To: Dr. Gregory Macfarlane

From: Dr. Gregory Macfarlane

Subject: Land Value Lab

Date: April 4, 2014

This lab requires students to examine spatial autoregressive models. In R, this is done with the `spdep` library. Additional plotting commands are available with the `ggmaps` and `ggplot2` libraries.

```
library(spdep)
library(ggmap)
library(apsrtable)
source("scripts/apsr_spdep.R")
```

```
## Creating a generic function for 'coef' from package 'stats' in the global environment
```

The data we will use for this analysis are provided on T-Square in the `HomePrices.Rdata` file. These data are evaluations of home prices in the central Atlanta in 2012, obtained from the Fulton County Tax Assessor's Office. The data are stored in a `SpatialPointsDataFrame` object that already contains projection data. These points are plotted in Figure 1.

```
load("./HomePrices.Rdata")
```

The dataset contains the following fields:

`X2012` The home's assessed value in 2012.

`X,Y` The home's coordinates in latitude and longitude.

`INCOME2010` The median income in the home's Census tract according to the 2010 Census.

`PCTWHITE` The percentage of white householders in the home's Census tract, according to the 2010 Census

`NEAREST_ STATION` The nearest MARTA station to the home, by Euclidean distance.[1]

`Acres` The acres of the home's property.

`EffAge` The home's effective age, considering improvements to the structure and plumbing or electrical systems.

`Rmtot` The number of rooms in the home.

`Rmbed` The number of bedrooms in the home.

`Fixbath` The number of bathrooms in the home.

1 Correlation in OLS

An important element of spatial models is a weights matrix, W, that captures the spatial relationship between each point in the dataset. Create a matrix that weights points by the inverse distance between them ($1/d_{ij}$), for the 50 nearest neighbors. Note that because some points are at the same location (condos), you will have to accomodate the fact that d_{ij} could equal zero; also note that the distances are given in kilometers.[2]

[1] Extra points to someone who can use R to efficiently calculate network distance.

[2] For details, see `?nbdists`.

```
# Plot stamen background map
Atlanta <- get_map(bbox(ClassPointsLL), source=stamen, maptype=toner)
# Create ggmap object to project points onto
AtlantaMap <- ggmap(Atlanta, extent="device")
AtlantaMap + geom_point(data = ClassPointsLL@data,
                        aes(x=X, y=Y, color=NEAREST_STATION), alpha=0.8)
```

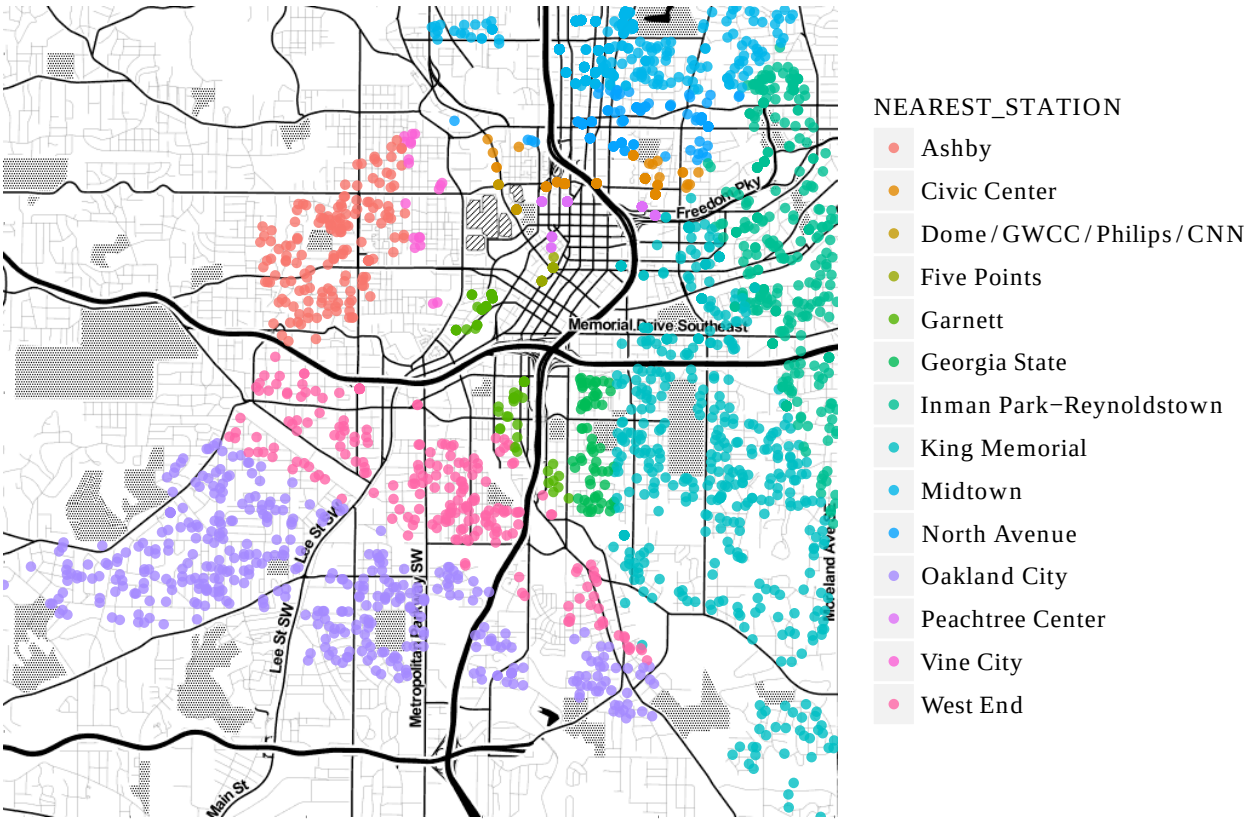

Figure 1: Homes in the analysis dataset, by nearest MARTA station.

```
# Calculate average neighbor distance for each observation
ClassPointsLL$meanNB <- unlist(lapply(dists, mean))
AtlantaMap + geom_point(aes(x=X, y=Y, color=log(meanNB)),
                        data = ClassPointsLL@data)
```

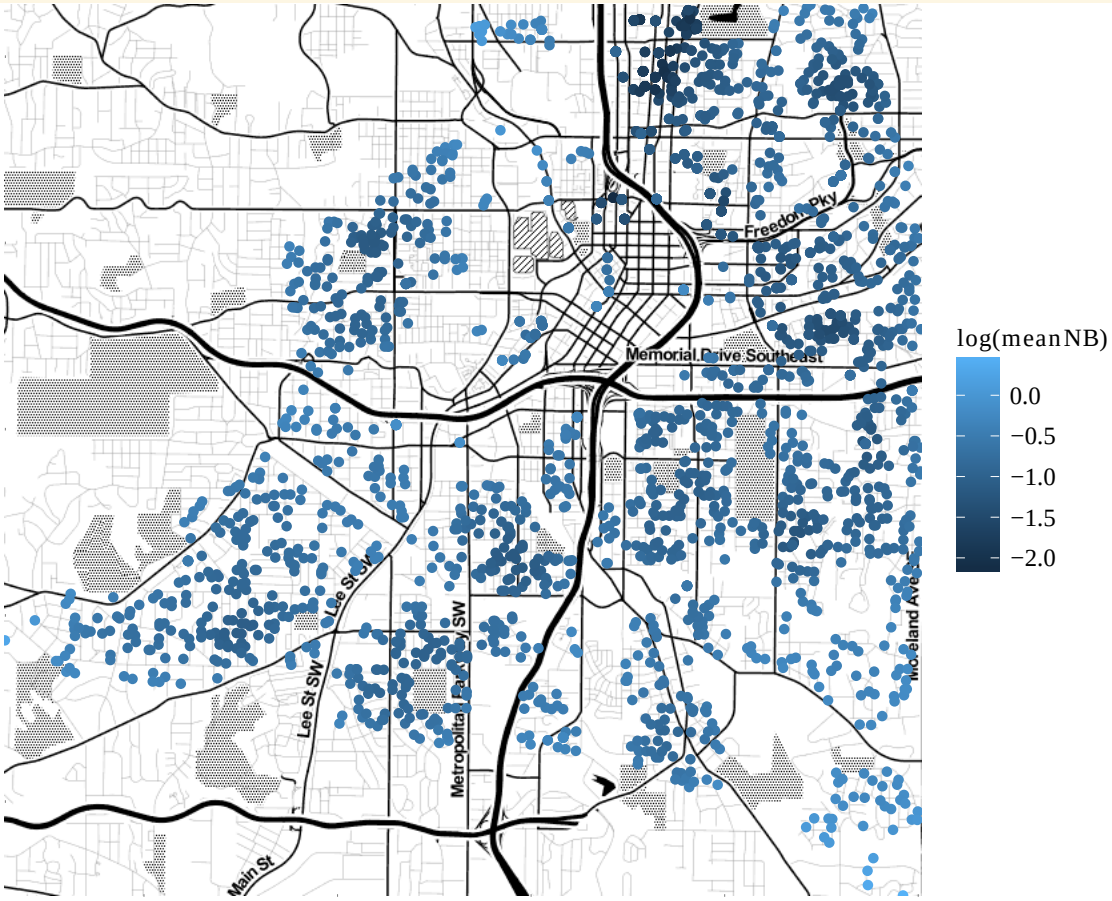

Figure 2: Observations by mean neighbor distance.

```
points.knn <- knn2nb(knearneigh(ClassPointsLL, k=50))

## Warning:  knearneigh:  identical points found

dists  <- lapply(nbdists(points.knn, ClassPointsLL), function(x) x)
dists.inv <- lapply(dists, function(x) 1/(x+0.01))
W <- nb2listw(points.knn, glist=dists.inv, style="W")
```

Provide a map of the points by average neighbor distance. This is given in Figure 2.

Run a least squares regression model ($y = X\beta$) to predict the price[3] of a home given its size, acreage, age, and neighborhood income and racial composition. Comment on the implications of this model for home price. Specifically, what features of a house lead to it having a higher value? Test, using the Moran's I autocorrelation statistic, whether the residuals from this model are spatially autocorrellated. Provide a map of the points by OLS residual.

```
price.ols <- lm(log(X2012) ~ Rmtot + Fixbath + log(INCOME2010) + PCTWHITE+
                log(Acres), data = ClassPointsLL@data)
```

[3]You should use a log transform here.

```
price.mI <- moran.test(price.ols$residuals, W)
price.mI
```

```
##
##  Morans I test under randomisation
##
## data:  price.ols$residuals
## weights: W
##
## Moran I statistic standard deviate = 50.1, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##         2.548e-01        -4.002e-04         2.595e-05
```

We reject (with a p-value of 0) the null hypothesis that there is no spatial correlation, and determine that a spatial model is necessary. The models for this analysis are given in Table 1; according to the OLS model, all our independent variables are significant predictors of home price. Further, all variables have an intuitive sign, with the possible exception of property acreage.

2 Autoregressive Models

Estimate SAR, SEM, and SDM versions of your OLS model. These models are

$$\text{SAR} : y = \rho W y + X\beta + \epsilon \tag{1}$$
$$\text{SEM} : y = X\beta + u, u = \lambda W u + \epsilon \tag{2}$$
$$\text{SDM} : y = \rho W y + X\beta + \gamma W X + \epsilon \tag{3}$$

Which model provides the best statistical fit? Plot the residuals of the SDM model, and comment on their spatial distribution. What happens to the explanatory variables you identified as being statistically significant in your OLS model when you use a spatial autoregressive model? Comment on the implications of this for research and practice.

```
price.sar <- lagsarlm(formula(price.ols), listw=W, data=ClassPointsLL@data)
price.sem <- errorsarlm(formula(price.ols), listw=W, data=ClassPointsLL@data)
price.sdm <- lagsarlm(formula(price.ols), listw=W, type="mixed",
                      data=ClassPointsLL@data)
```

Based on the log-likelihood of the models in Table 1, the SDM fits the data best. Also, given that not all $\gamma, \rho = 0$, we can reject that the OLS, SAR, or SEM models are adequate. Perhaps the most striking conclusion is that home size (represented by the number of rooms) seems to be a marginal predictor of home values, whereas it was highly significant in the OLS model. This could mean that home size is correlated with neighborhood attributes that affect home values such as school quality or crime rate, but that home sizes themselves are unimportant. Another important observation is that property size has a positive direct effect, but a negative indirect effect; it is good to have a large property but bad to have neighbors with large properties. The residuals are plotted in Figure **??**, and it is apparent that the major problems related to spatially correlated residuals have been counted for.

```
# Calculate average neighbor distance for each observation
ClassPointsLL$OLSresid <- price.ols$residuals
AtlantaMap + geom_point(aes(x=X, y=Y,
                            color=cut(OLSresid,
                                      breaks=quantile(OLSresid,
                                                      probs=seq(0,1,0.2)))),
                        data = ClassPointsLL@data) +
  scale_color_brewer(palette="RdBu", type="div", name= "Residual")
```

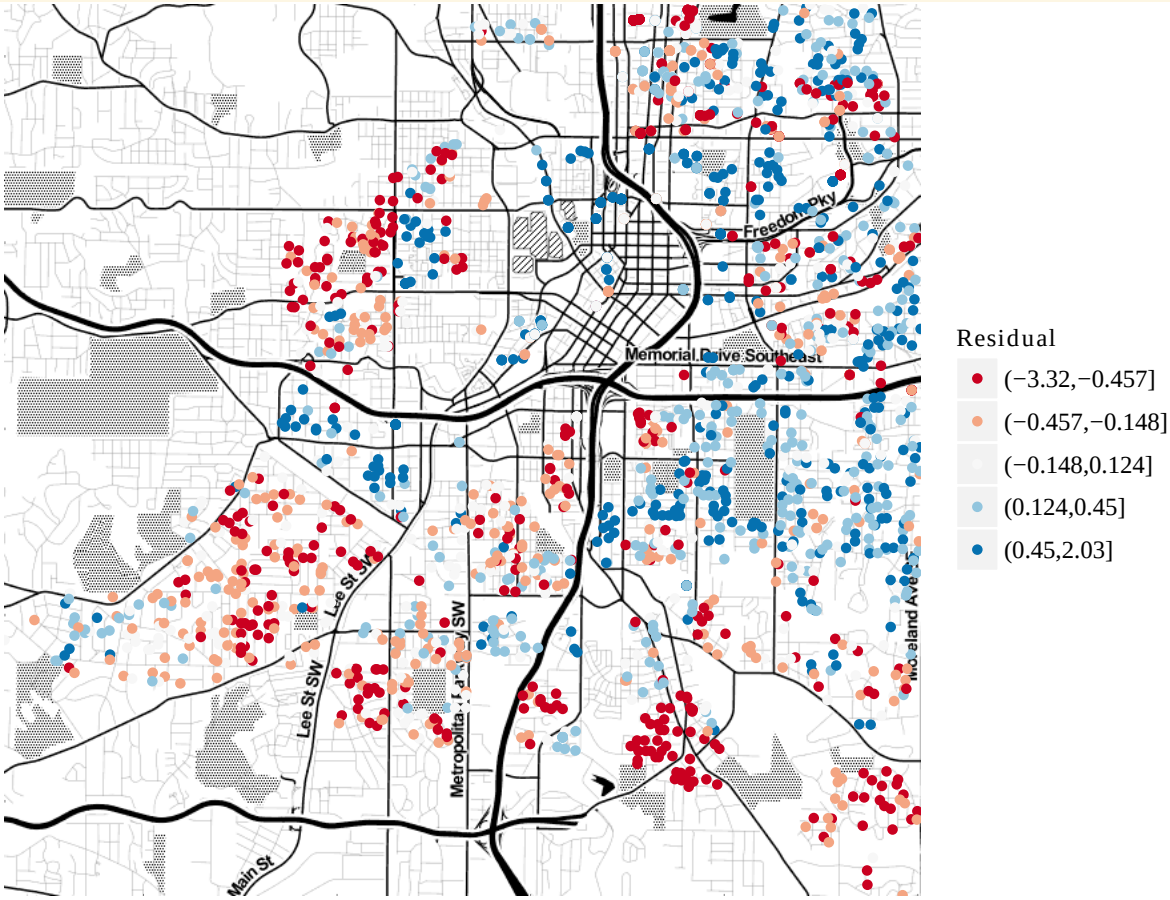

Figure 3: Spatial correlation in OLS residuals.

Table 1: Home Price Models

```
source("scripts/apsr_spdep.R")
apsrtable(price.ols, price.sar, price.sem, price.sdm,
          model.names=c("OLS", "SAR", "SEM", "SDM"),
          Sweave=TRUE, coef.rows=2, stars="default", digits=3)
                    ## Warning:  class of 'x' was discarded
```

	OLS	SAR	SEM	SDM
(Intercept)	5.575^{***}	0.940^{*}	9.379^{***}	0.342
	(0.429)	(0.374)	(0.631)	(0.485)
Rmtot	0.027^{**}	$0.014^{\dagger}$	0.005	0.010
	(0.010)	(0.008)	(0.008)	(0.008)
Fixbath	0.449^{***}	0.384^{***}	0.374^{***}	0.372^{***}
	(0.024)	(0.019)	(0.019)	(0.019)
log(INCOME2010)	0.367^{***}	0.054	0.178^{**}	0.157^{**}
	(0.041)	(0.033)	(0.059)	(0.060)
PCTWHITE	2.080^{***}	0.428^{***}	-0.545^{***}	-0.619^{***}
	(0.073)	(0.072)	(0.160)	(0.163)
log(Acres)	-0.064^{***}	0.036^{**}	0.146^{***}	0.147^{***}
	(0.015)	(0.012)	(0.017)	(0.017)
rho		0.802^{***}		0.708^{***}
		(0.019)		(0.032)
lambda			0.956^{***}	
			(0.009)	
lag.Rmtot				$0.042^{\dagger}$
				(0.024)
lag.Fixbath				-0.039
				(0.076)
lag.log(INCOME2010)				0.009
				(0.079)
lag.PCTWHITE				1.063^{***}
				(0.194)
lag.log(Acres)				-0.237^{***}
				(0.033)
N	2500	2500	2500	2500
AIC	4319.318	3324.959	3313.119	3170.718
$\log(\mathcal{L})$	-2152.659	-1654.480	-1648.560	-1572.359

Standard errors in parentheses

† significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

```
# Calculate average neighbor distance for each observation
ClassPointsLL$SDMresid <- price.sdm$residuals
AtlantaMap + geom_point(aes(x=X, y=Y,
                            color=cut(SDMresid,
                                      breaks=quantile(OLSresid,
                                                      probs=seq(0,1,0.2)))),
                        data = ClassPointsLL@data) +
  scale_color_brewer(palette="RdBu", type="div", name= "SDM Residual")
```

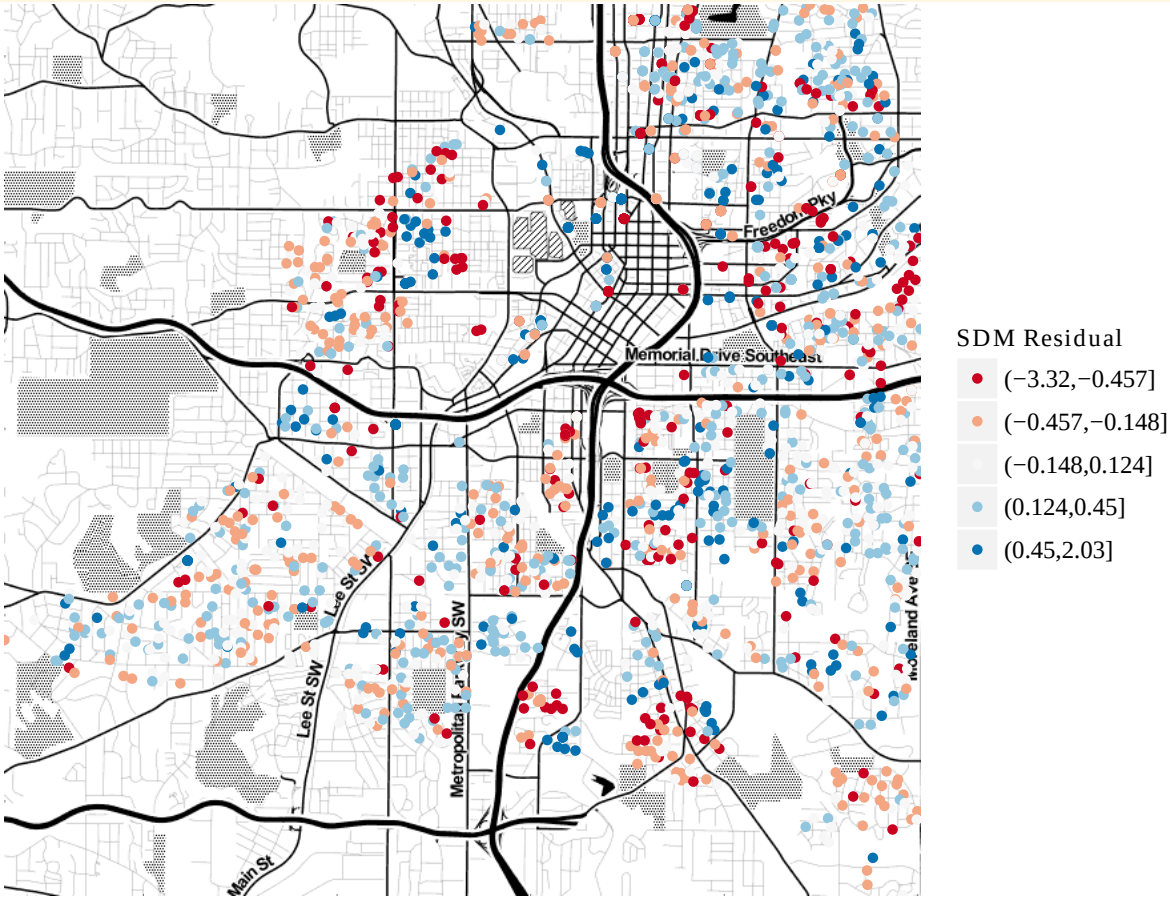

Figure 4: Spatial correlation in SDM residuals.