Michael Fivis
December 14, 2013
GTECH 705: Spatial Analysis

What Bars Indicate About New York Neighborhoods:
A Geospatial Modeling Study of Lower Manhattan Liquor Licenses

Introduction

Academic analyses of urban gentrification often use the word "frontier" to define the real or perceived demographic distance between incumbent residents versus the new, incoming individuals. The term has also been used to describe the mentality of entrepreneurs opening restaurants and bars in regions that have seen long periods of disinvestment. In New York City, nightlife development is considered a contentious byproduct of -- or the reason *for* -- a neighborhood becoming desirable. It's a signal of gentrification "happening" in any particular neighborhood. But just how well do the locations of bars and restaurants with liquor licenses correlate to real demographic gradients (income, race, poverty or other)?

"The Early Gentrifier: Weaving a Nostalgia Narrative on the Lower East Side," a paper by Richard E. Ocejo, identifies many stresses associated with a neighborhood's rapidly changing commercial identity. In a section titled "The Threat of Nightlife: Setting, Participants and Methods," Ocejo reported that the number of bars grew at incredible rates *and* in a particular pattern.

> The number of bars in the neighborhood more than doubled from 35 in 1985 to 76 in 1995. They doubled again to 144 in 2000 and then increased to 177 in 2005. In general, they followed gentrification's "frontier line" (Smith 1996) that gradually moved eastward. As of 2008, there were 692 establishments with an on-premise liquor license—that is, any place where an alcoholic beverage may be purchased and consumed—within the 1.8-square-mile area of the Lower East Side(...)

The detail Ocejo cites, Neil Smith's 1996 *The New Urban Frontier*, contains a

charged chapter titled “The Lower East Side as the Wild, Wild West” which makes the case that “homesteading” by way of building condominiums and nightlife is actually quite similar in purpose to historic American Western frontier homesteading in its political objectives. Instead of rich, natural Western lands, however, it’s urban dilapidation that is sought out for normalization. Whether or not that is true, frontiers may be defined using tract-by-tract demographic differences and how these relate to the locations of New York State liquor licenses can be analyzed.

Using Census Data and the Public License Query provided by the New York State Liquor Authority, it should be possible to study these questions. Cleaning will have to be performed on the geocoding of liquor licenses, and other complexities (like the square footage or frontage dedicated to liquor licensing) may be worked into the models using the City’s tax lot information.

Goals and Objectives

Using the locations of bars connected with the geographic information about their immediate surroundings, it is hoped that some connection to their surrounding residents’ demographics can be established. Bars are rather confounding as community features because, in practice, locals who are regulars are not the sole source of patronage. Nightlife districts, such as the Lower East Side (10002) and the East Village (10003/10009) are identified as destination places largely serving non-neighborhood residents. These neighborhoods generate more conflict between residents than others (the East Village and Lower East Side lodged the most commercial noise complaints to 311 in 2012 than any other city

neighborhood, according to the past year of NYC OpenData reporting) Incidentally these are some of the poorest neighborhoods of the study area.

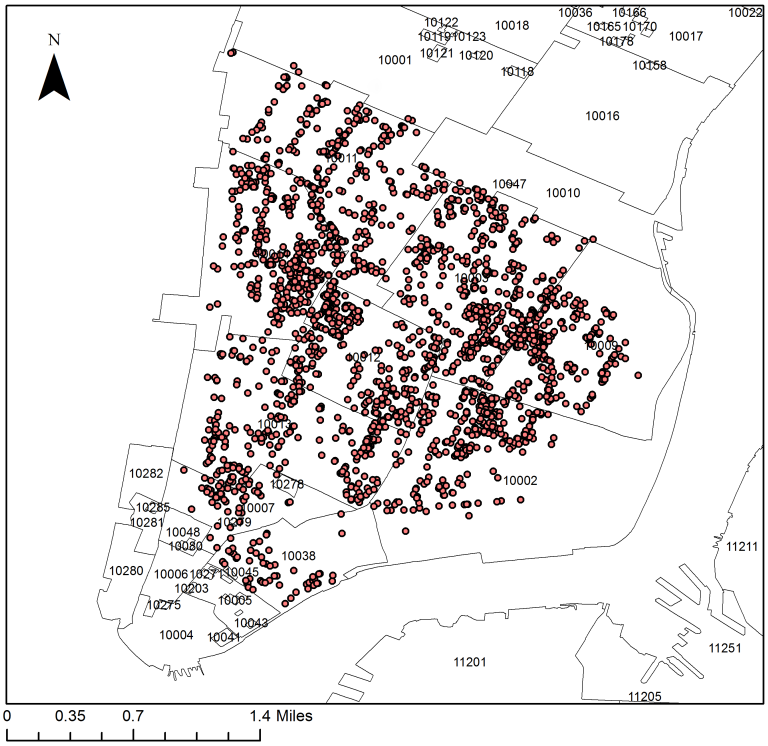
Additionally a geospatial difference between hard liquor licenses and those that permit only the consumption of beer and wine should be examined to see if its worth separating. A Beer and Wine license is more likely to be held by a restaurant whose primary focus is serving food and a full liquor license is more likely to be held by a bar whose focus is serving primarily cocktails. While this relationship is wrought with many exceptions (such as bars that only serve beer and wine with no food and the many restaurants that hold full liquor licenses), full liquor licenses - or "OP" licenses – still come under greater scrutiny in the application process for their attractiveness to minors and potential noise pollution.

Data and Study Area Description

On-premise liquor licenses from nine ZIP codes (2,469) encompassing most of lower Manhattan were obtained from the New York State Liquor Authority's online query system. The Liquor Authority's query tool returns little information other than the type of license (Beer and Wine Only or Full Liquor), serial number, address and a DBA (Doing Business As) string. The liquor licenses had coordinate sets attached to them using CartoDB's geocoder and were then cleaned of input errors and incorrect addresses.

The ZIP codes (*10003, 10009, 10002, 10012, 10013, 10038, 10007, 10011 and 10014*) span the full width of Manhattan and include a wide range of demographics. To get more local demographic information, the points were joined to 5-Year-Estimate ACS data at the census block-group level. A variety of demographic

attributes such as race, income age and housing data were attached to each related liquor license coordinate pair.



ZIP	10002	10003	10007	10009	10011	10012	10013	10014	10038	Total
# OP	199	288	39	130	217	198	195	241	39	1546
# All	308	487	69	237	317	319	287	379	66	2469

Methods

The data with several different model building methods as well as some other spatial diagnostics. Kernel density estimations will be made of the point pattern as a whole and measures of spatial autocorrelation for key demographic information will be made using Moran's I scatterplots. Ordinary Least Squares regression is used to determine the interesting terms to build models with and from there, spatially aware models like the Spatial Error and Spatial Lag models will be

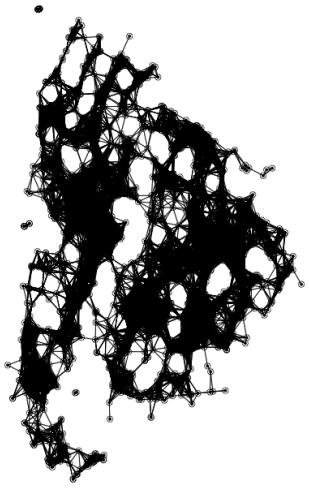
used to reduce clustering. The most interesting model terms will be incorporated into Geographically Weighted Regression models for further discussion.

In addition to these methods, the locations of bars will be studied both by attaching census data attributes to the coordinate pairs of bars, and by performing analysis on the census blocks themselves while staying aware of the number of liquor licenses contained within each census blocks. In this case, it becomes 241 observations (census block groups with the counts of liquor licenses contained in their geography). The first approach will allow the demonstration of some models predicated on the understanding that the only points of observations are the locations of the liquor licenses themselves. The second approach will allow the *Number of Liquor Licenses* to be a guiding response variable in model building.

To determine some elements of spatial autocorrelation, a neighbor list and weights matrix was constructed for both the point pattern of liquor licenses and the network of census block groups. For the point pattern, some "islands" of licenses occur when using the smallest maximum neighbor distance. These were left to be, given that some of the most noticeable disconnects are relatively unique locations such as waterfront restaurants and bars on Manhattan piers.

Additionally, given some missing data at the block-group level in the census data, models were only constructed on a subset of cases (241 cases instead of the original 264 block groups).

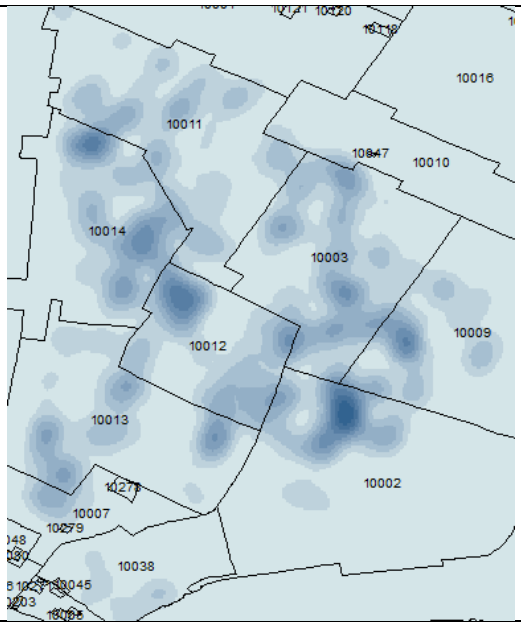
Block-group polygons study area.



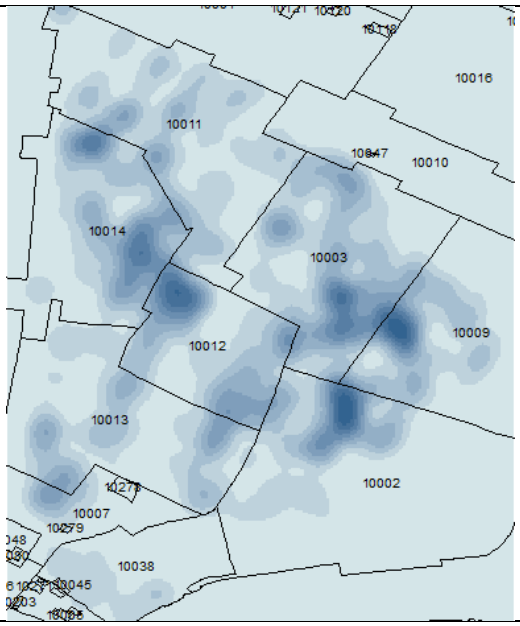
Spatial weights matrix of 2,469 liquor licenses, plotted.

Results and Discussion

The initial examination of the point pattern revealed heavier numbers of dense clusters on the Lower East Side in Manhattan, spread across ZIP codes 10002, 10003 and 10009. A kernel density surface with a search radius of approximately 75 meters revealed these dense clusters of licenses. At the same time, it was observed that the point patterns of OP licenses and the collection of all liquor licenses were not very that different in point distribution. In fact the mean centroids of the two point patterns were only 41 meters apart.



OP (Full Bar) Licenses

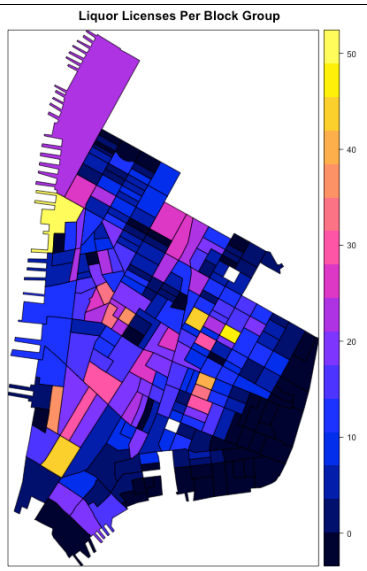


All Liquor License Types

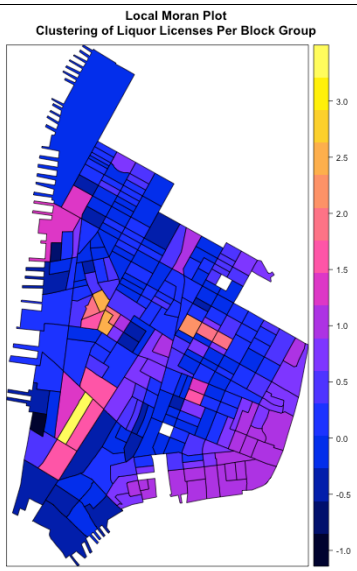
Within the study area of polygons, spatial autocorrelation analysis was used to determine the level of clustering that occurred with the number of liquor licenses found within each block group polygon. Moran's I found there to be clustering, with significance, at a rate much larger than expected given random data.

```
data:  point2$pntcnt
weights: lw  # Queen's Case

Moran I statistic standard deviate = 9.4607, p-value < 2.2e-16
alternative hypothesis: two.sided
sample estimates:
Moran I statistic       Expectation          Variance
      0.342539001      -0.004166667       0.001342986
```

It is important to see where a local Moran plot sees the clusters as occurring and to see if this aligns with what we've seen in the kernel density map generated. Moran's I values were attached to the block groups and replotted to ascertain this. In the graphic below, the East Village, the Lower East Side and certain pieces of the West Village again show up as extremely clustered in the numbers of liquor licenses.

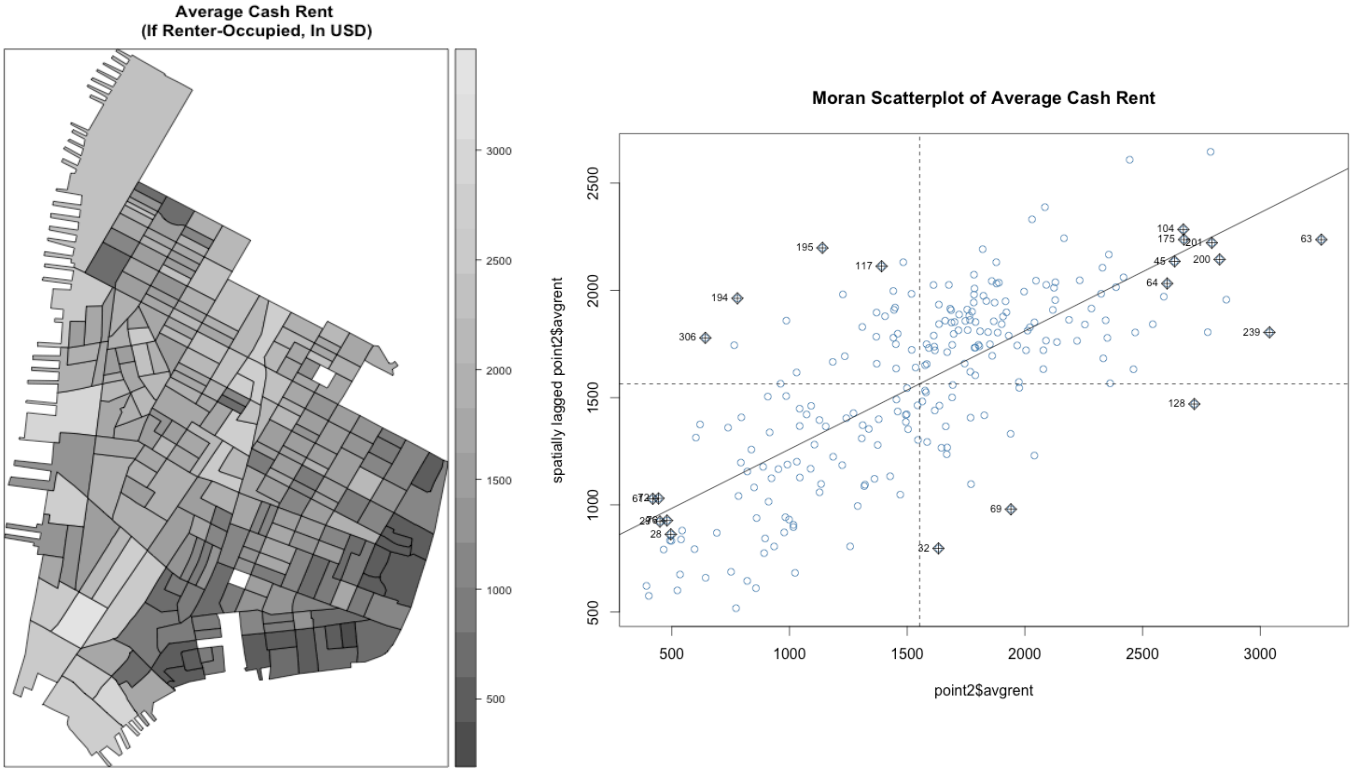


Liquor Licenses Per Block Group



Local Moran Plot of Clustering

One basic attribute that governs where liquor license holders locate themselves is nearby cash rent because, at the end of the day, bars and restaurants are not-so-unusual small businesses. In mixed use real estate, the storefront is simply another tenant in the building and rent may help us understand that that same clustering just as much as any demographic data.



Block group polygons occur in a highly irregular pattern, but these two measures (number of liquor licenses and Average Cash Rent) were both found to be clustered whether distance band weighting or queen's case adjacency was used. For the localness of the data being analyzed, adjacency will be adequate for modeling the effects of the conditions of one small group of Manhattan blocks versus another.

The appearance of clustering was much diminished when looking at many other demographic measures that could have associations with the locations of

liquor licenses, like Census respondents who classified themselves as White Alone or Median Household Income:

```
data:  point2$whitealone
weights: lw

Moran I statistic standard deviate = 15.5842, p-value < 2.2e-16
alternative hypothesis: two.sided
sample estimates:
Moran I statistic       Expectation          Variance
     1.049684e-01     -4.166667e-03      4.904121e-05
```

While still significant and different than the expectation, it may be more useful as an additive model term than as an observed clustering phenomenon itself.

Ordinary Least Squares Regression

Model building was started with some basic terms including *number of households, total population, individuals of race classified as White Alone, individuals of Hispanic descent, nonfamily households, average cash rent and median age.* These are used to predict the number of liquor licenses in every blockgroup.

An initial model proved weak, though through its analysis of variance report and summary it was determined which terms may matter much more than others.

```
> summary(lm1)

Call:
lm(formula = pntcnt ~ totalpop + medianage + whitealone + hispanic +
    households + nonfamilyh + avgrent, data = point2.df)

Residuals:
    Min      1Q  Median      3Q     Max
-18.655  -6.002  -1.024   3.068  36.074

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.0990417  5.5527144   3.980 9.21e-05 ***
totalpop     0.0009401  0.0027490   0.342 0.732675
medianage   -0.3631310  0.1043389  -3.480 0.000598 ***
whitealone   0.0008030  0.0022575   0.356 0.722377
hispanic    -0.0060469  0.0026360  -2.294 0.022682 *
households  -0.0055135  0.0095445  -0.578 0.564047
nonfamilyh   0.0038856  0.0083030   0.468 0.640238
avgrent      0.0021998  0.0016084   1.368 0.172717
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 9.029 on 233 degrees of freedom
Multiple R-squared:  0.2185,     Adjusted R-squared:  0.1951
F-statistic: 9.309 on 7 and 233 DF,  p-value: 3.683e-10
```

Though it had weak explanatory power, the p-value of the model itself was significant. The *medianage* and race terms proved to have higher explanatory power than the other terms.

```
> anova(lm1)
Analysis of Variance Table

Response: pntcnt
            Df  Sum Sq Mean Sq F value    Pr(>F)
totalpop     1    18.5    18.5  0.2264 0.6346867
medianage    1  3282.3  3282.3 40.2613 1.146e-09 ***
whitealone   1  1004.7  1004.7 12.3241 0.0005367 ***
hispanic     1   831.9   831.9 10.2040 0.0015950 **
households   1    11.4    11.4  0.1400 0.7086632
nonfamilyh   1    10.9    10.9  0.1332 0.7154203
avgrent      1   152.5   152.5  1.8707 0.1727174
Residuals  233 18995.5    81.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

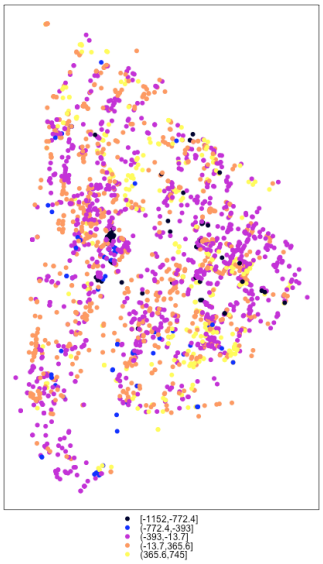
Flipping the model to see a model's strength at predicting demographics using the point pattern itself was also attempted. The samples are all taken at the coordinate points of observed liquor licenses. Other combinations of terms failed to improve the model beyond an R-squared value of about .25. This one below (*lm2*) substituted some of the terms for *income*, *number of individuals under 65*, and *college educated individuals over 25* and was then flipped to explain the average rent at the points where liquor licenses were located.

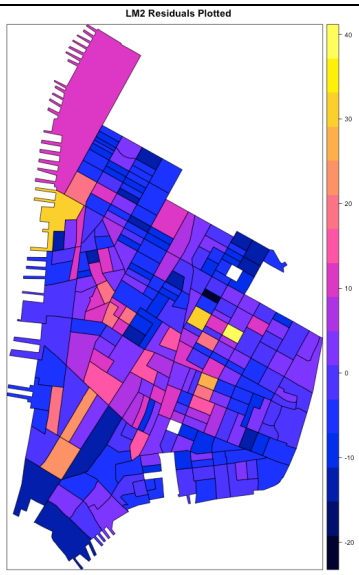
```
pntcnt ~ fampoverty + avghhsize + medianage + whitealone + hispanic +
pop25colle + totalunder + mhi + avgrent

Multiple R-squared:  0.2487 # Looking the block group polygons
```

The terms have a stronger connection when not directly in relation to the number of liquor licenses in each observation. When predicting one of the explanatory terms in the last model (*avgrent*), the model was able to explain with

significance more than twice as much of the variance, with most of the terms playing a strong part.

```
> summary(plm1)

Call:
lm(formula = avgrent ~ fampoverty + avghhsize +
    medianage + whitealone + hispanic + pop25college +
totalunder65 +
    mhi + avgrent, data = total1)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.079e+03  6.003e+01  34.639  < 2e-16 ***
medianfemale -1.123e+01  1.622e+00  -6.922 5.67e-12 ***
fampoverty   -6.588e-01  2.586e-01  -2.547 0.010914 *
avghhsize    -2.459e+02  2.664e+01  -9.234  < 2e-16 ***
medianage    -4.553e+00  1.921e+00  -2.371 0.017831 *
whitealone   -1.492e-01  3.805e-02  -3.922 9.00e-05 ***
hispanic     -5.565e-01  5.468e-02 -10.177  < 2e-16 ***
pop25college  4.629e-01  4.090e-02  11.318  < 2e-16 ***
totalunder65  1.041e-01  2.811e-02   3.704 0.000217 ***
mhi           4.590e-03  1.643e-04  27.940  < 2e-16 ***

Residual standard error: 302.3 on 2459 degrees of freedom
Multiple R-squared:  0.6577,  Adjusted R-squared:  0.6565
F-statistic:   525 on 9 and 2459 DF,  p-value: < 2.2e-16
```





Residuals from linar model predicting avgrent.

Residuals plotted for license count predction.

Though both models predict two different response variables and have vastly different strengths, there appears to be clustering occurring. In the point pattern model this is forgiven for the fact that the attributes attached to these points are identical within the boundaries of the block group they are associated with.

Spatially Aware Regression Models

In order to switch to improve upon the polygon analysis with a spatially aware model the Lagrange multiplier diagnostic is run to determine whether a Spatial Lag model or a Spatial Error model might be more appropriate. When considering the polygon observations and the second model used with them, still using queens case adjacency, the Spatial Lag model was chosen on account of slightly higher significance.

```
Lagrange multiplier diagnostics for spatial dependence

data:
model: lm(formula = pntcnt ~ fampoverty + avghhsize + medianage +
whitealone + hispanic +
pop25colle + totalunder + mhi + avgrent, data = point2.df)
weights: lw
LMerr = 31.5125, df = 1, p-value = 1.982e-08
LMlag = 41.165, df = 1, p-value = 1.399e-10
```

In fitting this data and these terms to a Spatial Lag model

```
Call:lagsarlm(formula = pntcnt ~ housingoccupied + avghhsize + medianage
+
    whitealone + hispanic + pop25colle + totalunder + mhi + avgrent,
    data = point2.df, listw = lw, tol.solve = 1e-21)

Residuals:
      Min        1Q    Median        3Q       Max
-18.67952  -5.61621  -0.73082   3.35157  35.60046

Type: lag
Coefficients: (asymptotic standard errors)
               Estimate  Std. Error z value  Pr(>|z|)
(Intercept)  1.6840e+01  6.2847e+00  2.6794  0.007375
fampoverty  -1.7046e-03  1.5682e-02 -0.1087  0.913446
avghhsize    3.3266e-01  1.9341e+00  0.1720  0.863436
medianage   -3.5097e-01  8.6360e-02 -4.0640 4.823e-05
whitealone  -6.2560e-04  2.6954e-03 -0.2321  0.816463
hispanic    -4.7166e-03  2.6024e-03 -1.8124  0.069919
```

```
pop25colle  -1.6514e-03  3.4568e-03 -0.4777  0.632850
totalunder   1.5185e-03  2.0774e-03  0.7309  0.464811
mhi          5.1739e-05  1.8790e-05  2.7535  0.005896
avgrent     -1.3908e-03  1.8798e-03 -0.7399  0.459371

Rho: 0.47151, LR test value: 32.16, p-value: 1.4198e-08
Asymptotic standard error: 0.076759
    z-value: 6.1427, p-value: 8.1132e-10
Wald statistic: 37.733, p-value: 8.1132e-10

Log likelihood: -848.7152 for lag model
ML residual variance (sigma squared): 64.283, (sigma: 8.0177)
Number of observations: 241
Number of parameters estimated: 12
AIC: 1721.4, (AIC for lm: 1751.6)
LM test for residual autocorrelation
test value: 0.54929, p-value: 0.45861
```

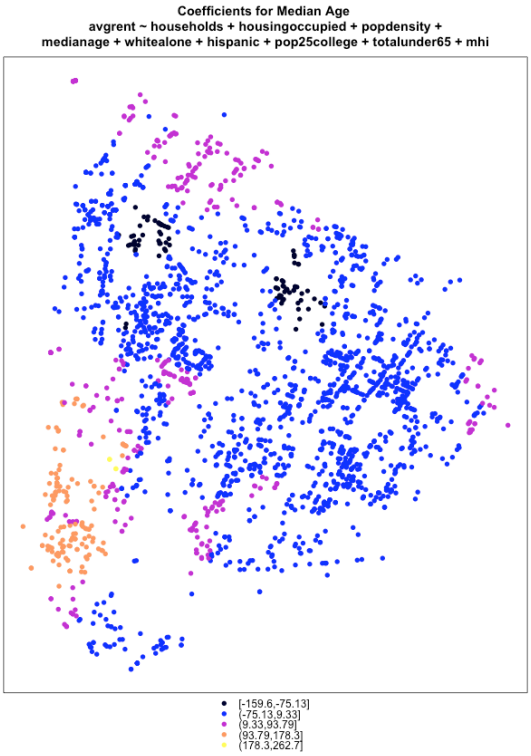
When taken spatial, few of the terms remained significant besides household income and medianage. Some of these weaker terms were substituted with others that may have been more informative given a spatially-aware model.

```
Call:lagsarlm(formula = pntcnt ~ households + housingocc + popdensity +
    medianage + whitealone + hispanic + pop25colle + totalunder +
    mhi + avgrent, data = point2, listw = lw, tol.solve = 1e-21)

Residuals:
      Min        1Q    Median        3Q       Max
-20.32999  -4.54233  -0.83238   3.23277  32.56036

Type: lag
Coefficients: (asymptotic standard errors)
               Estimate  Std. Error z value  Pr(>|z|)
(Intercept)  1.9397e+01  4.9822e+00  3.8934 9.886e-05
households   7.2287e-03  4.9927e-03  1.4478 0.1476597
housingocc  -1.5751e-02  4.4287e-03 -3.5566 0.0003757
popdensity  -3.4716e-05  1.0277e-05 -3.3779 0.0007303
medianage   -3.1113e-01  9.5162e-02 -3.2695 0.0010773
whitealone   2.0681e-03  2.4475e-03  0.8450 0.3981102
hispanic    -6.9832e-03  2.4415e-03 -2.8603 0.0042327
pop25colle  -5.0079e-03  4.2112e-03 -1.1892 0.2343601
totalunder   1.2930e-04  1.8059e-03  0.0716 0.9429184
mhi          7.6183e-05  2.0106e-05  3.7891 0.0001512
avgrent     -2.5761e-03  1.7624e-03 -1.4617 0.1438170

Rho: 0.40928, LR test value: 24.697, p-value: 6.7082e-07
Asymptotic standard error: 0.077176
    z-value: 5.3033, p-value: 1.1376e-07
Wald statistic: 28.125, p-value: 1.1376e-07

Log likelihood: -836.9486 for lag model
ML residual variance (sigma squared): 58.967, (sigma: 7.679)
Number of observations: 241
Number of parameters estimated: 13
AIC: 1699.9, (AIC for lm: 1722.6)
LM test for residual autocorrelation
test value: 0.38237, p-value: 0.53634
```

The term that included just how many housing units were occupied by their owners, as well as the population density terms, seemed to be productive additions to the Spatial Lag model, lowering the AIC value and maintaining a decent *Rho* value indicating each observations' impact on its neighbor (again, still based on Queen's case adjacency).

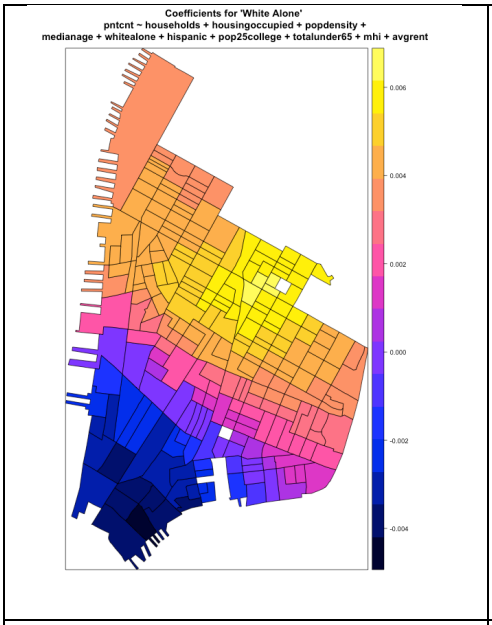
Geographically Weighted Regression

Geographically Weighted Regression offered an opportunity to examine the liquor license data in the most compelling way for this data, applying new coefficients for every point that would further our understanding of each term's relationship. While computing the GWR model was extremely computationally intensive in this data set, one "kitchen sink" model that sought to explain the *avgrent* term.

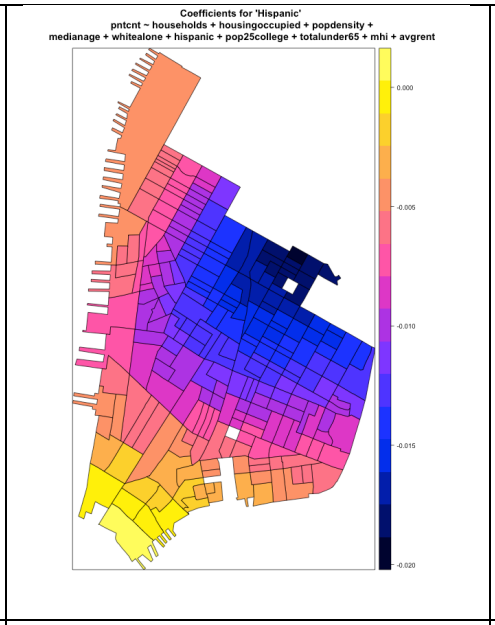


Because we are still relating the data, even only spatially, to the locations of nightlife destination, it is somewhat revealing to see that every year increase in MedianAge, in large swaths of the study area, may equal a (mild) decrease in the monthly rent for an apartment. Repeating a model with similar terms on the polygon study area revealed smooth, radial surfaces of the coefficients predicting the number of liquor licenses that might be found in a given Census block group.

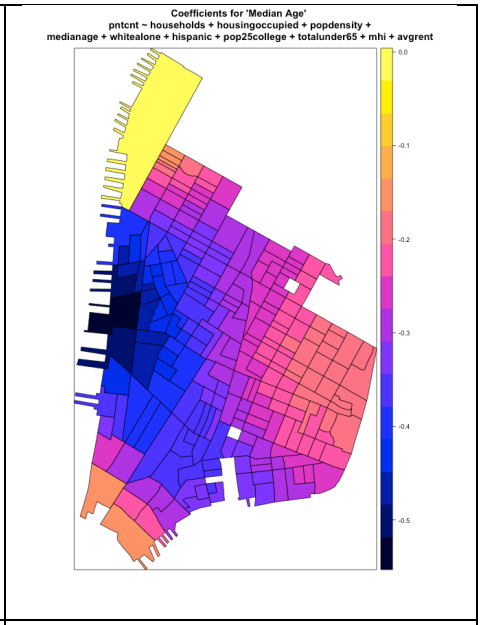
With regard to the quantity of liquor licenses it is conceptually more relevant and clear to perform GWR on the polygon study area itself, predicting the number of liquor licenses in each block group using a similar set of terms as before.



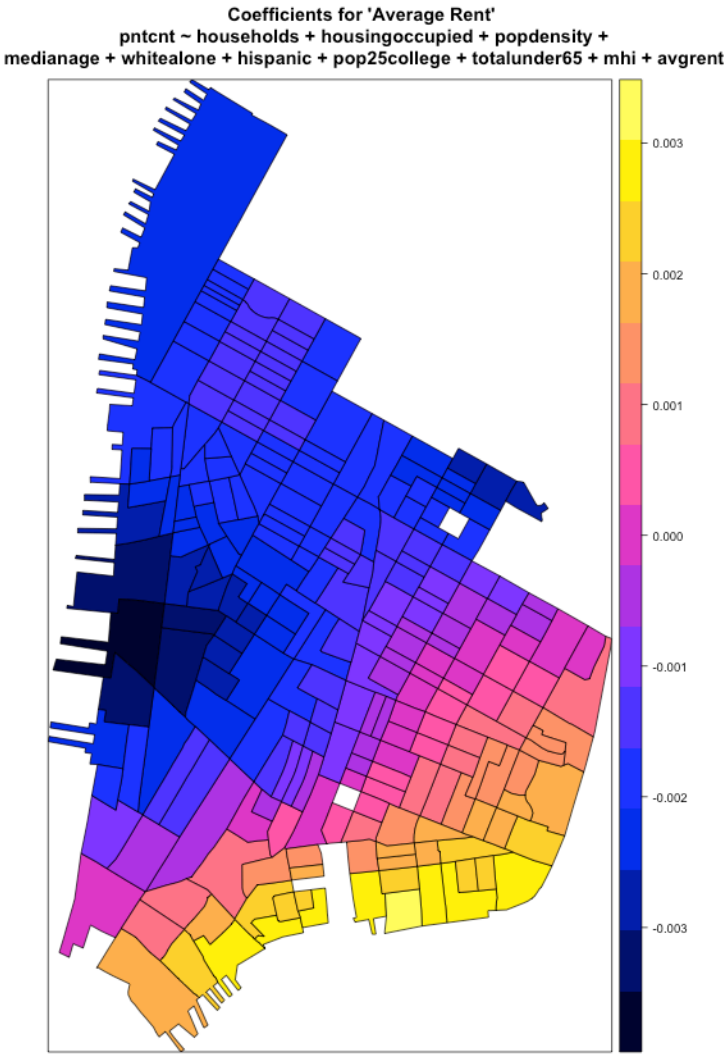
Coefficients for "White Alone"



Coefficients for "Hispanic"



Coefficients for "Median Age"



One of the most interesting relationships was, again, Average Cash Rent Paid. The curve it formed followed roughly the outline of where the point pattern of liquor licenses fell on the block group study area. In the steepest sections of the plot above, every dollar in average rent paid for an apartment is modeled to predict .003 more liquor licenses. While across the "frontier" additionally dollars appear to be causing the liquor licenses to decrease in count for every dollar, this could be adjusted for with a different type of modeling more appropriate to counting. Spatial

clustering that was a problem in Ordinary Least Squares regression here is alleviated with GWR here.

Conclusion

Geographically Weighted Regression is well suited for this not only hyper local, but relatively fuzzy estimation of community influences on response variables. The curves of data can be fitted to almost appear to be representing the social demographic edges of where bars, or liquor licenses, or any other phenomenon may be occurring. While it was surprising to find weaker relationships between any demographics, it is logical that Average Rent Paid would explain so much about the siting of liquor licensing given that it both explains the climate for small businesses in the area *and* describes nearby residents in an indirect way.