Banking Infrastructure and Inclusion: A Spatial Investigation of Physical Networks of Access and Measures of Inclusion

Louis Nix II

12/5/2019

1) Hypothesis:

I chose to investigate the relationship between physical portals of access to traditional finance and unbanked rates in NYC communities. In particular, I asked the following question: Do areas with higher densities of ATMs experience lower unbanked rates? Unbanked rates are defined as the percentage of the community without a bank account.

My hypothesis was that greater penetration of physical portals of access to financial institutions would reduce physical barriers to utilization of the services offered by those banks. While not a perfect operationalization of the independent variable, I used ATMs to model the network of physical portals. For my analysis, I used a measure of geographic penetration to operationalize the physical networks. However, with better data, I would prefer to redo the analysis and operationalize the networks as demographic penetration. In other words, the network would be modeled through ATMs per population. Regardless, geographic penetration should show the same results but with a potentially diluted magnitude.

From exploring the data, two maps helped identify areas of particular interest in the analysis. First, from maping out the variation in unbanked rates across NYC, it was clear that East Brooklyn, the Bronx, Lower Manhattan, Northern Queens, and the South Shore of Staten Island are at the extremes of rates. These areas were natural extremes for the regression to predict, later.

From the second graph, it was clear that ATM clustering was significant in the Bronx, East Brooklyn, and Lower Manhattan. After a realtionship was shown between ATM distribution and unbanked rates, the residuals in these three areas were of importance. In particular, they helped to identify areas where the relationship between ATMs and unbanked rates could be nonlinear. I discuss this more at the end of my analysis.

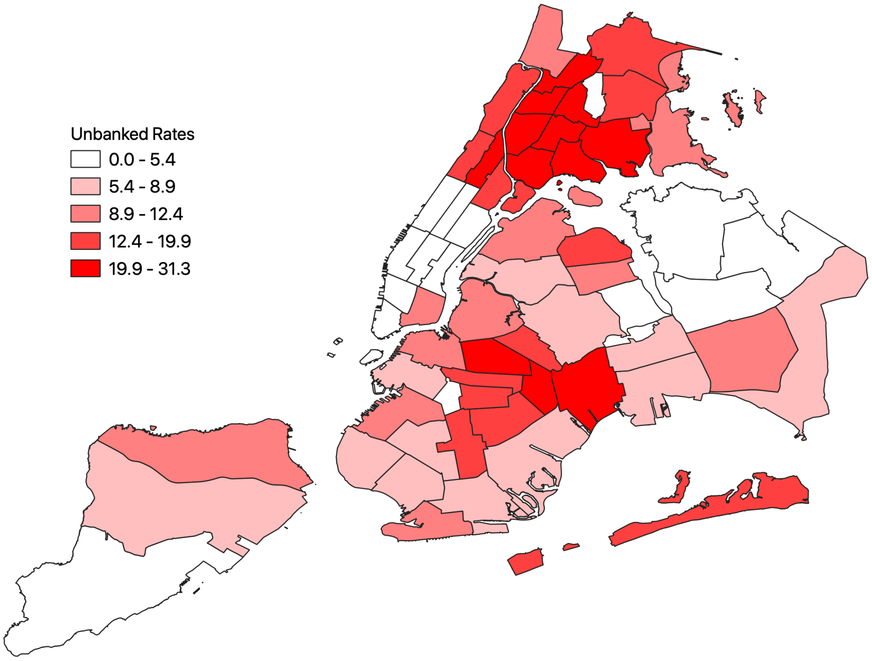

Figure 1: Unbanked rates by community district across NYC.

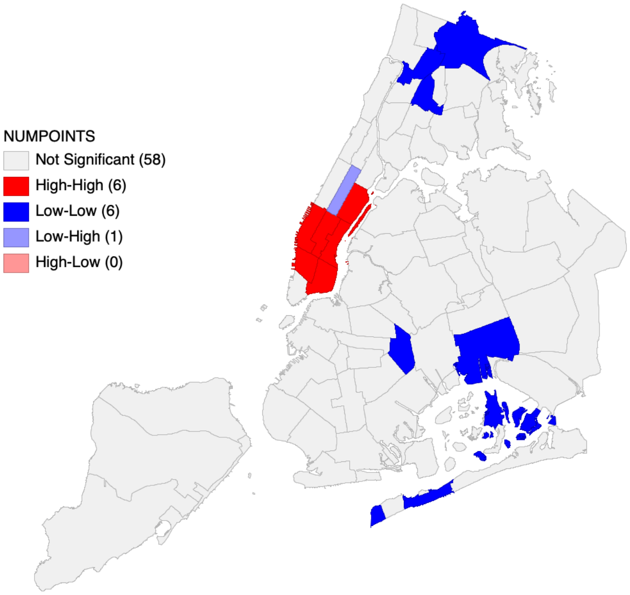

Figure 2: ATM clusters by community district in NYC.

2) Final Bivariate Analysis

The original bivariate analysis I conducted regressed *unbanked2013 on atmcount* and yielded a statistically significant effect with a coefficient of decent magnitude. However, when diagnosing for error and lag, the regression tested positive and highly statistically significant for both. When looking at the robust tests I found no evidence of error but the regression still suffered from lag. As a result, I chose to run a spatial lag model in my final bivariate analysis, the resutls of which are printed below.

```
# Queens weight matrix
list.queen <- poly2nb(shapefile, queen=TRUE)
# Create a weights matrix object
W <- nb2listw(list.queen, style="W", zero.policy=TRUE)

#run a spatial lag model
lag.model1 <- lagsarlm(unbanked_2013 ~ atm_count, data=shapefile@data, W, zero.policy = TRUE)
summary(lag.model1)
```

```
##
## Call:lagsarlm(formula = unbanked_2013 ~ atm_count, data = shapefile@data,
##     listw = W, zero.policy = TRUE)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -10.15566  -2.80763  -0.85187   2.76178  10.94440
##
## Type: lag
## Regions with no neighbours included:
##  36
## Coefficients: (asymptotic standard errors)
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  7.369946   1.541465  4.7811 1.743e-06
## atm_count   -0.097462   0.024802 -3.9296 8.509e-05
##
## Rho: 0.64362, LR test value: 36.374, p-value: 1.6288e-09
## Asymptotic standard error: 0.081652
##     z-value: 7.8825, p-value: 3.1086e-15
## Wald statistic: 62.134, p-value: 3.2196e-15
##
## Log likelihood: -177.5045 for lag model
## ML residual variance (sigma squared): 20.746, (sigma: 4.5548)
## Number of observations: 59
## Number of parameters estimated: 4
## AIC: 363.01, (AIC for lm: 397.38)
## LM test for residual autocorrelation
## test value: 1.1141, p-value: 0.29118
```

The model with a spatial lag showed a much smaller relationship between the variables, with the coefficient on *atm_count* reduced to about half of its former value in the OLS model. The coefficient was still highly statistically significant and indicated a negative relationship in which more ATMs is associated with a lower unbanked rate. Also, the likelihood ratio indicates that the addition of a spatial lag improved the model in a highly statistically significant way, with the probability of falsely rejecting the null (that the addition of the lag did not improve the model) at far less than 0.001%.

3) Regression w/ Control for Immigrant Status

To improve the model, I added a control for the percentage of each community district that is foreign born. My thought it that foreign born residents may either be newer or less integrated into the NYC economy, and thus less included in the financial system. As a result, the expectation was to see a negative relationship between the precentage of the community that was foreign born and the unbanked rate of the community.

```r
#start by running a linear regression
ols2 <- lm(formula = unbanked_2013 ~ atm_count + foreign_born, data = shapefile@data)
#print results
summary(ols2)
```

```
##
## Call:
## lm(formula = unbanked_2013 ~ atm_count + foreign_born, data = shapefile@data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.347  -4.570  -1.550   2.552  18.071
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.53677    3.13884   7.180 1.74e-09 ***
## atm_count    -0.19194    0.03491  -5.498 9.81e-07 ***
## foreign_born -0.11628    0.07388  -1.574    0.121
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.701 on 56 degrees of freedom
## Multiple R-squared:  0.3529, Adjusted R-squared:  0.3298
## F-statistic: 15.27 on 2 and 56 DF,  p-value: 5.101e-06
```

The results of the regression with the one control are quite interesting. As with the original OLS model, the coefficient on *atm_count* is quite large compared to the bivariate spatial lag model. The R^2 is also quite large, indicating that about 35.3% of the spatial variation is explained.

Unfortunately, the added control did not produce a statistically significant coefficient. The coefficient on *foreign_born* indicated that my updated hypothesis was correct in that areas with greater percentages of residents that are foreign born have higher unbanked rates. However, the probability of falsely rejecting the null is over 12%, indicating no support for the updated hypothesis.

I next decided that I needed to check and correct for spatial dependence since I knew that the previous model suffered extensively with spatial lag.

```r
# Spatial dependence using Moran's I
moran.lm <- lm.morantest(ols2, W, alternative="two.sided", zero.policy = TRUE)
print(moran.lm)
```

```
##
##  Global Moran I for regression residuals
##
## data:
## model: lm(formula = unbanked_2013 ~ atm_count + foreign_born, data
## = shapefile@data)
## weights: W
##
## Moran I statistic standard deviate = 4.8829, p-value = 1.045e-06
## alternative hypothesis: two.sided
## sample estimates:
## Observed Moran I      Expectation         Variance
##      0.431109544     -0.034675828      0.009099439
```

The Moran's I score is quite large and positive here with a very small p-value. As a result, I can conclude that there is statistically significant evidence for spatial clustering. The diagnostics for error and lag listed below are much more revealing of the problem.

```
# Diagnostics for spatial error and lag
LM <- lm.LMtests(ols2, W, test="all", zero.policy = TRUE)
print(LM)
```

```
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = unbanked_2013 ~ atm_count + foreign_born, data
## = shapefile@data)
## weights: W
##
## LMerr = 18.954, df = 1, p-value = 1.339e-05
##
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = unbanked_2013 ~ atm_count + foreign_born, data
## = shapefile@data)
## weights: W
##
## LMlag = 30.909, df = 1, p-value = 2.704e-08
##
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = unbanked_2013 ~ atm_count + foreign_born, data
## = shapefile@data)
## weights: W
##
## RLMerr = 0.096557, df = 1, p-value = 0.756
##
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = unbanked_2013 ~ atm_count + foreign_born, data
## = shapefile@data)
## weights: W
```

```
##
## RLMlag = 12.052, df = 1, p-value = 0.0005175
##
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = unbanked_2013 ~ atm_count + foreign_born, data
## = shapefile@data)
## weights: W
##
## SARMA = 31.006, df = 2, p-value = 1.85e-07
```

The initial diagnostics showed statistically significant evidence for both spatial error and lag. However, the robust measures were much clearer as the regression was clearly suffering from spatial lag and not error. The p-value for indicates high statistical significance for the measure estimated by the robust test for lag.

Given that I diagnosed the issed as one of lag, I then ran a spatial lag model for my final model with a control.

```
#run a spatial lag model
lag.model.final <- lagsarlm(unbanked_2013 ~ atm_count + foreign_born, data=shapefile@data, W, zero.polic
summary(lag.model.final)
```

```
##
## Call:lagsarlm(formula = unbanked_2013 ~ atm_count + foreign_born,
##     data = shapefile@data, listw = W, zero.policy = TRUE)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -10.29227  -2.96727  -0.38642   2.71827  11.14250
##
## Type: lag
## Regions with no neighbours included:
##  36
## Coefficients: (asymptotic standard errors)
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept)   8.777536   2.461446  3.5660 0.0003625
## atm_count    -0.101636   0.025140 -4.0428 5.282e-05
## foreign_born -0.032151   0.050203 -0.6404 0.5218991
##
## Rho: 0.63344, LR test value: 34.213, p-value: 4.9403e-09
## Asymptotic standard error: 0.083158
##     z-value: 7.6173, p-value: 2.5979e-14
## Wald statistic: 58.023, p-value: 2.5868e-14
##
## Log likelihood: -177.3081 for lag model
## ML residual variance (sigma squared): 20.728, (sigma: 4.5528)
## Number of observations: 59
## Number of parameters estimated: 5
## AIC: 364.62, (AIC for lm: 396.83)
## LM test for residual autocorrelation
## test value: 0.52835, p-value: 0.4673
```

The results of the final spatial lag model are quite interesting. As with the bivariate analysis earlier, the coefficient on *atm_ count* shrank quite a bit and maintained its statistical significance. Also, the effect was quite nearly the same as the estimate produced in my final bivariate model. The interesting part sathat the coefficient on *foreign_born* both shrank by a significant degree and lost even more statistical significance. Now, the probability of falsely rejecting the null of no relationship is over 50%. Having corrected for spatial lag, it was safe to conclude that there is no relationship between the percentage of a community that is foreign born and the unbanked rate of that community.

In conclusion, my hypothesis that greater penetration is associated with lower unbanked rate is upheld in the final analysis with controls. However, the competing hypothesis that foreign born populations are less financially included is not supported by the model.

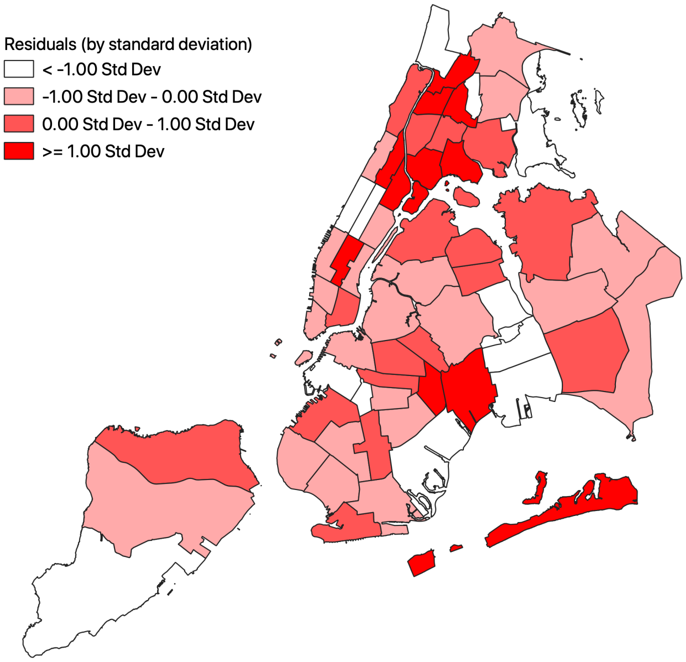
Residuals (by standard deviation)
< -1.00 Std Dev
-1.00 Std Dev - 0.00 Std Dev
0.00 Std Dev - 1.00 Std Dev
>= 1.00 Std Dev

Figure 3: Map of residuals from the final spatial lag model with a control variable.

Map of Residuals

The map of residuals above shows more extreme standard deviations for both the Bronx and Eastern Brooklyn. The standard deviations are closer to zero in Lower Manhattan, which was not as I had predicted. From the residuals and the lag model, it seems the relationship between ATM counts and unbanked rates may be dimenishingly exponential. That is to say, in areas with particularly low ATM counts and high unbanked rates (i.e. the Bronx and Eastern Brooklyn), the effect of one more ATM may be strong. At the same time, areas with high ATM counts and very low unbanked rates benefit much less from another ATM, bringing the true value of these areas closer to the mean estimate that works so well for the majority of the community districts. In other words, the residual map would indicate a diminishing decrease in unbanked rates as ATM counts rise.