1 beta2
## 3    no   yes    no
##         Metric
## ModelNum AIC GCV SCV LOOCV B0RMSE B1RMSE B2RMSE
##        1   0   0   2    11     12      0     11
```

```
##        2  10  10  12     6      0      2     26
##        3  80  78  49    68     26     70     25
##        4   1   2  11     5      3      5     33
##        5   7   7   6     4     24      2      0
##        6   0   1   2     1      1      4      1
##        7   1   2  16     4     31     12      2
##        8   0   0   1     1      3      4      1
##
##  true model = 4
## spatial variation...
##   beta0 beta1 beta2
## 4   yes   yes    no
##         Metric
## ModelNum AIC GCV SCV LOOCV B0RMSE B1RMSE B2RMSE
##        1   0   0   1     3      0      0      5
##        2  75  73  63    70     72      7     31
##        3   8   6  11     6      1     32     25
##        4  13  17  15    17     21     14     35
##        5   4   3   5     3      1      5      1
##        6   0   0   2     1      1      9      1
##        7   0   0   4     1      2     21      2
##        8   0   0   0     0      1     11      1
##
##  true model = 5
## spatial variation...
##   beta0 beta1 beta2
## 5    no    no   yes
##         Metric
## ModelNum AIC GCV SCV LOOCV B0RMSE B1RMSE B2RMSE
##        1   0   0   2    11     12     11      0
##        2   9   9  12     6      1     25      2
##        3   8   8   5     5     24      1      2
##        4   0   0   2     1      1      2      3
##        5  81  80  50    69     26     26     70
##        6   1   2  11     4      3     32      5
##        7   1   2  16     4     31      2     12
##        8   0   0   2     1      3      1      5
##
##  true model = 6
## spatial variation...
##   beta0 beta1 beta2
## 6   yes    no   yes
##         Metric
## ModelNum AIC GCV SCV LOOCV B0RMSE B1RMSE B2RMSE
##        1   0   0   1     3      0      5      0
##        2  75  73  65    71     73     31      8
```

```
##         3   4   3   5     3      1      1      5
##         4   0   0   1     1      1      1      8
##         5   8   6  10     6      1     26     34
##         6  13  17  15    17     20     34     13
##         7   0   0   3     1      2      1     20
##         8   0   0   0     0      1      1     11
##
##  true model = 7
## spatial variation...
##   beta0 beta1 beta2
## 7    no   yes   yes
##         Metric
## ModelNum AIC GCV SCV LOOCV B0RMSE B1RMSE B2RMSE
##         1   0   0   1     6     11      1      1
##         2   9   9  10     6      0      2      2
##         3  20  18  10    13     25     13      2
##         4   1   2   4     3      1     10      4
##         5  20  18  10    13     26      2     13
##         6   1   2   5     3      1      4     11
##         7  49  51  52    53     33     57     58
##         8   0   1   8     3      3     10     10
##
##  true model = 8
## spatial variation...
##   beta0 beta1 beta2
## 8   yes   yes   yes
##         Metric
## ModelNum AIC GCV SCV LOOCV B0RMSE B1RMSE B2RMSE
##         1   0   0   0     2      1      1      0
##         2  69  65  55    63     61      8      8
##         3   7   5   7     5      2     16      6
##         4   5   8   8     8     11     14     10
##         5   6   5   7     4      2      5     17
##         6   6   8   8     8     11      9     14
##         7   0   1   8     1      2     31     30
##         8   6   9   6     9     12     15     15
```

Let's visualize these results, starting with Model 1.

```
for (i in 1:8) {
  temp2 = which(mcOutput[,"True Model" ,] == i, arr.ind = TRUE)
  temp3 = mcOutput[8:14, c("Model Number"), unique(temp2[, 2])]
  temp4 = factor(temp3)

  newdata = data.frame(ModelNum = temp4,
                       Metric = factor(rownames(temp3), levels = rownames(temp3)))
```

```
  temptab = table(newdata)*100*7/sum(table(newdata))
testdf = round(temptab)
library(plotrix)
par(mar = c(0.5, 6, 6, 0.5))
testdf = round(testdf)
colnames(testdf)[5:7] = c("B0", "B1", "B2")
rownames(testdf) = c("GGG", "LGG", "GLG", "LLG", "GGL", "LGL", "GLL", "LLL")
cellcol<-color.scale(testdf, extremes = c("white", "blue"),
                     xrange = c(0, 100))
color2D.matplot(testdf,
                show.values = TRUE,
                axes = FALSE,
                xlab = "",
                ylab = "",
                vcex = 1,
                vcol = "black",
                border = NA,
                cellcolors = cellcol)
  axis(3, at = seq_len(ncol(testdf)) - 0.5, line = -1,
     labels = colnames(testdf), tick = FALSE, cex.axis = .9)
  axis(2, at = seq_len(nrow(testdf)) -0.5,
     labels = rev(rownames(testdf)), tick = FALSE, las = 1, cex.axis = 1.2)
  mtext("model selected", 2, line = 4)
  mtext(paste("true model is ", rownames(testdf)[i]), 3, line = 2.5, cex = 1.5)
  mtext("model selection metric", 3, line = 1)
  par(xpd = NA)
  rect(-1.3, 8-i, 0, 9-i, border = "blue")
}
```

true model is GGG

model selection metric

model selected	AIC	GCV	SCV	LOOCV	B0	B1	B2
GGG	0	0	8	72	6	6	7
LGG	28	28	29	7	3	23	24
GLG	37	36	22	8	22	4	24
LLG	0	1	5	1	5	3	33
GGL	34	33	22	8	23	23	3
LGL	1	1	5	2	5	36	2
GLL	0	1	8	1	32	4	4
LLL	0	0	1	0	4	2	3

true model is LGG

model selection metric

model selected	AIC	GCV	SCV	LOOCV	B0	B1	B2
GGG	0	0	1	4	0	4	4
LGG	89	90	82	87	94	32	33
GLG	5	4	6	3	1	3	24
LLG	0	1	2	1	1	1	33
GGL	5	4	6	3	1	24	4
LGL	0	0	2	1	1	34	1
GLL	0	0	1	0	1	1	1
LLL	0	0	0	0	0	1	1

true model is GLG

model selection metric

model selected	AIC	GCV	SCV	LOOCV	B0	B1	B2
GGG	0	0	2	11	12	0	11
LGG	10	10	12	6	0	2	26
GLG	80	78	49	68	26	70	25
LLG	1	2	11	5	3	5	33
GGL	7	7	6	4	24	2	0
LGL	0	1	2	1	1	4	1
GLL	1	2	16	4	31	12	2
LLL	0	0	1	1	3	4	1

true model is LLG

model selection metric

model selected	AIC	GCV	SCV	LOOCV	B0	B1	B2
GGG	0	0	1	3	0	0	5
LGG	75	73	63	70	72	7	31
GLG	8	6	11	6	1	32	25
LLG	13	17	15	17	21	14	35
GGL	4	3	5	3	1	5	1
LGL	0	0	2	1	1	9	1
GLL	0	0	4	1	2	21	2
LLL	0	0	0	0	1	11	1

true model is GGL

model selection metric

model selected	AIC	GCV	SCV	LOOCV	B0	B1	B2
GGG	0	0	2	11	12	11	0
LGG	9	9	12	6	1	25	2
GLG	8	8	5	5	24	1	2
LLG	0	0	2	1	1	2	3
GGL	81	80	50	69	26	26	70
LGL	1	2	11	4	3	32	5
GLL	1	2	16	4	31	2	12
LLL	0	0	2	1	3	1	5

true model is LGL

model selection metric

model selected	AIC	GCV	SCV	LOOCV	B0	B1	B2
GGG	0	0	1	3	0	5	0
LGG	75	73	65	71	73	31	8
GLG	4	3	5	3	1	1	5
LLG	0	0	1	1	1	1	8
GGL	8	6	10	6	1	26	34
LGL	13	17	15	17	20	34	13
GLL	0	0	3	1	2	1	20
LLL	0	0	0	0	1	1	11

true model is GLL

model selected	AIC	GCV	SCV	LOOCV	B0	B1	B2
GGG	0	0	1	6	11	1	1
LGG	9	9	10	6	0	2	2
GLG	20	18	10	13	25	13	2
LLG	1	2	4	3	1	10	4
GGL	20	18	10	13	26	2	13
LGL	1	2	5	3	1	4	11
GLL	49	51	52	53	33	57	58
LLL	0	1	8	3	3	10	10

true model is LLL

model selected	AIC	GCV	SCV	LOOCV	B0	B1	B2
GGG	0	0	0	2	1	1	0
LGG	69	65	55	63	61	8	8
GLG	7	5	7	5	2	16	6
LLG	5	8	8	8	11	14	10
GGL	6	5	7	4	2	5	17
LGL	6	8	8	8	11	9	14
GLL	0	1	8	1	2	31	30
LLL	6	9	6	9	12	15	15

The results of the previous tables must be taken with a grain of salt, as there are frequently 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.

Some patterns clearly emerge.

- The AIC and GCV metrics *never* select Model 1, even when it is the actual model.

- There are several occasions where the model/bandwidth combination with the smallest RMSE is not the "correct" model.

3.1 Bandwidth Size

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.

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.