Point Process Models for Household Distributions within Small Areal Units*

Zack W. Almquist[1,**] and Carter T. Butts[1,2]

[1] *Department of Sociology; University of California, Irvine*
[2] *Institute for Mathematical Behavioral Sciences; University of California, Irvine*
[**] *To whom correspondence should be addressed. Email: almquist@uci.edu.*

September 1, 2011

Abstract

Spatio-demographic data sets are increasingly available worldwide, permitting ever more realistic modeling and analysis of social processes ranging from mobility to disease transmission. The information provided by these data sets is typically aggregated by areal unit, for reasons of both privacy and administrative cost. Unfortunately, such aggregation does not permit fine-grained assessment of geography at the level of individual households. In this paper, we propose to partially address this problem via the development of point process models that can be used to effectively simulate the location of individual households within small areal units.

Keywords: Spatial Demography, Household Distribution, Areal Units, Point Processes, Simulation

*This work was supported in part by ONR award #N00014-08-1-1015 and National Science Foundation (NSF) award BCS-0827027.

1 Introduction

Spatio-demographic data sets are increasingly available worldwide, permitting ever more realistic modeling and analysis of social processes ranging from mobility to disease transmission. The information provided by these data sets is typically aggregated by areal unit (e.g., the State, County, Tract, Block Group, and Block hierarchy of the U.S. Census), for reasons of both privacy and administrative cost. Unfortunately, such aggregation does not permit fine-grained assessment of geography at the level of individual households, a scale that is potentially important for accurate modeling of micro-social processes such as transmission of disease between households, daily mobility patterns, or patterns of interpersonal contact. While the potential to model such phenomena across large geographical areas thus exists, efforts are hampered by a lack of data on household location.

In this paper, we propose to partially address this problem via the development of point process models that can be used to effectively simulate the location of individual households within small areal units. Given basic information such as number of households, general pattern of land use, and/or population of neighboring units, our objective is to identify a probability distribution over household locations within a polygonal region whose average spatial properties reflect the corresponding properties of the unobserved true household distribution in that region. Examples of targeted properties include standard point process descriptives (Diggle, 2003; Ripley, 1988), such the mean nearest neighbor distance, measures of spatial clustering (e.g. the $\mathcal{F}$ and $\mathcal{G}$ functions), mean $\mathcal{K}$ function value, et cetera. While the resulting distributions will not reproduce household locations with perfect fidelity, the approximations may nevertheless prove adequate for modeling of basic social processes.

While this problem can be approached in many ways, our focus within this paper is on the application of *simple, scalable* models that require no extra information (beyond areal unit and household count) from the analyst. Such models can be employed in virtually any setting, and are a natural starting point for any more complex modeling effort. To that end, we begin with two baseline models – a constant-intensity N-conditioned Poisson process, and a low-discrepancy sequence model – that incorporate only population density. We then extend the density-based models by incorporating additional information from the areal units themselves, using an inhomogeneous Poisson framework in which households are more likely to be found near polygonal borders (a common phenomenon in the observed data). To evaluate these simple point process models, we compare their behavior with observed household location distributions from three different communities. Test samples consist of household location data from Portland, OR, Deschutes County, OR, and Irvine, CA[1], with areal units given by the 2000 U.S. Census. All modeling is performed in `R` (R Development Core Team, 2010). Our test cases include examples of urban, suburban,

[1]Data from Deschutes County GIS office; City of Portland, OR GIS office, and Irvine, CA GIS Office.

and rural settings, with varying spatial scale and levels of population density.

Evaluation of the suggested point processes on our three communities suggest that that simple models can provide quite reasonable approximations to household location distributions for small areal units. Performance degrades substantially for larger units, although the inhomogeneous model shows some potential within more urbanized regions. Practical suggestions are given for the use of these and related point processes within large-scale simulations, and for applications of this technique to settings beyond the U.S. (and the developed world more generally).

2 Human Settlement Patterns and Baseline Models

Human settlement patterns play an important role in shaping human interaction and the demographic processes which result. A classic example is that of marriage in Western societies: couples in such societies rarely marry without prior meeting and extensive face-to-face interaction, and marriage is thus disproportionately propinquitous (Bossard, 1932). Many demographic processes, such as mortality, fertility, and mobility are also influenced by human settlement patterns (see, e.g. Binka et al., 1998; Freeman and Sunshine, 1976; Guilmoto and Rajan, 2001); however, exploiting such geographical information is frequently limited due to difficulty of acquisition. For example, in the United States information on population within aggregate areal units is largely available (e.g., via the US Census), but the coordinates of individuals and households are unavailable due to privacy concerns. There is thus a distinct need for a methodology to generate household (or individual) distributions over small scale areal units such as US Census geography, so as to inform statistical models agent-based simulations, and the like.

Adding to the difficulty of this problem is the need for plausible models to be easily computable. For instance, the year 2000 US census reports population in over 8 million blocks, themselves organized into well over 50,000 tracts (U.S. Census Bureau, 2001). Applying household location models at national or regional scales thus requires simulation of location distributions for large numbers areal units, making efficiency an important concern. In addition to computability, models to be used in a range of settings should be simple, robust, and require minimal information inputs on the part of the analyst. (For instance, a household location model requiring detailed street maps may be of limited use in historical applications, or in countries for which such maps are not readily available.) Such concerns motivate the initial consideration of highly minimal models, that employ as little information as possible, and that can be easily simulated for large numbers of areal units. Following Mayhew (1984), we regard *baseline models* (and minor extensions thereof) as a natural starting point. By beginning with basic, readily available information such as counts of households and areal unit boundaries, we first construct models that treat household placement as

conditionally uniform, subsequently modifying this assumption by introducing higher "evenness" in placement, and then by allowing household location probability to be affected by the geometry of the areal unit in which it resides. To the extent that the resulting models produce household distributions whose properties approximate those observed in real settings, we regard them as adequate proxies with respect to those properties. Where these simple models fail, they may nevertheless be used as a starting point for building more complex models (e.g., models with inter-point interaction, or additional covariates) for particular applications.

3 Background

Increasingly, large scale archival data sets containing administrative borders and population or household counts are available to demographic researchers (e.g., IPUMS: Minnesota Population Center, 2011; U.S. Census Bureau, 2001). This form of data, and the study thereof, is often known as spatial demography or the formal demographic study of areal aggregates (Voss, 2007). Most such data sets, however, rarely contain point location for individuals or households because of privacy and safety concerns. This is not a problem for many macro-level analyses such classic demographic projection; however, for more micro-social processes such as transmission of disease between households, daily mobility patterns, or patterns of interpersonal contact require more detailed knowledge of household placement. Given that it is often difficult and sometimes impossible to obtain exact household locations, an alternative approach is necessary. One solution is to develop a series of point process models that simulate the individual household distribution for these small areal units with known statistical properties. Given basic information such as number of households, general pattern of land use, and/or population of neighboring units, the general objective is to identify a probability distribution over household locations within a polygonal region whose average spatial properties reflect the corresponding properties of the unobserved true household distribution in that region.

3.1 Spatial Data

Spatial information associated with spatio-demographic data includes, but is not limited to, points (single locations, e.g., a house), lines (e.g., a road), and polygons or areal units (e.g., a Census Block). Typically, Geographic Information Systems (GIS) are employed for handling and performing analysis on a myriad of spatial data (Reibel, 2007); in particular, this includes linking spatial coordinates to socio-economic and demographic data. For the present problem, the two most important spatial units are those of the point and the polygon. A point consists of X and Y coordinates (e.g., longitude and latitude, or a projection thereof into the plane) and a polygon represents a series of line segments (again in either latitude/longitude or planar coordinates) identifying a closed region on the Earth's surface. Because of

the curvature of the Earth's surface, most map-based and related calculations are based on points and polygons that have been *projected* onto a plane; the choice of map projection can have non-trivial effects for such important measures as interpoint distances and polygonal areas, and thus must be chosen carefully. Fortunately, when working with small areal units such as those employed in this paper, distortions due to projection are easily overcome (e.g., by using orthonormal projections about the centroid of the areal unit). More details on choice of projection and coordinate system can be found in Snyder (1987).

3.2 Household Distributions

There exist a plethora of reasons to be interested in the distribution of human populations over space, and particularly the location distribution of human households. Humans have since prehistory gathered together in small groups (often kin groups) to manage their livelihoods (McC. Netting et al., 1984), and we loosely refer to a group of persons residing at the same location and sharing resources a *household*. In the modern context households are often studied as units of decision making (e.g., Davis, 1976), used in the study criminology (e.g., Hipp et al., 2011; Short et al., 2010), as well as units for disease and information spread (Salathé and Jones, 2010), et cetera. Here, we focus on the household as our basic unit of interest. The study of household activities over spatially diverse contexts has been performed primarily through the concatenation of administrative data (e.g., censuses) and spatial data (e.g., surveys or sensors) to make various predictions, forecasts and simulations for scientific and public policy reasons (Jefferson Fox and Mishra, 2003). It is common to use spatial data at a largely aggregate level (e.g., a US census tract), and this has allowed for much scientific progress; however, reliance on aggregate data raises concerns regarding the risks of fallacious ecological inference (Gibson et al., 2000) and the modifiable areal unit problem (Openshaw, 1984). Another issue with aggregate data is that it does not allow for certain types of analysis necessary for social science, public health or demographic research. Here, we are particularly concerned with the situations where one cannot conduct one's analysis without household-level spatial information, such as modeling of transmission of disease between households, daily mobility patterns, or patterns of interpersonal contact. Because administrative and archival data often lacks individual or household locations, we propose in this research to use point process probability models to simulate household distributions over administrative polygons which maintain key statistical properties of interest.

4 Point Process Models and Simulation

A point process is defined mathematically as a random element whose values are *point patterns* on a set S. This can be defined more technically, but for our purposes it

is sufficient to think of a point pattern as a countable subset of S that has no limit points.

The most important and basic point process model is the *spatial Poisson process*. The following development follows that outlined in Diggle (2003) and is one of the standard descriptions of Poisson or planar Poisson processes. A spatial Poisson process is equivalent to a standard (or temporal) Poisson process with some known rate function where one is associating each event a random vector $(x, y) \in S$ sampled from some probability density function. More formally a *homogeneous planar Poisson process* may be defined under the following conditions:

i) For some $\lambda \geq 0$, and any finite planar region S, $N(S)$ – the number of events with corresponding vectors in S) follows a Poisson distribution with mean $\lambda|S|$, where $|S|$ is the area of S. (Note that, here λ is called the *intensity* of the process.)

ii) Given $N(S) = n$, the n events in S form an independent random sample from the uniform distribution on S.

It is worth pointing out that $\lambda|S|$ is the integral of λ over S, and thus to acquire an inhomogeneous planar Poisson process, one need only replace the constant λ by a spatially dependent intensity $\lambda(x, y)$ – replacing condition i with "for some $\lambda(x, y) \geq 0$, and any finite planar region S, $N(S)$ follows a Poisson distribution with mean $\int_S \lambda(x, y)dxdy$."

The homogeneous Poisson process is only one of a wide range of point processes that may be employed to simulate household location distributions. Here, we employ three variant processes, the first of which is an application of the uniform or homogeneous process (conditioned on region boundaries and observed population), the second of which is a deterministic low-discrepancy process that behaves much like a uniform distribution (but tends to place households away from one another), and the third of which is an inhomogeneous Poisson process whose intensity function (λ) depends on proximity to unit boundaries. We now consider each of these processes in turn.

4.1 Constant-intensity N-conditioned Poisson Process Model (Uniform)

The constant-intensity N-conditioned Poisson process model (from here on referred to as the *Uniform* model) is a maximum entropy solution in which households (or individuals) are placed uniformly at random subject to known geographical constraints (e.g., tract borders). This is commonly known as Complete Spatial Randomness (CSR) or Spatial Poisson process and is the most basic point process model (an example may be seen in Figure 5(c)).

4.2 Low-discrepancy Sequence Model (Quasi-random)

The low-discrepancy sequence model (henceforth referred to as the *Quasi-random* model) is a near-minimal entropy solution in which households are placed in an extremely even, "grid-like" manner using a two-dimensional Halton sequence. A Halton sequence is a deterministic sequence of points that "fills" space in a uniform manner, while also maintaining a high nearest-neighbor distance. The result (sometimes called a "quasi-random" distribution) is similar to a set of draws from the uniform distribution, but substantially more evenly placed (see Gentle, 1998, for algorithmic details). An example may be seen in Figure 5(d).

4.3 Inhomogeneous Poisson Process Model (Attraction)

The inhomogeneous Poisson process model (henceforth referred to as the *Attraction* model) is one in which assume points are distributed such that they cluster around polygon boundaries. This is controlled by a given *point potential* function defining the intensity $\lambda(x, y)$. We consider two forms for the potential function, which are defined as follows. Let $\mathbb{Z}$ be a collection of line segments (indicating boundaries of the areal unit, internal polygons (such as subsidiary unit boundaries), or elements such as roads), and let $d((x, y), z)$ for $z \in \mathbb{Z}$ be the minimum distance between the point (x, y) and the line segment z.

$$\lambda(x, y) = \max_{z \in \mathbb{Z}} (1 + \left| \frac{d\left((x, y), z\right) - o)}{s} \right|^{e}, \tag{1}$$

where s is a scale factor, o is an "optimum" distance, and e is an exponent. Generally the parameters s, o, and e are selected so that $s > 0$, $o \geq 0$, and $e < 0$. Intuitively, the resulting point potential attracts points to polygon boundaries (or, more generally, the elements of $\mathbb{Z}$), with maximum intensity occurring when one is at distance o from a line segment. This definition is motivated by the frequent use of roads, waterways, or other similar physical elements as boundaries of areal units: housing units are often located along such features, but are frequently offset by some amount. For an example of this process see Figure 5(b).

Although the parameters of λ may potentially be inferred from data via likelihood-based methods, we are interested here in the heuristic setting in which the potential must be employed with limited fine-tuning. Given this, we set $o = 0$ and used a crude grid search to find values of s and e that produced highest average p-values over all aggregated cases in the test data (described below). This resulted in parameter values of $s = 0.00015$ and $e = -1.5$ (with the former in angular units. Experimentation suggested that the results reported here are reasonably robust to these settings, and minor changes do not greatly change the resulting point patterns.

5 Standard Statistical Measures for Point Processes

In order to compare the distribution of household locations arising under our models to those empirically observed, we require appropriate descriptive statistics. Here, we describe several standard descriptives from the point process literature, that may be employed to assess the extent to which simulated household distributions do or do not deviate from their empirical counterparts.

5.1 Ripley's $\mathcal{K}$ Function

Ripley's $\mathcal{K}(s)$ function (sometimes called the reduced second moment measure) is a tool for analyzing completely mapped spatial point process data (Diggle, 2003). These are usually events recorded in two dimensions, but they may be locations along a line or in multidimensional space (e.g., households within a city block). Intuitively, the K function expresses the degree of spatial clustering among points, at multiple scales – more specifically, the tendency for other points to appear within distance s of an arbitrary realized point.

5.1.1 Theoretical $\mathcal{K}$

The $\mathcal{K}$ function is defined as:

$$\mathcal{K}(s) = \frac{1}{\lambda} E\left[\text{number of other events within distance } s \text{ of a randomly chosen event}\right], \quad (2)$$

where λ is the density (number per unit area) of events; thus, $\mathcal{K}$ describes characteristics of a point process at different distance scales. Note that many alternative standard measures such as the nearest neighbor methods (see Section 5.2) do not have this property. $\mathcal{K}$ is generally the preferred characterization of spatial point process by statisticians and geographers (see, e.g. Diggle, 2003), and we use it as the basis of our empirical investigation in Section 8.

5.2 Nearest Neighbor Measures

In addition to the variation in conditional density through space, one can also consider point processes in terms of their nearest-neighbor properties. Here, we comment on two functions of this sort that are of potential utility in assessing point pattern adequacy.

5.2.1 $\mathcal{G}$ Function

The $\mathcal{G}$ function measures the distribution of the distances from an arbitrary event to the nearest other event (see, Diggle, 2003). Usually these distances are denoted $d_i = \min_j\{d_{ij}\ \forall j \neq i\}$, $i = 1, \ldots, n$, so that the $\mathcal{G}$ function is is

$$\mathcal{G}(r) = \frac{\#\{d_i \ : \ d_i \leq r_i, \ \forall \ i\}}{n}, \tag{3}$$

where the numerator is the number of elements in the set of distances that are lower than or equal to d, and n is the total number of points.

5.2.2 $\mathcal{F}$ Function

The $\mathcal{F}$ function measures the distribution of all distances from an arbitrary point of the plane to the nearest realized event (see, Diggle, 2003). Bivand et al. (2008) notes that this function is often called the *empty space* function because it is a measure of the average space left between events. (Note the contrast with $\mathcal{G}$, in which the focal point is itself a realized event.) The $\mathcal{F}$ function of a stationary point process X is the cumulative distribution function $\mathcal{F}$ of the distance from a fixed point in space to the nearest point of X. Under CSR, $\mathcal{F}$ is:

$$\mathcal{F}(r) = 1 - \exp\left(-\lambda \cdot \pi \cdot r^2\right). \tag{4}$$

6 Comparison Data: US Census Geography and Household Parcel Lots

To evaluate the above models, we seek to compare their resulting simulated household distributions from those encountered in realistic settings. Although household location data is difficult to obtain, we are able to employ parcel data from three US communities for testing purposes. While not representative of all communities worldwide, we view these three cases as a "proof of concept" for the wider use of settlement pattern imputation from simulation models like those employed here.

6.1 US Census Geography

Our basic source of geographical information is the year 2000 US census. "The United States Census is a decennial census mandated by the United States Constitution. The population is enumerated every 10 years and the results are used to allocate Congressional seats (congressional apportionment), electoral votes, and government program funding" (U.S. Census Bureau, 2001). The data collected in the decennial census has since 2000 been made available to the public as spatial polygon data broken down into three key designations: Tract, Block Group, and Block, each representing different levels of human population aggregation. The Block represents household or individuals aggregated at the level of city block (if the population density is sufficient not to

jeopardize an individuals privacy) or larger unit; Block Groups represent an aggregation of Blocks, and Tracts represent an aggregation of Block Groups (U.S. Census Bureau, 2001). This data is made available through the US Census website[2], and through statistical software such as the `UScensus2000` `R`-package (Almquist, 2010).

6.2 Household Distribution Data in the US

There is limited access to household data in the United States, and this can be even more difficult in other countries. In some cases, however, household-level geospatial data may be acquired from cities and counties across the US that is collected for purposes of local or state property tax administration. This household data available is known as *parcel data*, and is either maintained as Shapefiles or simple longitude/latitude point files; typically this data is difficult and time consuming to acquire when available.[3] To provide an empirical comparison set for our point process models, we have acquired three different sets of parcel data within the US: an urban setting (Portland, OR), a suburban setting (Irvine ,CA), and a rural setting (Deschutes County, OR). For an example see Figure 5(a). Although a more general, representative sample of parcel data is not available at this time, the range of urbanization in our three cases provides some suggestion of how model performance might vary across similar communities in the United States or other countries with comparable settlement patterns.

6.3 Urban, Suburban, and Rural Classification

The US Census classifies areas as either *urban* or *rural*. Urban areas are broken into two classifications: *Urbanized Areas (UA)* – a continuously built-up area with a population of 50,000 or more; and, *Urban Places Outside of UAs* – an urban places is any incorporated place or census designated place (CDP) with at least 2,5000 inhabitants. The rural designation is defined as follows: A territory, population, and housing units that the Census Bureau does not classify as urban are classified as rural (U.S. Census Bureau, 2001).

We extend the US Census Urban/Rural classification to include a notion of *suburban*. "Suburban areas are typically considered to be regions of lower density residential land use at the urban fringe, and are often thought to be synonymous with sprawl, but there is no standard quantitative definition" (Theobald, 2004). The notion of Suburbia is old and is found in the Sociology literature as far back as 1943 (Harris, 1943). In this case we use the concept of suburb to represent a city which is less dense than urban center, not a proper Metropolitan Statistical Area (MSA) into itself (e.g.,

[2] `www.uscensus.gov`

[3] This data may also be expensive, because it is created by local area governments and then sold to local area development firms.

Los Angeles MSA; U.S. Census Bureau, 2001), and that is contained near a large Metropolitan Area.

With this classification in mind, we briefly consider our three cases in turn.

6.3.1 Urban: Portland, OR

Portland, Oregon is a city with an estimated population of 529,121 people and estimated household population of 223,737 (U.S. Census Bureau, 2001). The local city government of Portland has parcel data for 248,325 households[4]. Portland is the largest city in Oregon and represents the economic center of the state. The city also contains the largest University in Oregon, and its suburbs include the large business such as Nike and Intel. The US Census classifies Portland as *urban* (see Table 1: U.S. Census Bureau, 2001). A visual portrayal of the household distribution of Portland overlaid on US Census Blocks, Block Groups and Tracts may be seen in Figure 1.

	Portland Oregon
Urban:	527,255
Rural:	1,866
Total:	529,121

Table 1: Portland, Oregon Urban/Rural classification by the US Census in 2000.

6.3.2 Suburban: Irvine, CA

Irvine, California is a city with an estimated population of 143,072 people and estimated household population of 51,199 (U.S. Census Bureau, 2001). The local city government of Irvine has parcel data for 49,002 households[5]. The US Census classifies Irvine as *urban* (see Table 2: U.S. Census Bureau, 2001). For the purposes of this research we classify Irvine as a *suburban* city, as it is less dense than Portland, does not represent an MSA and is close in proximity to the significant MSA of Los Angeles. A visual portrayal of the household distribution of Irvine overlaid on US Census Blocks, Block Groups and Tracts may be seen in Figure 2.

6.3.3 Rural: Deschutes County, OR

Deschutes County, Oregon is a county with an estimated population of 115,367 people and estimated household population of 45,595 (U.S. Census Bureau, 2001). The local

[4]Note this is the population we employ here; due to demographic changes, the parcel data contains more households than were present in the 2000 census.

[5]Note this is the population employed here, and is slightly smaller than the household count in the 2000 census.

	Irvine California
Urban:	143,011
Rural:	61
Total:	143,072

Table 2: Irvine, California Urban/Rural classification by the US Census in 2000.

county government of Deschutes has parcel data for 70,293 households[6]. The US Census classifies Deschutes County as mix of *rural* and *urban* (see Table 3 U.S. Census Bureau, 2001). The *urban* portion of the county is Bend, OR (and few outlying areas around Bend) a city of 52,029 in 2000 (see Table 3: U.S. Census Bureau, 2001). Deschutes County is used primarily for it rural nature. A visual portrayal of the household distribution of Portland overlaid on US Census Blocks, Block Groups and Tracts may be seen in Figure 3.

	Deschutes County Oregon
Urban:	72,554
Rural:	42,812
Total:	115, 367

Table 3: Deschutes County, Oregon Urban/Rural classification by the US Census in 2000.

7 Comparison Measure

The evaluation of our proposed household location models involves the comparison of two point distributions: that of the observed household distribution and that of the simulated household distribution. The literature in applied spatial analysis has tended to focus on the comparison of point distributions over two (or more) time points rather than the comparison of two different point processes. The most common examples are in the ecological literature, especially dealing with trees (for a good review see, Perry et al., 2006). However, as we are comparing two different point distributions (i.e., not emanating from a temporal process) we apply Diggle and Chetwynd's (1991) recommendation of using the sum of normalized difference of Ripley's $\mathcal{K}$ statistic at m breaks.

[6]This population substantially larger than the 2000 count, likely due to considerable growth in Bend, OR (the largest city in the county) between 2000 and 2010.

$$D(s) = \mathcal{K}_1(s) - \mathcal{K}_2(s)$$
$$D = \sum_{k=1}^{m} \frac{D(s_k)}{\text{var}(D(s_k))} \tag{5}$$

The numerator is sometimes known as *Diggle's D*. To test whether the two distributions are different we apply *Monte Carlo (MC) tests for spatial patterns* (Besag and Diggle, 1977).

A MC test consists of ranking the value of a statistic computed on observed data amongst a corresponding set of statistic values generated by random sampling from a null distribution. In this case the null distributions are our three proposed models (Uniform, Quasi-random, and Attraction), with our aim being to assess the extent to which the distributions of D under these models cover the D values of the observed data.

Note that under mild conditions this test determines an exact significance level and that the number of simulations, k, can be quite small.[7] We call the resulting p-value, an *MC-pvalue*. In this research we will not be interested in the MC-pvalue in the traditional sense, but in its inverse. In other words, we are interested in the case when the two distributions are not strongly distinguishable. We will therefore use a standard α level of 0.05 (or really 0.025 for a two tailed test) to determine whether the two point processes are sufficiently different to be considered effectively distinct.

8 Analysis and Results

To evaluate our proposed models, we simulate distributions for samples of polygons from each of our three cases, comparing those distributions against the observed data via the MC test of the D statistic (as shown above). Here, we briefly describe software and procedural issues, before turning to our findings.

8.1 Software

All code for this paper was written in the `R` statistical programming language (R Development Core Team, 2010). `R` is among other things a powerful GIS tool (see, Bivand et al., 2008). To perform the analysis functions from `spatstat` (Baddeley and Turner, 2005), `networkSpatial` (Butts and Almquist, 2011), `splancs` (Rowlingson and Diggle, 1993), `rgdal` (Keitt et al., 2009) and `UScensus2000`-suite of packages Almquist (2010) were employed.

[7] Due to computational complexity of this problem $k = 40$ for this research.

8.2 Comparison of Point Distributions

For computational reasons, we chose to perform our Monte Carlo D test on a population-weighted subsample of areal units from each level for each test case. Th sample size for each level/case combination was 100, if 100 units were available; otherwise, all units in the specified level/case combination were used.

For each polygon in each sample, we perform an MC D test for each of the three proposed models. For each such test, we regard the observed data as adequately covered by the model if the D statistic lies within the central 95% simulation interval produced by the model in question.[8] To assess overall adequacy, we then examine the fraction of areal units for which coverage is adequate. We note that this is a fairly demanding standard of "adequacy," in that a simulated distribution may prove to be a reasonable approximation of the observed data, while still being statistically distinguishable from it. (We return to this issue below.)

8.2.1 Model Adequacy for the Test Data

Tables 4, 5, and 6 provide the fraction of areal units in each test region for which D does not differ significantly from each of the three proposed models. Looking across the three regions, we observe immediately that model performance is substantially better for block-level data than for block groups or tracts. This appears to result from the fact that block groups and tracts are not only much larger than blocks, but also substantially more heterogeneous; to reproduce D within a block group or tract requires the model to correctly reproduce the very considerable variation in population densities observed at the block scale, a feat for which none of the three models are well-prepared. On the other hand, we also see that, of the three models, the Attraction model substantially outperforms its peers on larger areal units. This is because the Attraction model can use boundary information as "clues" about where dense clusters of points might reside, thus recovering some of the underlying heterogeneity. Nevertheless, none of models approach perfect performance for larger areal units.

For small areal units, on the other hand, performance is quite good: in both Irvine and Deschutes County, approximately 87% of sampled blocks did not differ significantly from the simulated data. Even in Portland, where performance was lowest, the majority of blocks were not statistically distinct from the Attraction model. This suggests that, where one needs a proxy for household location data at the block level, even a very simple model may prove adequate for many applications.

8.2.2 Qualitative Comparison

While the Monte Carlo test provides a strict criterion for model adequacy, it is also useful to consider the extent to which the K distributions produced by the three

[8] Note that cases containing fewer than two points were removed from consideration.

Portland, Oregon			
	Quasi-random	Uniform	Attraction
Tract	0.00	0.02	0.13
Block Group	0.00	0.14	0.19
Block	0.38	0.56	0.58

Table 4: Proportion of blocks non-significant under the MC test performed on the D statistic.

Irvine, California			
	Quasi-random	Uniform	Attraction
Tract	0.04	0.06	0.22
Block Group	0.13	0.22	0.25
Block	0.73	0.86	0.87

Table 5: Proportion of blocks non-significant under the MC test performed on the D statistic.

proposed models qualitatively approach the observed data. As a basic point of comparison, we consider the average squared correlation (R^2) between the distribution of K functions for the simulated household distributions and the observed K function. Given the monotone nature of the K function, all R^2 values tend to be high (mean apx 0.98 for Tract and Block Group units, and 0.5 for Blocks), but we may directly inspect "typical" cases by selecting the areal unit in each location and scale class for which the R^2 is at or closest to the median. The resulting curves are shown in Figure 4.

As the figure shows, the qualitative fit of the median case to the data is excellent in Portland, OR at all scales. Although this may seem surprising in light of the findings of Table 4, we note that the two procedures involved answer distinct questions: the MC test tells us that deviations from the model are detectable in the Portland case, but the qualitative examination shows that the behavior of the curves in question are otherwise quite close. By contrast, the fit to the other two cases is less good; while the overall shape of each curve tracks the data, the magnitudes are plainly off for larger areal units. At the Block level, the figure underscores the point that there is considerable variability in the associated distributions, thus contributing to the lack of significant deviations. Taken together with the adequacy results, these results seem to suggest that the proposed models may be good proxies for large-unit behavior in urban areas (even where they are statistically distinguishable), and Block-level behavior in most areas.

Deschutes County, Oregon			
	Quasi-random	Uniform	Attraction
Tract	0.00	0.00	0.00
Block Group	0.01	0.04	0.07
Block	0.87	0.86	0.87

Table 6: Proportion of blocks non-significant under the MC test performed on the D statistic.

8.2.3 Case Study

Finally, to get additional insight into the simulation processes under study we provide a closer examination of simulated and observed data for a Tract in Portland, Oregon. We begin by considering the point plot of the observed data and the simulated pattern of each of the three baseline models: Uniform, Quasi-random, and Attraction Models (Figure 5). We then proceed to visually compare the $\mathcal{K}$, $\mathcal{G}$, and $\mathcal{F}$ functions.

Figure 6 makes it visually apparent that in the chosen Tract the Attraction model performs significantly better than the other two baseline models. In Figure 7 we see that none of the models capture the fine details of the observed data, although the Attraction model does capture the basic pattern of inhomogeneity in population density throughout the tract. Lastly, we see that in Figure 8 that the Attraction model performs the best on the $\mathcal{F}$ statistic.

9 Conclusion and Discussion

In this paper we have set forth an important problem that exists because of the aggregation of areal units for large scale administrative data such as the US Census. The placement of households (or individuals) is important for many social and demographic processes and the ability to map households over a given polygon boundary is potentially important for micro-social processes such as transmission of disease between households, daily mobility patterns, or patterns of interpersonal contact. When dealing with processes that require modeling interaction directly (e.g., social networks) one often has need of a specific location for individuals or households. For example, consider the simulated spatial networks of Butts et al. (forthcoming); Carter T. Butts (2011).

We demonstrate that at the Block level all three models perform reasonably well, but the Attraction model typically outperforms the Quasi-random and the Uniform model in Tract and Block Group levels (sometimes by as much as 16 percent). Since the Attraction model performs as well or better than the other two models, we advocate that for household simulation one should in general use the Attraction model. The Attraction model has the advantage of being able to take into account macro-level

patterns such as roads or waterways, unlike the Uniform and Quasi-random models (Figure 5(b)).

We want to emphasize here that the statistical test employed to assess model adequacy is a quite stringent one, and thus the simulated distributions may be sufficiently good approximations to meed research needs even where distinguishable in terms of the D statistic from the empirical household distribution. Take, for example, a median areal unit from any of the three test case (Figure 4) where we can see that the simulated point process appear to capture the general trend of the observed $\mathcal{K}$ function.

References

Almquist, Z. W. (2010). US Census spatial and demographic data in R: The UScensus2000 suite of packages. *Journal of Statistical Software*, 37(6):1–31.

Baddeley, A. and Turner, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42.

Besag, J. and Diggle, P. J. (1977). Simple monte carlo tests for spatial pattern. *Journal of the Royal Statistical Society: Series C.*, 26(3):327–333.

Binka, F. N., Indome, F. and Smith, T. (1998). Impact of spatial distribution of permethrin-impregnated bed nets on child mortality in rural northern ghana. *The American Journal of Tropical Medicine and Hygiene*, 59(1):80–5.

Bivand, R. S., Pebesma, E. J. and Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with* R. New York, NY: Springer.

Bossard, J. H. S. (1932). Residential propinquity as a factor in marriage selection. *American Journal of Sociology*, 38(2):219–224.

Butts, C. T., Acton, R. M., Hipp, J. R. and Nagle, N. N. (forthcoming). Geographical variablity and network structure. *Social Networks*.

Butts, C. T. and Almquist, Z. W. (2011). *networkSpatial: Tools for the Generation and Analysis of Spatially-embedded Networks*. R package version 0.6.

Carter T. Butts, R. M. A. (2011). Spatial modeling of social networks. In: : Timothy Nyerges, Helen Couclelis, R. M. (ed.), *The Sage Handbook of GIS and Society Research*. Thousand Oaks, CA: SAGE Publications, pages 222–250.

Davis, H. L. (1976). Decision making within the household. *Journal of Consumer Research*, 2(4):241–260.

Diggle, P. and Chetwynd, A. (1991). Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, 47(3):1155–1163.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. Oxford, UK: A Hodder Arnold Publication, second edition.

Freeman, L. C. and Sunshine, M. H. (1976). Race and intra-urban migration. *Demography*, 13(4):571–575.

Gentle, J. E. (1998). *Random Number Generation and Monte Carlo Methods*. New York, NY: Springer.

Gibson, C. C., Ostrom, E. and Ahn, T. K. (2000). The concept of scale and the human dimensions of global change: a survey. *Ecological Economics*, 32(2):217 –239.

Guilmoto, C. Z. and Rajan, S. I. (2001). Spatial patterns of fertility transition in indian districts. *Population and Development Review*, 27(4):713–738.

Harris, C. D. (1943). Suburbs. *American Journal of Sociology*, 49(1):1–13.

Hipp, J. R., Faris, R. W. and Boessen, A. (2011). Measuring [']neighborhood': Constructing network neighborhoods. *Social Networks*, In Press:–.

In: Jefferson Fox, Ronald R. Rindfuss, S. J. W. and Mishra, V. (eds.) (2003). *People and the Enviroment: Approaches for Linking Household and COmmunity Surveys to Remote Sensing and GIS*. New York, NY: Kluwer Academic Publisher.

Keitt, T. H., Bivand, R., Pebesma, E. and Rowlingson, B. (2009). *rgdal: Bindings for the Geospatial Data Abstraction Library*. `http://CRAN.R-project.org/package=rgdal`. R package version 0.6-21.

Mayhew, B. H. (1984). Baseline models of sociological phenomena. *Journal of Mathematical Sociology*, 9:259–281.

In: McC. Netting, R., Wilk, R. R. and Arnould, E. J. (eds.) (1984). *Households: Comparative and Historical Studies of the Domestic Group*. Berkeley, CA: University of California Press.

Minnesota Population Center (2011). *Integrated Public Use Microdata Series, International: Version 6.1 [Machine-readable database]*. University of Minnesota, Minneapolis.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem*. Norwich: Geo Books.

Perry, G. L. W., Miller, B. P. and Enright, N. J. (2006). A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. *Plant Ecology*, 187(59–82).

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org`. ISBN 3-900051-07-0.

Reibel, M. (2007). Geographic information systems and spatial data processing in demography: A review. *Poulation Research and Policy Review*, 26:601–618.

Ripley, B. D. (1988). *Statistical Inference for Spatial Processes*. Cambridge: Cambridge University Press.

Rowlingson, B. S. and Diggle, P. J. (1993). Splancs: Spatial point pattern analysis code in s-plus. Technical report, Lancaster University, Lancaster, UK.

Salathé, M. and Jones, J. H. (2010). Dynamics and control of diseases in networks with community structure. *PLoS Comput Biol*, 6(4):e1000736. `http://dx.doi.org/10.1371%2Fjournal.pcbi.1000736`.

Short, M. B., Brantingham, P. J., Bertozzi, A. L. and Tita, G. E. (2010). Dissipation and displacement of hotspots in reaction-diffusion models of crime. *Proceedings of the National Academy of Sciences*.

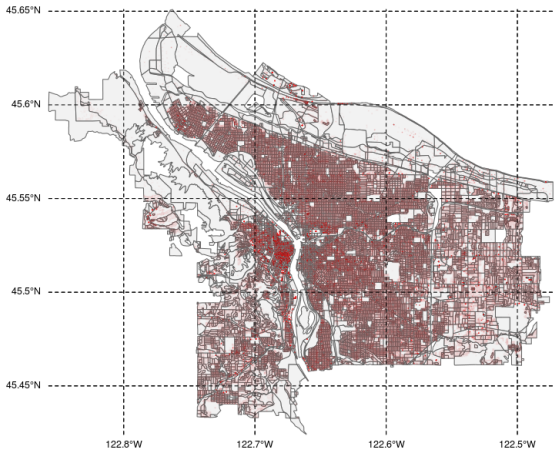
Snyder, J. P. (1987). Map projections – a working manual. U.s. geological survey professional paper, Unite States Government Printing Office, Washington, D.C.

Theobald, D. M. (2004). Placing exurban land-use change in a human modification framework. *Frontiers in Ecology and the Enviroment*, 2(3):139–144.

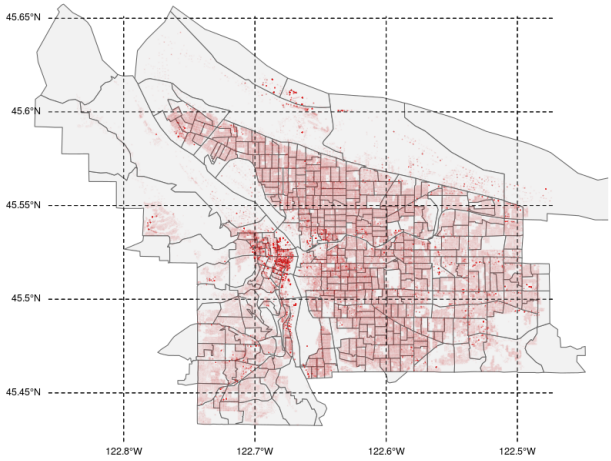
U.S. Census Bureau (2001). Census 2000 summary file 1 united states/prepared by the u.s. census bureau. Technical report, US Census Bureau.

Voss, P. R. (2007). Demography as a spatial social science. *Poulation Research and Policy Review*, 26:457–476.

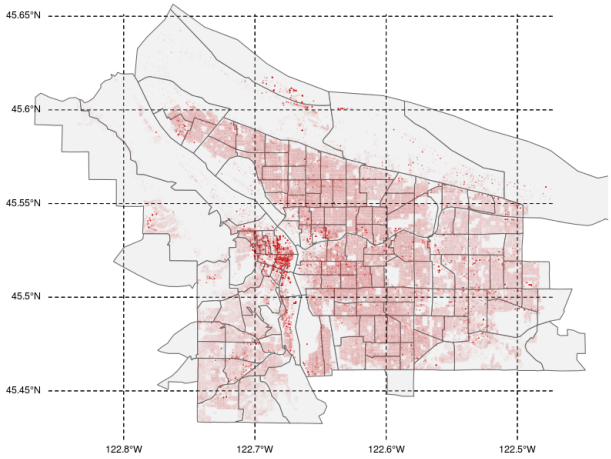
Portland, Oregon



(a) Parcel data & US Census 2000 Blocks of Portland, OR.



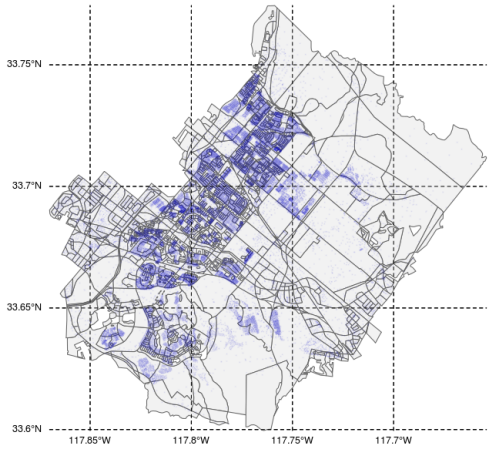
(b) Parcel data & US Census 2000 Block Groups of Portland, OR.



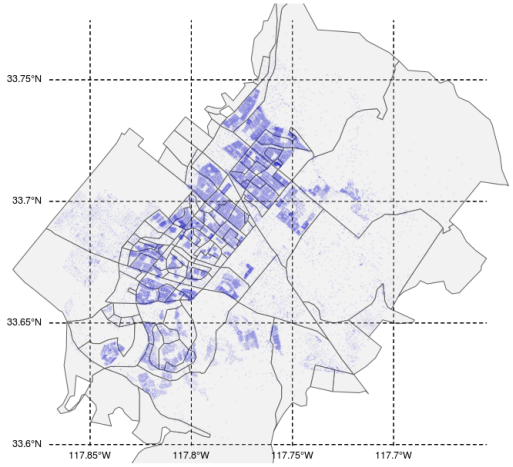
(c) Parcel data & US Census 2000 Tracts of Portland, OR.

Figure 1: Portland, Oregon Households and polygons (Blocks, Block Groups, and Tracts).

Irvine, California



(a) Parcel data & US Census 2000 Blocks of Irvine, CA.

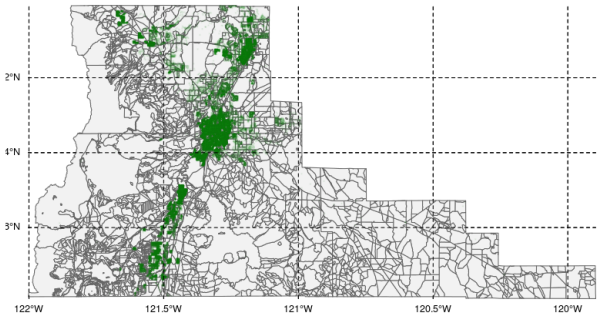


(b) Parcel data & US Census 2000 Block Groups of Irvine, CA.

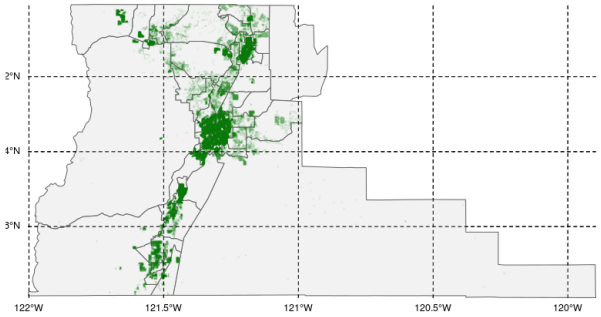
(c) Parcel data & US Census 2000 Tracts of Irvine, CA.

Figure 2: Irvine, California Households and polygons (Blocks, Block Groups, and Tracts).

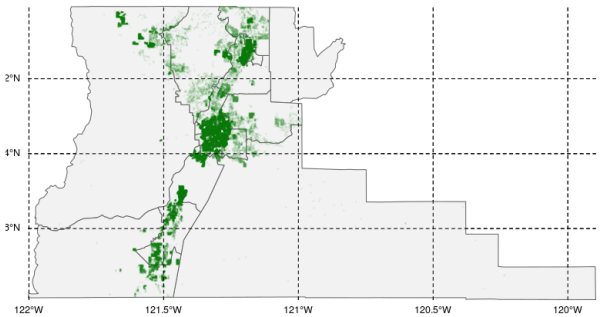
Deschutes County, Oregon



(a) Parcel data & US Census 2000 Blocks of Deschutes County, OR.

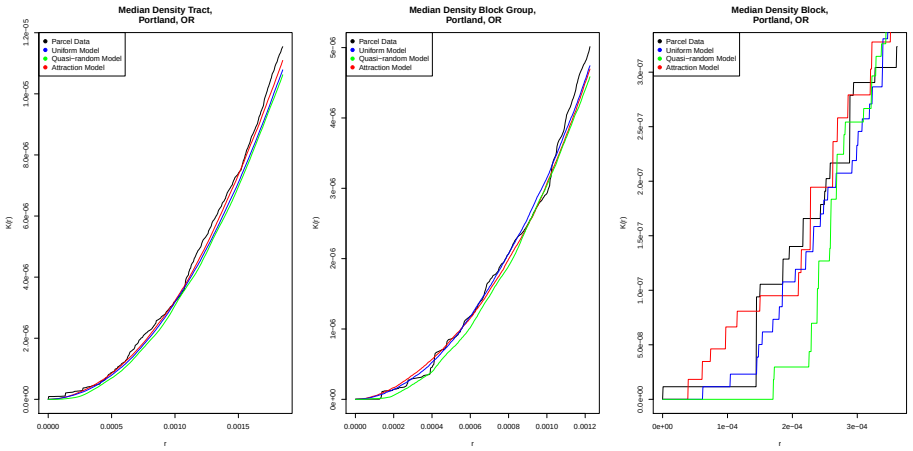


(b) Parcel data & US Census 2000 Block Groups of Deschutes County, OR.

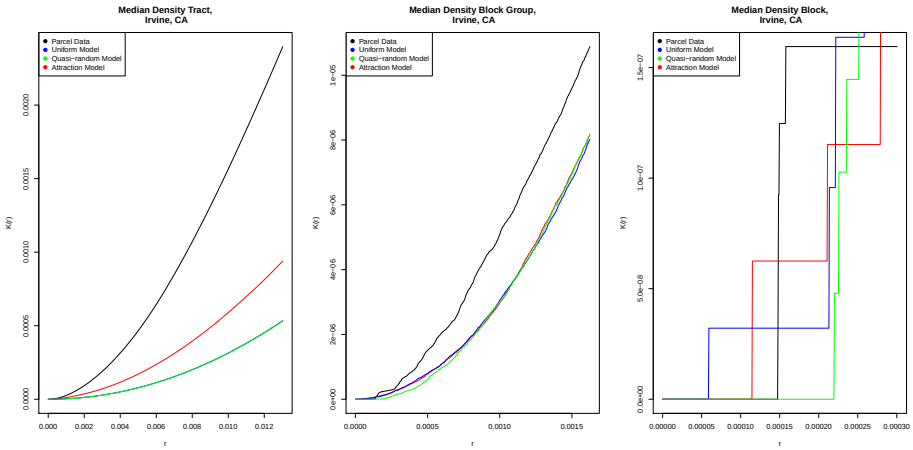


(c) Parcel data & US Census 2000 Tracts of Deschutes County, OR.

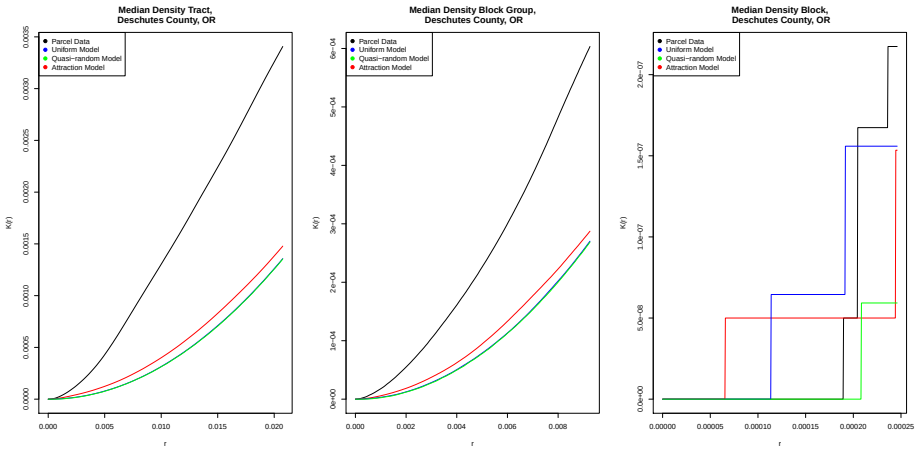
Figure 3: Deschutes County, Oregon Households and polygons (Blocks, Block Groups, and Tracts).



(a)

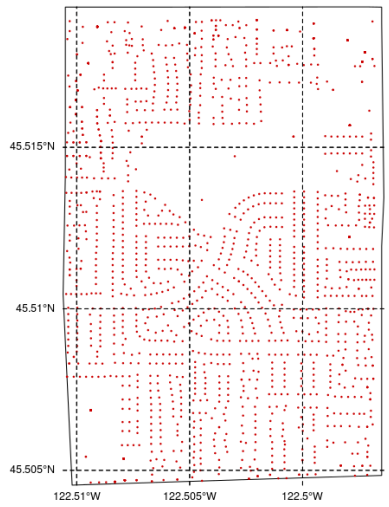


(b)

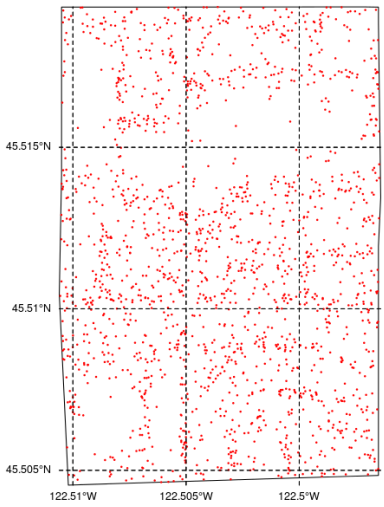


(c)

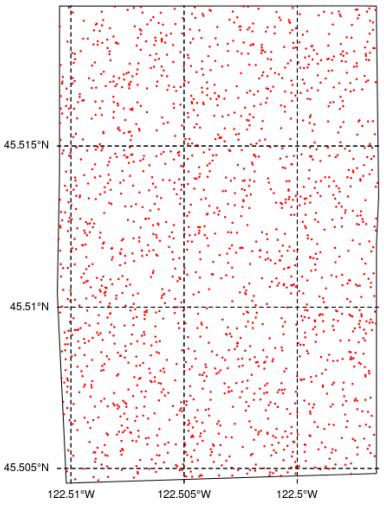
Figure 4: $\mathcal{K}$ function for the median Tract/Block Group/Block geography for Portland, OR (a); Irvine, CA (b); and Deschutes County, OR (c).



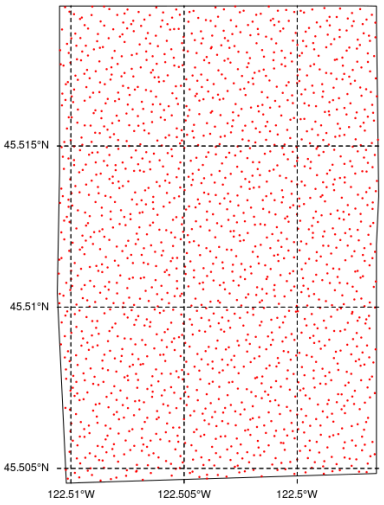
(a) Parcel Data.



(b) Attraction Model



(c) Uniform Model



(d) Quasi-random Model.

Figure 5: Observed and simulated point distributions over tract "009701" in Portland, Oregon for the three baseline models considered in this paper.

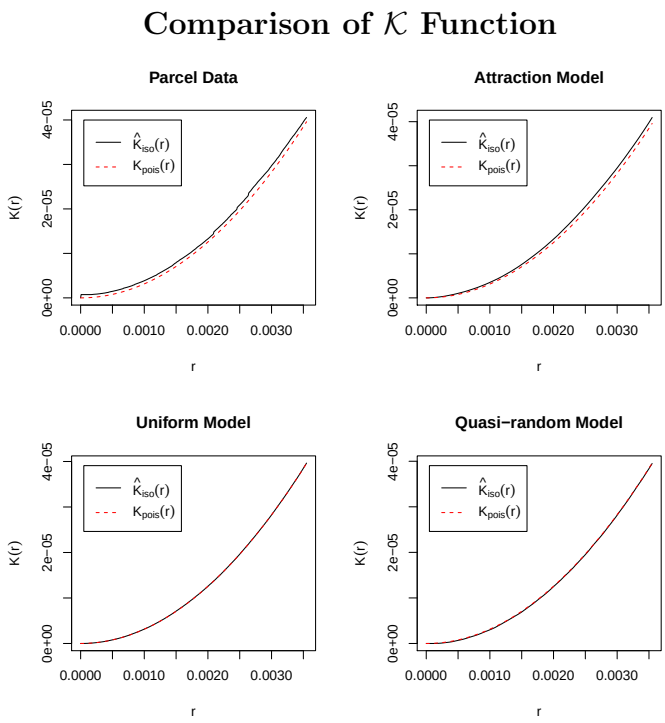


Figure 6: Comparison of the three baseline models and the observed distribution of $\mathcal{K}$.

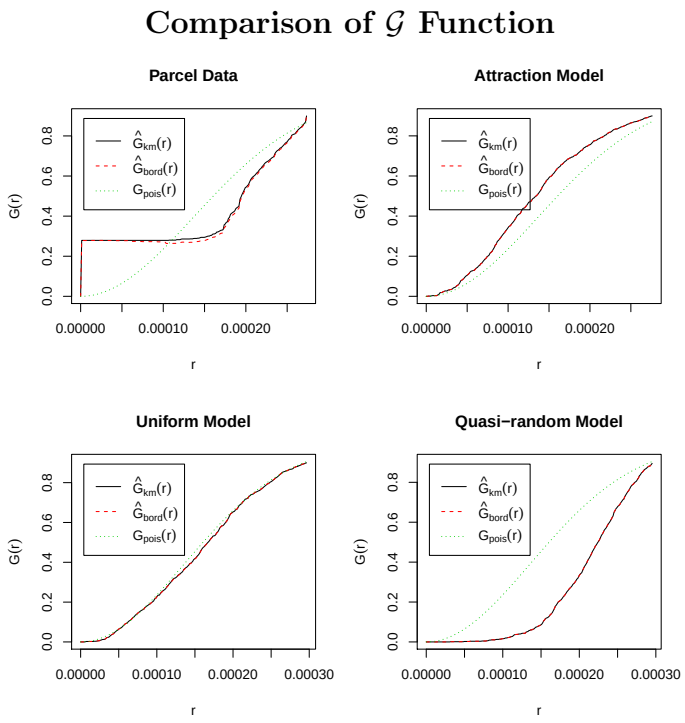


Figure 7: Comparison of the three baseline models and the observed distribution of $\mathcal{G}$.

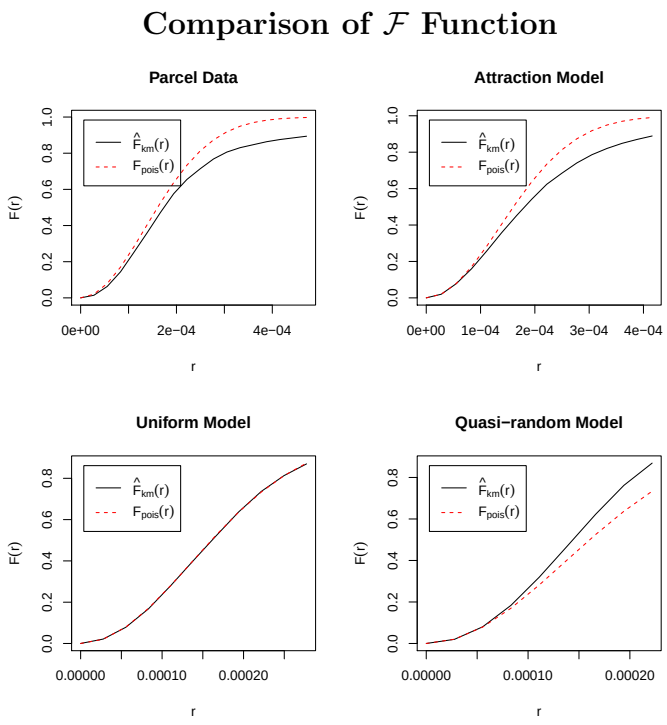


Figure 8: Comparison of the three baseline models and the observed distribution of $\mathcal{F}$.