Robust point pattern inference from spatially censored data

Stuart H. Sweeney
and
Kevin J. Konty

Department of Geography
University of California, Santa Barbara

– forthcoming –

Environment and Planning A

May 27, 2004; 0800PST

Abstract

Administrative data sources are increasingly used for spatial analysis and policy formation. For example, welfare-to-work programs have stimulated demand for spatial mismatch studies using ES-202 employment files. The increased resolution gained by geo-coding the address records in administrative files can be of enormous research value when the process under study resolves over small distances. Yet the resulting point-referenced data is problematic for inferential analysis. In particular, administrative data typically represents a sample of convenience thus posing serious validity problems for statistical inference. This paper proposes a robust estimation method for spatial pattern inference based on spatially censored data. The performance of the estimator is explored using simulated data and is also demonstrated on ES-202 data from North Carolina.

Key Words: Administrative data, point pattern analysis, spatial censoring, address geocoding.

1 Introduction

Spatially-referenced data is becoming increasingly important to progress in the social sciences. The U.S. Census Bureau and other federal data providers continue to unleash a bounty of spatially-referenced data products that are readily ingested by Geographic Information Systems. Though the data released through official statistics agencies is voluminous, we have only just begun to witness the potential torrent of new information that could become available through access to administrative records. The critical distinction between the two is that the former is based on sampling designs tailored for inference at a particular spatial scale whereas the latter is essentially commandeered for spatial analysis. As social science researchers bring these new data sources into service, we should heed the warnings from Alonso (1968) regarding imperfect data. He questioned, "...whether we have arrived at the design of skyscrappers but...have only lumber for construction materials." Similarly in the emerging spatial data environment we should be searching for robust methods of inference, in Alonso's analogy "..build low to the ground..", especially when inference will be used to guide policy intervention. This paper proposes a method for spatial point pattern inference that is robust to data imperfections stemming from spatial censoring.

We should note that the increased supply of spatial data is only one of several forces behind the spatial awakening in the social sciences. Indeed, the increasing importance of spatial data and methods, and the increasing role of 'space' and 'place' in social science theorizing is evident in institutional forces that have the power to shape the social sciences. The U.S. National Science Foundation (NSF) has demonstrated an increasingly strong interest in promoting what it now terms *spatial social science*. Significant long-term investments include the *National Center for Geographic Information and Analysis - NCGIA* (1988-1996) and the *Center for Spatially Integrated Social Science - CSISS* (1999-2004) (Goodchild 2000, Goodchild et al. 2000).[1] In both instances, especially the latter, the funding explicitly encourages the diffusion of spatial concepts, methods, and technologies to the social sciences. Another major social science funder, the National Institute of Child Health and Human Development, identifies *spatial demography* as one of the target funding areas in its long-range plan (National Institute of Child Health and Human Development 2002). Though impacts of funding initiatives are notoriously hard to measure, we can point to CSISS training workshops targeting graduate students and junior faculty that are filled to capacity each year with considerable excess demand (Janelle 2002).

Foundations have been actively promoting spatial analytic tools and increased access to spatially-referenced data to affect change in a different realm. Specifically, the *Ford Foundation*, *Fannie Mae Foundation*, and *Annie E. Casey Foundation* have been actively involved over the past several years in building the institutional momentum for the development of community indicators and eventually a more formally structured Community Statistical System. Both the U.S. Department of Housing and Urban Development and the U.S. Census Bureau have been involved in this movement as an outlet for some of their data products. The *Urban Institute* is probably at the forefront of the *community indicators* movement that seeks to capture local data and use it in grass-roots community planning and policy analysis. It is the cavalier use of spatial data in this realm, in particular, that would have alarmed Alonso.

[1] Luc Anselin and Michael Goodchild were the major forces behind the creation of both centers. NCGIA was awarded while both were Professors of Geography at the University of California at Santa Barbara. Both were also the principal investigators for CSISS, though Anselin is now a professor at University of Illinois, Urbana-Champaign. U.C. Santa Barbara also recently was awarded another significant NSF grant to support *SPACE* - Spatial Perspectives on Curriculum Enhancement - which targets undergraduate education and curriculum in the social science.

As the spatial awareness envelope has continued to expand within the social sciences a fundamental tension has emerged due to an inverse relationship between spatial resolution and data quality. In particular, the primary method – address geo-coding – that allows one to refine spatial scale and observe processes operating over small distances, also degrades the data quality. The tension is summarized aptly by Longley et al. (2001):

> ... there is an emergent tension within the socio-economic realm, for there is a limit to the domains of inference that can be made from conventional, scientifically valid data sources which are frequently out-of-date, zonally coarse, and irrelevant to what is happening in modern societies. Yet the alternative of using new rich sources of marketing data is profoundly unscientific in its inferential procedures. (pp. 135)

The core issue with spatial scale is one of construct validity. Given a particular social science process under study, that process will naturally resolve at one or more spatial scales. To study the process our empirical measures should match, as near as possible, the spatial scale of the process. If the spatial scales do not match between the process and the measure the resulting inference has a classic construct validity problem (Sweeney and Feser 2002). For example, studies of inter-firm productivity spillovers should attempt to measure those interactions over small distances to remain consistent with economic theories of the process. With spatial data, the problems are even more complex due to a host of other well-known validity issues related to scale-dependence (aggregation bias) and frame-dependence (modifiable areal unit problem) (Longley et al. 2001, Oppenshaw 1984, Yule and Kendall 1950).

Social science researchers are generally well aware of the construct validity problem though somewhat less attuned to the other issues with spatial data. The research dilemma presents two alternative paths when the process under study resolves at a finer resolution than is provided by the areal units of standard data sources from official statistics agencies. One path is to disregard the validity issues inherent to lattice data and proceed with the analysis. This is an appealing choice because the access to training and software needed for lattice based models is widely available. [2] The second path is to find data that resolves at a satisfactory spatial scale. It is at this juncture that social scientists will increasingly use address geo-coding to resolve data at the point scale. Even if the address geo-coding is perfect in the sense of positional accuracy, which will rarely be the case, the resulting data will be problematic if address information is missing for some records. It will also generally be true that the data was never intended for use at the point scale, thus any inference is outside of the intended use of the sample design. If the data is from administrative records it is highly likely that a sampling design does not exist. As a rule, then, the data quality will be worse at the point scale compared to the quality of lattice data. The tension, noted above, results because one set of validity issues (construct) are traded off against another set of validity issues (data quality).

The data quality issues, and more generally issues related to uncertainty in spatial data, are well documented in the Geographic Information Science literature (Longley et al. 1999, Longley et al. 2001). For example, Longley et al. (1999) note that, "...much GIS practice continues to proceed as if data were perfect. Results of GIS analysis...rarely show estimates of confidence, or other indicators of the effects of data quality." There are generally two threads within this

[2]Indeed, CSISS sponsored graduate workshops and long-standing ICPSR (Interuniversity Consortium for Political and Social Research) workshops by Luc Anselin have trained an army of social scientists in spatial econometrics. Anselin's GeoDa software provides an extremely intuitive platform for exploratory and confirmatory analysis of spatial data. Though it does include some functionality for point patterns, the bulk of tools are for lattice data.

literature. One develops taxonomies of sources of error and uncertainty and their impact on the resulting features of a map (Veregin 1999, Fisher 1999). For example, positional accuracy is a concern. The other strand examines the implications of error propagation. That is, given an input map with a certain level of uncertainty in its features, they seek to ascertain the nature of the error in a map that derives from map algebra operations on the first map (Heuvelink 1999).

Our concern in the present paper is with the impact of unrepresentative samples on inference from spatial point patterns. [3] We are particularly concerned with instances where pattern inference is used in a process control setting such as disease early warning detection (Aylin et al. 1999). Incorrect inference in that setting could result in extremely costly mistakes. The potential research value of spatially-referenced administrative records, 'marketing data', and health registry data is great, but inference based on the data must satisfy the usual scientific validity requirements. A major limitation of many administrative data sources is that the de facto sampling design is a sample of convenience. Though administrative records samples are often large, upwards of 50 percent, the potential bias imparted from the lack of a formal sampling design needs to be carefully evaluated. There is a growing body of research that documents spatial bias in administrative data sources. [4] An unresolved issue is whether valid statistical inference for spatial patterns can result from geo-coded administrative records.

This paper proposes a robust method for spatial pattern inference when the quality of the point data is suspect. The basic idea is to use stratified resampling to pool information across spatial scales (lattice and point) thus providing unbiased estimates of spatial pattern from spatially censored point data. The term *censored* is apt in this case because the spatial information is removed systematically, due to some censoring rule, as opposed to missing completely at random. The same terminology is used in the statistics literature dealing with incomplete data (Little and Rubin 1987). The paper proceeds with two primary divisions. The first describes the proposed correction method and presents results from a simulation study of the estimator. The second provides an application to a representative large administrative data source, the national ES-202 file, that has undergone address geo-coding. The application includes a discussion of the sources of bias and the magnitude of bias in the data and complications with the robust estimator. The paper concludes with a discussion of the results and poses some unresolved research issues. Though the application uses the ES-202 data, the robust estimation method applies broadly to all spatially point-referenced administrative data. Specifically, there are major potential application areas within public health, urban economics, and criminology.

2 Robust estimation: stratified random thinning

The robust estimation method proposed in this section derives from the simple observation, noted above, that the quality of the data degrades as the spatial resolution of the data increases. Thus there are at least two levels of spatial information. We assume that there exists an aggregate scale, a set of lattice counts, with no censoring and a point scale with censored spatial

[3] As such, our interest is closer to the second strand in the 'GIS and uncertainty' literature but we are not interested in a map as the final product. Also, though the first strand includes 'completeness' as a source of error, it is caste within an enumeration paradigm characteristic of cartography. That is, cartography is disinclined towards portraying a sample of map features. It is routine to portray census tract data on a map where the underlying data is based on a representative 12 percent sample of the population within each tract. It is not routine to display a randomly selected 12 percent of the census tracts.

[4] For example see Feser and Sweeney (2004) for a discussion of the ES-202 data or for a review of the extensive literature on address geo-coding of public health see the discussion and references in Rushton and Armstrong (1997).

identifies. It is easiest to think in terms of a standard rectangular data array with observations comprising the rows and variables as the columns. Assume that each observation represents an individual or firm that we would like to represent as a unique point in space because of our research question; the construct validity concern. The variables include spatial identifiers at the point scale, a unique x- and y-coordinate for each individual, and a non-unique identifier at the lattice scale indicating that the individual lies within a particular zone. We assume spatial censoring occurs such that the x- and y-coordinates are missing for several observations whereas the lattice membership is recorded for all observations. To make inference to spatial patterns at the point scale it seems prudent to treat the more aggregate data as an auxiliary source of information which can be used to adjust the estimates from the censored point data. [5] The logic of the approach, pooling information across ecological scales, is standard in the incomplete data literature (Little and Rubin 1987). It is also a standard notion in the small area estimation literature where incidence rates based on small counts are smoothed toward a global rate or macro-regional rates (Rao 2003).

To restate the problem more formally, we define a partially observable point pattern P generated by some underlying point process N in a region A. The observable portion we label Q. We also have some auxiliary information for the points in P that are not in Q, say $P \setminus Q$. The auxiliary data could be in the form of lattice cell membership or could be informed by joint membership in a lattice and any other informative strata. For example, in the application section of the paper we have information on lattice membership (county) and industry sector. We are interested in estimating some measure F that summarizes essential properties of the underlying process. We suspect that $\hat{F} = f(Q, P \setminus Q)$ will improve over $\hat{F} = f(Q)$. The central issues revolves around the specification of $f(Q, P \setminus Q)$. There are at least three specifications that would utilize both levels of information:

1. *Imputation.* A standard approach in the incomplete data literature is to impute, either explicitly or implicitly, missing values by conditioning on the auxiliary information. Explicit imputation relies on a generic predictive function, $P = f(\beta; Q, P \setminus Q)$, to assign values for observations with missing coordinate information. This could be accomplished directly, for instance, using a Poisson process within each lattice cell to assign missing values. Or indirectly, the β parameters would be estimated from the data and auxiliary information, and then the model would be used to predict missing values. For example, we could estimate parameters that predict a continuous surface such that the height at any location is the relative likelihood of observing a point at that location. Missing data would be recovered by sampling the likelihood surface. The problem with this approach is that deriving a likelihood surface that conditions on the auxiliary data is complex. Likelihood surfaces are interpretable as giving, for each location, the relative chance of observing one additional point. Additional points are dependent on the previously placed points and so the surface would need to be updated after each imputation. A solution to the updating problem might be possible using the expectation-maximization algorithm (McLachlan and Krishnan 1997). Any parametric solution will likely involve a complex likelihood specification requiring Markov Chain Monte Carlo to derive the parameter estimates. The discrete event simulation framework developed by Wolpert and Ickstat

[5] We should note that Baddeley (1997, 1999) has done excellent work on the effects of censoring and bias resulting from the placement of a bounding box defining a study region. As such, methods for partially observed patterns focus on unobserved observations lying outside the study region boundary. In contrast, our focus is on missing observations within the study region. As such, our definition of censoring differs from his definition of censoring.

(1998) would seem to provide a promising approach for the explicit imputation problem. An alternative is implicit imputation. In this case the data and auxiliary information are combined to provide a direct estimate of the measure of interest; F, in our case. Here we are thinking of a non-parametric approach akin to survival analysis methods used for censored data (Cox and Oakes 1984). Baddeley (1997), for instance, has proposed point pattern methods based directly on survival analysis but has only addressed censoring as it applies to edge correction.

2. *Weighting.* This approach treats the auxiliary information as though it describes a sampling scheme. Given the population counts on the lattice, the goal would be to weight the sample of observed points to match the known population of each lattice cell. A lattice cell with a large amount of censoring would receive a large sampling weight attached to each of the observed points. Weighting schemes are commonplace in the survey literature but the approach is problematic in point pattern analysis. Assigning weights effectively places multiple points in a single location. This is equivalent to letting the known distribution of inter-event distances serve as a proxy for missing interevent distributions. Yet we know that other points exist in the lattice cell, in different locations, and that this should inform our distribution of interevent distances. The information is especially valuable for shorter distances. The weighting approach essentially magnifies the missing data problem by adding the most weight to locations where information is the weakest. Also note that the weighting approach can be cast as a naive form of imputation in which missing points are fractionally distributed among known point locations.

3. *Subsampling.* The final approach constructs estimates by resampling the observed points at rates derived from the auxiliary data. The idea is that each subsample is more representative of the underlying point pattern, P, than the observed points, Q. As discussed above, events are assumed to not share the same point. Therefore, the usual bootstrap method of sampling with replacement, thereby sampling from the empirical distribution, does not work. Instead we take successive subsamples of the data such that the original relative proportions among strata, derived from the auxiliary information, are maintained. The resampling method is akin to those used for variance estimation for complex survey designs (Rao and Shao 1999, Krewski and Rao 1981). In that literature, subsamples are defined to keep the strata balanced following the original sampling design. In our problem setting we assume that there is no sampling design. Subsampling is particularly appealing in the pattern inference context given recent research showing that it provides robust estimates in a wide range of problem settings (Politis et al. 1999).

In this paper we implement the subsampling approach and intend to pursue the imputation approach at a latter date. Valid point pattern inference based on subsampling requires two assumptions. First we require that the point process, P, is stationary. This is a standard assumption in the analysis of second-order properties of point patterns. Second, we assume that the point pattern within each strata is a random thinning with respect to the subset of point pattern, P, that occurs within a given lattice cell. In contrast, we assume that Q is not a simple random thinning of the overall point pattern, P. Given those assumptions, the robust subsampling estimator is defined by the following algorithm:(1) Compute the sampling rate p_x for the strata in P that is most under represented in Q, (2) Generate a series of subsamples, $R_{[s]}$ where $s = \{1, ..., S\}$, such that every strata in Q is sampled at the same rate $p_r < p_x$, (3) Compute and store the values of the measure $\hat{F}_{[s]}$ derived from each subset $R_{[s]}$, and (4) Calculate $\hat{F}_R$ as the mean value of the set $\{\hat{F}_{[1]} ... \hat{F}_{[S]}\}$. Confidence intervals can be recovered

from the same information set used in step (4) as in Monte Carlo simulation (Diggle 1983). Since the algorithm essentially extracts a stratified random sample on each iteration, we term $\hat{F}_R$ the Stratified Random Thinning (SRT) estimate.

Our conjecture is that $\hat{F}_R$ should perform better than the *raw* estimate $\hat{F}_Q$ based directly on the spatially censored data Q. In the sections below we show that not only is the conjecture confirmed, but the SRT estimate also improves slightly over the raw estimate when Q represents a simple random thinning of the complete data P. The improvement is in the sense of reducing bias, specifically $|E(\hat{F}_R) - F| < |E(\hat{F}_Q) - F|$. Since each subset $R_{[s]}$ is smaller than Q, the bias improvement is at the expense of efficiency. In this regard note that each subset $R_{[s]}$ can only be as large as $p_r{*}N_Q$. Also, if subsamples are selected at the rate p_r, R will always contain every point from the most under represented class. If the points in that strata of Q are unrepresentative of the same strata in P then the SRT estimate will suffer.

3 Simulation results for the SRT estimator

One way to evaluate our conjecture is to choose a particular specification of F and then examine the behavior of $\hat{F}_R$ and $\hat{F}_Q$ for simulated point patterns under alternative spatial censoring schemes. Two common measures for describing and comparing point patterns that are invariant under random thinning are the K-function (Ripley 1977) and the D-function (Diggle and Chetwynd 1991). The K-function has been used, and formal tests developed, to compare patterns of events to the Poisson distribution (Ripley 1976, Diggle 1983, Getis 1984, Bailey and Gatrell 1995). The D-function simply extends this to allow the comparison of arbitrary point patterns.

K is defined by:

$$\lambda K(t) = E[N(t)] \tag{1}$$

where λ denotes the intensity, or mean number of points per unit area, and $N(t)$ is the number of points within distance t of an arbitrary point in the region A. Empirically, K can be estimated by

$$\hat{K}(t) = \frac{|A|}{n(n-1)} \sum_r \sum_{t>r} I_t(x_r x_t) w(x_r x_t) \tag{2}$$

where $|A|$ is the area of region A, n is the number of points, $w(a,b)$ is the reciprocal of the proportion lying in A of the circle with center a and radius $||a-b||$, and I_t is an indicator for $||a-b|| < t$. Thus, $\hat{K}$ is based on the distribution of all interevent distances with some weighting in cases where the events occur near the boundary of A.

For any given spatial point pattern, we can calculate the measures $\hat{K}_P$ from the complete data, P, and both $\hat{K}_R$ and $\hat{K}_Q$ from any spatially censored subset of the complete data, Q. Following Diggle (1977, 1978) we simulate from a Matern(λ, r) which yields a closed-form expression for $K_P(t)$, as

$$K(t) - \pi t^2 = \begin{cases} \frac{\left(\frac{2t^2}{r-2r}\right)\cos^{-}1\left(\frac{t}{2r}\right)+\pi r - t\left(1+\frac{t^2}{2r^2}\right)\sqrt{\frac{1-t^2}{4r^2}}}{\pi\lambda r} & \text{if } 0 \le \text{t} \le \text{2r} \\ \frac{1}{\lambda} & \text{if t} > \text{2r} \end{cases} \tag{3}$$

where r is the radius of the clusters, λ is the Poisson parameter for the number of clusters, and πt^2 is the K-function for a Poisson process (complete spatial randomness). Given equation (3)

we can calculate the exact population value of K_P for any Matern process and directly evaluate the conjecture that $|E(\hat{K}_R) - K_P| < |E(\hat{K}_Q) - K_P|$.

The Matern process is a doubly stochastic Poisson process (Matern 1971). Realizations are simulated in two steps: (1) parents are Poisson distributed in a region according to intensity λ, and (2) N offspring are independently, uniformly distributed inside circles of radius r centered on each parent. The point pattern defined by the N offspring is a single realization. Our baseline complete data, P, is a single 1,000 point realization on a unit square from a Matern(29,.061); see the left panel of Figure 1.[6]

If spatial censoring occurred randomly over the unit square, we would have a classic random thinning. For example, Figure 2 contains the means and simulation envelopes for 200 random thinnings with the subsets from P equal to 100 points (dashed lines) and 700 points (solid lines). In both cases the means should closely approximate the population statistic, K_P, since E($\hat{K}$)=K; the invariance under random thinning property. Also note that the precision degrades, the confidence envelopes widen, as the subsets decrease in size.

If spatial censoring occurs disproportionately in particular quadrants we will call it an *unbalanced thinning*. For example, the right panel of Figure 1 contains a 755 point subset, Q, with points deleted only from quadrants 2 and 4 ($P_x > 0.5$). For auxiliary information, we assume that we know the quadrant for each point in $P \setminus Q$. Using the information in Q we can calculate $\hat{K}_Q$ or use Q in combination with our auxiliary information, $P \setminus Q$, to calculate the SRT estimate, $\hat{K}_R$. The population statistic and the two estimates, $\hat{K}_Q$ and $\hat{K}_R$, are shown in Figure 3. For small distances $\hat{K}_Q$ follows K_P closely but underestimates it severely for large distances. In contrast, when we calculate $\hat{K}_R$ based on subsets of the largest possible size, in this case 578, then $\hat{K}_R$ closely matches K_P at each distance. As noted above, the confidence envelopes for $\hat{K}_R$ should be comparable to those derived from randomly thinned 578 point subsets of P.

3.1 SRT estimator for unbalanced subsets

The results in Figure 3 only represent a single unbalanced thinning. Using the same basic framework described above, we can also generalize the results to examine the performance of the SRT estimator over the entire domain of unbalanced subsets of P where $p_x > 0.5$. The domain contains 989,000 possible 755 point subsets in which each region has at least 50 percent of its original points. The degree of unbalance for each subset can be characterized by Goodman and Kruskal's τ, a measure of nominal association, as

$$\tau = \frac{\sum_i \sum_j (\pi_{ij}^2 / \pi_{i+}^2) - \sum_j \pi_{+j}^2}{1 - \sum_j \pi_{+j}^2} \tag{4}$$

where π is a proportion from a contingency table. In our application, τ measures the proportional reduction in *variance* of the counts by strata when conditioning on membership in Q. Or expressed differently, it measures the probability of incorrectly guessing the lattice membership of a missing point. If the sample Q is a random subset of P then conditioning on membership in Q does not reduce this variance; that is, for $\tau(Q) = 0$, Q is a balanced thinning. If membership in Q determines the strata then the variance reduction is total and $\tau(Q) = 1$. The single instance of Q shown in Figures 1 and 2 has a $\tau(Q) = 0.06252$ which places it in the 86th percentile of the domain of 989,000 subsets.

[6]The parameters are taken from previous discussions of this process (Diggle 1977, Diggle 1978) where the point pattern data were 62 redwood seedlings.

To measure bias over the range of τ we need to choose a set of metrics since our estimate, $\hat{K}$, is a function of distance rather than a single point estimate. The most obvious measurement would be the area between the curves $\hat{K}_Q$ and K_P. In practice $\hat{K}(t)$ is estimated at a finite number of distances, t. Rather than use a single metric we report six different errors in Table 1: (1) mean absolute error, (2) mean percent error, (3) mean error, (4) error at t=0.075, (5) error at t=0.15, and (6) error at t=0.225. Thus, for any Q, we can measure $\tau(Q)$, the errors in $\hat{K}_Q$, and the errors in $\hat{K}_R$.

The errors reported in Table 2 are derived from 2,000 randomly selected unbalanced subsets from the domain of 989,000. The percent error indicates an 11.5 percent positive bias for the raw estimate versus 3 percent for the SRT estimate. As a whole, the SRT estimate provides an approximately 70 percent improvement over the raw estimate. Also note that the performance of the SRT estimate relative to the raw estimate improves as distance increases.

Another characterization of error for unbalanced subsets is presented in Figure 4. In this case we select a single measure, mean absolute error, and compare the raw estimate to the SRT estimate over the range of $\tau(Q)$. The solid line (raw) and dashed line (SRT) are Loess smooths of the underlying scatter of results. For relatively balanced subsets, $\tau(Q)$ near zero, the SRT estimate is no worse than the raw estimate. As $\tau(Q)$ increases the raw estimate quickly degrades while the SRT error remains relatively constant. This suggest that in any case where the spatial sample is suspected of being unbalanced, $\hat{K}_R$ is always the recommended estimate. The only intervening consideration is the degree of efficiency lost using $\hat{K}_R$.

3.2 SRT estimator for simple random thinning subsets

In cases where the subsample is a random thinning, we would expect that the procedure to produce estimates similar to the direct estimate. In fact, the SRT estimate also improves over the raw estimate based on random thinnings. The results reported in Table 3 and Figure 5 are based on 100 randomly thinned 755 point subsets. Though the relative improvement of the SRT estimate is less than that shown for unbalanced subsets, the SRT estimate still improves over the raw estimate.

Thus, even though $E(\hat{K}_Q) = K_P$ when Q is a randomly thinned subset, in a finite simulation some randomly thinned patterns will result in unbalanced subsets by strata and SRT performs a correction. This is simply the result of having auxiliary information available which tells us that, even though the subset is random, it is unbalanced and we can improve our estimates slightly. A slight improvement is made with the correction on average, but more importantly the correction procedure does not damage estimates when no bias is present.

4 Application: Robust pattern inference for industrial location analysis

As noted in the introduction, a major new source of spatial data that is increasingly used for socio-economic applications is administrative records. In some cases the use has been stimulated by legislation. For instance, the Workforce Investment Act (1998) created new demand for spatially referenced employment data. In particular, there has been a major effort by the U.S. Census Bureau's Center for Economic Studies, the U.S. Bureau of Labor Statistics' Office of Federal-State Programs, as well as individual states' employment security divisions to use the raw Unemployment Insurance Wage Record administrative files to build longitudinally- and spatially-referenced data (Stevens 1994a, Stevens 1994b, Stevens 1994c, Pivetz et al. 2001, Spletzer 1997).

The community indicators movement also has fueled interest in administrative records. For example, the U.S. Department of Housing and Urban Development recently compiled their record level administrative data for housing research. There are also several cases where universities or non-profits have worked in conjunction with state and local agencies to compile detailed record-level data sets that span human services, health, crime, education, and other realms for particular metropolitan areas or regions (for example see Hillier and Culhane (2003)).

Our focus in this section is on the national ES-202 files, a subset of the Unemployment Insurance Wage Record files. The national ES-202 file provides a representative example of the sources, structure, and scale of censoring bias that may be present in administrative records more generally. The ES-202 series derives from a quarterly enumeration of establishments in which the objective is to collect information related to employee payroll taxes. The series is administered through the U.S. Bureau of Labor Statistics' Office of Federal-State programs; individual state employment agencies undertake the primary data collection and the national office then pools the individual state files into a national file. The resulting national file contains establishment level records with basic information such as the number of employees, total and taxable payroll, and the industry sector. As an artifact of the data collection process the file also records an administrative address (for example, of an accountant that provides information on the establishment) and the actual address of the establishment (the "physical" address in the parlance of the ES-202 file).

At the national, state, or even metropolitan scale, the ES-202 provides exceptional coverage of the economy. However, coverage issues arise when the data is used for detailed spatial analysis. There are two primary sources of censoring that eliminate records from the geo-coded file. First, the quality of the physical address information in the national ES-202 ultimately depends on the aggressiveness and savvy of individual state agencies, who perform the data collection. Some states have historically expended minimal effort to either record physical address information or to confirm its veracity. Second, even if the physical address is recorded in the file, an establishment record may still be censored due to insufficiencies in either matching algorithms or the underlying street location information derived from the Census Bureau's Tiger Files.

The tabulations in Table 2 provide a rough indication of the scale of censoring as one attempts to work with the file at increasingly detailed spatial scales. The top panel of the table records the number and percent of establishments in which the county, ZIP code, and street address are present in the file. At the county level, for instance, the file consistently contains over 95 percent of the records with complete information. At the more detailed physical address and postal code scales information was only recorded for 22 percent of establishments in 1989 though coverage had dramatically improved by 1997 when 61 percent of records in the file contained complete address information. Of course, an address is only useful if it can be matched to geographic coordinates. The lower panel in Table 1 records the address-matching rates and resulting sampling rate for observations in the 1997 data. Postal codes are relatively easy to match whereas individual street address match-rates are near 80 percent. At the point-referenced spatial scale the effective national sampling rate is 55 percent. In other words, the resulting address geo-coded data constitutes a 55 percent sample of convenience.

A 55 percent sample is extremely large relative to the survey sampling rates used in social science. By contrast the decennial census long-form is a 20 percent sample and the largest Public Use Microdata Sample (PUMS) derived from the long-form is a 5 percent sample. The difference is that the PUMS file is designed as a representative sample whereas the derived spatially-referenced file from the national ES-202 file is a sample of convenience which emerges from several layers of unintentional censoring. Another paper that focuses specifically on the

quality of the geographic identifiers in the ES-202 file reveals that the censoring is not random but instead varies by the employment size of establishments, the rurality of address location, and other strata (Feser and Sweeney 2004). Spatial pattern inference based on the 55 percent sample is likely to produce biased results akin to those from the simulation study.

Thus, the ES-202 provides a perfect setting to apply the robust SRT estimation methods. For the application we use data for a single state, North Carolina, rather than work with the entire national file. Figure 6 is a map of the 4,692 points locations of manufacturing establishments (SIC codes 20 to 39) residing in 32 counties in North Carolina. [7] Research based on the address geo-coded North Carolina ES-202 data is reported in Sweeney and Feser (1998) and Feser and Sweeney (2002a, 2002b). In that research spatial pattern tests are used to identify whether a particular industry group exhibits spatial concentration relative to the general pattern of concentration evident in the settlement system. [8] In the present example we use the 2-digit industry classification (SIC) as the group indicator and we wish to compare the pattern of the locations of its establishments (the *cases*) to the pattern of all other manufacturing establishments (the *controls*). A simple test can then be constructed by differencing the two K-functions and constructing a test using random re-labelling. This is identical to Diggle's and Chetwynd's (1991) D-function methodology for inhomogeneous point processes. The manufacturing establishment data has the exact type of spatial censoring described in section two of this paper. We observe complete records for every establishment but only a subset of the establishment have complete point coordinate values derived from address geo-coding.

The North Carolina data presents an interesting complication for the SRT estimator. The lattice scale aggregation in this case is the county. As noted above, we include 32 counties in our study area but some of the counties contain very low match rates for the address geo-coding. The overall effective sampling rate for the data is 73.8 percent; significantly better than the match rate for the national file. However, across SIC codes the sampling rate ranges from 53 percent to 88 percent and across counties from 0 percent to 100 percent. Recall that the SRT estimator relies on the minimum sampling rate across strata to determine the re-sampling rates. Whereas the simulation results above allowed us the luxury of defining a $p_x > 0.5$, the real data imposes 100 percent censoring for some areal units making a direct application of the SRT estimator impossible.

One approach that makes SRT feasible is to nest the SRT algorithm in a county agglomeration algorithm. For each iteration of SRT we can combine counties to construct regions such that a minimum sampling rate threshold is satisfied. The modified procedure is as follows: (1) Choose a sampling threshold, (2) Calculate the minimum sampling rate, (3) If below the threshold, randomly merge that county with a neighbor, thereby generating a lattice with fewer cells and higher cell counts, (4) Recalculate the neighbors of the new lattice and the associated minimum sampling rate, and (5) If the rate is still below the threshold then repeat step 3.

Note that the procedure presents a balance between bias and precision. By combining all counties we can increase the threshold to the overall sampling rate (73.8 percent) and any re-sampling will return the naive estimate in which bias is suspected. By combining subsets

[7] SIC abbreviates Standard Industrial Classification. The two-digit SIC codes ranging from 20 to 39 contain the manufacturing industry sectors.

[8] Note that point pattern analysis of retail and manufacturing industries has a long history and new work continues to be published on the topic. Rogers' (1969a, 1969b) research using quadrat analysis to analyze retail establishments is exemplary of early work. K-functions were first used by Barff (1987) and recent work using point pattern methods to study industry location includes (Sweeney and Feser 1998, Feser and Sweeney 2000, Duranton and Overman 2002, Feser and Sweeney 2002a, Feser and Sweeney 2002b, Sweeney and Feser 2002, Marcon and Puech 2003)

of counties to achieve a threshold less than the overall sampling rate we degrade precision but move towards greater bias correction. Since the SRT procedure attempts to correct for selection bias, the threshold chosen should provide enough regions so that there is no evidence of selection bias within each aggregate region. As such, choosing a sampling threshold involves a statistical test of unbiasedness at each threshold as well as a consideration of its precision given the sub-sample size. Concern for mitigating bias argues for higher thresholds while concern for increased precision argues for lower thresholds. It may be possible to derive an optimal threshold by using a loss function to explicitly characterize the trade-offs. We leave that as a topic for a later paper. At this time we simply suggest that the threshold choice be informed by judgement and experimentation.

Given the foregoing discussion, we can proceed with the North Carolina industrial location analysis. As a first step we evaluate the county agglomeration effects of four different sampling thresholds. Table 4 contains the average number of regions for each two digit industry based on the sampling threshold indicated at the top of each column. For example, moving from a threshold of 0.3 to 0.6 for Textile Mill Products results in a substantial increase in granularity of our auxiliary information: 23 versus 8.1 counties. Next we calculate the SRT estimate for each industry and sampling threshold. Table 5 indicates the difference between the raw estimate and the SRT estimate for industry specific K-functions expressed as a percentage of the raw estimate. In general the difference increases as the sampling threshold decreases. That is, as expected, the more refined or stratified the spatial auxiliary information the more the SRT estimate diverges from the raw estimate. For a few industries, such as Fabricated Metal products, the largest difference is at the 0.5 threshold. These latter results suggest a redundancy in the information gained through spatial disaggregation of the auxiliary information.

As noted above, the D-function can be used to examine clustering relative to an inhomogeneous background pattern. Figure 7 shows the confidence envelopes for the SRT estimate and a raw estimate for the *control* K-functions and the *case* K-functions for the Printing and Publishing industry using a sampling threshold of 0.5. The SRT envelopes (solid lines) are formed by the minimum and maximum of the SRT estimates calculated using 10 different regionalizations and 30 stratified thinnings for each regionalization. The main finding is that the raw estimates for both the *cases* and *controls* lie either near the edge, or almost completely outside the envelopes for the SRT estimate. Thus, if the auxiliary information is to be trusted, our results suggest that inference should be based on the SRT estimate. The only unresolved issue is the selection of the best threshold value.

5 Conclusions

This paper has argued two main points. One has to do with the increasing relevance of spatial data in social science and the other has to do with circumspection in the application of that data. The advent of plentiful geo-coded data is continuing to transform the methods and research questions of social science. Indeed the relevance of many social science questions increases directly with spatial resolution. Tests of spatial mismatch, localization economies, and other theories are conceptually specified at spatial resolutions below the census block level. As such, just as the advent of microdata and panel data on individuals pushed research in labor economics in the 1980s (for example see Killingsworth (1983), Killingsworth and Heckman (1986)) spatially refined data allows for more direct tests of theories in urban economics and other areas of the social sciences where human decision-making plays out over a refined space-time setting.

Yet, as relevance increases directly with spatial resolution, the inverse is usually true of

data quality. Historically, space in the social sciences was operationalized as areal units with binary weight matrices indicating connectivity. For many research questions, lattice data and spatial econometric research methods are an entirely appropriate way to proceed. Indeed, spatial econometric tools such as Anselin's *GeoDa* software package are now widely used and represent an important step forward for social science research. For other research questions it makes no sense to proceed with spatial aggregates if the social science concepts refer to processes operating at a finer spatial resolution (Sweeney and Feser 2002). In fact, the prevalence of spatial analytic methods appropriate to lattice data is largely an artefact of historical data collection and dissemination techniques along with a cultural obsession with confidentiality. As point-referenced data sources become available the theories and analytic methods will likely adapt to the new setting. Both the simulation results and the real data suggest that spatial censoring does indeed result in biased estimates of spatial pattern and an alternative estimator is needed. Stratified random thinning is simple and it works in complex data situations. Importantly, it corrects estimates when bias is present and does not damage estimates when bias is absent.

The paper opens more research avenues than it resolves. We suggest three general approaches for combining the information from a sample Q and auxiliary data $P \setminus Q$. Stratified random thinning is one approach and the other two should be studied in future work. Inference for the SRT estimate is also somewhat problematic. Intuitively it seems that the resampling envelopes from the SRT estimator should be similar to those from randomly thinned sub-samples of the same size. Future work should examine that conjecture and whether methods from complex survey variance estimation would apply in this setting. Lastly, in the application we had to introduce a sampling threshold to randomly aggregate counties but we do not provide a solution to the optimal threshold selection problem.

References

Alonso, W.: 1968, Predicting best with imperfect data, *Journal of the American Institute of Planners* **34**, 248–255.

Aylin, P., Maheswaran, R., Wakefield, J., Cockings, S., Jarup, L., Arnold, R., Wheeler, G. and Elliot, P.: 1999, A national facility for small area disease mapping and rapid initial assessment of apparent disease clusters around a point source: the UK Small Area Health Statistics Unit, *Journal of Public Health Medicine* **21(3)**, 289–298.

Baddeley, A.: 1999, Spatial sampling and censoring, *in* O. Barndorff-Nielsen, W. Kendall and M. van Lieshout (eds), *Stochastic Geometry: Likelihood and Computation*, Chapman and Hall, CRC, p. 37.

Baddeley, A. and Gill, R.: 1997, Kaplan-Meier estimators of distance distributions for spatial point processes, *Annals of Statistics* **25(1)**, 263–292.

Bailey, T. and Gatrell, A.: 1995, *Interactive Spatial Data Analysis*, Longman, Essex.

Barff, R.: 1987, Industrial clustering and the organization of production: A point pattern analysis of manufacturing in cincinnati, ohio, *Annals of the Association of American Geographers* **77**, 89–103.

Cox, D. and Oakes, D.: 1984, *Analysis of Survival Data*, Chapman and Hall, London.

Diggle, P.: 1977, Discussion of Dr Ripley's paper, *Journal of the Royal Statistical Society B* **39**, 196–197.

Diggle, P.: 1978, On parameter estimation for spatial point processes, *Journal of the Royal Statistical Society B* **40**, 178–181.

Diggle, P.: 1983, *Statistical Analysis of Spatial Point Patterns*, London Academic Press, London.

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

Duranton, G. and Overman, H.: 2002, Testing for Localization Using Micro-Geographic Data. *CEPR Discussion Paper No. 3379*, Department of Geography and Environment, London School of Economics.

Feser, E. and Sweeney, S.: 2000, A test for coincident economic and spatial clustering among business enterprises, *Journal of Geographical Systems* **2**, 349–373.

Feser, E. and Sweeney, S.: 2002a, Spatially binding linkages in manufacturing product chains, *in* R. McNaughton and M. Green (eds), *Global Competition and Local Networks*, Ashgate, p. 111.

Feser, E. and Sweeney, S.: 2002b, Theory, methods, and a cross-section comparison of business clustering, *in* P. McCann (ed.), *Industrial Location Economics*, Edward Elgar, p. 222.

Feser, E. and Sweeney, S.: 2004, On the state of the geography in the U.S. BLS Covered Wages and Employment (ES-202) series, *Working Paper* .

Fisher, P.: 1999, Models of uncertainty in spatial data, *in* P. Longley, M. Goodchild, D. Maguire and D. Rhind (eds), *Geographical Information Systems, Volume 1: Principles and Technical Issues, 2nd edition*, John Wiley, p. 191.

Getis, A.: 1984, Interaction modeling using second-order analysis, *Environment and Planning A* **16**, 173–183.

Goodchild, M.: 2000, New horizons for the social sciences: Geographic Information Systems, *Sciences for a Digital World: Building Infrastructure and Databases for the Future.*, Organisation for Economic Cooperation and Development, p. 163.

Goodchild, M., Anselin, L., Appelbaum, R. and Harthorn, B.: 2000, Toward spatially integrated social science, *International Regional Science Review* **23(2)**, 139–159.

Heuvelink, G.: 1999, Propagation of error in spatial modelling with gis, *in* P. Longley, M. Goodchild, D. Maguire and D. Rhind (eds), *Geographical Information Systems, Volume 1: Principles and Technical Issues, 2nd edition*, John Wiley, p. 207.

Hillier, A. and Culhane, D.: 2003, Predicting housing abandonment with Philadelphia's Neighborhood Information System, *Urban Affairs* **25(1)**, 91–105.

Janelle, D.: 2002, *Center for Spatially Integrated Social Science, Annual Report to the National Science Foundation, Year 3 (1 July 2001 - 30 April 2002)*. University of California, Santa Barbara, April 2002.

Killingsworth, M.: 1983, *Labor Supply*, Cambridge University Press, Cambridge.

Killingsworth, M. and Heckman, J.: 1986, Female labor supply: a survey, *in* O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics, Volume II*, Elsevier, p. 103.

Krewski, D. and Rao, J.: 1981, Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods, *Annals of Statistics* **9**, 1010–1019.

Little, R. and Rubin, D.: 1987, *Statistical Analysis with Missing Data*, John Wiley, New York.

Longley, P., Goodchild, M., Maguire, D. and Rhind, D.: 1999, *Geographical Information Systems, Volume 1: Principles and Technical Issues, 2nd edition*, John Wiley, New York.

Longley, P., Goodchild, M., Maguire, D. and Rhind, D.: 2001, *Geographic Information Systems and Science*, John Wiley, New York.

Marcon, E. and Puech, F.: 2003, Evaluating the geographic concentration of industries using distance-based methods, *Journal of Economic Geography* **3(4)**, 409–428.

Matern, B.: 1971, Doubly stochastic Poisson process in the plane, *Statistical Ecology* **1**, 195–213.

McLachlan, G. and Krishnan, T.: 1997, *The EM Algorithm and Extensions*, John Wiley, New York.

National Institute of Child Health and Human Development: 2002, emphGoals and Opportunities: 2002-2006, Demographic and Behavioral Sciences Branch, Center for Population Research, National Institute of Child Health and Human Development.

Oppenshaw, S.: 1984, The modifiable areal unit problem. Concepts and Techniques in Modern Geography 38. Norwich, UK: GeoBooks.

Pivetz, T., Searson, M. and Spletzer, J.: 2001, Measuring job and establishment flows with BLS longitudinal microdata, *Monthly Labor Review* **124(4)**, 13–20.

Politis, D., Romano, J. and Wolf, M.: 1999, *Subsampling*, Springer, New York.

Rao, J.: 2003, *Small area estimation*, Wiley, New York.

Rao, J. and Shao, J.: 1999, Modified balanced repeated replication for complex survey data, *Biometrika* **86**, 403–415.

Ripley, B.: 1976, The second-order analysis of stationary point processes, *Journal of the Applied Probability* **13**, 255–266.

Ripley, B.: 1977, Modeling spatial patterns (with discussion), *Journal of the Royal Statistical Society B* **39**, 172–212.

Rogers, A.: 1969a, Quadrat analysis of urban dispersion: 1. theoretical techniques, *Environment and Planning* **1**, 47–80.

Rogers, A.: 1969b, Quadrat analysis of urban dispersion: 2. case studies of retail systems, *Environment and Planning* **1**, 155–171.

Rushton, G. and Armstrong, M.: 1997, *Improving Public Health through Geographic Information Systems: An Instructional Guide to Major Concepts and their Implementation*. CD-ROM, Department of Geography, The University of Iowa.

Spletzer, J.: 1997, Longitudinal establishment microdata at the Bureau of Labor Statistics: Development, uses, and access, *Proceedings of the Section on Survey Research Methods*, American Statistical Assoc.

Stevens, D.: 1994a, *Confidentiality and the design of a National Wage Record Database*. Washington, DC: Division of Occupational and Administrative Statistics, Bureau of Labor Statistics, U.S. Department of Labor.

Stevens, D.: 1994b, *Research Uses of Wage Record Data: Implications for a National Wage Record Database*. Washington, D.C.: Division of Occupational and Administrative Statistics, Bureau of Labor Statistics, U.S. Department of Labor.

Stevens, D.: 1994c, *The use of UI wage records for JTPA performance management in Maryland*. Baltimore: Office of Employment Training, Maryland Department of Labor, Licensing, and Regulation.

Sweeney, S. and Feser, E.: 1998, Plant size and clustering of manufacturing activity, *Geographical Analysis* **30(1)**, 45–64.

Sweeney, S. and Feser, E.: 2002, Spatial externalities: theoretical and measurement issues, *in* M. Goodchild and D. Janelle (eds), *Spatially integrated social science: examples in best practice*, Oxford University Press, p. 239.

Veregin, H.: 1999, Data quality parameters, *in* P. Longley, M. Goodchild, D. Maguire and D. Rhind (eds), *Geographical Information Systems, Volume 1: Principles and Technical Issues, 2nd edition*, John Wiley, p. 177.

Wolpert, R. and Ickstadt, K.: 1998, Poisson/gamma random field models for spatial statistics, *Biometrika* **85(2)**, 251–267.

Yule, G. and Kendall, M.: 1950, *An Introduction to Statistics*, Hafner, New York.

Table 1. Average errors for random unbalanced subsets

	Uncorrected	Corrected	Improvement
Errors			
Mean Absolute Error	0.00175	0.00039	77.7%
Percent Error	11.46%	3.02%	73.6%
Mean Error	0.00576	0.00117	79.7%
Error at D=.075	0.00158	0.00058	63.2%
Error at D=.15	0.00447	0.00130	70.9%
Error at D=.225	0.00986	0.00199	79.8%

* Errors are calculated for 2,000 randomly selected unbalanced subsets from the domain of 989,000 with p_x <0.5.

Table 2. Average errors for random thinnings

Measure	Uncorrected	Corrected	Improvement
Mean Absolute Error	0.00041	0.00035	15.0%
Percent Error	3.04%	2.68%	12.0%
Mean Error	0.00124	0.00103	17.2%
Error at D=.075	0.00053	0.00051	3.6%
Error at D=.15	0.00125	0.00114	8.3%
Error at D=.225	0.00220	0.00182	17.3%

* Errors are calculated for 100 random thinning subsets of 755 points.

Table 3. ES-202 geography, all sectors

A. Records with non-missing spatial identifiers, 1989 and 1997

	1989		1997	
	Units	Pct	Units	Pct
Total	5,739,712	100.0	7,228,085	100.0
With county	5,592,297	97.4	6,967,605	96.4
With zip code	1,280,402	22.3	4,402,816	60.9
With street address	1,267,575	22.1	4,401,632	60.9

B. Censoring and effective sampling rates, 1997

	Postal codes		Street addresses	
	% match	% sample	% match	% sample
United States	99.6	59.0	82.1	48.6
California	99.7	53.4	85.5	45.8
Florida	99.6	53.9	79.3	42.9
North Carolina	99.6	85.0	76.9	65.6
Pennsylvania	99.4	49.6	86.6	43.2
Wisconsin	99.4	91.1	75.5	69.2

All contiguous states but Massachusetts and Wyoming. Based on 3d quarter files.

Table 4. Average number of county conglomerates by threshold

	Threshold			
Industry	0.3	0.4	0.5	0.6
Food & Kindred Products	17.8	15.9	11.2	8.1
Textile Mill Products	23	19.4	14.3	8.1
Apparel & Other Textile Products	24.5	20.8	15.8	9.4
Furniture & Fixtures	20.8	19.2	15.4	7.5
Paper & Allied Products	21.3	16.7	12.7	8.9
Printing & Publishing	23.2	20.8	15.9	9
Chemical & Allied Products	21.5	19.7	15.6	9.6
Rubber & Miscellaneous Plastics Products	25.3	20	15.6	9.6
Stone, Clay, & Glass Products	20.7	18.4	12.3	7.7
Fabricated Metal Products	24.6	21	15.5	9.7
Industrial Machinery & Equipment	23.9	19.6	16.5	10.4
Electronic & Other Electric Equipment	21.3	18.2	13.5	9.7
Transportation Equipment	18.7	16.2	11.1	7.2

Table 5. Change in estimate by sampling threshold

	Threshold			
Industry	0.3	0.4	0.5	0.6
Food & Kindred Products	19.0%	14.6%	12.1%	4.0%
Textile Mill Products	9.3%	9.4%	6.6%	3.0%
Apparel & Other Textile Products	13.9%	8.9%	8.4%	2.2%
Furniture & Fixtures	2.7%	3.2%	2.7%	1.2%
Paper & Allied Products	2.5%	2.0%	1.8%	1.4%
Printing & Publishing	3.8%	5.6%	5.8%	5.1%
Chemical & Allied Products	3.1%	2.9%	5.2%	3.4%
Rubber & Miscellaneous Plastics Products	4.7%	3.4%	1.4%	0.8%
Stone, Clay, & Glass Products	15.0%	11.8%	4.8%	1.4%
Fabricated Metal Products	4.5%	6.6%	8.5%	6.3%
Industrial Machinery & Equipment	5.6%	6.3%	5.1%	2.7%
Electronic & Other Electric Equipment	2.8%	5.5%	6.0%	2.9%
Transportation Equipment	9.7%	4.2%	7.2%	2.1%

* regionalizations=10, stratified thinnings=30.

Figure 1. Matern(29,.061) process realization

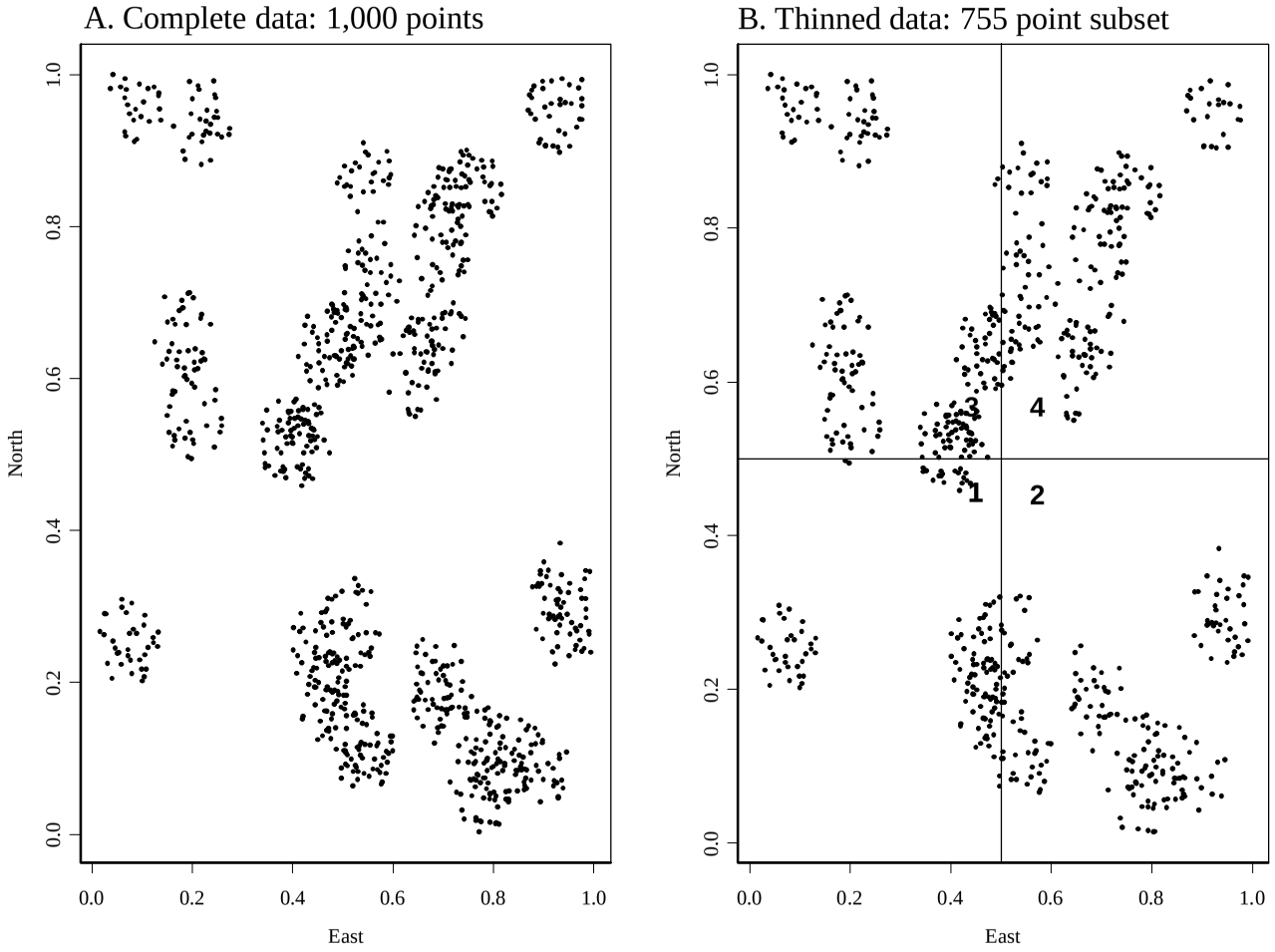


Figure 2. K function means and confidence envelopes for Matern(29,.061)

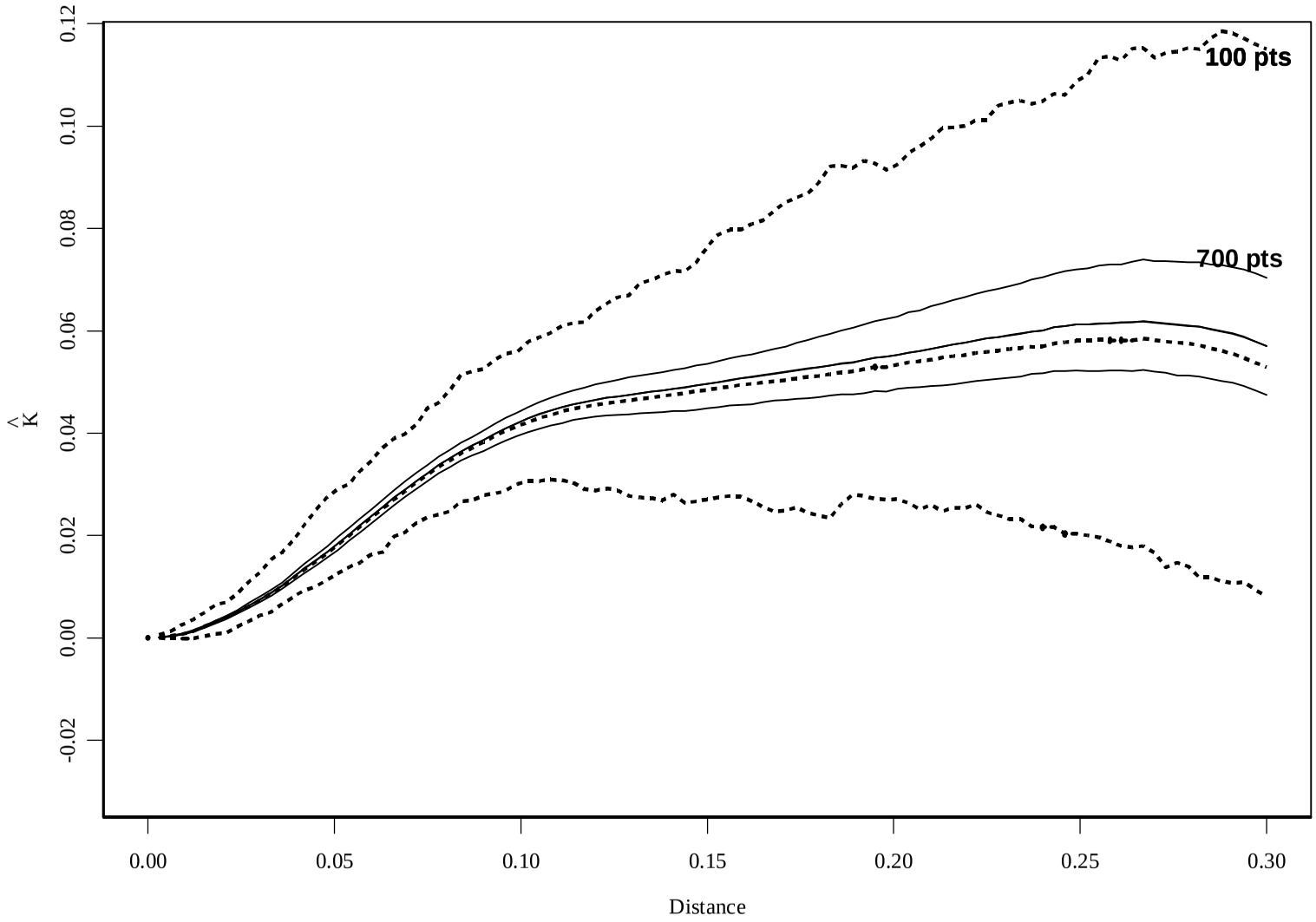


Figure 3. K-functions: true value (K_p), raw estimate (K_q) and SRT estimator (K_R)

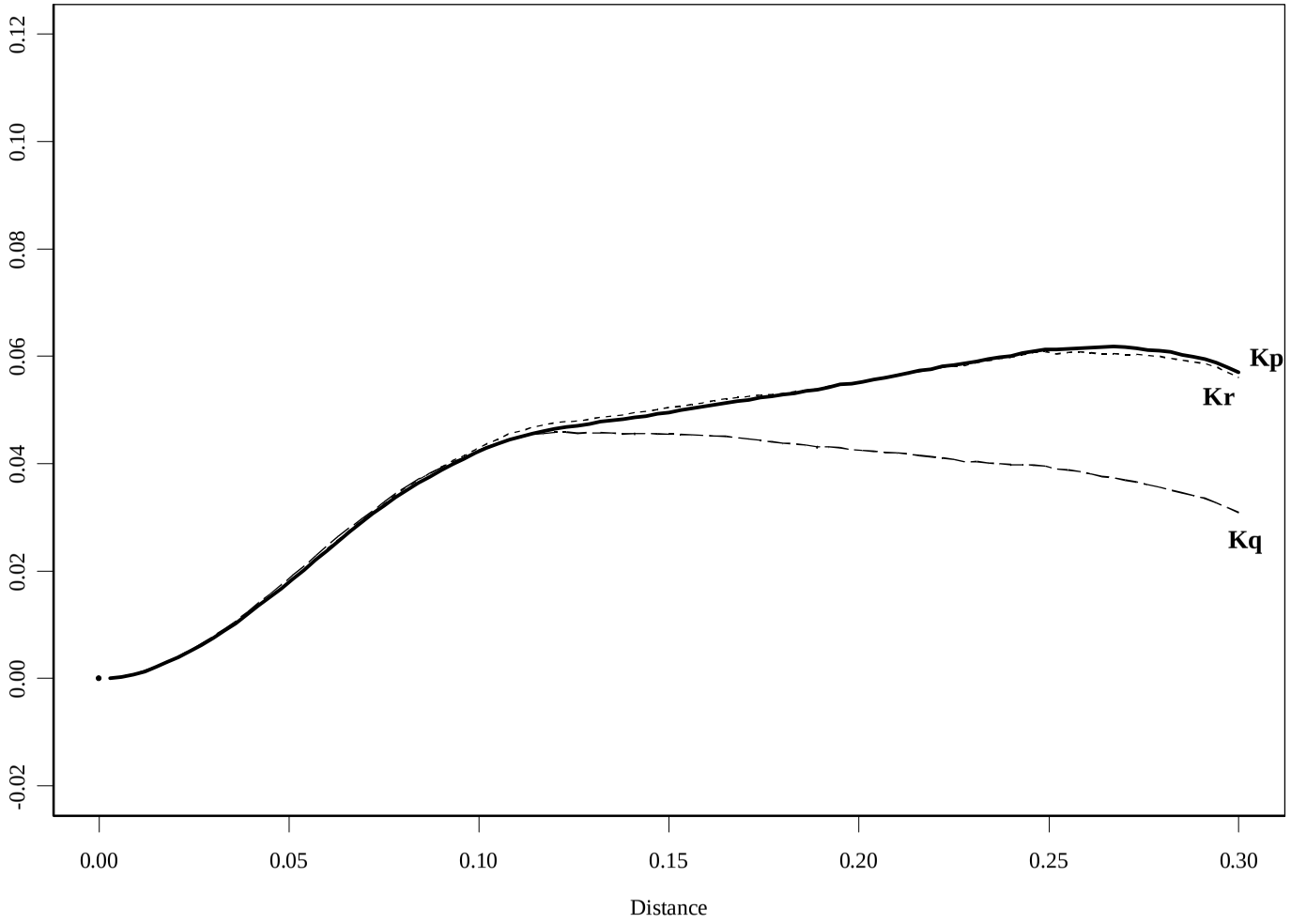


Figure 4: SRT versus raw estimator for unbalanced subsets

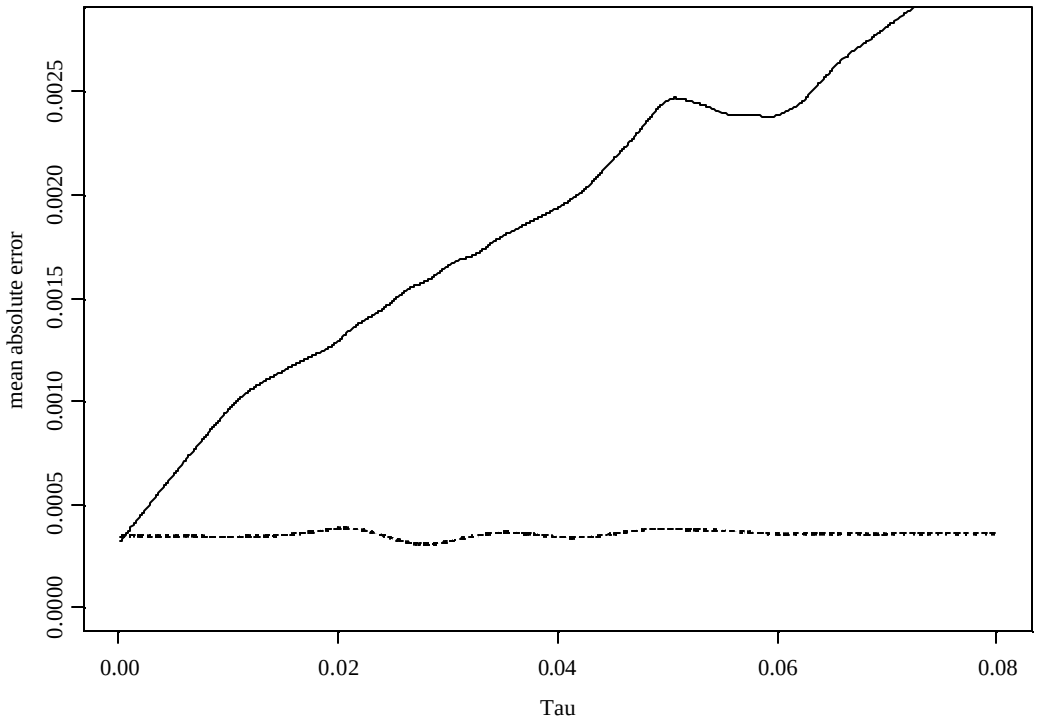


Figure 5: SRT versus raw estimator for random thinning subsets

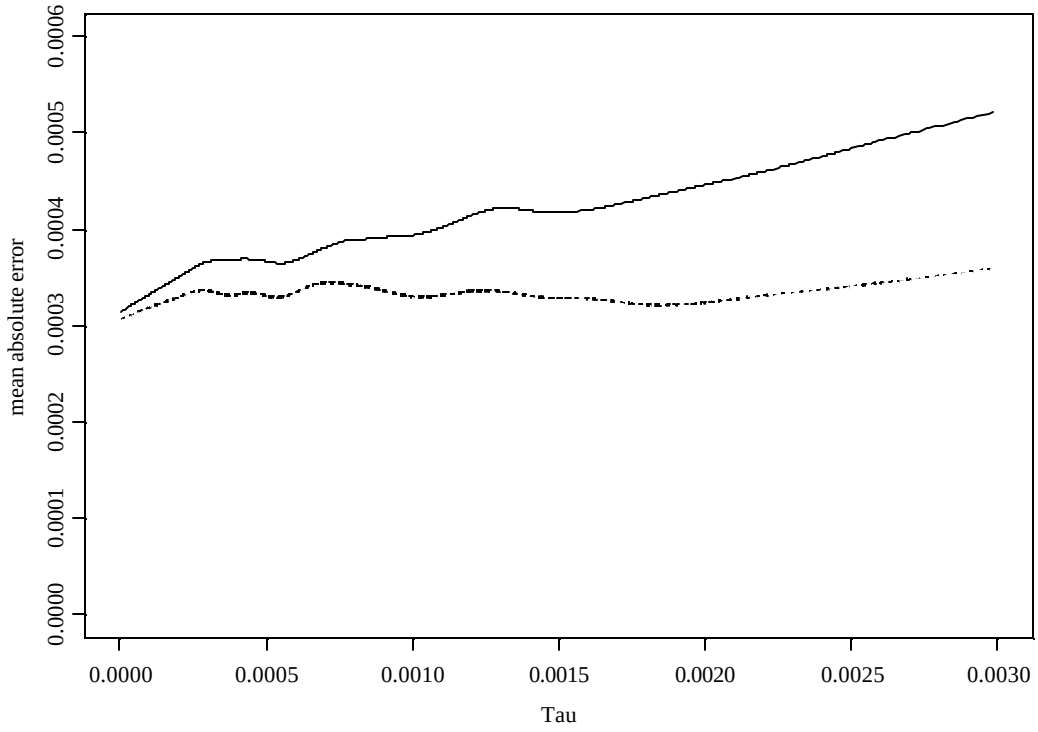
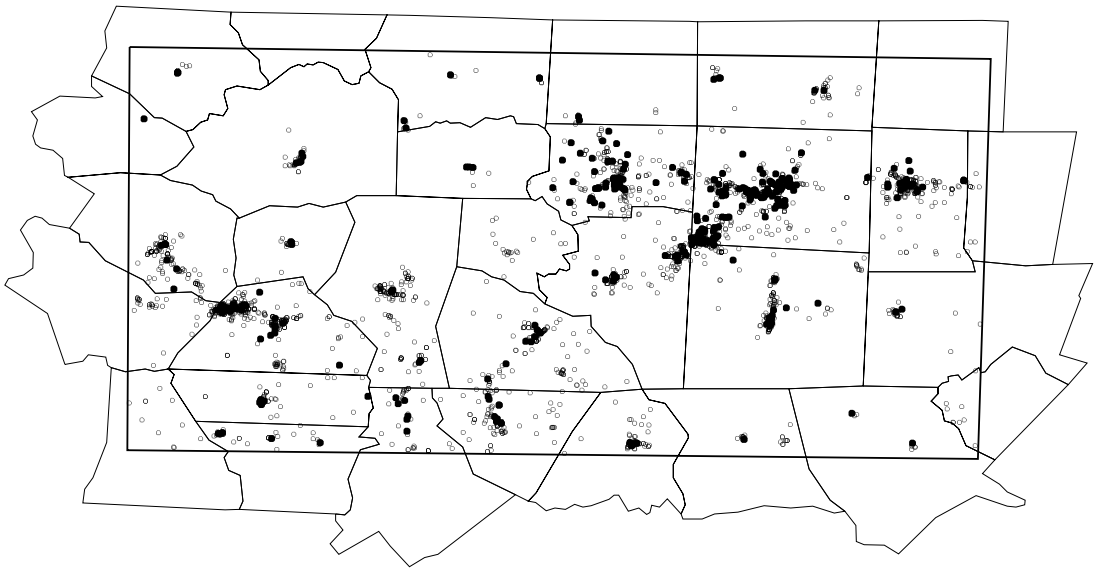


Figure 6: North Carolina manufacturing establishments



Note: Data is from North Carolina, Employment Security Commission, ES-202 files. Point coordinates are identified using address geo-coding. The solid circles indicate establishments in SIC 27 and the open circles represent all other manufacturing establishments.

Figure 7. Raw estimates and SRT confidence envelopes for North Carolina manufacturing establishments