# Thinking inside the box: mapping the microstructure of urban environments (and why it matters)

**Seth E. Spielman, David Folch, John Logan, Nicholas N. Nagle**

**ABSTRACT:** Cartographers and human geographers have long relied on data sources which, for reasons of confidentiality, aggregate information to administrative units such as census tracts. The borders of these administrative units sometimes correspond to the jurisdictional boundaries of towns and counties, but most often they follow visible features of the landscape like streets, rivers, or coastlines. Because disaggregated data is rarely available the internal structure of these administrative units are poorly understood. These micro-scale patterns have important methodological and theoretical implications for cartography. This paper develops methods to explore the microstructure of the urban environment using a unique probability-based sample of streets in the American city of Chicago, Illinois. We examine six aspects of the urban environment having to do with visible signs of disorder, the use of public space, and vehicular traffic. We find that most aspects of the urban environment vary substantially over very small distances. The spatial autocorrelation of the studied aspects of the urban environment seems to have a characteristic range of 2-3 blocks- that is beyond two or three blocks streets bear little resemblance to their neighbors. These findings suggest a poor correspondence between the scale of census administrative units (like tracts) and the structure of the urban environment. These findings also raise important questions about the fidelity of maps based on administrative units.

## Introduction

Choropleth maps are commonly used to communicate the geographical distribution of statistical information such as population density, average household income, or the median year of building construction. Choropleth maps show statistical variation among enumeration units. In general, there are two types of choropleth maps: classed choropleth maps divide enumeration units into a set of discrete categories (classes) based on their values. Each class is associated with both a range of values and a color on the map. Unclassed choropleth maps assign a range of colors to the enumeration units' range of values. In all but their most esoteric forms choropleth maps show only the mean or some other measure of central tendency for each enumeration unit. These measures of central tendency can communicate important information about variation between enumeration units, however, they mask variation within enumeration units.

Variation between enumeration units can be understood through the framework of "composition" and "configuration" (Boots 2003). Where composition refers to the overall distribution of the means (or other values) for the set of enumeration units and configuration refers to the arrangements of those values in space. Accurate representation of both the composition and configuration of the data requires balance. Fidelity to the composition of the data may require a choropleth map with many classes however, as the number of categories increases the configuration of the data can become obscured. To allow readers to understand the configuration, or patterns present in the data, it may be necessary to constrain the number of classes on the map. Downward pressure on the number of categories creates intra-class variation in maps, each class on the map may represent a wide range of values and therefore may mask variation among enumeration units. Stewart and Kenneley (2011) propose addressing the problem of intra-class variation by using "Illuminated" maps, that use the third dimension to represent intra-class variation.

This paper explores a different problem, intra-enumeration district variation. Whether classed or unclassed, choropleth maps present statistical summaries for enumeration units. Enumeration units are used to statistically summarize the things that occur or reside inside of them – households, people, businesses, crime, etc. Typically, information about the things internal to the census units is unavailable, individuals or events are aggregated into a single district-level statistic. The utility of enumeration district level statistics depends, in large part, upon the internal composition of the elements being summarized. For example, consider an enumeration unit with 1000 residents, 500 of whom earn, on average, $15,000/year ($\sigma = $7500$) and the other half earn, on average $85,000 a year ( $\sigma=$20000$). This enumeration unit might have a median income around $40,000 and a mean income around $50,000 (figure 1, solid line is the mean, dashed line is the median). These measures of central tendency describe the living conditions of very few actual residents of the enumeration district. The value presented on the map is a poor representation of the living conditions of the residents of the area, most earn much more or much less than the tract-level mean.

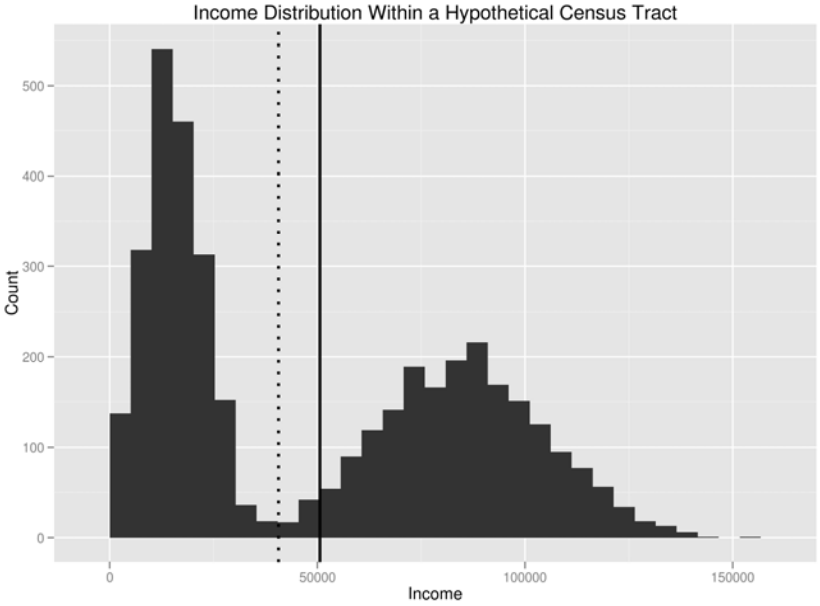


FIGURE 1: Income distribution within a hypothetical census tract. Mean indicated by solid line, median by the dashed line. Measures of central tendency, like the mean, are poor descriptors of this tract's population.

The internal dynamics of enumeration units affects the fidelity of maps to the variables and the places they describe. However, these internal dynamics are not typically disclosed by statistical agencies. Since disaggregated social statistics are seldom made available at a spatial resolution that would allow examination of the composition and configuration of enumeration zones, little is known about the “ground truth” of choropleth maps based on census data.[1]

Concerns about the fidelity of enumeration level summaries to the real world are often dismissed through reference to the ecological fallacy. The ecological fallacy is a logical fallacy, developed by Robinson (1950). The essence of the fallacy is assuming that the properties of a population at one level of aggregation apply at other levels of aggregation. One should not attribute the characteristics of an enumeration zone (like a tract) to the residents of the zone because such attribution constitutes a logical fallacy. In cartographic practice, the magnitude of the ecological fallacy is related to the magnitude of variation in the characteristic of the persons, events, or buildings that occur within the enumeration district. For example, if only the lower income group in figure 1 resided in the tract, the tract level mean of $15,000 would be a reasonable description of the tract.

Whereas the internal composition of enumeration units relates to the ecological fallacy, the internal configuration values raises questions about the design of the zones used to report census information. It is perhaps an overstatement to say that little is known about variation within census tracts. The US Census Bureau publishes detailed cross tabulations for many variables. These cross tabulations provide substantial insight into the composition of tracts but shed little light on the arrangement (configuration) of groups within tracts. Both the composition and configuration of census tracts are poorly understood and have important implications for both the ground truth of maps and design of zones used in choropleth maps.

There is some evidence that important social indicators are highly variable at the sub-tract scale. In 1985, 1989 and 1993, the American Housing Survey did a special analysis of one percent of the dwelling units in the core urban sample. This one percent sample was selected as seeds for “micro-neighborhoods.” For each household in the one percent sample, its ten nearest neighbors were also interviewed, forming a micro neighborhood of proximal households. These highly clustered samples provide a picture of micro-neighborhood composition. Hardman and Ioanniedes (2004) examined these micro-neighborhoods and found that the coefficient of variation on household income was .87 and .85 in 1985 and 1993, respectively. For example, in 1985, across all micro-neighborhoods, the mean income was $29,755 and the standard deviation was $25,937. In earlier work Ioanniedes (2002) established that these AHS micro-neighborhoods were a nationally representative sample. The AHS data contains little geographic information (other than the MSA of the micro-neighborhood). Nonetheless, the data provide evidence of significant intra-tract variability in both the composition and configuration.

This paper exploits a unique data source collected from a random sample of census tracts in the Chicago region in the mid-1990s. This data source provides block-face level information about the built and social environment in Chicago and allows us to examine

[1] The U.S. Census Bureau does provide the Public Use Microdata Sample (PUMS), which contains detailed data on individuals. However, the corresponding enumeration areas are home to at least 100,000 people as opposed to census tracts that represent approximately 4,000 residents.

intra-tract heterogeneity. We explore the micro scale spatial structure of six variables including:

- Amount of trash visible on the block face (ordinal scale): Presence of trash is coded in six levels: none (12.5% of observations), very light (36.5%), light (25.1%), moderate (15.4%), heavy (5.8%), and very heavy (3.8%).

- People visible on the block face (yes/no): The presence of people on the street is a dichotomy. Just over half of streets (53.2%) did not have people visible, 46.3% had people visible.

- Beer bottles visible on the block face (yes/no): The presence of empty beer bottles on the street is coded as a dichotomy. 24.6% of streets had visible empty beer containers and 74.1% did not.

- Cigarette butts visible on the block face (ordinal scale): Cigarette butts visible on the street is an ordinal variable coded in four levels: none (28.6% of block faces), “yes, few” (54.7% of block faces), “yes, fair number” (12.6% of block faces), and “yes, everywhere” (3.1% of block faces).

- Amount of traffic visible on the block face (ordinal scale): The traffic variable is coded in six levels: none (35% of observations), very light (26.6%), light (16.1%), moderate (13.8%), heavy (5.2%), and very heavy (2.2%).

- Condition of the street (ordinal scale): Street condition is coded in five levels: under construction (1.5%), very poor (10%), fair (53.7%), moderately good (25.9%), and very good (8.1%).

These variables represent key aspects of the neighborhood including visible signs of disorder, street life, and the built environment.

## Data

The Project on Human Development in Chicago Neighborhoods was a longitudinal study of how neighborhoods, families, and schools affect children’s health and development. A large scale Structured Social Observation (SSO) was conducted as part of this study. During the summer and fall of 1995 80 neighborhoods in Chicago were systematically observed by a 4 person team (a videographer, 2 observers, and a driver) traveling in a vehicle at low speed down each street in the sampled neighborhoods. The data collection teams worked 14 hours a day over a 4 month period, observing a total of 23,816 city blocks, we refer to the resulting data as the SSO.

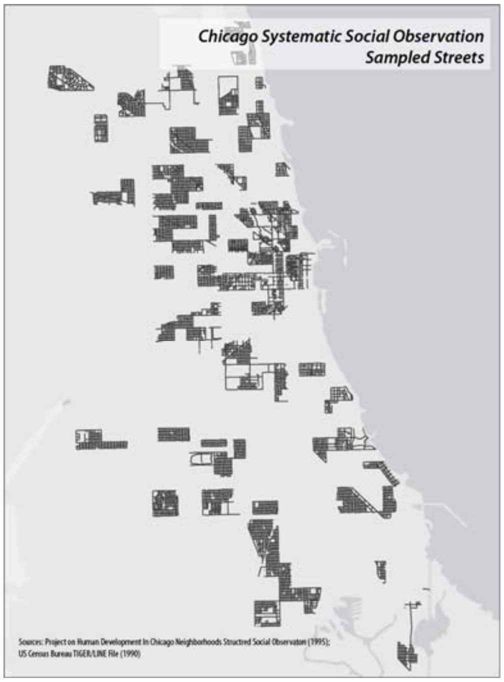
The data analyzed here were collected by the two observers traveling in the vehicle. As the vehicle traveled a street each observer, by looking out of a different side of the vehicle, was able to observe a face of a different block. These observers recorded measures describing land use, the condition of streets (e.g. presence of litter), pedestrian and vehicular traffic, and behavior of people visible on the street. Complete observation logs are available for 22,418 block faces.

These data have not previously been mapped, and we began by creating a procedure to do this. The coded observer logs include tract and block-level census identifiers. The logs indicate the tract and block that was respectively on the left and right side of the vehicle as it drove down the street, but they do not identify the street name, the address range of the face block or the cross streets at its ends. To deduce both the street and direction the vehicle was traveling we joined the observations logs to the 1990 raw census T.I.G.E.R. Line files that define "edges" – the basic building block for all census geographic data products. An edge is a line that can be defined by either visible features on the landscape (a street, a river) or invisible features (a county boundary). For the 1990 census a single "all edges" file includes the definition of all streets, tracts, blocks, counties, water bodies, etc. In addition to a geometric description of the coordinates (nodes) defining the edge, the file identifies the census geographic units on either side of the edge. This information allows us to merge the SSO observation logs to the raw census data. In a PostgreSQL spatial database we merged the observation logs and the 1990 census edges file by creating a common unique identifier that encoded both the census tract and block on the left and right of a given edge and whether the observer's right/left corresponded to the right/left in the all edges files. The end result of the mapping is a file that describes all of the streets that were observed in the SSO using two pieces of information. A single "edge" (street centerline) contains information about the block face on its left and right side.

We have analyzed a set of 6 variables selected from the 1995 Chicago SSO. Figure 2 is a map showing the SSO study areas in Chicago and illustrating the sampled streets.

## Methods:

The overall goal of the methods used is to illustrate the configuration and composition of Chicago block faces at both the disaggregate and aggregate level. Census tract-level composition is compared to the disaggregate composition using six variables. The configuration of block face characteristics is examined with two novel visualizations. No single method allows the description of both composition and configuration at multiple geographic scales. Therefore, we apply a suite of methods to characterize the micro-scale variability in the built and social environment.



Chicago Systematic Social Observation
Sampled Streets

Sources: Project on Human Development In Chicago Neighborhoods Structred Social Observaton (1995); US Census Bureau TIGER/LINE File (1990)

FIGURE 2: Streets Sampled by the Project on Human Development in Chicago Neighborhoods Structured Social Observation.

Spatial analysis techniques for qualitative data, like the Chicago SSO, are generally underdeveloped (Boots 2003). The Chicago SSO variables are coded as either ordinal or binary. To facilitate analysis the ordinal variables are conceptualized as representing discrete measurements of continuous phenomena. For example, the variable describing the number of cigarette butts on the street has four levels, ranging from “none” to “yes, everywhere”, based on the belief that the number of cigarettes butts on a street is in fact a continuous variable we compute tract-level means of both the ordinal and binary variables from the SSO.

### *Composition: Block Face and Tract-level Means*

Composition refers to the distribution of values within a tract, it is not a spatial concept, in that we are not concerned where values occur within a tract. To illustrate composition tract level histograms and means are computed for each variable using the block face data. In figure 3 a random sample of 10 tracts is drawn from all tracts that contained at least 50 block faces. Each of the sampled tracts is described using a series of six plots, each showing the distribution of the block face data for one variable. Each subplot contains the tract level mean for each variable (figure 3).

### *Configuration: Similarity as a function of distance*

Configuration refers to the arrangement of values within a tract. It is explicitly spatial, tracts with similar composition could have very different geographic configurations of observed values. The composition of a tract is related to its boundaries, changing the boundaries of a tract can change its composition. However, the amount of change is related to the configuration of block face values. If block faces with similar values are geographically clustered than boundary changes can have a profound impact on composition. On the other hand, if block face values are randomly distributed in space changes to the boundaries of tracts will have little impact on composition. Composition affects the fidelity of choropleth maps to conditions on the ground, configuration speaks the possibility of defining enumeration units where the composition is well described by a measure of central tendency.

To measure configuration the characteristics of proximal streets are examined at a variety of scales scale profiles. If $Z$ is a property of some spatially distributed set of observations and observations $Z_i$ and $Z_j$ are separated by some distance $h$ we can think of the spatial autocorrelation at the scale $h$ as the similarity between $Z_i$ and $Z_j$. Classical measures of spatial autocorrelation such as Moran's I (Anselin 1995) are not designed for categorical data. Some efforts have been made to develop indicators of spatial association for categorical data (e.g. Boots 2003, Paez et al 2011) however these methods become cumbersome for polychotomous variables.

Conceptualizing measures of spatial autocorrelation and questions about the configuration of values in space, as questions about similarity simultaneously expands and simplifies the concept. There is broad literature on both the ontology and measurement of similarity. In an influential paper Tversky (1977) argues that measures of similarity should have several properties, among them matching and monotonicity. The idea of matching is that as the degree of similarity between two objects corresponds to the number of "matching" characteristics. One can measure agreement on a single quantitative dimension, but in the fullest sense of the word similarity should account for the degree of agreement among multiple dimensions. Monotonicity is the idea that as the number of matching elements increases the similarity among two objects increases. Gower (1971) developed a widely used measure of similarity that satisfies both the matching and monotonicity criteria, we have implemented this metric as a measure of spatial autocorrelation but due to space constraints we focus on two simple intuitive statistical visualizations which relate only to Tversky's matching criterion.

## Topological Distance Profiles

Topological Distance Profiles (TDPs) are a qualitative analogy to the geostatistical variogram (Isaaks and Srivastava, 1990). TDPs are graphs that show the similarity of each street segment to its neighbors at various distances, where distance is measured by the number of intersections between block faces. We use the term 1st order neighbors to describe block faces that are 1 intersection away from a street, 2nd order neighbors to refer to block faces that are two intersections away, and so on up to the 10th order. The 2nd order neighbors are not inclusive of the 1st order neighbors (i.e., the 1st and 2nd order neighbors are non-overlapping sets).

To construct TDPs 1000 streets are randomly selected. Then for each of these street segments we identified all neighboring block faces from the 1st to the 10th order. We calculated the average value at each scale (treating ordinal scales as though they were interval). The smallest scale, the focal street itself, consists of a single block face; as the scale increases the number of block faces averaged tends to increase until around the 6th order, then typically decreases due to the irregular nature of the sample (figure 2). Each TDP is composed of many blue lines and a small number of summary red lines (figure 4). Blue lines represent sampled street segments and their neighbors. The position on the vertical axis at each order of distance is the average value of a variable for all neighbors of the sampled street. The red lines are the average value of all of the blue lines at each order of distance. Hence, TDPs present central tendency (the red line) and variance (the blue lines).

### Random Walks

Whereas the Topological Distance Profiles present the average of a large set of connected street segments, random walks represent one possible path through a set of 10 connected street segments. We construct random walks by randomly sampling 300 streets, each of which forms the seed for a walk. The walk progresses randomly but is subject to some constraints, a walk cannot visit the block face it visited in the previous two steps. TDPs average over a large area and thus mask the block-to-block variability the random walks are meant to illustrate.

## Results

Figure 3 shows the distribution of values and the means for each variable for a sample of 10 tracts. Each row in the plot represents a census tract, each column is a variable.

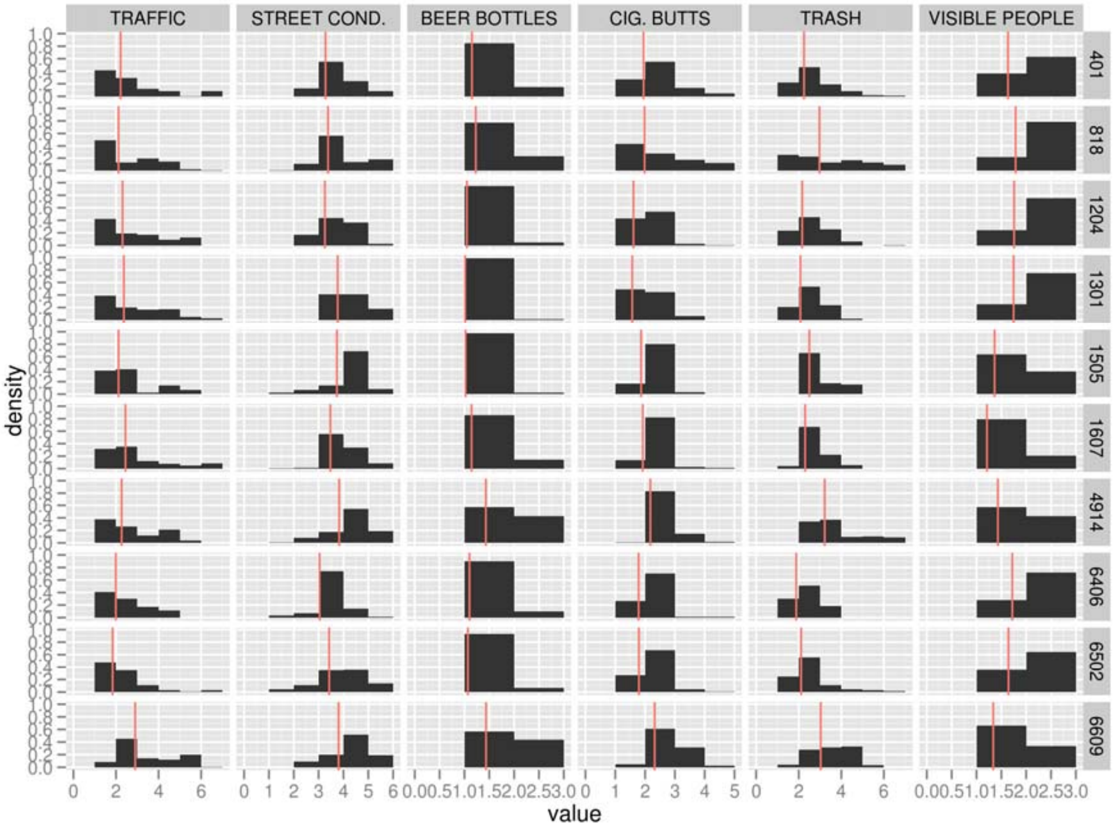


FIGURE 3: Distributions of variables and tract means for six SSO variables. Each row corresponds to a tract, each column a variable.

It is apparent in figure 3 that many tracts have long tailed and/or bimodal distributions. For example, the amount of visible trash in a tract varies substantially from street to street. For tract #401 and #818 (top two rows) there is a broad spread around the mean, making the mean a poor representation of the amount of trash on a randomly selected street. However, in tracts #1505 and #1607 the mean seems to be a reasonable summary of the distribution, with most streets no more than one level on the ordinal scale away from the mean. Reading horizontally across figure 3 it seems some tracts, such as #818 are characterized by high levels of variability across all variables and others, such as #6406, have relatively little variation. Based on figure 3 (and many other similar figures) it is hard to generalize about variance around tract-level measures of central tendency. In some places and for some variables the mean provides a fair picture of tract characteristics, in other places it does not.

Figure 3 focuses strictly on the composition of tracts, the Topological Distance Profiles (TDPs, figure 4) and Random Walk Charts (figure 5) aim to characterize the configuration of block face variables. We have not found a summary statistic that can capture the patterns on these visualizations, and therefore we have to interpret the patterns as shown in the figures.

Figure 4 illustrates the configuration of two variables at the block face level using a topological distance profile. Every blue line represents a sampled street segment, in the top plot each sub-plot represents a random sample of street segments that have been observed to have “none,” “very light,” “light,” “moderate,” “heavy,” and “very heavy” amounts of trash visible. The position on the vertical axis is the average value on this neighborhood characteristic of all of the block faces at each order of distance. The red lines are the average value of the blue lines at each order of distance, hence we are presenting a central tendency and a variance around it by distance.

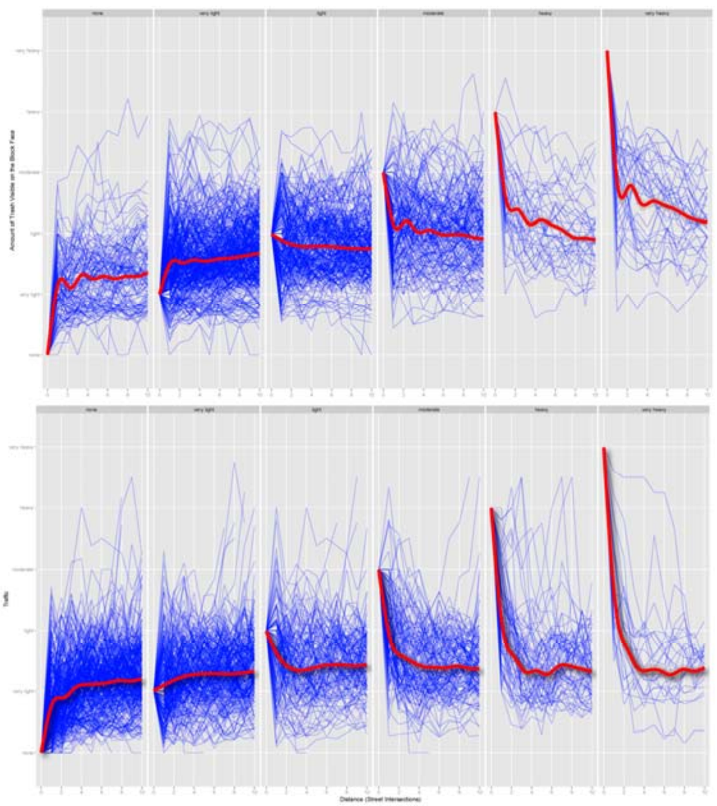


FIGURE 4: Topological Distance Profiles (TDPs); amount of trash on the block face (top) and amount of vehicular traffic (bottom). Each panel on the top and bottom plots correspond to block faces with different levels of trash and traffic, respectively. Vertical axis in each plot corresponds to the average value for the neighbors of each street at each scale. Red line shows the overall average. See text for a full description.

Figure 4 presents only two of the six variables we examined. However, they are representative of the two types of TDPs we observed. In both top and bottom of figure 4, there is a tendency for curves to flatten out as lags increase, but in the top there is a clear relationship between the starting and ending values, which is not as apparent in the bottom of the figure. In the top graph the slope of the curves flattens after two blocks. Overall, the amount of trash on a block face (figure 4, top) seems to be associated with the amount of trash 10 blocks away–a lower start point corresponds to a lower end point 10 blocks away, and vice versa. In contrast the amount of traffic seems to be local, the amount of traffic on a block face is associated with the amount of traffic on the street’s first and second order neighbors, but after two blocks most curves level off at the same mean regardless of starting level. There are patterns over small scales but these patterns attenuate quickly.

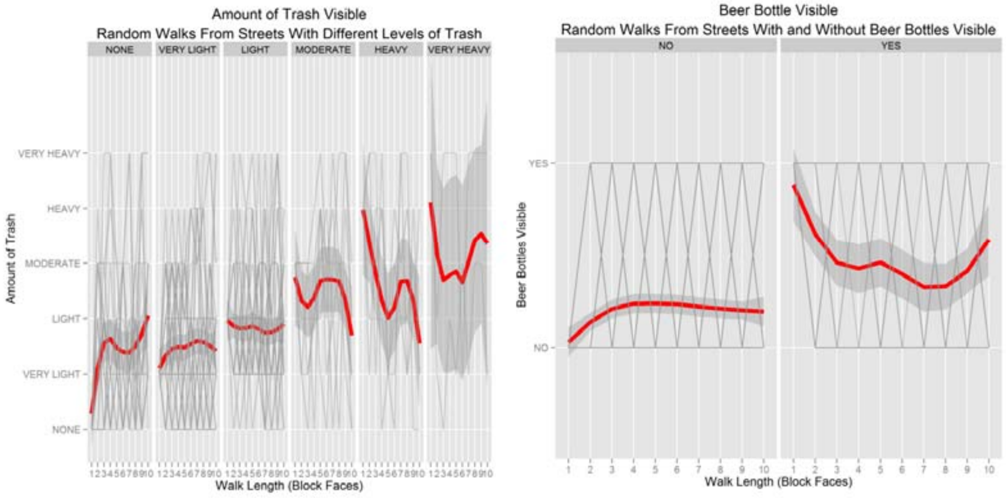


FIGURE 5: Random walks on the Chicago SSO block faces. Amount of trash visible (left) and the presence of empty beer bottles on the block face (right). Grey lines represent random walks (n=300) red lines are the average each at leg of the journey. Walks sorted into sub-plots based on the level of each variable at the initial (seed) block face.

Figure 5 shows the results of 300 random walks. Each walk is represented by a light grey line in the figure. The vertical axis represents the value of the trash (left) and beer bottles (right) variables, the horizontal axis shows the walk length. The red line on each plot represents the mean of all the walks at each scale, the ribbon (gray region) around the red line shows the 95% confidence interval around the mean. The random walk figures show more variability but are in general consistent with the patterns observed in the topological distance profiles.

## Discussion

In general, these graphs suggest three trends. First, we notice that in figures 4 and 5, in most sub-plots, the characteristics of streets very quickly level out, on average the characteristics of block faces that are 4 intersections away from a focal street tend to be similar to those that are 10 blocks away, this effect is most pronounced in figure 4. For most variables this flattening occurs after two blocks. We expect that with a larger sample size we might observe a similar signal in the random walks. This is an interesting result and suggest micro-scale structure at scales much smaller than the typical census unit.

The second general trend we observe is that some variables attenuate with distance while others do not. For non-attenuating variables, such as the amount of trash or presence of beer bottles, the level achieved after 2-3 blocks is determined by the condition of the initial street. That is, a street with beer bottles visible is more likely to be near other streets with bottles visible, even after many blocks. On the other hand attenuating variables are insensitive to initial conditions. Traffic is an attenuating variable, the amount of traffic on a block face seems to be related to the amount on neighboring block faces but only over small distances, after a few blocks the neighbors of a street with light traffic are the same as the neighbors of one with very heavy traffic. Some variables have a spatial echo, the attributes of streets persist over distance, for others they do not.

The third and final observation is that there significant variation in the composition of tracts, some tracts have more variance in the SSO observations than others. While these findings are limited to a small set of characteristics and a single city the patterns that emerged were quite strong. In both the compositing and configuration of variables there is significant micro-scale heterogeneity. There seems to be a characteristic scale for the variation in visible aspects of block faces in Chicago. This empirical regularity is striking, after a short distance the studied characteristics of the urban environment level-out. This "leveling-out" means that changes in the visible aspects of the urban environment happen quickly over one or two block and then change little (on average) over greater distances.

Some tracts have compositions that are not well summarized by measures of central tendency however, the configuration of block faces values suggests that defining more homogenous tracts would be difficult. Within our limited dataset the configuration of the block face data seem to correspond poorly to existing census geographies (though as figure 3 shows the problem is more pronounced in some places).

Unfortunately, SSO's are quite rare and data like those generated by PHCDN exist in few other places. There is significant variation in observable aspects of the built and social environment within census tracts and block groups. While we cannot generalize our findings beyond Chicago these results raise concerns about the fidelity of urban choropleth maps and are a reminder to remain wary of the ecological fallacy when using such maps.

## References

Anselin, L. (1995), Local Indicators of Spatial Association—LISA. Geographical Analysis, 27:93–115

Boots, B. (2003), Developing local measures of spatial association for categorical data. Journal of Geographical Systems, 5(2):139-160

Gower, J.C. (1971), A general coefficient of similarity and some of its properties. Biometrics, 27:857-74.

Hardman, A. and Ioannides, Y.M. (2004), Neighbors' income distribution: economic segregation and mixing in US urban neighborhoods. Journal of Housing Economics, 13(4):369-382

Ioannides, Y.M. (2002), Residential neighborhood effects. Regional Science and Urban Economics, 32(2):145–165

Isaaks, E. and Srivastava, M., (1990) An Introduction to Applied Geostatistics. Oxford University Press

Páez, A., Ruiz, M,, López, F. and Logan, J. (2012) Measuring ethnic clustering and exposure with the q statistic: an exploratory analysis of Irish, Germans, and Yankees in 1880 Newark. Annals of the Association of American Geographers, 102(1).

Robinson, W. S. (1950) Ecological correlations and the behavior of individuals. American Sociological Review, 15:351-357.

Stewart, J. and Kennley, P.J. (2010) Illuminated Choropleth Maps. Annals of the Association of American Geographers, 100(3):513-53.

Tversky, A. (1977) Features of similarity. Psychological Review, 84(4):327-352.

**Seth E. Spielman**, Assistant Professor, Department of Geography, University of Colorado, Boulder, CO 80309. Email <seth.spielman@colorado.edu>