A Monte Carlo Investigation of Locally Weighted Regression

Aaron Swoboda and Sam Carruthers

November 2, 2012

1 Research Overview

What do we want to know about LWR? The goal of this work is to address the following questions:

- If there is no spatial relationship, will LWR default back to global OLS?
- Are there systematic differences in the bandwidth size selected by different techniques? How do LOOCV, Standardaized CV, Generalize CV, and the AICc compare?
- Which CV strategy tends to result in the most accurate coefficient estimates?
- What sort of spatial variation in the coefficients is necessary relative to the error to need LWR?
- What happens when the LWR model is misspecified?

2 Simulation Overview

This article simulates data under known conditions and examines the results of Locally Weighted Regression to better understand the accuracy and reliability of the technique. To do this we replicated simulations under multiple parameter combinations and ran some simulations, varying the sample size of the data set, the standard deviation of the error term in the model and the degree of spatial variation in the model coefficients.

Each simulation was conducted as follows:

1. Determine the number of simulation replications and a set of values for each of the four simulation parameters:
   (a) sample size
   (b) variance in model error term
   (c) degree of spatial variation in β_1
   (d) degree of spatial variation in β_2
2. Select a value for each parameter from the set of values.
3. Generate the data according to the chosen parameters.
4. Select a bandwidth for the Locally Weighted Regression (the number of observations to receive positive weights in the regression equation) from the set of all acceptable bandwidths (ranging from 5 to the sample size).

5. Estimate a Locally Weighted Regression model using the selected bandwidth for each observation in the dataset.
6. Calculate a number of metrics for the given bandwidth (cross validation scores, pseudo R^2, Root Mean Squared Errors, etc.).
7. Return to Step 4 and choose another bandwidth
8. Repeat Steps 4) - 7) until all bandwidths have been implemented.
9. Replicate the simulations in Steps 3) - 8) until the desired number of replications has been reached.
10. Repeat Steps 2) - 9) until all combinations of simulation parameter values have been achieved.

We chose the following sets of simulation parameter values and replicated each simulation 100 times (for a total of 2.4×10^4 different simulated data sets and simulations.[1]):

- sample size $\{50, 200, 500, 1000\}$
- variance of the error term $\{2^2, 4^2, 6^2\}$
- degree of spatial variation in β_1 $\{0, .1, .2, .3\}$
- degree of spatial variation in β_2 $\{0, .1, .2, .3\}$

We kept track of the following model performance metrics, the pseudo R^2 of the model results, the correlation between the $\hat{\beta}$ and the true β, the percent of the observations for which we can reject the null hypothesis that $\hat{\beta} = \beta$, cross validation scores (leave one out, generalized, and standardized according to Paez), lastly the AIC score.

2.1 Data Generation Process

The Data Generation Process (DGP) is a modified version of a two variable linear DGP. In particular, rather than,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + error, \tag{1}$$

where β_0, β_1, and β_2 are constants, and $error \sim n(0, \sigma^2)$, we construct a dataset according to Equation (2).

$$Y = \beta_0 + \beta_1(location) X_1 + \beta_2(location) X_2 + error \tag{2}$$

The difference is that β_1 and β_2 are a function of location rather than constant. Each observation is located within a geographic coordinate system $(east, north)$ where both $east$ and $north$ values are $\sim u(0, 10)$. The functions determining β_1 and β_2 are :

$$\beta_1(east, north) = 1 + Bsv_1 * north - 5 * Bsv_1 \tag{3}$$
$$\beta_2(east, north) = 1 + Bsv_2 * east - 5 * Bsv_2 \tag{4}$$

The above equations show that when Bsv_i is 0, the βs used to generate the dependent variable are constant across space. However, non-zero values of Bsv_i imply that the marginal impact of the independent variables will differ over space. The βs used in our simulations are visualized in Figure 2.1.

Given the different combinations of coefficient spatial variation parameters, we have data generation processes in which:

1. neither coefficient varies over space ($Bsv_1 = 0$ & $Bsv_2 = 0$)
2. both coefficients vary over space ($Bsv_1 \neq 0$ & $Bsv_2 \neq 0$)

[1] 100 replications, 5 different sample sizes, 3 different error term variances, and 4 different degrees of spatial variation for both β_1 and β_2.

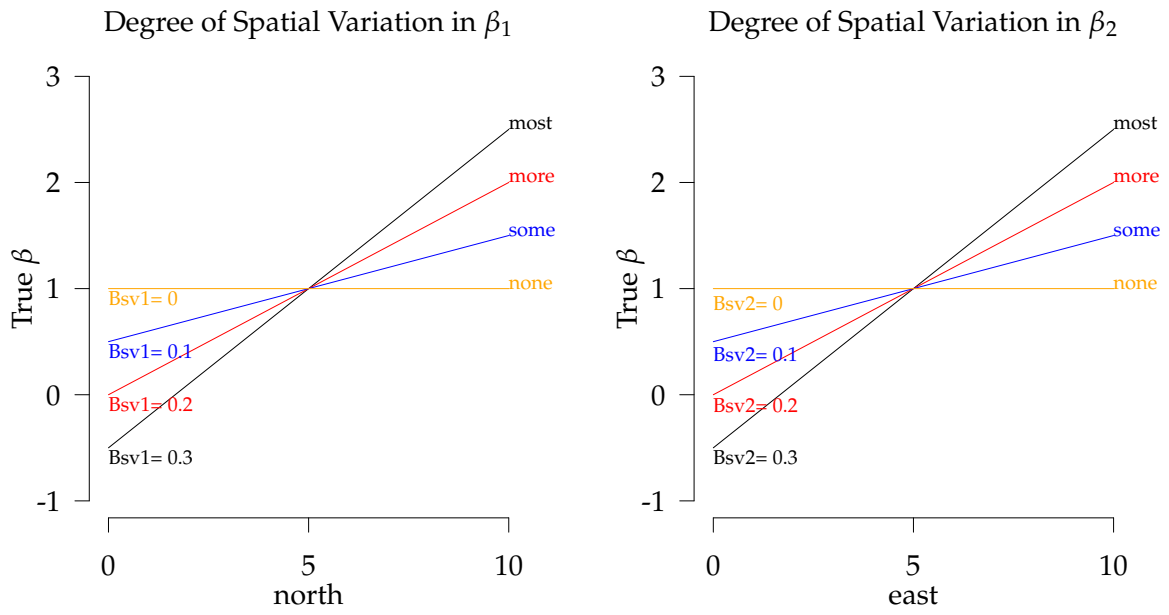

Figure 1: The above images depict the different degrees of spatial variation in our DGP coefficients. Horizontal lines indicate that the coefficient is constant over space, while the steeper sloping lines indicate DGPs with higher degrees of spatial variation.

3. only one coefficient varies over space ($Bsv_1 = 0$ & $Bsv_2 \neq 0$ OR $Bsv_1 \neq 0$ & $Bsv_2 = 0$)

One of the goals of our research is to better understand the results of Locally Weighted Regression in the presence of these various underlying DGPs.

3 Locally Weighted Regression Description

After generating the data, we implemented Locally Weighted Regression with the data and calculated numerous diagnostics in order to measure the performance of the regression technique and answer our research questions.

Locally Weighted Regression (LWR) is an estimation strategy allowing non-stationary model parameters. Specifically, a vector of regression parameters is estimated using Equation (5) for *each location within the dataset*,

$$\hat{\beta}_{location_i} = (X^T W_{location_i} X)^{-1} X^T W_{location_i} Y, \tag{5}$$

where X is the standard n x m data matrix, Y the n x 1 vector of dependent variable values, and $W_{location_i}$ is an n x n weights matrix. We construct the weights matrix for a given $location_i$ to give positive weights to the k-nearest data points to $location_i$, with weights $\in [0, 1]$ and inversely related to distance between data observations and $location_i$. Specifically, we create the weights matrix *for* $location_i$ with zeros on the off-diagonal and calculate the jjth diagonal element as,

$$w_{jj} = \begin{cases} \left[1 - \left(\frac{d_{ij}}{d_{ik}}\right)^2\right]^2 & \text{if } d_{ij} \leq d_{ik} \\ 0 & \text{if } d_{ij} > d_{ik} \end{cases} \tag{6}$$

where d_{ij} is the distance between observations j and $location_i$, and d_{ik} is the distance to the kth nearest observation to observation i. Thus, one of the additional parameters necessary for the implementation of LWR is selecting k, the local regression bandwidth.

Theory does not provide guidance as to how many observations should receive positive weights in the local regression and must be determined by the researcher for the problem at hand. Typically, the k parameter is determined by implementing LWr with several different bandwidths and then selecting the k value that minimizes a model performance metric (usally a cross-validation score). This research aims to systematically compare the performance of four different cross-validation metrics used in LWR research under different, but known, data generation processes. Does choosing the LWR bandwidth through these four strategies yield similar results? If there are differences, are there patterns in how they are different?

3.1 Bandwidth Selection Metrics

We compare the performance of four different bandwidth selection metrics. In particular, we examine:

1. Leave-One-Out Cross-Validation
2. Generalized Cross-Validation
3. Standardized Cross-Validation
4. Akaike Information Criterion

3.1.1 Leave-One Out Cross-Validation

$$\sum (y - \hat{y}_{-i})^2 \tag{7}$$

3.1.2 Generalized Cross-Validation Score

$$n * \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{(n - v_1)^2}, \tag{8}$$

where y_i is the dependent variable value, $\hat{y}_i$ is the predicted dependent variable value for observation i, and v_1 is the "effective number of model parameters."[2] In an LWR model, the number of parameters to be estimated is no longer equal to the number of variables included because we allow the regression coefficients to vary over space. The GCV score calculates the "effective" number of model parameters, v_1, and penalizes the model for increasing the number of parameters without sufficient reduction in model accuracy. Taking the square root of Equation (8) and rearranging yields,

$$\sqrt{GCV} = \sqrt{\frac{n}{n - v_1}} \sqrt{\frac{\text{Sum of Squared Residuals}}{n - v_1}}, \tag{9}$$

which approaches $\hat{\sigma}$ as v_1 approaches m for large n. Henceforth, throughout the paper we report the square root of (8) because of its similarity to $\hat{\sigma}$.

3.1.3 Row Standardized Cross-Validation

Something about Paez, who wanted a CV score that was more robust to outliers.

$$\frac{\sum (y - y_{-i})^2}{\sum y} \tag{10}$$

[2] v_1 =tr(**S**), where the matrix **S** is the "hat matrix" which maps y onto $\hat{y}$,

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

and each row of **S**, r_i is given by:

$$r_i = X_i (X^T W(location_i) X)^{-1} X^T W(location_i).$$

3.1.4 Akaike Information Criterion

$$2 * n * ln(\hat{\sigma}) + n * ln(2 * \pi) + n * \frac{n + v_1}{n - 2 - v_1} \tag{11}$$

4 Which Bandwidths Do Selection Metrics Suggest?

In this section we report on the simulation results. Specifically, we present data on the bandwidths selected by the different metrics under the different simulation conditions. The Cross Validation (CV), Generalized Cross Validation (GCV), and Akaike Information Criterion (AICc) metrics suggest strikingly similar bandwidths, while the Standardized Cross Validation metric sometimes chooses different bandwidths.

4.1 Overall

Table 1 displays the basic summary statistics of the bandwidths selected by the four different metrics across all of our simulations. Comparing the bandwidths at this level of aggregation is of limited use because we expect the selected bandwidth to vary based on the simulation conditions. For instance, the bandwidth is constrained to be smaller than the sample size of the dataset, and we expect the bandwidth selected to be a function of the degree of spatial variation in the underlying data generation process. However, the table begins to show some interesting results.

	min	Q1	median	mean	Q3	max	sd
CV	10	40	70	112	135	999	134
GCV	5	40	70	111	135	999	134
SCV	10	45	90	162	235	955	149
AICc	15	45	80	121	145	999	134

Table 1: Summary statistics for the bandwidths selected by each metric in our simulations.

4.2 Bandwidth Distribution by Simulation Sample Size

Table **??** breaks down the results from Table 1 by simulation sample size. The differences between the bandwidths selected by the Standardized Cross-Validation (SCV) metric and the other three become starker with the larger sample sizes. In general, the bandwidths selected via SCV tend to be more tightly clustered than the other three metrics. The SCV metric has the smallest standard deviation of selected bandwidths across all sample sizes. F-tests allow us to reject the null hypothesis that the variances of the selected bandwidths are equal between the SCV metric and the AICc metric (the metric with the second smallest standard deviation) for all sample sizes but 50.[3]

```
## [1] "Sample Size = 50"
##      min Q1 median mean Q3 max sd
## CV    10 20     30   31 40  49 12
## GCV    5 20     30   30 40  49 12
## SCV   10 25     30   29 35  49  8
## AICc  15 30     35   37 45  49  9
## [1] "Sample Size = 100"
##      min Q1 median mean Q3 max sd
## CV    10 30     45   50 65  99 24
```

[3] $P(F_{99,99} > \frac{22^2}{13^2} = 2.86) < .001$

```
## GCV    5 30     45   49 65  99 24
## SCV   15 45     50   51 60  99 13
## AICc  20 40     55   58 71  99 22
## [1] "Sample Size = 200"
##      min Q1 median mean  Q3 max sd
## CV    20 50     70   81 100 199 46
## GCV   15 45     70   80 100 199 46
## SCV   40 80     90   93 105 199 21
## AICc  30 55     80   90 110 199 44
## [1] "Sample Size = 500"
##      min  Q1 median mean  Q3 max  sd
## CV    30  85    120  151 170 499 107
## GCV   35  85    120  151 170 499 107
## SCV  100 190    215  217 235 480  38
## AICc  45  95    130  161 180 499 105
## [1] "Sample Size = 1000"
##      min  Q1 median mean  Q3 max  sd
## CV    55 130    185  246 260 999 210
## GCV   45 130    185  246 260 999 210
## SCV  255 380    420  421 455 955  62
## AICc  65 140    195  257 270 999 209
```

Figure 4.2 visually presents the bandwidths selected by the different metrics and sample sizes of the simulations. Notice that the distributions of selected bandwidths are similar for the CV, GCV, and AICc metrics, while the SCV metric distribution stands out, especially at higher bandwidths. Additionally, note that most distributions have a cluster of selected bandwidths near the sample size. Given that one simulation parameterization included no spatial variation within the data generation coefficients, it is promising to see a cluster of large bandwidths (the more data that are considered to be "local," the closer the model is to Ordinary Least Squares regression).

4.3 Bandwidths by Degree of Coefficient Variation

We now proceed to examine the distributions of bandwidths by degree of spatial variation in the model coefficients. That is, let's take a single image from Figure 4.2 (Sample Size = 50) and decompose it into the 16 different plots representing the different combinations of spatial variation in the DGP coefficients (4 different levels of spatial variation for both β_1 and $\beta_2 = 16$ combinations).

Figures 4.3 and 4.3 display these results for sample sizes of 50 and 1000. The images are striking. The images are arranged with results of simulations containing no spatial variation in DGP coefficients in the upper left part of the page and simulations containing the most spatial variation in both DGP coefficients in the lower right part of the page. For both the small sample size (50) and the large sample size (1000), the largest possible bandwidth tends to be selected when there is no spatial variation in the DGP coefficients. As spatail variation in the DGP coefficients increases (moving down or to the right, or both), all metrics tend to select smaller bandwidths. The changes in bandwith size is most dramatic for the simulations with sample size of 1000 observations. Whereas the decrease in average bandwidth size is smooth as we moved from the upper-left to the lower-right in Figure 4.3, the average bandwidth decreases substantially as soon as we introduce *any* spatial variation in the DGP coefficients in Figure 4.3.

4.4 Bandwidth and Error Term Variances

None of the previous analysis has examined the role of the Data Generation Process error term variance. We ran each simulation combination of sample size, degree of spatial variation in β_1 and degree of spatial variation in β_2 with three different DGP error term variances, $\sigma^2 \in \{2^2, 4^2, 6^2\}$. This section

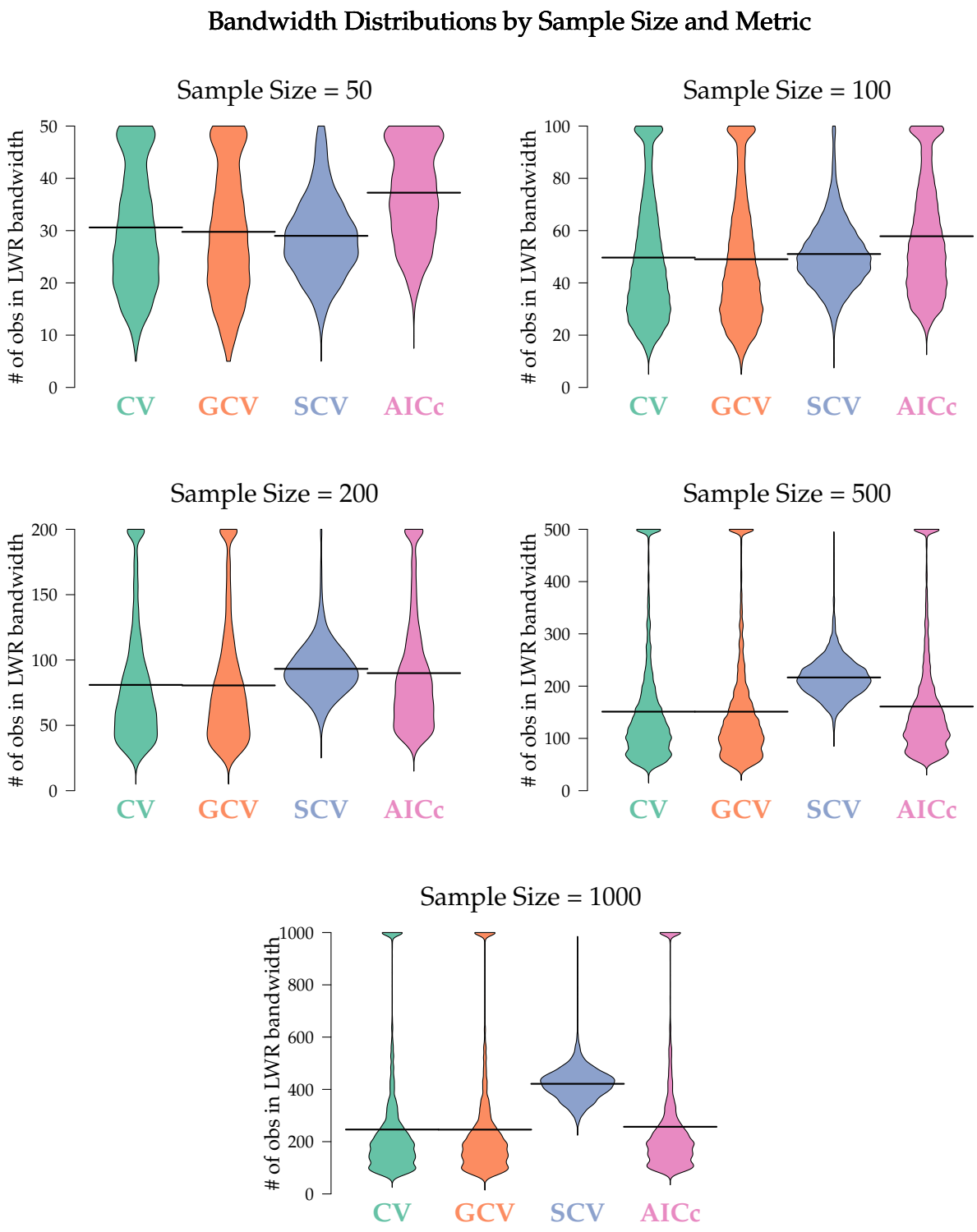

Figure 2: The "beanplots" above show the distribution of bandwidths chosen by each of the four metrics across all simulation sample sizes. Wider areas in the graph represent more data. Note that the data for the Standardized Cross Validation metric (SCV) tend to be more tightly clustered than the others. The horizontal black lines show the mean bandwidth for the given metric.

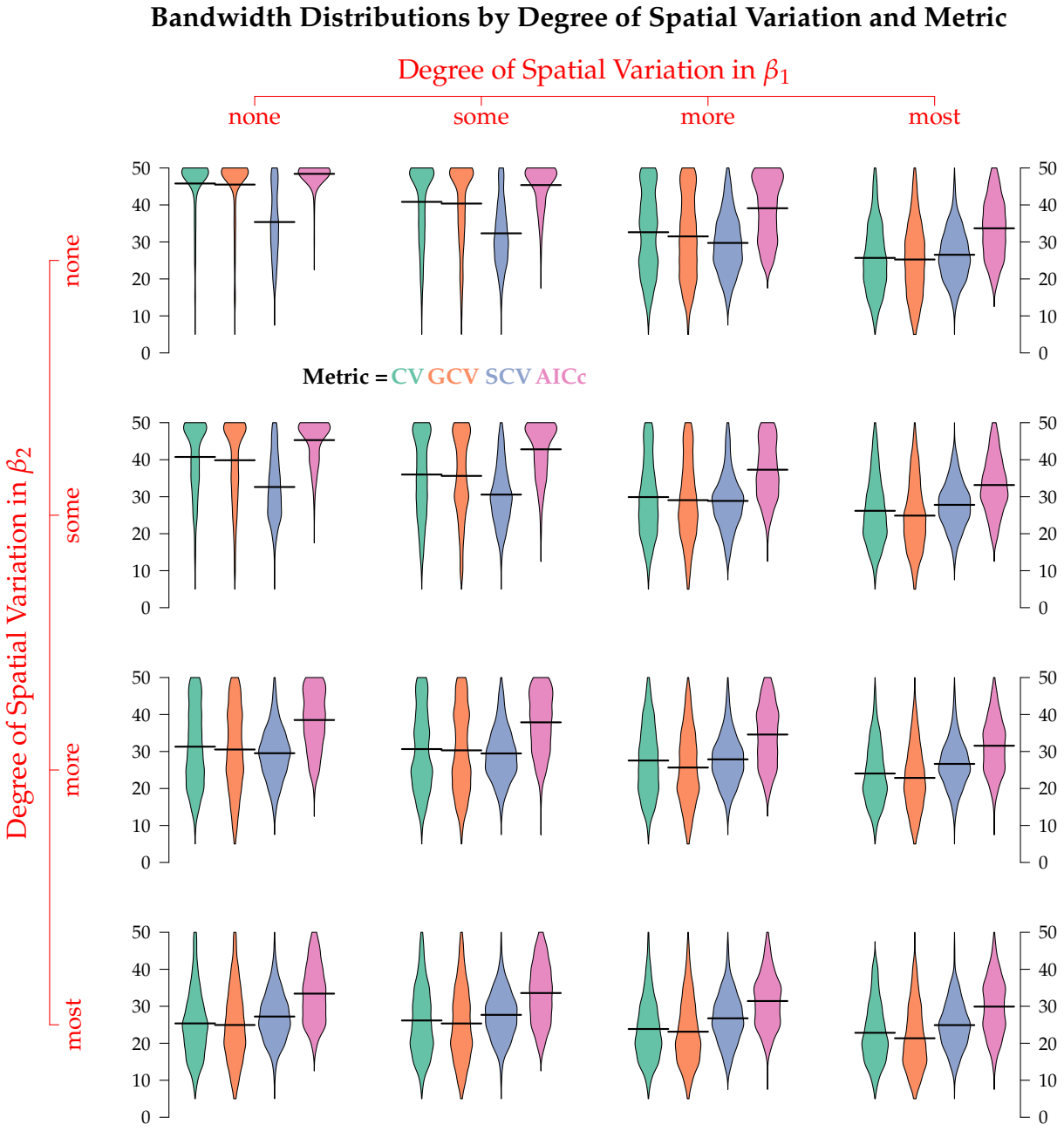

Figure 3: These images display "beanplots" of the bandwidths selected by each metric for each combination of β_1 and β_2 spatial variation for a sample size of 50 data points. Note how the distributions change from the upper left (no spatial variation in either coeffient) to the lower right (the most spatial variation in both coefficients). The size of the selected bandwidth tends to be inversely related to the degree of spatial variation in the coefficients.

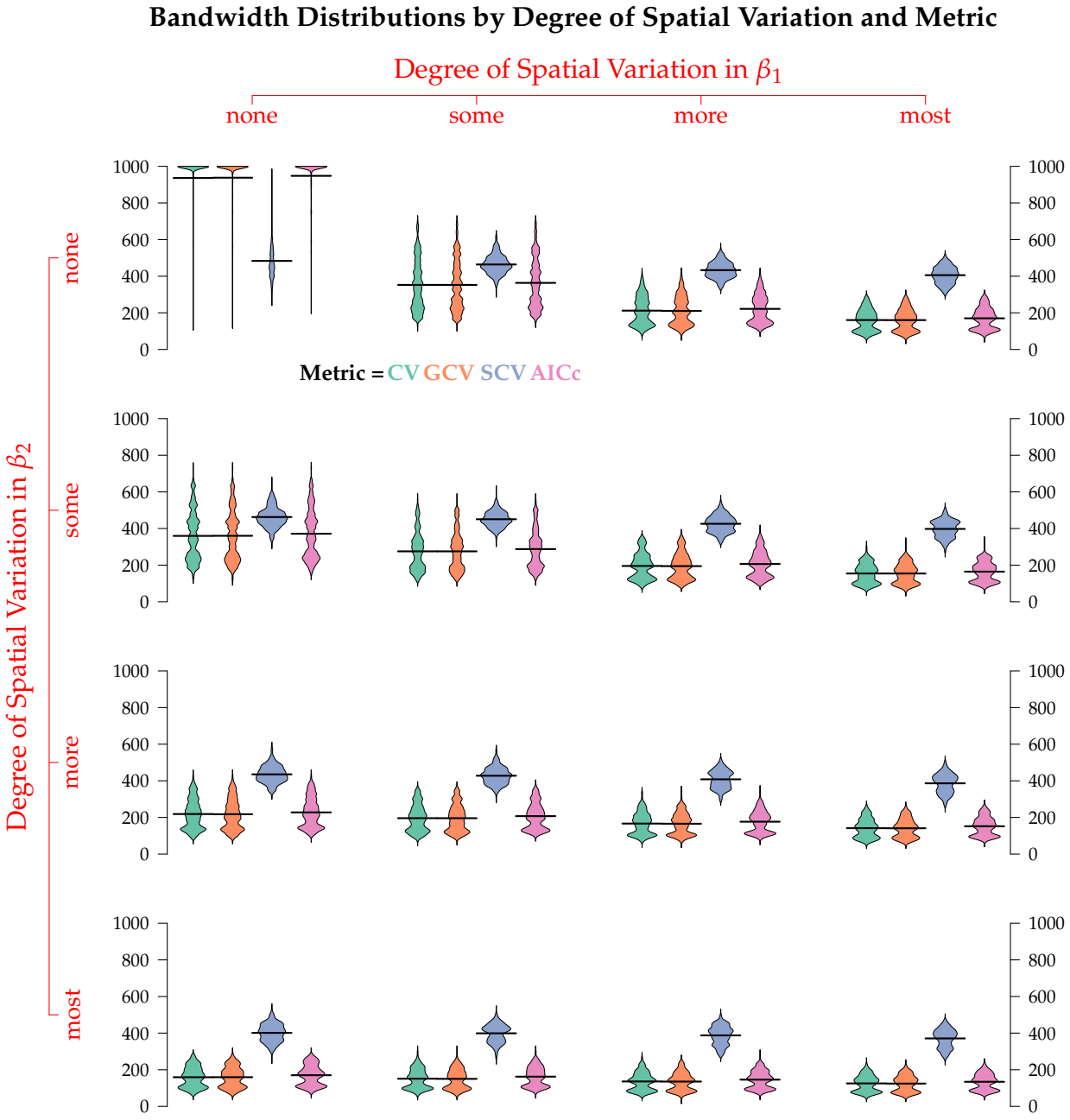

Figure 4: These images display "beanplots" of the bandwidths selected by each metric for each combination of β_1 and β_2 spatial variation for a sample size of 1000 data points. Note how the distributions change from the upper left (no spatial variation in either coeffient) to the lower right (the most spatial variation in both coefficients). The size of the selected bandwidth tends to be inversely related to the degree of spatial variation in the coefficients.

describes the bandwidths selected by the four different selection metrics across these differing DGPs. Larger error term variances serve to "hide" the underlying model in noise. We therefore hypothesize that, all other things equal, larger error term variances will lead to higher bandwidths selected, as the Locally Weighted Regression will have a more difficult time "finding" spatial variation in the regression coefficients, and it will therefore select a larger bandwidth.

Figures 4.4 and **??** are similar to Figures 4.3 and 4.3, with a few important differences. First, rather than plotting the entire bandwidth distribution density for a given simulation scenario, we make a line plot showing the middle 90 percent of the distribution and denote the mean selected bandwidth. This decision helps to fit the large amount of data into Figures 4.4 and **??**.

5 How Accurate are the LWR Coefficient Estimates?

Rather than just looking at the bandwidths selected, researchers are probably more interested in the accuracy of the model predictions, specifically with regard to the model coefficients. In particular, does LWR tend to overfit the data by choosing small bandwidths and spurious coefficients? In this section we compare the estimated model coefficients to the true model coefficients to better understand the reliability of the LWR procedure.

6 What Happens When the Model is Misspecified?

In previous sections we assumed that the model to be estimated using LWR was properly specified. That is, bothe variables (X_1 and X_2) are included and their coefficients are allowed to vary over space to reflect the true data generation process. This section relaxes the assumption of a perfectly specified model and omits one variable in the regression. Our new regression equation becomes:

$$y = \alpha(location) + \beta_1(location)X_1 + error \tag{12}$$

An important question to consider in these circumstances is, "What happens when the omitted variable had a spatially varying coefficient, but the included variable coefficients are stationary?" Does LWR choose a large bandwidth and reflect the stationarity of the included model parameters? Does LWR select a small bandwidth and estimate spatially varying intercept terms? If so, what are the impacts on our estimates of the stationary parameter?

7 Simulation Code

This section includes the code used to run our simulations.

The Data Generation Process is achieved using the `DataGen` function, the code for which is given below.

```
source("../SimFunctions.R")
DataGen

## function (sample.size, error.sd, B1.spatial.var, B2.spatial.var)
## {
##     n = sample.size
##     east = runif(sample.size) * 10
##     north = runif(sample.size) * 10
##     indep.var1 = runif(sample.size) * 10
##     indep.var2 = runif(sample.size) * 10
##     trueB0 = 0
```

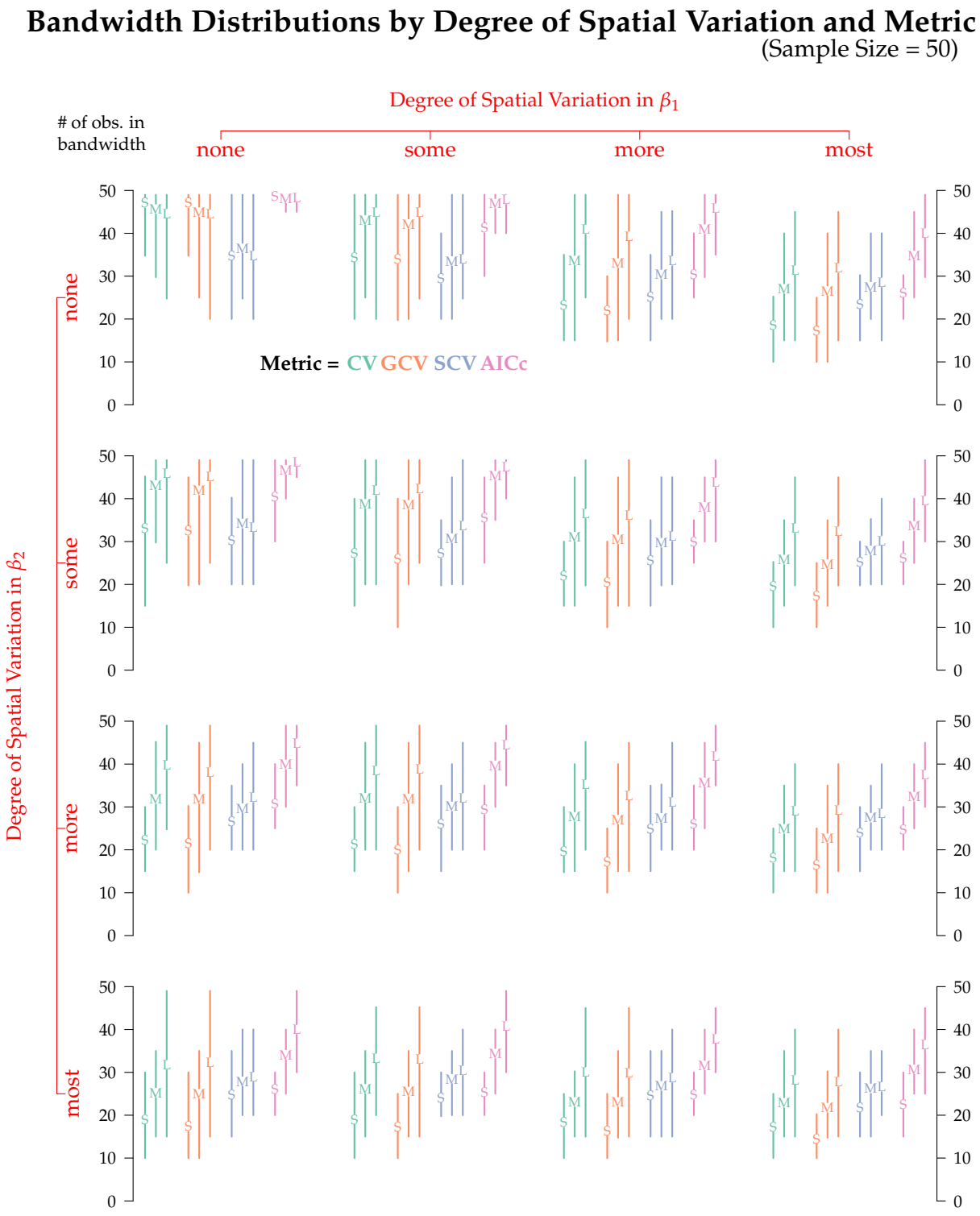

Figure 5: These images display the bandwidths selected by each metric for each combination of β_1 and β_2 spatial variation and DGP error variance for a sample size of 50 data points. Note how the distributions change from the upper left (no spatial variation in either coeffient) to the lower right (the most spatial variation in both coefficients). The size of the selected bandwidth tends to be inversely related to the degree of spatial variation in the coefficients.

```
##     trueB1 = B1.spatial.var * north + 1 - 5 * B1.spatial.var
##     trueB2 = B2.spatial.var * east + 1 - 5 * B2.spatial.var
##     error = rnorm(sample.size, 0, error.sd)
##     dep.var = trueB0 + indep.var1 * trueB1 + indep.var2 * trueB2 +
##         error
##     output = data.frame(dep.var, north, east, indep.var1, indep.var2,
##         trueB0, trueB1, trueB2, error)
##     output
## }
```

The simulations were run and initial metrics calculated in a recent run of `uberScript.R`. It contains the following code:

```
# set our simulation parameters
Replications = 100
sample.size = c(50, 100, 200, 500, 1000)
error.sd = c(2, 4, 6)
B1.spatial.var = c(0, .1, .2, .3)
B2.spatial.var = c(0, .1, .2, .3)
# now march through the different parameter combinations running the simulations
for( i in 1:meta.sim.num) {
  start = Sys.time()
  simRepOut = simulationReplicator(Replications, sim.parameters[i, ], MC = TRUE)
  simOut = simRepReorganizer(simRepOut)

  R2Output[as.character(sim.parameters[i, "sample.size"]),
           as.character(sim.parameters[i, "error.sd"]),
           as.character(sim.parameters[i, "B1.spatial.var"]),
           as.character(sim.parameters[i, "B2.spatial.var"]), , ] = simOut[[1]]

  MetricOutput[as.character(sim.parameters[i, "sample.size"]),
               as.character(sim.parameters[i, "error.sd"]),
               as.character(sim.parameters[i, "B1.spatial.var"]),
               as.character(sim.parameters[i, "B2.spatial.var"]), , , ] = simOut[[2]]
  end = Sys.time()
  print(paste("For loop", i,"of", meta.sim.num))
  print(round(difftime(end, start, units = "m"), 2))
  save(R2Output, MetricOutput, file = "SpecificationSims/uberScriptOutput.RData")
}
```

I'm not going to run that code here (it took almost a month to run on the R Server), but let's load up the results and start to look at them. Or at least come up with some questions to ask of the data and a plan for the future.

```
load("../Data/uberScriptOutput20120919.RData")
dimnames(MetricOutput)

## $ss
## [1] "50"   "100"  "200"  "500"  "1000"
##
## $error.sd
## [1] "2" "4" "6"
```

```
##
## $B1sv
## [1] "0"   "0.1" "0.2" "0.3"
##
## $B2sv
## [1] "0"   "0.1" "0.2" "0.3"
##
## $simNum
##   [1] "1"   "2"   "3"   "4"   "5"   "6"   "7"   "8"   "9"   "10"  "11"  "12"  "13"
##  [14] "14"  "15"  "16"  "17"  "18"  "19"  "20"  "21"  "22"  "23"  "24"  "25"  "26"
##  [27] "27"  "28"  "29"  "30"  "31"  "32"  "33"  "34"  "35"  "36"  "37"  "38"  "39"
##  [40] "40"  "41"  "42"  "43"  "44"  "45"  "46"  "47"  "48"  "49"  "50"  "51"  "52"
##  [53] "53"  "54"  "55"  "56"  "57"  "58"  "59"  "60"  "61"  "62"  "63"  "64"  "65"
##  [66] "66"  "67"  "68"  "69"  "70"  "71"  "72"  "73"  "74"  "75"  "76"  "77"  "78"
##  [79] "79"  "80"  "81"  "82"  "83"  "84"  "85"  "86"  "87"  "88"  "89"  "90"  "91"
##  [92] "92"  "93"  "94"  "95"  "96"  "97"  "98"  "99"  "100"
##
## $optimized
##  [1] "AICc"     "corB0"    "corB1"    "corB2"    "CV"       "GCV"      "R2"
##  [8] "RMSE.B0"  "RMSE.B1"  "RMSE.B2"  "SCV"      "ttest%B0" "ttest%B1" "ttest%B2"
##
## $metric
##  [1] "bandwidths" "B0.cor"     "B1.cor"     "B2.cor"     "B0.RMSE"    "B1.RMSE"
##  [7] "B2.RMSE"    "B0.t.perc"  "B1.t.perc"  "B2.t.perc"  "GCV"        "SCV"
## [13] "CV"         "AICc"       "R2"
```

```
dimnames(R2Output)
```

```
## $ss
## [1] "50"   "100"  "200"  "500"  "1000"
##
## $error.sd
## [1] "2" "4" "6"
##
## $B1sv
## [1] "0"   "0.1" "0.2" "0.3"
##
## $B2sv
## [1] "0"   "0.1" "0.2" "0.3"
##
## $simNum
##   [1] "1"   "2"   "3"   "4"   "5"   "6"   "7"   "8"   "9"   "10"  "11"  "12"  "13"
##  [14] "14"  "15"  "16"  "17"  "18"  "19"  "20"  "21"  "22"  "23"  "24"  "25"  "26"
##  [27] "27"  "28"  "29"  "30"  "31"  "32"  "33"  "34"  "35"  "36"  "37"  "38"  "39"
##  [40] "40"  "41"  "42"  "43"  "44"  "45"  "46"  "47"  "48"  "49"  "50"  "51"  "52"
##  [53] "53"  "54"  "55"  "56"  "57"  "58"  "59"  "60"  "61"  "62"  "63"  "64"  "65"
##  [66] "66"  "67"  "68"  "69"  "70"  "71"  "72"  "73"  "74"  "75"  "76"  "77"  "78"
##  [79] "79"  "80"  "81"  "82"  "83"  "84"  "85"  "86"  "87"  "88"  "89"  "90"  "91"
##  [92] "92"  "93"  "94"  "95"  "96"  "97"  "98"  "99"  "100"
##
## $R2
## [1] "OLS" "LWR"
```