Ryan Marinelli

Professor Unwin

Spatial Statics for GIS

02 December 2019

Analysis of Agricultural Production in Puerto Rico

Introduction

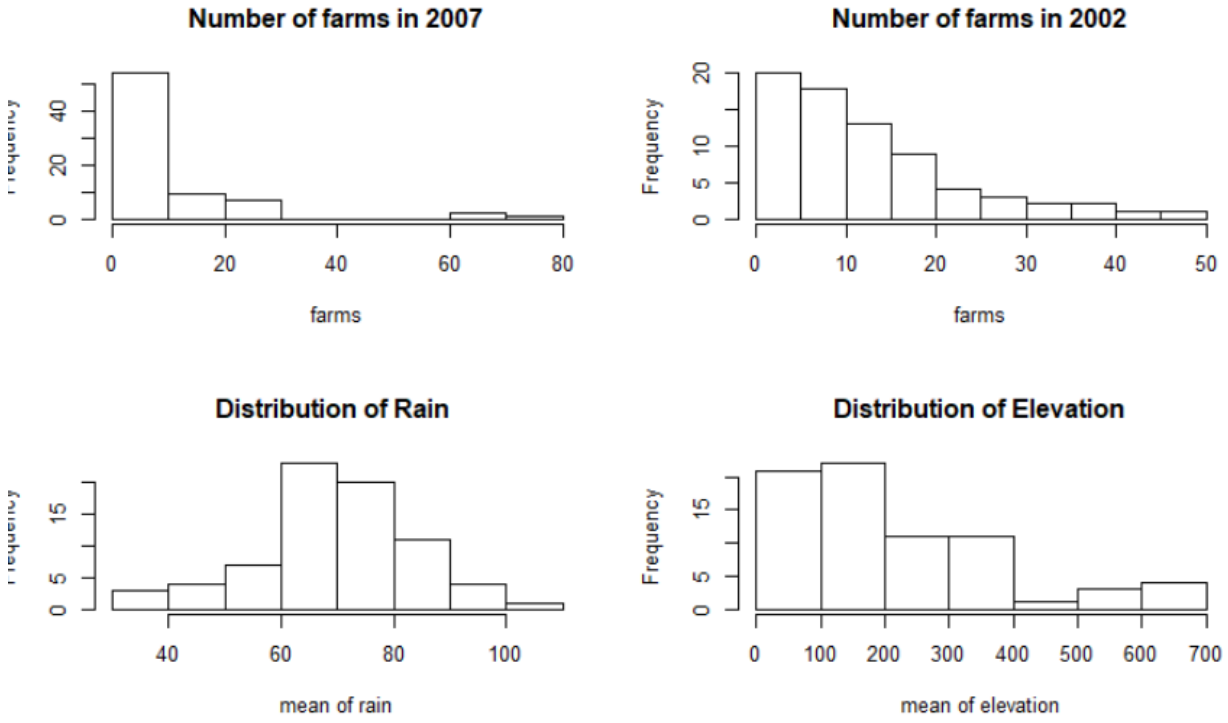
The goal of this research initiative is to review the relationships between agricultural production and spatial features in Puerto Rico. The data consists of information on the number of farms producing on the municipality level. The data is annual data from 2007 and 2002 is utilizing only farms with irrigation in the sample. An initial review of the data is conducted to observe peculiarities with the data and to provide framing to the research. The most significant portion of this endeavor emphasized estimation procedures. In typical econometric methodology, the standard technique for estimation is Instrumental Variable regression. The goal of this technique is to apply exogenous variables as an instrument to facilitate causal arguments. However, this approach is not optimal when estimating spatial parameters. The issue is spatial autocorrelation. This correlation essentially violates assumptions when conducting modeling. Namely, homoscedasticity is violated, and model parameters are fundamentally correlated with the error term when observing spatial relationships. Given spatial autocorrelation, assumptions are unable to hold, and bias is introduced to estimates. If IV regression were applied, it would be unable to effectively adjust estimates to be robust to this source of error. It is too difficult to select an instrument that would be exogenous. Thus, unique modeling procedures are warranted to minimize bias from this source. In this analysis, a simultaneous autoregressive model is utilized to lessen bias in estimates to study the relationship of agricultural production and the spatial factors of Puerto Rico.

Data

The data used in this study is comprised of agricultural data from 2002 and 2007. Since a time-dimension is present, the data is cross-sectional in nature. The variables of interest are the number of farms in both years on the municipality level, the mean of rain fall, and mean elevation. While standard deviation would have also provided informative analysis, it would likely be problematic to include in the model specification. Since the underlying distributions of standard deviation are likely to be affected by phenomenon that are not generated through spatial processes, bias could be introduced unnecessarily. For instance, consider how construction would affect measurements. To build, it might be nessary to make the ground of specific areas more level which would add more leverage to the tails of the distribution of standard deviation of elevation while also distorting the relationship of the amount of farms as more land is reserved for construction. A similar argument could be made for clearing land for farming as well. Since, there are underlying relationships that would be difficult to control; it is wiser to use the mean for this analysis.

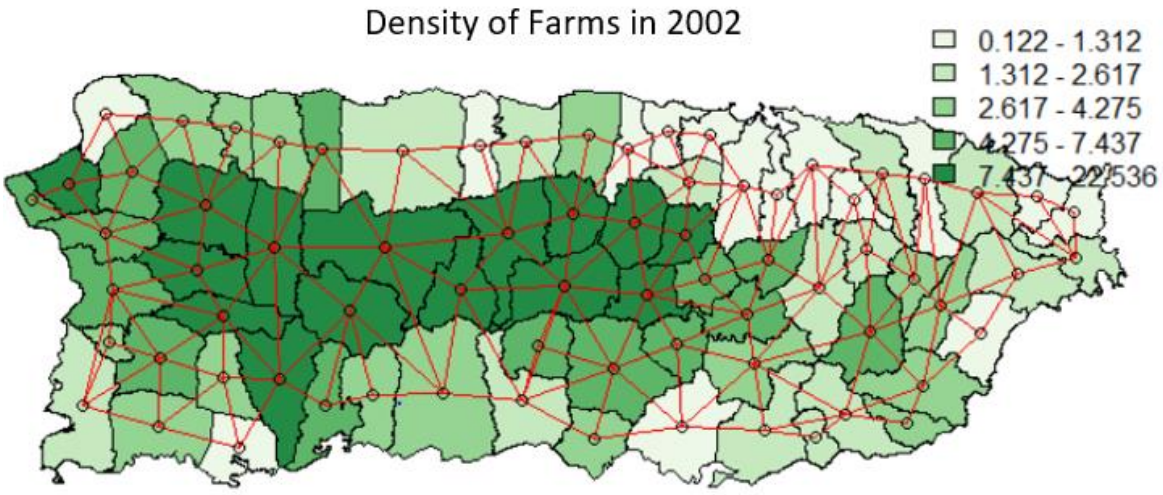
In the below are the distributions for those variables. The number of farms is significantly right skewed. Both distributions are leptokurtic. The kurtosis coefficient associated for the

number of farms in 2002 is 1.1 and 9.5 for the distribution for farms in 2007. This suggests that the normality assumptions should be loosened and should be understood as a weakness that is acknowledged. However, while this would introduce bias, it would not change the consistency of estimation procedure. Thus, it is marginal in how the lack of normality detracts from this analysis generally speaking.



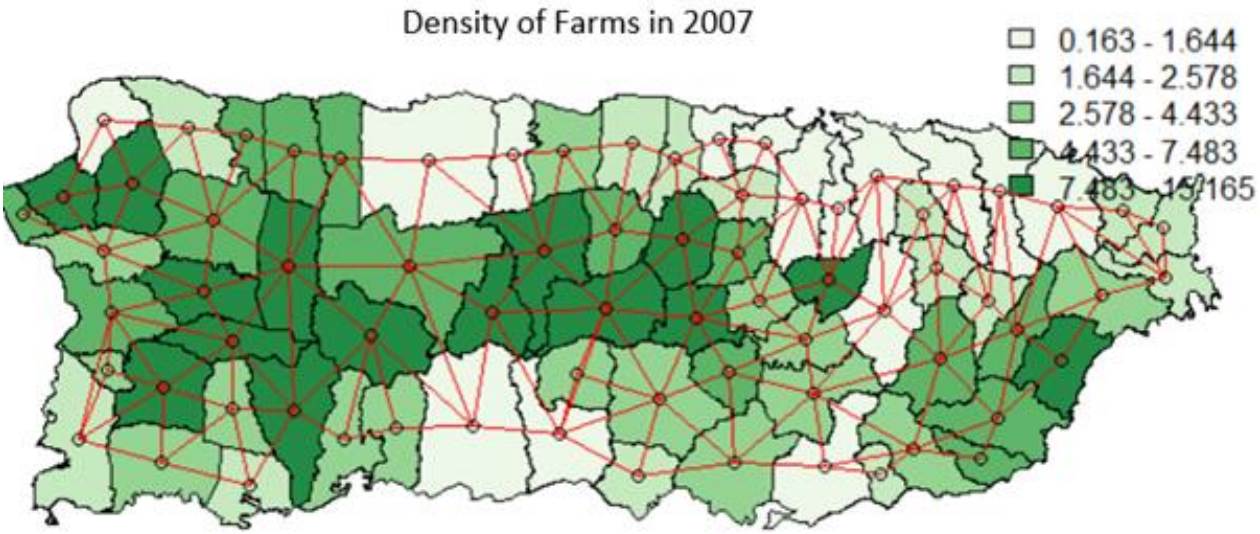
Visual Inspection

The map in the below represents the density of farms in specific regions. The density of the farms is signified by the hue of green associated with the region. The red dots are the centroids of each of the regions. It is apparent that there is considerably more farming in the



center of Puerto Rico and in the western part of the territory. This is likely why the data is significantly skewed. There is also intuitive reasoning that can be made. Firstly, the eastern part of the country is where the capital is located. There are likely more jobs available, so there likely is less farming in exchange for increased industrialization.

In 2007, the pattern in the orginal 2002 plot is largely maintained. However, there are some notable trends. In the southeastern portion of the territory, the density of farms decline. Most of the portions in this quadrant are now a lighter shade of green. There are also numerous other sections that demonstrate the same trend. For instance, in the northwestern area, there were light-green areas that are now white. It appears that this trend of losing farm desity is largely represented overall. This is demonstrated by the difference in mean of the densities. In 2002, the mean density was 4.8 farms per acre. However, in 2007 this fell to 4.5 farms per acre. This difference though was not statistically significant, as a difference of means test did not yield a sufficently small p-value. The differences though are visually dectectable as the differences are likely more pronouced within regions rather than across the data as a whole.



Model Specification

$$\widehat{\text{Number of Farms}} = (\text{I} - \sigma\text{W})(\widehat{\beta_0} + \widehat{\beta_1\text{rain}} + \beta_2\widehat{\text{elevation}} + \widehat{\beta_1\text{city}}) + \sigma\text{W}\widehat{\text{Number of Farms}} + \varepsilon$$

In this specification, a simultaneous autoregressive (SAR) model is utilized. The dependent variable is the number of farms in a particular region. W is a connectivity matrix, and I is an identity matrix. Row is an estimated parameter of spatial autocorrelation with elision representing the error term. Rain is the mean of the rainfall in a particular region in a particular year. Elevation is meters above sea level. City is a factor variable encompassing San Juan, Arecibo, Mayaguez, Ponce, and Caguas. These areas are distributed around Puerto Rico and act as an indicator variable as to represent regional effects.

The reason for the selecting this model has been aforementioned. However, it is nessary to lean upon this model given typical econometric tools are insufficient to encapsulate spatial autocorrelation. This is captured by row and the weighted connectivity matrix.

Results

```
Type: lag
Coefficients: (asymptotic standard errors)
                Estimate Std. Error z value  Pr(>|z|)
(Intercept)    26.671432   7.364364  3.6217 0.0002927
pr.f$rain_mean -0.308766   0.094738 -3.2592 0.0011175
pr.f$elev_mean  0.011231   0.006981  1.6087 0.1076747
admArecibo      3.935048   3.294704  1.1944 0.2323389
admMayaguez     3.410564   2.989521  1.1408 0.2539366
admPonce        1.210054   3.544950  0.3413 0.7328433
admCaguas       3.014206   3.081871  0.9780 0.3280525

Rho: 0.24545, LR test value: 1.9, p-value: 0.16808
Asymptotic standard error: 0.15094
    z-value: 1.6262, p-value: 0.1039
Wald statistic: 2.6446, p-value: 0.1039
```

This first table describes the estimates generated from the model within the 2002 cross-section. Rain has a negative association. This could be detecting flooding as elevation is nearly significant with a positive association as well. In terms of how different cities are performing, none of the other cities seem to be significantly outperforming the capital although all the coefficients are all positive.

```
Type: lag
Coefficients: (asymptotic standard errors)
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    23.5403989  8.7356764  2.6947 0.007044
pr.f$rain_mean -0.2988402  0.1176738 -2.5396 0.011099
pr.f$elev_mean  0.0027326  0.0087409  0.3126 0.754571
admArecibo      9.5095620  4.4123701  2.1552 0.031146
admMayaguez     3.8104814  3.8915307  0.9792 0.327495
admPonce       -1.9106519  4.5359452 -0.4212 0.673591
admCaguas       1.9587689  3.9655239  0.4939 0.621342

Rho: 0.47122, LR test value: 7.1109, p-value: 0.0076615
Asymptotic standard error: 0.12525
    z-value: 3.7622, p-value: 0.00016839
Wald statistic: 14.155, p-value: 0.00016839
```

This second table describes the estimates generated from the model within the 2007 cross-section. It would not be useful to interpret the coefficients in most cases. For instance, the relationship with rain is negative. However, since there is a fractional coefficient, it is not easily understood. The relationship could be detecting the propensity of storms in the region. Likewise,

this could elaborate on the positive coefficient of elevation, although it lacks a significant p-value. In terms of the cities, the analysis is a bit more interesting. The intercept term represents San Juan. It appears that only Arecibo is preforming more strongly. The other cities are not significant in their losses or gains as compared to San Juan.

After reviewing these two cross-sections, it appears the most interesting results is that Arecibo is significantly outperforming San Juan in terms of the number of farms. Also, rain seems to have a generally negative relationship to the number of farms. This could be due to flooding or if rain is plentiful, people may have their own smaller farms that were not counted. However, this remains an interesting finding.

Methodology

This section will provide an essential walkthrough of the code used to produce results and graphics. In conducting this analysis, *Spatial Statics & Geostatics* by Chun & Griffth is learned upon. Through accessing data provided from Sage publishing, it was possible to recover the data that had been abstracted.

```
library(rgdal)
library(classInt)
library(spdep)
library(RColorBrewer)
library(car)
library(spatialreg)
library(e1071)
```

Rgdal, spdep, and spatial are simply spatial packages. E1071 has a function to calculate kurtosis. The other packages were loaded to use styling procedures.

```
setwd("C:/Users/Ryan/Desktop/Fall 2019/Spaital Statistics/PuertoRico_data")
pr<-readOGR("PuertoRico_SPCS.shp")
pr.f <- read.csv(file="PR-farm-data.csv")
pr.nb <- read.gal("PuertoRico.gal")
```

This process is reading data from the ESRI shape file and ID information.

```
plot(pr)
summary(pr)
points <- coordinates(pr)
plot(pr, lwd = .5)
plot(pr.nb, points, add=T, col="red")
farm.den07 <- pr$nofarms_07/pr$area
farm.den02 <- pr$nofarms_02/pr$area

plot(farm.den02)
mean(farm.den02)
```

```
mean(farm.den07)
```

The goal of this code was to conduct exploratory analysis and to make more meaningful variables. It is mostly a guiding process in the analysis.

```
pal.gray <- gray.colors(4)
pal.green <- brewer.pal(4,"Greens")
q5.den <- classIntervals(farm.den07,5,style="quantile")
cols.den <- findColours(q5.den, pal.green)
plot(pr, col=cols.den)
brks.den <- round(q5.den$brks,3)
leg.txt  <- paste(brks.den[-6], brks.den[-1], sep=" - ")
legend("topright", fill=attr(cols.den,"palette"), legend=leg.txt ,bty="n")
plot(pr.nb, points, add=T, col="red")
```

```
pal.gray <- gray.colors(4)
pal.green <- brewer.pal(4,"Greens")
q5.den <- classIntervals(farm.den02,5,style="quantile")
cols.den <- findColours(q5.den, pal.green)
plot(pr, col=cols.den)
brks.den <- round(q5.den$brks,3)
leg.txt  <- paste(brks.den[-6], brks.den[-1], sep=" - ")
legend("topright", fill=attr(cols.den,"palette"), legend=leg.txt ,bty="n")
plot(pr.nb, points, add=T, col="red")
```

These blocks made the green maps of Puerto Rico. In these blocks, the code is essentially mapping a color to a variable, then choosing to how finely to divide the data up and map the divided data into hues of a selected color. The legend provides additional styling. The plot overlays connectivity. Green was used to represent the data as it is agricultural in nature and the red would provide a visual contrast.

```
par(mfrow=c(2,2))
hist(pr.f$irr_farms_07, main = "Number of farms in 2007", xlab = "farms")
hist(pr.f$irr_farms_02, main = "Number of farms in 2002", xlab = "farms")
hist(pr.f$rain_mean, main = "Distribution of Rain ", xlab = "mean of rain")
hist(pr.f$elev_mean, main = "Distribution of Elevation", xlab = "mean of elevation")
```

```
test <- log(pr.f$irr_farms_07)
kurtosis(pr.f$irr_farms_02)
kurtosis(pr.f$irr_farms_07)
t.test(pr.f$irr_farms_02, pr.f$irr_farms_07)
summary(pr.f$irr_farms_07)
summary(pr.f$irr)
```

```
mean(farm.den02)
```

```
mean(farm.den07)
t.test(farm.den02,farm.den07)
```
This generated the histograms of the variables of interest. The goal of performing a t-test was to determine if there is a statistical difference between the two means.

```
pr.listw <- nb2listw(pr.nb, style="W")
pr.listb <- nb2listw(pr.nb, style="B")
```
This segment of code reads data into the weighted connectivity matrix.

```
# Is an even distribution of larger points
# San Juan is east and capital
# Archibo is northwest
# Mayagues is southwest
# Ponce is south center
# Caguas is south east
adm <- factor(pr.f$ADM, levels=1:5, labels= c("San Juan", "Arecibo", "Mayaguez", "Ponce",
"Caguas"))
```

This segment creates factor variables to represent different areas in Puerto Rico

```
# Running a complete SAR Model
sar.chi<-lagsarlm(pr.f$irr_farms_07~ pr.f$rain_mean + pr.f$elev_mean + adm, data = pr.f,
pr.listw)
summary(sar.chi)
sar.chi.2 <- lagsarlm(pr.f$irr_farms_02~ pr.f$rain_mean + pr.f$elev_mean + adm, data = pr.f,
pr.listw)
summary(sar.chi.2)
```

This code performs SAR for the given variables of interest.

Conclusion

There are statistically significant differences in how the number of farms in Puerto Rico are affected by geography. Namely, rainfall is a significant factor. It is suggested that the risk of flooding and storms could be a determining factor in the number of farms being generating in a particular region. Additionally, it appears that Arecibo and the north-western segments of Puerto Rico are leading the way in creating the number of farms. This appears to be a more recent development as in 2002, they were not performing significantly more than San Juan. However, this changed in 2007. This is demonstrated through visual inspection of the density of farms per region as well.

While these finds may be interesting, it is nessary to concede the weaknesses of this analysis. Firstly, the distributions of variables are non-normal. This is a violation of the modeling assumptions and welcomes bias into the model. The other was the selection of only included

irrigated farms into the sample. While this weakens claims on the relationship of rain and how it affects farms, it is a required trade-off. If all farms were included, there would be more heterogeneity in the sample. If a farm has irrigation, it is more likely to be a standard farm. If a farm lacks it, it might not be a standard farm. It could be a garden or could fit any of number of classifications that could fit within the label. Thus, it was a modeling choice to constrain the sample. These points made with proposed relationship are substantiated by the positive relationship with elevation. This provides additional support to proposed relationship. However, it is at least clear that geography is strongly related with the number of farms overall in Puerto Rico as well as the density of farms. While, the extent of relationships remain unclear, additional study could provide additional definition and address the weaknesses and finding explored here.

Works Cited

Chun, Yongwang, and Daniel Griffth. *Spatial Statics & Geostatics*. PDF, London, Sage, 2013.