1 Background

Sociology is a discipline of latent variables. We never directly observe a class, a political party, an organization, but only behavior or traces of behavior that we believe to be expressions of a hidden principle. In this sociology is no different than any empirical science. But unlike some sciences, sociology has had a hard time finding ways to measure our most important latent variables so that we believe these measurements to be reliable and valid.

I believe that this has not been lack of ability, but to lack of data, and that in this moment of unparalleled and creepy recording of traces of human behavior, we can do better. The main challenge we face, the one we have always faced is how to convince ourselves about something that we cannot see.

This work on neighborhood boundaries hopes to be evidence for that hope. First, I will learn a set of neighborhood boundaries for the 20 largest American cities using a collection of labels that people have applied to locations and existing theory of neighborhoods. Second, I will evaluate whether the derived neighborhood boundaries correlates with differences in demographics, as we would predict. Finally, I will evaluate whether incorporating information about derived neighborhood boundaries improves our predictions about spatial, social processes, particularly crime rates.

This particular paper is about work on the first step, learning neighborhood boundaries.

2 Data

I have a nightly updated database of geocoded Craigslist apartment rentals, sublet, and roommate listings. For most of these listings, the poster entered some text in the "Specific Location" field. From this data, we want to create a map of the probabilities that some point in a city will be labeled by a neighborhood name.

2.1 Pre-processing

The first task must be to filter and normalize the "Specific Location" labels into canonical neighborhood names. Aside from variations of spelling of neighborhood names, some labels consist of address or intersection information or claims that the listing is near a train station, park, beach, or other generic landmark. We want to discard such labels, and keep only labels that refer to a neighborhood or some unique landmark, such as a named park or university.

Ideally, we would like to learn this filter and normalization mapping, perhaps by creating a metric that combines a Levenshtein and Euclidean distance metric. Alternatively, if we had sufficient data, we would likely do well just to discard all labels that appear less than some threshold. However for the case of Chicago, I have built up a set of handcoded rules that works well.

For some listings, the poster enters more than one neighborhood in the "Specific Location" field, i.e. "Lakeview/Lincoln Park". For such listings, we create a new location for each neighborhood in the label, so, for the example above, it is as if we saw two listings at the same location, one labeled "Lakeview" and the other "Lincoln Park." Some other scheme of credit assignment might be worth exploring.

After all this processing, we make sure that every location-label pair is unique. Almost all duplicates are due to the same poster reposting the same listing, and such reposts contain no more information than the first post. We throw away some information here, and we would be a better off if we could only discard postings that we know are really from the same poster.

Finally, we reproject the latitude and longitude encoded coordinates to a State Plane Coordinate System, particularly NAD83 / UTM zone 16N. We will be using other geocoded inormation, and this is a good common projection for Chicago.

3 Density Estimation

Now, we want to make a kernel density estimate for each neighborhood. For now, we will only estimate the density of a subset of neighborhoods on the near north side where Craigslist postings are concentrated.

As our kernel density estimates will depend upon the variance of the data, we attempt to find and remove outliers using the Hampel test.[1]

3.1 Bandwidth Selection

The most important choice in kernel density estimation is selecting the bandwidth, and I examined a number of multivariate bandwidth selectors: drop in plugins, smoothed cross validation, cross validated nearest neighbors, and cross validated adaptive nearest neighbors.

The most plausible looking results came from a drop-in plugin from the 'ks' library[2]. This method depends upon assuming, at one point, that the true density is Gaussian. I would like to implement the bandwidth selection procedure described by Botov et.al., which makes no such assumption and which seems to make more efficient estimates for distributions which are not Gaussian.[3]

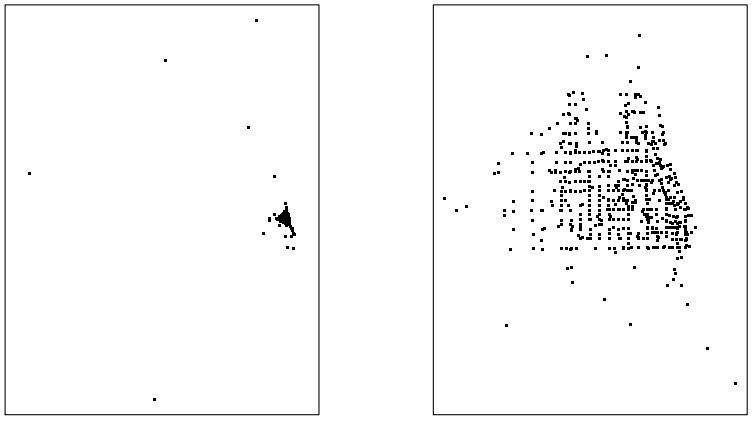
4 Evaluation and Discussion

After estimating the density for every neighborhood, we can plot them, and the decision boundaries between them. In addition to calculating the boundaries

[1] Lauire Davies and Ursula Gather. "The Identification of Multiple Outliers." *Journal of the American Statistical Association*, 1993, (88:423), p.782.

[2] `http://cran.r-project.org/web/packages/ks/`

[3] Botev, Grotowski, Kroese, "Kernel Density Estimation Via Diffiusion." *The Annals of Statistics* 2010, (38:5), p. 2916-2957.



(a) (b)

Figure 1: (a) All "Lakeview" locations; (b)"Lakeview" locations with outliers removed, CV = 0.9

between neighborhoods, I define a threshold value, such that if the probability that a point belongs to any neighborhood is below that threshold I assign that point to a 'no neighborhood' class.

In order to get a sense of plausibility of those boundaries we will also plot the official boundaries of the Chicago Community Areas, the river, and large city parks. We should expect that the geographic features should make up many of the boundaries. At the median values of neighborhood labels, I plot the neighborhood name and the number of points that the density was estimated with.

The resulting decision boundaries follow our expected borders remarkably well, particularly where we have many observations. The decision boundary between Lakeview and Lincoln Park follows Diversey. Roscoe Village and North Center are separated from Lakeview by the railroad tracks. Humboldt Park and Wicker Park are divided from Logan Square and Bucktown by North Avenue. Almost no neighborhood crosses the river, and where a decision boundary does cross the water it is for neighborhoods where we have relatively little support.

Interestingly, both Ravenswood and Andersonville lay across the tracks, and this does not seem to be an approximation error.

It seems like we could definitely benefit from incorporating beliefs that neighborhood boundaries fall along certain kinds of geographic features, but we will need to allow neighborhoods to cross these lines if the data strongly suggest that they do.

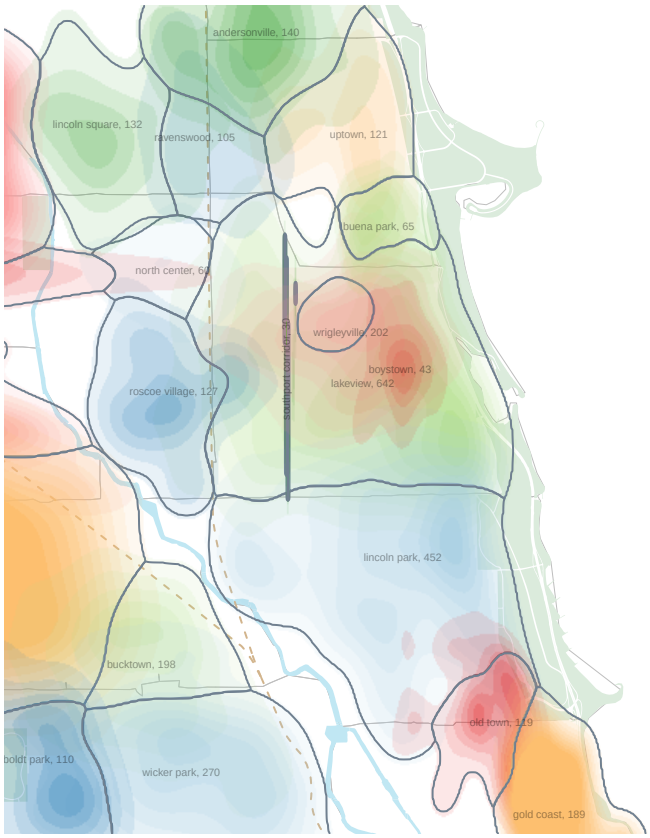


Figure 2: Density and decision boundaries for North Side neighborhoods