SSOAR

Open Access Repository
www.ssoar.info

The case for spatially-sensitive data: how data structures affect spatial measurement and substantive theory

Chan-Tack, Anjanette M.

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Chan-Tack, A. M. (2014). The case for spatially-sensitive data: how data structures affect spatial measurement and substantive theory. *Historical Social Research*, *39*(2), 315-346. https://doi.org/10.12759/hsr.39.2014.2.315-346

Nutzungsbedingungen:
Dieser Text wird unter einer CC BY Lizenz (Namensnennung) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier:
https://creativecommons.org/licenses/by/4.0/deed.de

Terms of use:
This document is made available under a CC BY Licence (Attribution). For more Information see:
https://creativecommons.org/licenses/by/4.0

gesis
Leibniz-Institut
für Sozialwissenschaften

Mitglied der
Leibniz-Gemeinschaft

Diese Version ist zitierbar unter / This version is citable under:
https://nbn-resolving.org/urn:nbn:de:0168-ssoar-384875

The Case for Spatially-Sensitive Data: How Data Structures Affect Spatial Measurement and Substantive Theory

Anjanette M. Chan-Tack [*]

Abstract: *»Raumsensible Daten: Wie Datenstrukturen räumlich-geographisches Messen substantielle Theorie beeinflussen«.* Innovations in GIS and spatial statistics offer exciting opportunities to examine novel questions and to revisit established theory. Realizing this promise requires investment in *spatially-sensitive* data. Though convenient, widely-used administrative datasets are often spatially insensitive. They limit our ability to conceptualize and measure spatial relationships, leading to problems with ecological validity and the MAUP – with profound implications for substantive theory. I dramatize the stakes using the case of supermarket red-lining in 1970 Chicago. I compare the analytical value of a popular, spatially insensitive administrative dataset with that of a custom-built, spatially sensitive alternative. I show how the former constrains analysis to a single count measure and aspatial regression, while the latter's point data support multiple measures and spatially-sensitive regression procedures; leading to starkly divergent results. In establishing the powerful impact that spatial measures can exert on our theoretical conclusions, I highlight the perils of relying on convenient, but insensitive datasets. Concomitantly, I demonstrate why investing in spatially sensitive data is essential for advancing sound knowledge of a broad array of historical and contemporary spatial phenomena.

Keywords: Spatial regression, spatially-sensitive data, spatial measurement, ecological validity, Modifiable Areal Unit Problem (MAUP), retail red-lining, supermarket access, neighborhood effects.

1. Introduction

1.1 The Spatial Revolution

In recent years, "space" has received renewed attention as a category of analysis across the social sciences; from sociology, anthropology, and history, to demography, criminology and health studies (Goodchild and Janelle 2004).

[*] Anjanette M. Chan-Tack, Department of Sociology, University of Chicago, 1126 East 59th Street, Chicago, IL 60637, USA; amc75@uchicago.edu.

This explosion of interest in "spatial thinking" and spatial methods as well as new calls for "spatially integrated social science" stands on the shoulders of a technological revolution which has made it possible to manipulate spatial information with unprecedented ease. It also builds on new bodies of statistical theory and user-friendly software packages, which have made spatial modeling accessible to the wider scholarly community (Anselin and Rey 2010).

The new "spatial turn" thus offers exciting opportunities. First, it has opened up *vast new data vistas*. In a strong sense, the spatial revolution is a data revolution. The technological advances on which the spatial turn is built have made old data that was previously difficult to handle far more accessible to researchers. Researchers have not only gained a new capacity to process old data stored in truculent forms through tools like map digitization, but also new ways to manipulate and operationalize data; to measure with greater accuracy; and to bring different kinds of data from multiple sources into unprecedented conversation. These breakthroughs in data accessibility and manipulability have prompted demand for even more *geo-referenced data*, gathered at ever-finer levels of granularity, and also dynamically, in real time (Entwistle 2007).

Innovation in statistical theory and methods is the second major engine driving the contemporary spatial turn in social science. The challenges of incorporating spatial structure, geographic information and locational data into statistical models – and the pitfalls of excluding spatial effects in model estimation – have long been recognized by quantitative geographers and statisticians (Cliff and Ord 1984; Openshaw and Taylor 1979). However, widely accessible statistical methods for handling the challenges that spatial data represent have become available only relatively recently. Consequently, in addition to a vast expansion of our data vistas, the spatial turn in social science offers at least two other important opportunities:

1) the ability to more precisely estimate parameters of interest in spatial phenomena,
   and
2) the ability to explore the substantive roles that space itself plays as a determinant of our outcomes of interest.

1.2 The Spatial Revolution and Urban Sociology

Until recently, spatial methods and theory have seen limited use in urban sociology. Their sparse use has not been for want of either potential or need. Urban sociology heavily thematizes space, place and ecology. It thus stands to reap enormous theoretical gains from spatial theory and from new spatial modeling techniques. This is especially so in two related lines of urban sociological research: the residential segregation and neighborhood effects literatures. Space and spatial concepts are ontologically fundamental to these sub-fields. *Sociological interest in residential segregation* is undergirded by the belief that:

1) social characteristics are spatially patterned;
2) spatial distance reflects social distance;
3) that specific patterns of spatial concentration beget particular kinds of social and ecological relations; and that
4) these relations are consequential for a broad range of social processes.

Interest in residential spatial patterning began with Robert Park (1926) and continues today. It is implicit in the neighborhood effects literature – the dominant stream of research within urban sociology. This literature contends that spatial isolation by race and class is itself an important force contributing to racial inequality. For example, it is believed that minority deficits in safety, health, education, and employment arise from the historical and contemporary segregation of neighborhood spaces (Briggs et al. 2010; Sampson 2012).

Both residential segregation and neighborhood effects research use "the neighborhood" as their primary unit of analysis. From a substantive viewpoint, it is obvious that processes occurring in one neighborhood typically affect similar processes occurring in surrounding areas. Thus, quantitative analyses of neighborhood-level processes should *always* consider the effects of *spatial interdependence*. Starting in the 1990s, a handful of adventurous senior scholars began experimenting with spatial methods, and to a lesser extent, spatial concepts (Sampson, Morenoff and Earls 1999; Tolnay, Deane and Beck 1994; Reardon and O'Sullivan 2004).[1] Thanks to their pioneering efforts, spatial methods are now beginning to take off within urban sociology.

The early adopters were excited by the novel ways to theorize spatial social processes and by the analytical precision that spatial regressions offered. In the terminology of spatial methodologists, they were especially concerned to deal with *spatial autocorrelation* (Netrdová and Nosek 2014, in this HSR Special Issue) between predictors and outcome variables, both analytically and theoretically. They spent less time thinking about spatial theory. In particular, their attention to the crucial issue of *ecological validity* – an issue which sits at the intersection of spatial ontology and quantitative spatial analysis – was almost non-existent (but see Downey 2003, 2005, 2006). Among those who developed spatial theory and spatial statistics, ecological validity has long been considered a central issue in research design (Openshaw 1984).

Careful thought about ecological validity ought to guide the logic of variable construction and the choice of spatial analytic units. As spatial methods diffuse widely through the urban sociology and residential segregation research communities, scholars need to think more carefully about how to construct analyses that are appropriate to the research question at hand. The issue is particularly

[1] Doriean (1980, 1981, 1982) introduced spatial regression methods to the sociological research community much earlier. Sociologists began to apply them just over a decade later.

urgent for scholars using *historical data*, as these data offer far less flexibility in operationalizing spatial concepts.

2. Theoretical and Methodological Issues

2.1 The Question of Ecological Validity: Urban Sociology and the Modifiable Areal Unit Problem (MAUP)

While the spatial revolution offers great promise for urban sociology, the question of ecological validity and the attendant *Modifiable Areal Unit Problem (MAUP)* present quantitative urban sociologists with a fundamental theoretical and methodological challenge. The term *ecological validity* refers to a situation in which the spatial units used in an analysis are relevant to the social processes under examination.

The MAUP tells us that if the spatial units chosen lack ecological validity, the results of the analysis are likely to be highly misleading. Specifically, the MAUP tells us that the magnitude, sign, and significance of variables in regression equations change unpredictably as the scale and boundaries of the spatial units are altered (Cliff and Ord 1981; Openshaw 1984). In all regression analyses, we want our estimates to provide us with reliable information about the significance and strength of the relationship between predictors and outcomes. The MAUP therefore tells us that for regression analyses of spatial phenomena to have any reliable substantive meaning, the choice of ecologically valid spatial units is essential.

The MAUP offers a particularly serious challenge to areal analyses, such as those common in the neighborhood effects literature. Most regression analyses in this tradition use *administrative constructs* such as census tracts, zip codes and block groups as proxies for real *social* neighborhoods. These administrative spatial constructs have been created by states for purposes other than the analysis of social processes of interest to researchers. For example, zip codes are areal units created by the US Postal Service to ensure the efficient delivery of mail along the street grid. Census tracts and census blocks are constructs of the US Census. Their boundaries are drawn in ways that attempt to maximize the homogeneity of the populations within them (with respect to income, race, population size, etc.), in ways that make inter-censal comparisons easier.

When analysts use zip codes, tracts and blocks as proxies for real social neighborhoods, they implicitly make a *number of assumptions*. Namely:

1) that the social processes and relationships that constitute real neighborhoods start, end, and are *homogeneous within the boundaries of these administratively defined spatial units*,

2) that the *scale* and *boundaries* of these spatial units accurately represent the spatial extent over which the variety of phenomena that constitute real social neighborhoods actually occur.

In reality, the correspondence between the social networks, institutional affiliations, etc. which constitute real neighborhoods and the diverse array of administratively-defined spatial units is at worst dubious, and at best, an open question. The correspondence between administrative spatial units and other social processes – such as the spatial dispersion of crime, and the spatial distribution of amenities – is similarly open to debate.

Most sociologists who have adopted spatial methods have spent little time thinking about how ecological validity should shape the spatial units they use in their analysis, or about the ways that it should guide their variable construction techniques. Not surprisingly, they have also avoided considering the implications of the MAUP for the validity of their findings.

There have been exceptions. Within residential segregation research (O'Reardon and O'Sullivan 2004) have discussed the MAUP in terms of "zoning effects" (Openshaw's "boundary problem") and "scale effects" (Openshaw's "areal unit size"). Within neighborhood effects research on crime, Hipp (2007) attempted to ascertain how regression results change when data aggregated to different officially-available census spatial units (from blocks to tracts) are used. Downey (2005, 2006), who examines neighborhood effects on exposures to environmental inequalities, has been the only researcher to use questions about ecological validity to guide him in his variable construction techniques, and in his choice of spatial units of analysis. These exceptions have been few and far between. Moreover, they have almost exclusively focused on analyses of contemporary phenomena. Little attention has been paid to how attention to ecological validity might influence historical urban research.

2.2 The MAUP: Challenges for Historical Urban Research

The spatial revolution offers great promise (through new methods and theory) and significant challenges (through the MAUP and the question of ecological validity) to the conduct of urban sociological research. Seen in one light, the MAUP is eminently soluble. One need only construct variables and select units of analysis that are appropriate to the research question. However, this requires *spatially flexible datasets* – datasets which allow researchers to build spatial units and measures that are tailored to the processes that they are studying.

Unfortunately for historical researchers, most readily available spatial datasets were gathered before the development of contemporary GIS capabilities. While all quantitative researchers examining historical spatial phenomena face the problem of finding spatially flexible data that will allow them to address the MAUP, the acuteness of the problem varies with the historical period and the

nature of the spatial process. A key distinction is whether the process is *spatially continuous* or *spatially discontinuous*.

For example, historical demographers examining the 1st and 2nd demographic transitions deal with data that was collected before the existence of nation states, or before the maturation of state bureaucracies' data-collection procedures (e.g. Bocquet-Appel and Jakobi 1998).[2] The data they work with are typically collected over vast areas, and are often riddled with missing values. Luckily, most demographers work on *spatially continuous, or "smooth" phenomena*. These are phenomena whose features are assumed to vary gradually over space. Those who study with spatially continuous processes have a number of techniques (such as *kriging* and *wombling*) [3] which they can use to reconstruct sparse data. These techniques allow researchers to impute values for any point on the smooth surface, allowing great flexibility in modeling space (Stein 1999).

Quantitative historical urbanists also have to contend with major data limitations. Historically, government data-collection agencies have aggregated to phenomena into spatial units (such as counties) that are too large to plausibly be considered neighborhoods. In addition, the existence of residential segregation means that social types (the rich, the poor, immigrants, ethnic majorities and minorities) are clustered, and display uneven spatial distribution. For research on urban social phenomena, *spatial discontinuity* is thus the operating ontological assumption. Methods for estimating values for spatially continuous processes cannot be applied to most urban sociological questions (Anselin 2002 and Section 5.3.1, below). At the same time, dealing with the boundary problem in urban historical research is a particular challenge because neighborhood boundaries are *themselves* products of spatial social processes.

The dynamics of racial residential in America illustrate this well. U.S. urban history is littered with instances where White-controlled city governments placed physical barriers, such as major highways and rail-lines, between White

---

[2] The 1st and 2nd demographic transitions are terms used to describe major changes in the nature of population growth, which began in 18th Century Europe, and in the West in the late 20th Century. The 1st transition saw dramatic declines in mortality, and a fall in the fertility rate to replacement levels. The 2nd transition saw a continued decline in fertility to sub-replacement levels (see Lesathaeghe 2010).

[3] *Kriging* and *wombling* are computational methods that can be used to estimate the properties of spatially continuous, or "smooth", surfaces. When spatially continuous surfaces suffer from missing data, analysts use *kriging* to estimate the values of the missing data based on the spatial distribution of known values (Stein 1999). *Wombling* is used to identify zones of major change on a continuous surface. It does so by first measuring the *rate of change* of values between contiguous points on the surface. Zones of major change are those whose gradients are at a maximum relative to the gradient distribution on the whole surface (Barbujani et al. 1989). Both kriging and wombling methods rely on a "field" statistical approach to spatial phenomena. See Section 5.3.1 of this paper for more information on "field" vs. "object" statistical approaches.

and non-White neighborhoods in ways that were intended to exacerbate their social separation (Hirsch 1998; Hunt 2010). This segregation generated racial inequalities in a host of subsequent social processes, including the allocation of tax-bases to municipalities, children to schools, individuals to employment opportunities, as well as differential rates of crime, wealth, poverty, amenities, housing and land values to neighborhoods (Wilson 1978, 1987; Sharkey 2012). Processes that generate observed neighborhood-level inequalities are thus *endogenous* to neighborhood boundaries. The MAUP tells us that for urban studies to have ecological validity, the *scale* and *shape* of spatial units used to represent neighborhoods must accurately reflect these social processes. Unfortunately, the socio-physical processes embedded in the urban built environment are rarely captured by administratively constructed spatial units.

The MAUP thus presents quantitative scholars of urban history with a particularly thorny set of problems. Fine-grained geographic data on contemporary phenomena are widely available from sources like Google Maps, as well as from for-profit GIS firms. But public and private bodies have put little effort into creating spatially flexible datasets for historical phenomena.[4] For individual researchers, building spatially flexible datasets is time-consuming and expensive. In addition, the MAUP is also not widely known among urban scholars. In this situation, it is tempting to simply pretend it does not exist. It is therefore not surprising that urban researchers investigating historical subjects often resort to pre-made administrative spatial units, despite their poor ecological validity.

While researchers' reluctance to build spatially flexible datasets is understandable, my paper goes against this grain. I argue that the harder work of creating appropriate datasets through archival research is necessary. I demonstrate the value of creating spatially-flexible, ecologically-valid datasets through an empirical examination of racial inequalities in amenity access in 1970 Chicago. For tractability, I focus on one important amenity: supermarkets, a widely desirable neighborhood establishment that affects communities' physical and economic health (Morland, Diez-Roux and Wing 2006; Porter 1995). I perform parallel analyses on two datasets which offer contrasting levels of spatial sensitivity and permit distinct constructions of "access".

The *spatially insensitive dataset* is count-based and is typical of official datasets most readily available to researchers. The "*spatially sensitive" dataset* is an original census of point-locations. Through this paper, I show:
1) how constructs arising from these datasets differ in their ecological validity;

[4] There are a handful of exceptions, see for example, the Center of Population Economics at the University of Chicago Booth School of Business <http://research.chicagobooth.edu/cpe/about-cpe>, as well as John Logan's work <http://www.s4.brown.edu/S4/P-Logan.htm>.

2) how these contrasting spatial constructs lead to distinct regression techniques; and
3) how these distinct measures and models affect our conclusions about the role that race and class play in distributing disadvantage in American urban space.

Below, I first give a brief sketch of the substantive value of this particular empirical question. I then introduce the datasets, and concentrate my discussion on the issue of ecological validity, and spatial variable construction. I present the two models and their results. I conclude by discussing the implications of their divergent results on substantive theory, and by calling for more funding to be devoted to the creation of spatially flexible datasets for historical researchers.

2.3 The Substantive Issue: Amenities, Inequality and Neighborhood Effects Research

Examining the role that neighborhoods play in distributing and reproducing racial and class-based disadvantage has been a key line of American sociological research since at least the late 1960s. A good deal of this work was organized around the seminal debate between William Julius Wilson (1978) and Douglas Massey (Massey and Denton 1993). These authors offered competing claims about the role that race and class play in distributing disadvantage across neighborhoods. Wilson (1978) argued that in the wake of the Civil Rights Movement class, not race, would be the chief determinant of neighborhood inequalities from the 1960s onward. In contrast, Massey argued that racial discrimination was likely to persist, and that both race and class would remain persistent determinants of neighborhood inequality.

The Wilson-Massey debate inspired a vibrant tradition of quantitative "neighborhood effects" research. Recently, scholars have turned their attention to the role that organizations might play in mediating neighborhood disadvantage. Studies in this line have found that neighborhood amenities, such as child care centers and supermarkets, can improve the mental and dietary health, and reduce mortality among residents (Small 2009; Browning, Wallace, Feinberg et al. 2006; Morland, Diez-Roux and Wing 2006). These findings have prompted attempts to systematically examine the determinants of neighborhood inequalities in amenity access (Small and McDermott 2006).

Small and McDermott's (2006) study is illustrative of the blind-spots many urban sociologists still have with respect to quantitative spatial analysis. While wide-ranging – the authors examine access to more than ten amenities across all U.S. Metropolitan Statistical Areas (MSAs) – Small and McDermott (2006) fail to use spatial statistical techniques for an obviously spatial phenomenon. In addition, the authors fail to think carefully about the ecological validity of the

variables and the spatial units used in their analysis. Given the MAUP, this leaves the validity of their results open to question.

A second gap in urban sociological research on amenity access is the absence of an historical angle. Urban sociologists, policy-makers and the wider public are deeply concerned about contemporary patterns of inadequate amenity access in poor and minority neighborhoods. But ethnographic accounts suggest that in the 1950s to the 1970s, these very same poor and minority neighborhoods were rich in amenities. What accounts for this dramatic change? The only way to answer this question is to take an historical approach. To understand how neighborhoods that were amenity-rich in 1970 became amenity-poor today, we need to: (1) understand how these neighborhoods changed, and (2) untangle which changing neighborhood characteristics were the main drivers of a decline in amenity access. The empirical example I present in this paper using data for 1970 is part of a larger project, in which I attempt to close these gaps by tracing neighborhood change in Chicago from 1970 to 2000, in order to understand the changing relationships between race, class, population composition and amenity access (Chan Tack, forthcoming).

3. Spatially-Sensitive Data, Ecological Validity and Spatial Variable Construction

3.1 Constructing Spatial Variables: Discrete vs. Continuous Measures

In the past, studies investigating inequalities across social groups in their access to social goods often failed to take the spatial structure of these distributions into account (Downey 2003, 2005, 2006). Access has typically been defined either in terms of a *simple count* of a given social good (such as parks, playgrounds, hospitals or other facilities), or as a *dummy variable* indicating the presence of absence of any such facility located within each case.[5] Dummy and simple-count measures of access are problematic because they ignore important spatial patterns in the data.

[5] While research on amenity access is new to sociology, these studies have been conducted in the fields of environmental studies and urban planning. With the explosion of GIS technology over the last decade, quantitative geographers who were cognizant of the MAUP began to contribute to the urban planning literature (see Talen and Anselin 1998 for a review). The use of spatial methods has been slower to diffuse to sociology, where there has been remarkably little interest in using these techniques to more accurately test explicitly spatial sociological theory. Downey (2003)'s examination of neighborhood inequalities in exposure to polluting factories was the first to introduce these methodological issues in spatial variable construction to sociologists. Few have followed Downey's (2003) lead.

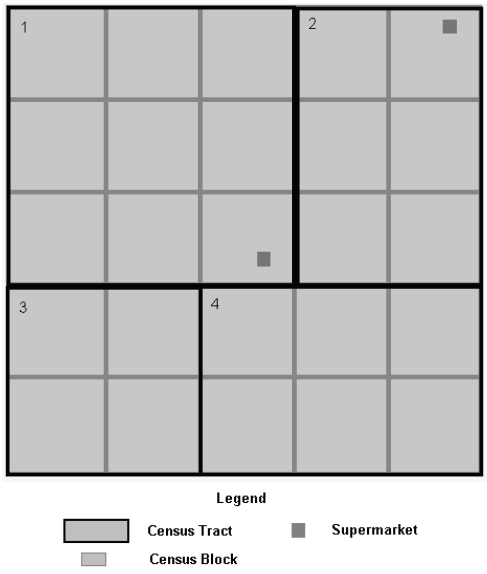
As Downey (2003, 2006) points out, these measures adopt a discrete notion of access and parcel space into distinct, unrelated containers. They ignore the effects of spatial spillovers or spatial externalities between neighboring units. Dummy and count measures treat facilities located near areal unit boundaries as though they

1) have no effect on people residing in adjacent units and
2) as though they affect every square inch of their home units equally.

This "container" treatment of space also makes implicit assumptions of the logic of facility supply. Namely, that social goods are only allocated to the residents of the spatial unit that contains them, and that the areal unit itself lacks an internal spatial structure.

Figure 1 illustrates these problems. The bold lines indicate census tract boundaries and the thinner lines delineate census blocks. The square-dots are supermarkets. Using a count measure, tracts 1 and 2 would have access to supermarkets, while tracts 3 and 4 would have no access. However, it is clear that certain blocks in tract 1 are further from a supermarket than are certain blocks in tracts 3 and 4. Clearly, count measures do not accurately capture spatial access. Distance-based measures are obviously preferable.

Figure 1: Count vs. Distance Measures of Access



3.2 Choosing Spatial Units

In analyses of "access" the cases used are invariably some kind of geographic unit, whether zip-codes, census tracts, wards, counties or metropolitan statistical areas (MSAs). From a spatial analytic perspective, many of these units are problematic because they separate space arbitrarily, without regard to the social uses and meanings of place. As discussed, the MAUP tells us that choosing ecologically appropriate spatial units is not simply an issue of ecological precision, as it can affect the sign, size and significance of regression estimates in unpredictable ways. The MAUP tells us that if we fail to choose ecologically valid spatial units, our substantive conclusions are highly questionable.

Small and McDermott (2006)'s examination of amenity access is illustrative of the tendency of sociologists to dismiss the question of ecological validity in their research designs. They use simple count data for their measure of access, zip-codes as proxies for neighborhoods, and fail to take spatial auto-correlation between neighboring areal units into account in their regression analyses. Among the numerous administrative spatial units on which the state collects administrative data, the zip code is often a popular choice among researchers as a proxy for real social neighborhoods. A consideration of the US postal zip-codes limited ecological validity in terms of its scale and its boundaries serves to illustrate why state administered spatial units often fit awkwardly with the needs of social science researchers.

At the national level in the US context, zip-codes vary widely in size, ranging from 45 to more than 105 square miles (Small and McDermott 2006). The variability in scale of zip-codes calls into question their comparability as real types of social neighborhoods. Moreover, at more than 100 square miles, their scale on the higher end is an implausible scale for the social processes operating in real neighborhoods. Since, zip-codes are merely areal units defined by the post-office for the efficient delivery of mail, their boundaries are entirely arbitrary with respect to neighborhood-level social processes. Thus arbitrary administrative spatial units like the zip-code violate ecological validity requirements in numerous ways.

3.3 The Pitfalls of Defaulting to Ready-Made Data

Small and McDermott (2006)'s paper illustrates the pitfalls of defaulting to ready-made data sources in quantitative analyses of spatial processes. The authors sourced their data from "Biz-zip", the United States Census' Zip Code Business Patterns dataset. "Biz-zip" has been collected by the US Census since 1964.[6] It is a commonly used resource for researchers examining amenity ac-

[6] See the U.S. Census documentation pages for the Zip Code and County Business Patterns Datasets <www.census.gov/econ/cdp>.

cess. Biz-zip is preferred by researchers, because it saves them the work of collecting and assembling a dataset themselves. Unfortunately, Biz-zip – and many officially gathered datasets like it – lacks spatial flexibility. Biz-zip provides only count data (# of amenities per spatial unit), aggregated to the zip-code level. It provides no information about how those amenities are distributed within zip-codes, and it offers researchers no flexibility to operationalize either access, or the boundaries of spatial units in alternative ways.

The foregoing discussion about the limitations of zip-code aggregated count-data for satisfying the MAUP tells us that scholars interested in *neighborhood effects* on amenity access who choose to utilize such ecologically inappropriate, spatially inflexible datasets run the risk of producing highly misleading results. One key argument advanced in this paper is that it is preferable, from both a methodological and theoretical perspective, to assemble datasets in which information about amenity distribution is stored as point-locations. I argue that point-location datasets are vastly superior to their highly aggregated count-based counterparts because they allow researchers the maximum amount of flexibility for both spatial variable construction and for the choice of spatial units. Point-location data thus give researchers the strongest opportunities to generate ecologically valid spatial constructs, and to generate sound, and reliable spatial regression estimates.

Small and McDermott (2006)'s analysis focused on amenity access in 2000. By this time, the authors could have readily purchased a dataset of point-locations from alternative sources for address-specific (i.e. point-location) spatial information, such as the Bradstreet and Dunn (B&D) or the RefUSA electronic business databases.[7] Scholars of historical urban process face stiffer challenges. For historians, databases like RefUSA and B&D have limited value. RefUSA lacks historical data before the 1990s. B&D has historical data through 1970, but the quality and completeness of its historical data is questionable, because, at the time, data-collection focused on only the largest businesses in the zip-code, rather than a full census of amenities (Bradstreet and Dunn 1980).

Historians interested in examining amenity access and seeking point-location data must thus resort to archival materials. Such data exists, but it is stored in non-electronic archival formats, such as hard copy, or micro-film. The task of assembling a spatially sensitive dataset for historical work is thus a more pain-staking, time-intensive affair. Faced with such a task, it is understandable that many urban historians succumb to the temptation to use pre-assembled readily-available, Biz-Zip data, despite the fact that the data are spatially-insensitive and of questionable ecological validity. This paper seeks to convince researchers that, rather than relying on ready-made state administra-

[7] For RefUSA, see <http://www.referenceusa.com> and for the Bradstreet and Dunn Directories, see <http://www.dnb.com>.

tive datasets that offer spatially insensitive aggregates of our outcomes of interest the only option, it is essential to build spatially flexible point-location databases for the phenomena that interest us.

In the next section, I present a unique, self-assembled dataset of point-locations for supermarkets in 1970 Chicago. I discuss how the dataset was created and illustrate its flexibility for spatial variable construction and for the selection of spatial analytic units. I then compare the results of regression analyses for supermarket access in Chicago that use:

1) data assembled into the spatially inflexible count-based, zip-code aggregated format, and
2) data constructed from spatially flexible point-locations.

4. The Value of Spatially-Flexible Data

4.1 Building a Spatially-Sensitive Dataset: A Historical Census of Grocery Stores in 1970 Chicago

To overcome the problems inherent in widely available official data sources, I constructed a unique census of all supermarkets for the city of Chicago for 1970. The census was composed by combining supermarket lists from the Yellow Pages with an exhaustive combination of lists from industry directories for the years in question, including the Bradstreet & Dunn Regional Market Area Business Directory, the Directory of Supermarket, Grocery & Convenience Store Chains, The Directory of Single Unit Supermarket Operators, the Directory of Retail-Owned Cooperatives, Wholesale-sponsored Voluntaries, Wholesale Grocers and Service Merchandisers, the Directory of Wholesale Grocers, and the Thomas Grocery register. This choice of sources had the advantage of providing address-specific locations (i.e. point-locations) for all stores, while maximizing the completeness of the census (Marsden et al. 1990; Bader et al 2010). As I show below, the high granularity of the point-location allows for invaluable flexibility in developing measures of access and units of analysis that are ecologically appropriate to the research question.

4.2 Spatial variable construction: Operationalizing "Access" and Selecting a Unit of Analysis

I chose the census tract as the spatial unit of analysis for the regression procedures. While census tracts do not perfectly accord with real neighborhood boundaries, sociologists find them the most ecologically reasonable proxy for neighborhoods because they approach real neighborhoods in area, population size and composition (see Small and Stark 2005, 1018-9). My approach to variable construction for the outcome variable "access", takes advantage of my

unique address-specific store data and the capabilities of GIS technology to create a *block-level, tract-averaged minimum distance* measure of access. This measure improves on three existing types of measures:

1) commonly-used count and dummy variable measures,
2) the aggregate minimum distance, and
3) Downey (2006)'s "rasterized" tract-averaged minimum distance access measure.

The minimum distances computed here are network distances. They are operationalized as the shortest path traveled along the city street grid from the areal unit centroid to the closest supermarket. This "network distance" measure improves upon the more widely utilized Euclidean, straight-line distance because it takes the structure of the transportation network into account, thereby improving the measure's ecological validity.

Measuring the distance between points (in this case, the precisely geocoded supermarkets) and areal units (in this case, census tracts, which are proxies for neighborhoods) is not straightforward. Once the path length (network vs. Euclidean) is chosen, the next challenge is to define a point within the areal unit from which to measure distance. Typically, geographers solve this problem by using the geometric center or "centroid" of the areal unit. Downey (2003) calls this measure the *aggregate minimum distance*. This operationalization of access assumes, implausibly, that residents are clustered at the geometric centers of census tracts. Given the research question at hand, the true desired measure of access is the average distance that *residents* in a census tract have to travel to access a supermarket.

The ideal way to create a population-weighted measure of access would be to:

1) measure the distance between every household and the closest grocery store along the street network and
2) take the average of this value across all households in the tract.

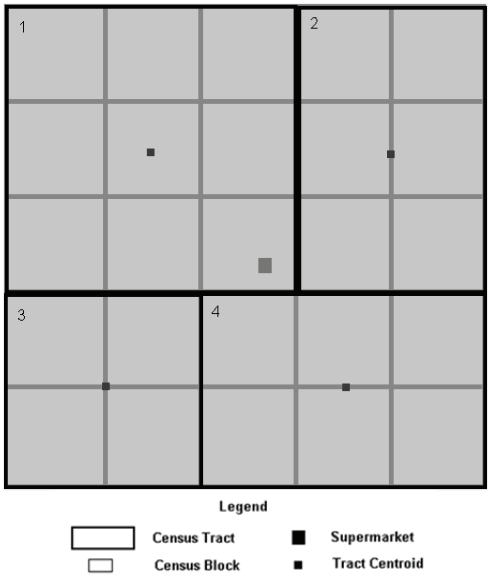
The result would be a household-level *average minimum distance*. The US Census does not release household-level location and demographic information due to privacy concerns, so researchers must resort to creative methods to approximate the household-level average minimum distance measure.

Downey (2003, 2006) offered an ingenious "rasterized" *average minimum distance* measure. If we followed Downey's (2003) "rasterizing" approach, we would lay a grid of 25m by 25m squares over the census tracts in his analysis. We would compute the average of these values for all grid squares within a census tract.

Downey (2003) demonstrates at length the improved precision of his rasterized *average minimum distance* measure over the *aggregate minimum distance*. To understand the difference between an Average and Aggregate Minimum Distance measures, consider Figures 2 and 3. Figure 2 reproduces Figure 1, with the exception of the small dots, which are the geometric center of the

census tract. Under the aggregate measure, access is measured from the center of the census tract to the closest supermarket. Using Downey's (2003) grid-based average distance measure, the distance from the geometric center of every 25m by 25m square to the nearest supermarket is measured, summed and averaged to create an access measure for the entire tract.

Figure 2: Aggregate vs. Average Minimum Distance Demonstration



Neither of these measures fully account for real population distribution within a neighborhood. In reality, households are spread unevenly across a neighborhood's total area, based on their lot size, and the configuration of neighborhood streets. Downey's (2003) grid-based average distance measure is nevertheless more ecologically realistic than the tract aggregate measure. It is more realistic to assume that households are evenly spaced within a neighborhood than it is to assume that all households in a neighborhood are piled on top of each other at the neighborhood's geometric center.

While Downey's (2003) grid-based average distance measure improves on the widely used tract aggregate distance measure, his rasterizing procedure has numerous disadvantages: 1) the dimensions of the grid square must be determined, and this choice is ecologically arbitrary; 2) the grid squares have no relationship to any meaningful social or physical geography, such as population distribution or city streets. It is thus vulnerable to the MAUP.

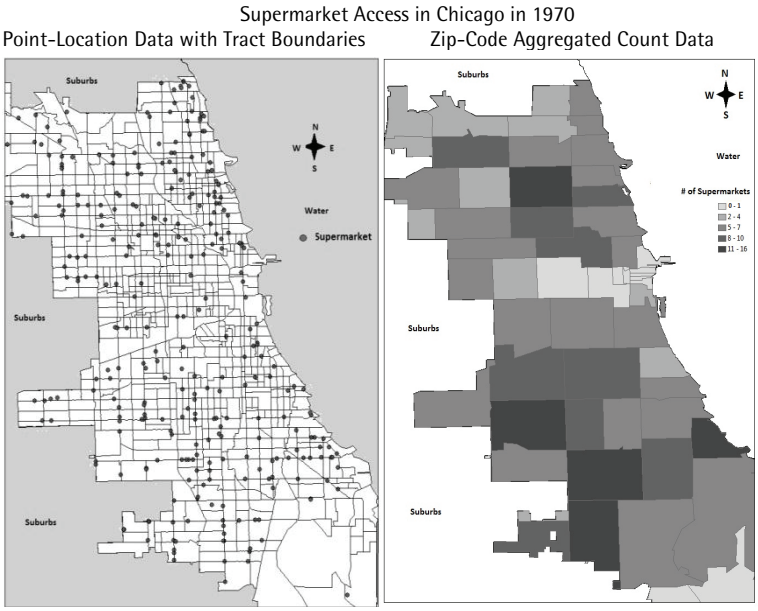
The best way to moderate the MAUP is to choose spatial units that have a meaningful relationship to the phenomenon being studied. Given these consid-

erations, I selected census blocks for the 1970 census as the most appropriate spatial unit available to create an *average minimum distance* measure. Census blocks are the smallest spatial units for which data is collected. They are bounded on all sides by real physical features, such as streets, roads, parks and rivers. Such physical features are also reasonable proxies for the spatial distribution of houses and households (Tatian and Cornelius 2003). They thus have greater ecological validity than Downey (2003)'s arbitrarily-scaled square grids.[8]

Census blocks improve on Downey's (2003) grid because they have meaningful relationships to the underlying social and physical geography that is of interest to us, in particular the transportation network.

4.3 A Brief Re-Capitulation

Figure 3: Comparing the Spatial Granularity of the Point-Location and Biz-Zip Datasets



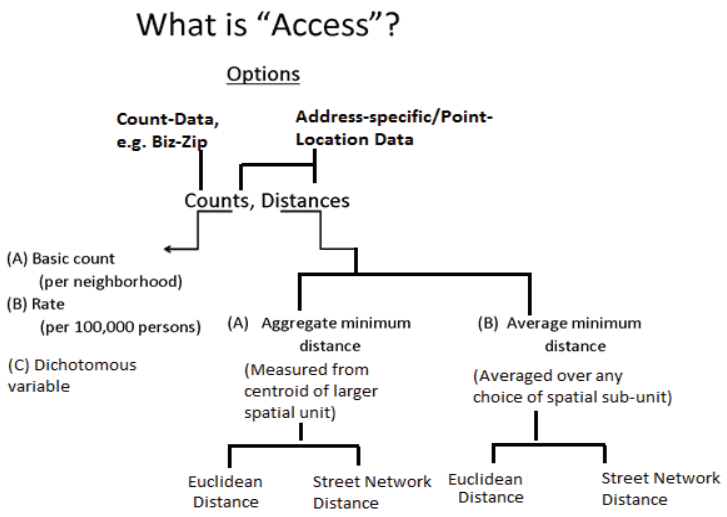
In the next section, I will compare the regression analyses that are the logical consequence of the two contrasting datasets. First, let us recap how the differ-

[8] Luc Anselin supports my assertion that census blocks improve on Downey's (2003) raster grid (personal communication). Downey has reflected that his square grids have questionable ecological validity, and that census blocks improved on the raster grid (personal communication).

ences between the datasets translated into differences in variable construction and choice of spatial unit. For the "spatially inflexible" dataset, we have count data of supermarkets aggregated to the zip-code level. Because this Biz-Zip Dataset provides no further information, we are constrained to the zip-code as the unit of analysis and the count variables as our measure of access. As discussed, zip codes are poor approximations of neighborhoods, and count variables distort the spatial reality of access. In contrast, the "spatially flexible" dataset of supermarket point locations offer us numerous advantages.

We are free to choose an appropriate spatial unit, and we have multiple options for operationalizing access. With point-location data, we could operationalize access as either (1) a count or a distance, (2) Euclidean or Network distance, (3) tract-aggregate or tract-averaged. In addition, we had a range of choices for computing the tract-averaged measure. We could have (4) used Downey's "rasterize" approach or my census-block approach. As Figure 4 shows, where the spatially insensitive dataset allowed only three, relatively similar options for operationalizing the outcome variable, the spatially flexible dataset allowed me to considered $2^4 = 16$ possible operationalizations. The flexibility of the dataset allowed me to select a variable construction technique that maximized the ecological validity of my access construct.

Figure 4: Comparing Options for Operationalizing Access in Point-Location vs. Areally-Aggregated Count Data



In fact, the options for variable construction could be much broader, depending on the research question. Regardless, the notable observation is that point-

location data allows researchers incredible flexibility to tailor their variables and spatial units to the social process under investigation. Figure 3 illustrates the differences in spatial granularity for the two data sources. Figure 4 summarizes and contrasts the range of options for spatial variable construction afforded by spatially inflexible count data, compared with spatially flexible point-location data. As we will see below, point-location data also allows researchers to more readily take spatial processes into adequate account in their choice of regression models.

5. Analytical Approach

5.1 Predictors

Table 1: Descriptive Statistics at the Tract and Zip-Code Level

	Tract-Level		Zip-Code	
	Mean	SD	Mean	SD
Predictors				
% Persons in Poverty	15.3	(12.0)	13.3	(10.6)
% Non-Hispanic White	57.6	(40.4)	68.4	(31.3)
% Non-Hispanic Black	30.8	(42.3)	23.0	(30.8)
% Hispanic	8.5	(14.2)	6.0	(7.1)
Population	3980	(2610)	63,255	(35,524)
Area (sq. mile)	0.25	(0.37)	1.12×10^7	(6.61×10^5)
Population density (persons per sq. mile)	23,100	(16,300)	19,500	(11,040)
% Foreign Born	11.1	(9.2)	11.0	(6.5)
% Residential Stability (Persons in the same house 5 years ago)	52.7	(14.7)	52.4	(12.6)
Distance from the City Center (miles)	6.3	(3.1)	7.2	(3.8)
Presence of Public Housing (0,1)	0.24	(0.83)	0.53	(0.5)
Outcome Variables	Avg. Minimum Distance (miles)		# Supermarkets per Zip-Code	
Supermarket Access	0.539	(0.305)	5.7	(3.4)

The predictors of interest for the models are standard for neighborhood effects research: *Neighborhood Poverty Rate* and *Neighborhood Ethnic Composition,* represented by *% Poor*, *% of Non-Hispanic Black* and *% of Hispanics.* The census only created fully mutually exclusive categories for Non-Hispanic Blacks, Non-Hispanic Whites and Hispanics in 1980, and fully mutually exclusive groups for Asians and non-Asians in 1980. As a result, mutually-exclusive categories for these groups had to be constructed for the 1970 and 1980 census years. I did so by following Timberlake and Iceland (2007)'s procedures. The controls used are also standard to quantitative research on neighborhood ef-

fects. These were: *% of Foreign Born*, *Residential Stability*, or the % Residential of persons over age five, who had the same residence five years ago, *Population Density (Logged),* and *Public Housing Project Presence,* a dichotomous variable, which was controlled for in order to distinguish its effect from that of neighborhood poverty. To create this predictor, I obtained address-specific data for multi-family unit public housing for the four decades from the Chicago Housing Authority (CHA) through the Freedom of Information Act. Table 1 gives the summary statistics for the variables used in the analysis.

5.2 Comparing Estimation Strategies

The "spatially insensitive" and "spatially sensitive" datasets have given us two, quite different, outcome variables. These imply distinct estimation strategies. The Biz-Zip data provides us with count-data for supermarkets aggregated to the zip code level. The outcome variable from the Biz-Zip data is the number of supermarkets per zip-code. This calls for a model appropriate to count-data. Possibilities include the Poisson, the Negative Binomial Regression, and others.[9] While spatial Poisson models exist, it is difficult to run them on the aggregated count data. This is for several reasons. Key among them: (1) Since the data is aggregated, we have no information on how these supermarkets are located relative to each other, either within or between zip-codes, making it impossible to use traditional spatial Poisson techniques (2) Spatial models for aggregated count processes are not well developed. In contrast, the "spatially sensitive" dataset has given us a continuous, tract-averaged measure of access. Spatial models for continuous data are of long standing (see Anselin 2010 for an overview). Thus, a range of spatial models – such as spatial lags and spatial autocorrelation – for continuous outcomes, spatial lags and autocorrelation can be run relatively easily and are accessible to a wide range of researchers.[10]

[9] In all regression analyses, the technique selected must be appropriate to the data type. In the case of this paper, we are dealing with two kinds of outcome variable: a continuous measure of distance and a count measure of the number of supermarkets in an areal unit. Continuous measures are usually treated as having a normal distribution, for which Ordinary Least Squares (OLS) regression is appropriate (Allison 1998). For count data, regression techniques such as the Poisson, Negative Binomial and Gamma Model can be used. The choice between count models depends on which model best describes the observed distribution of the data. Analysts use a range of statistical tests to help them choose the count model that is most appropriate to the data (Long 1997).

[10] Footnote 12 provides a glossary of these terms.

5.3 The Modeling Constraints on Aspatial, Aggregated Count Data

5.3.1 Objects vs. Fields

As mentioned, explicitly spatial models for analyzing discrete or count data exist. However, it is difficult to model the count-aggregated aspatial data presented in these studies using these models. This is so for three reasons. This is because of (1) the nature of the substantive process under study and (2) the limitations of aggregated aspatial count data for modeling phenomena as a spatially *continuous* process, and (3) discrete spatial models for aspatial aggregated count data, are still under development by statisticians, and are not yet available for widespread use. To understand why, we need to know a little bit about the intellectual history of spatial statistics.

Similar to the history of GIS, spatial statistics emerged independently, scattered across a diverse set of research domains. Estimation methods were devised to solve immediate problems within an application context. This developmental trajectory has resulted in two major divisions in spatial statistical practice. Broadly construed, the division is between regional scientists, and spatial econometricians on the one hand, and geo-statisticians, pure statisticians and bio-statisticians.

The distinction between these two schools of spatial modeling is often described in terms of "object" vs. "field" views, or "lattice vs. continuous/point-patterns" (Diggle 2010). As Anselin (2002, 254-6) notes, since the two perspectives represent distinct approaches to statistical inference and sampling, the choice between these frameworks for statistical inference has far-reaching implications. The object view and associated lattice data perspective are best suited to the study of discontinuous phenomena, while the field view treats observations as sample points from a continuous surface. In the lattice approach, the observed spatial distribution of *all points* is considered one sample point in *a universe of all possible spatial distributions of all points*. In the field approach, the points selected for analysis are considered a sample of *a larger population of points*.

The object view is especially appropriate for studying discrete social and economic processes, e.g. neighborhood composition, land-use values, business location decisions etc. Spatial econometricians and regional scientists thus work through the lattice perspective. The field view (used by demographers, as I discussed above), is particularly useful for the examining spatially continuous processes. The most widely known and thoroughly developed family of spatial models for count data was developed within the field perspective (Lambert et al. 2010).

5.3.2 Limitations on a Field Perspective for Aggregated Count Data

While the field view is not ideal for economic and social processes, there are certain models which could be feasibly applied to non-aggregated socio-economic point data. In the field view, inferences are made by examining by the pairwise association between the sample points, expressed as a function of the distance that separates them (Cressie 1993).[11] To apply these estimation strategies, we need information about *the spatial relationship between points*. Even if we could assume that spatially discontinuous neighborhoods in an urban area constitute a smooth or continuous surface, we could not apply these models to Biz-Zip data, because the data are aggregated counts, with no information about the spatial relationships between the points (supermarkets) located within the zip-codes.

5.3.3 Limitations on an Object Perspective for Aggregate Count Data

As I have detailed, an object-approach to the analysis of count data generated from spatially discontinuous socio-economic processes is theoretically preferable over a field approach. The literature on field-oriented spatial Poisson models in the field perspective is far more mature than that within the object perspective (see Smirnov 2010, for a recent review).

Spatially distributed data present researchers with distinct analytical issues, including those of spatial heterogeneity, spatial autocorrelation (i.e. spatial dependence), spatial lags and spatial heteroskedasticity.[12] The choice of spatial model must be guided by substantive knowledge and theory of the process in question. With respect to supermarket location, firm location theory strongly emphasizes the importance of modeling spatial lags (Weber 1929; Guimarães et al. 2003, 2004). Several spatial econometric approaches are available to

[11] Cressie (1993), Chapter 8 covers many examples of models for spatial point data.

[12] This footnote gives a brief definition of quantitative spatial terms used in this paper (see also Netrdová and Nosek 2014, in this HSR Special Issue). Spatial data typically exhibits *spatial autocorrelation* – i.e. the values of y and x in a spatial unit *i* are affected by the values of y and x among *i*'s neighbors. In the case of this study, the presence of a supermarket in one neighborhood affects the number of supermarkets in surrounding neighborhoods due to market competition. The *Moran's I* is a common statistical test for spatial autocorrelation. If spatial autocorrelation is significant, then the data violate the independence assumption of standard regression, and spatial regression techniques, namely the *spatial lag* or *spatial error* models, are necessary. The choice between a spatial lag or error model is guided by the goals of the analysis. *Heteroskedasticity*, in aspatial statistics, refers to a situation where the variance of an outcome variable y is not constant across the range of a second variable x, that predicts it. *Spatial heteroskedasticity* is the spatial corollary to this idea. Namely, the variance of y is not constant over its spatial range. A process exhibits *spatial heterogeneity* when its mean (or "intensity") varies over space. Readers should consult Fortheringham and Rogerson (2009) for further information.

model spatially correlated disturbances in count models, but structurally consistent count models incorporating spatial lag autocorrelation are still undergoing active development (Lambert et al. 2010). I therefore present the results for the aspatial Poisson model. This is the model most commonly applied by researchers examining count data in the amenity access literature.

6. Results

6.1 Results from the Spatially-Sensitive Analysis

The first two models in Table 2 present the aspatial Ordinary Least Squares (OLS) and the spatially lagged regression of the continuous access measure on the predictors. A lag model was selected because of the theoretical importance of lag effects in location decision theory (Weber 1929). In moving from an aspatial OLS to a spatial model, one must specify the weights matrix. Various weighting schemes are available, including the rook and queen contiguity, distance decay etc.[13] The weights matrix is usually chosen through a combination of substantive and theoretical logic, as well as through testing of model fit. Substantively, we assume that a focal neighborhood is affected by all those that border it, suggesting a queen contiguity criterion. First-order queen contiguity was tried. The first-order queen assumes that only the immediate neighbors are relevant. This failed to eliminate the spatial autocorrelation (indicated by Moran's I) in the data.[14] A second-order queen, which takes the neighbors of neighbors into account, was then tried (Wang 2010).

As Table 2 shows, the use of second order queen weights to create spatially lagged controls eliminates problems with spatial autocorrelation present in the simple OLS model. The global Moran's I is used to measure the extent of spatial autocorrelation. Moran's I takes values ranging from 0 to 1, with 1 representing the highest degree of spatial autocorrelation. Inspection of Table 2

[13] Weighting schemes are used to model the structure of spatial autocorrelation in the data using a *weights matrix*. Application of an appropriate weighting scheme eliminates spatial autocorrelation and enables the accurate estimation of regression coefficients. A variety of weights matrices exist, including first and second order rook, and queen contiguities, as well as distance decay and gravity models. If we imagine a spatial unit *i*, surrounded by neighboring units *j*, each weighting scheme postulates a particular spatial conformation through which the values of x and y in spatial units *j* affect the value of x and y in spatial unit *i*. In other words, the weighting scheme tells us which neighbors matter and how they matter. Weighting schemes are chosen such that they are consistent with existing theories of spatial influence, and such that they eliminate observed spatial autocorrelation in the data. For more information, see Wang (2010).

[14] See Netrdová and Nosek 2014 (in this HSR Special Issue) or footnote 12 for an explanation of Moran's I.

reveals that in general, the degree of spatial autocorrelation in the simple OLS models is substantial and highly significant, at 0.31 ($p>0.0001$). Once spatial lags are introduced, the Moran's I falls to well below 0.01, and is no longer statistically significant.

Turning to the effects of predictors on supermarket access, we find that as a neighborhood's distance from the central city increases, access to supermarkets declines. A one-mile increase in distance from the center city translating into a decline in access ranging from 0.023 miles in 1970 ($p<0.001$). Notably, the magnitude and the significance of the central city predictor increase as we move from the simple OLS model to the model with spatial lag. Population is associated with an increase in supermarket access. Importantly, the spatial lag variable is also statistically significant. In the lag predictor, a unit increase in the log of population density in both the main and lag variables is associated with an increase in access ranging from a minimum of 0.125 miles ($p<0.001$). The spatial model also shows that the log population density of surrounding census tracts exerts effects that are substantial – they are about half as strong as that exerted by the individual census tract.

Supermarkets tend to locate further away from residentially stable neighborhoods, but the stability of surrounding neighborhoods has no significant effect. A 10% increase in residential stability is associated with a 0.028 mile greater travel distance. The residential stability spatial lag was not significant in any models studied and did not improve fit, so was dropped in the final model presented here. While supermarkets are desirable entities, they also bring traffic and noise, and can disrupt the aesthetics of neighborhood areas. Since supermarkets and the large retail developments that contain them are both desirable amenities, and possible neighborhood nuisances, it is likely that neighborhoods would resist having supermarkets within their boundaries, but would welcome supermarkets in abutting neighborhoods. Logan and Molotch (1987) have argued that highly organized, residentially stable neighborhoods are more likely to successfully resist the entry of disruptive retail developments within their neighborhood boundaries. The positive relationship between residential stability and distance to the nearest supermarket supports this idea.

The influence of % Hispanic residents within the neighborhood or in surrounding tracts on grocery store access was not significant. This is most likely due to the fact that the Hispanic population in Chicago in 1970 was small, with Hispanics comprising only 8% of the average tract population, and with many Hispanics spatially concentrated in only a few tracts. % Foreign born had no significant effect on supermarket access in 1970. Immigrant enclave theory (Portes and Rumbaut 1996) would predict that increases in the % of foreign born residents raise the tract's saturation with local ethnic businesses, making them potentially less welcoming to chain stores. The absence of a large Hispanic minority immigrant community in Chicago in 1970 probably explains the % foreign born did not exert statistically significant, independent effects on store

access at this time. The final control variable, the presence of public housing density was also statistically insignificant and did not improve model fit.

Finally, we turn to the effect of our variables of interest on supermarket access. These results are quite interesting. The second-order queen contiguity models in Table 2 suggest that while the poverty rate of a given tract did not play a significant role in influencing access in 1970, the average poverty rate of *surrounding tracts* did. As the average poverty rate in the surrounding tracts increases by 10%, distance to the nearest grocery store also increases by 0.085 miles in 1970 ($p<0.001$). These results are consistent with the idea that supermarkets seek locations that maximize their access to consumers with expendable incomes across reasonably wide areas.

Table 2: Model Estimations from "Spatially Sensitive" and "Spatially Insensitive" Datasets

Outcome Variable	Continuous: Tract-Averaged Minimum Distance		Count: Supermarkets per Zip-code	
Estimation Strategy	OLS	Spatial Lag	Poisson β_i (r.s.e.)	IRR (r.s.e.)
Predictors				
Intercept	1.32***	1.46***	-21.74	
% Poor	0.0034**	-0.000027	-0.03839	0.962
% Black	0.00018	0.0010	-0.00641	0.994
Spatial Weights for Predictors				
% Poor-lagged				
% Black-lagged		-0.0016*		
Controls				
% Hispanic			-0.01779	0.982
% Foreign born			-0.03955*	0.962*
% Residential Stability	0.0029***	0.0028***	-0.01389	0.986
Log (Population Density)	-0.11***	-0.089***	0.9927**	2.699**
Distance from Central City	0.014**	0.023**	-0.04235	
Public Housing			-0.2937*	0.745*
Spatial Weights for Controls				
Residential Stability-lagged				
Log(Population Density)-lagged			-0.046***	
Moran's I	0.31***	0.0015		

*p<0.05, ** p<0.01, ***p<0.001

The relationship between the % of African American residents in a tract, and the average % of African Americans in the surrounding tracts on supermarket access is even more interesting. In 1970, the main effect of an increase in Black residents by 10% was to *increase* the distance to the nearest store by 0.010 miles ($p<0.05$), while the lag effect of a 10% increase in Black residents was to

decrease the distance by 0.16 miles ($p<0.05$). Research on residential mobility has shown that living in a neighborhood with a high % of black residents makes white households more likely to move. However, if those neighborhoods are surrounded by areas that also have high percentages of black residents, then the household is less likely to move (Crowder and South 2008). This is because most residential moves occur over short distances. For white households, living in a neighborhood surrounded by other high % Black areas eliminates the value of a residential move.

We see a similar pattern here with supermarket location decisions. In 1970, we see a positive association between % Black and distance to the nearest supermarket. Thus, the higher the proportion of African American residents in a neighborhood, the further away a supermarket is likely to locate. This effect is weakly significant. Conversely, we see that increased % Black in surrounding neighborhoods makes supermarkets locate *closer* to the focal neighborhood. This is a statistically significant effect. Both patterns suggest that firms are engaging in retail-redlining. All things equal, a supermarket would distance itself from a neighborhood with higher % Black residents. But if surrounding areas also had high % Black residents, the value of moving would be diminished. Overall, the results show evidence of retail-redlining directed specifically at black neighborhoods in 1970.

6.2 Results from the Spatially-Insensitive Analysis

I now present the results of the aspatial Poisson model. Results from this model are presented because it is the most commonly used model among urban sociologists conducting historical and contemporary amenity research. The logic of data-collection tells us that the data is neither zero-inflated nor zero-truncated. The Poisson model is best used when the mean is equal to the variance. With a mean of 5.7 and a variance of 11.9, the Biz-Zip supermarket data for Chicago displays slight over-dispersion. Despite evidence of over-dispersion, the Chi-Squared statistic and other checks suggested that alternative models, such as the as the Negative Binomial and the Gamma model, did not substantially affect the results.[15]

[15] The Chi-Squared Statistic is the numerical value that is produced after one performs a Chi-Squared Test. The term Chi-Squared Test refers to a broad family of tests that are widely used in statistical analyses. They are commonly used to tell if two things (models, observed distributions of data, etc.) are statistically different from each other. In the case of this paper, I use the Chi-Squared Test to determine whether the Negative Binomial, or Gamma models explain the data better than the Poisson model I present. The Chi-Squared Statistic that was generated from this test tells me that these models do not explain the data better than the Poisson model. It thus reassures me that I have good reason to proceed with the Poisson model. For more on the Chi-Squared Test, and its use in statistical analyses, qualitative readers should refer to introductory texts in regression analyses, such as Allison (1998).

The aspatial Poisson model regresses the aggregate count-measure on the same predictors as those used for aspatial OLS and the spatially lagged regressions of the continuous tract-averaged measure of access. Since the size of zip-codes varies widely, I use the square area of the zip-code as the exposure window variable. Over-dispersion does not affect the accuracy of the point estimates, but it does exaggerate the precision of regression results. This can be tempered if we use robust-standard errors, which reduces the likelihood that we will mis-attribute statistical significance to any of the predictors. Table 2 presents both the β_i and the Incidence Rate Ratios (IRRs) for the Poisson regression results.[16]

It is worth noting that while the outcome variable is a count, neither the Poisson, Negative Binomial nor the Gamma distributions fit the data well. The Poisson model had a Chi-Squared Statistic of 63.2 ($p = 0.030$). As is the case for all standard statistical techniques, these distributions assume independence between observations. The data is, however, *highly* spatially auto-correlated. The situation cannot be corrected without using an appropriate spatial Poisson model. Unfortunately, this is precisely the kind of analysis that Biz-Zip's spatially insensitive, aggregate count model forecloses.

A glance at Table 2 reveals stark differences between the results of the spatially sensitive lag regression on the continuous tract-averaged outcome, and those of the spatially insensitive Poisson on the zip-level aggregated count outcome. Except for the effects of the Log of Population Density, the final models disagree entirely on which predictors are statistically significant. Whereas the Spatial Lag model found % Black, % Poor, Distance from the Central City, and Residential Stability to have statistically significant effects, the Poisson model found % Foreign Born and Public Housing to be statistically significant.

Specifically, the Poisson model estimated that a unit increase in Log Population Density approximately tripled the incident rate of supermarkets within a zip-code (IRR = 2.699, $p<0.01$). The presence of Public Housing reduced the incident rate by approximately 25% (IRR = 0.745, $p<0.05$); and a 1% increase in the Foreign Born population within a zip-code was associated with a 3.8% decrease in the incident rate (IRR = 0.962, $p<0.05$). Thus, a 10% increase in the foreign born population would cause a 33% decrease in the incidence rate.[17]

---

[16] β_i is the coefficient of the Poisson regression. It represents the increase in the log odds of the outcome variable (in this case, supermarket access) associated with a one unit increase of the predictor. Incident Rate Ratio (or IRR) is simply another way of presenting Poisson regression results. The IRR is the anti-log of βi. In the case of this analysis, the IRR is thus a multiplier of the expected number of supermarkets associated with a one-unit increase in the predictor. While IRRs and the βi communicate exactly the same information, IRRs are often preferred when presenting results as they are more intuitively interpretable. For more information, see Long (1997).

[17] The IRR is the ratio of the incident rates, for a unit-change in the predictor X. For a continuous predictor, an *S*-unit change alters the incident rate by exp ($S\beta_i$). Thus a 10% increase

The significance of all predictors in the Poisson model was, overall, much weaker than in the Spatial Lag model. % foreign born ($p = 0.047$) and public housing ($p = 0.037$) were weakly significant at the 0.05 level.

The findings suggest that population density is a powerful driver of supermarket access, while poverty and race have no effect. While we would expect that supermarkets would want to locate in high-density areas, it is surprising that a wealth measure such as poverty would have no effect on supermarket density. The finding that the race variable % Black was not significant would also be quite surprising for urban sociologists, and would warrant careful examination. The idea that an increasing presence of foreign-born populations is associated with declining supermarket density would accord with ethnic enclave theory (Portes and Rumbaut 1996). Overall, however, when we compare the substantive results generated from the aspatial and spatial regression models, the findings of the spatial model would be far more intelligible and convincing to urban sociologists when considered in the context of both other empirical research, and substantive theories about urban social process.

7. Concluding Remarks

This paper argues for the importance of gathering point-location data in both historical and contemporary research on spatially-distributed social phenomena. It focuses on historical research because ready-made, spatially flexible historical datasets are hard to come by. This is particularly an issue for researchers studying spatially discontinuous social phenomena, such as neighborhoods, firm locations, political configurations and land values. While many researchers may be tempted by the convenience of readily-available historical data from sources like National Census Bureaus, these datasets are usually spatially-inflexible, and offer only count-data aggregated to inappropriate spatial units. In this paper, I argue that constructing spatially-sensitive datasets of point-locations not only makes methodological sense, but is essential for drawing accurate substantive conclusions. I demonstrate this argument using the case of supermarket access in 1970 Chicago.

In making my argument, I introduce a major issue facing all studies of spatially distributed data: the issue of ecological validity and the associated problem of the MAUP. The MAUP tells us that the sign, significance and magnitude of regression coefficients changes unpredictably as the boundaries of the spatial unit change. When spatial units and measures lack ecological validity, regression estimates will give misleading information about the mechanisms

in the foreign-born population is associated an (exp 10 x -0.0396), or 33% decrease in the incidence of supermarkets within the zip-code.

driving the process of interest. The only way to avoid the pitfalls of the MAUP is to ensure that spatial units and measures are ecologically appropriate to the research question (Openshaw 1984).

I dramatize these issues by discussing the logic of spatial variable construction and spatial unit choice for the analysis of amenity access. I introduce two datasets: a spatially-sensitive dataset of supermarket point-locations constructed from archival research, and a spatially insensitive dataset of aggregate counts from Biz-Zip, the US Census Zip-Code Business Patterns data. The latter is widely used in historical and contemporary amenity access studies, as researchers prefer the convenience of ready-made datasets. I demonstrate that point-location data allows researchers enormous flexibility in the choice of spatial unit, and in the construction of outcome variables.

In addition, I discuss the constraints that spatially inflexible datasets place on spatial modeling. In the amenity access case, point location data allows for the construction of ecologically appropriate continuous measures of access, which are easily amenable to spatial lag regression models – precisely the type of model consistent with classic amenity location theory Weber 1929). Areally-aggregated count data not only suppresses information about the real spatial distribution of amenities, it is also very difficult to model spatially, from both object and field statistical perspectives. I contrast regression results for the spatial lag model on the continuous dependent variable, and those from the aspatial Poisson. The parallel analyses give highly divergent results – and demonstrate the dramatic differences in substantive results that can arise from datasets and analyses that allow researchers to measure and model space in a manner that is appropriate to the research question.

Overall, my analysis supports the value of eschewing readily-available spatial data when that data is structured in ways that make it impossible to choose ecologically meaningful spatial units and to construct spatial measures appropriate to the phenomenon of interest. Beyond concerns about distortion and/or inconsistency in regression estimators, this paper demonstrates the restrictions that spatially inflexible datasets place on substantive-theory building in another way. The underlying structure of spatial data contains important information about the *mechanisms* underlying the processes under study. Spatial parameters cannot be thought of as merely nuisance parameters, whose "noise" must be purged from the "true" aspatial signal.

The Spatial Lag model presented above eloquently illustrates this point. The lag variables show how the social composition of surrounding neighborhoods influence supermarket location decisions in the focal neighborhood. In this case, the results provide evidence of racially-based retail red-lining in the supermarket industry, and lend support to Massey (1993)'s argument that racial discrimination remained relevant in the United States in 1970, in the wake the Civil Rights period.

This paper is also an argument for collaborative data-gathering initiatives to create spatially-flexible datasets for historical researchers. Building spatially flexible datasets of point locations is time-consuming. It requires archival research, extensive transcription, and geo-coding. It can become expensive if the researcher outsources these tasks to GIS professionals. A handful of historical GIS (HGIS) data-building projects exist, and some of them are collecting point-location data.[18] But the multiplication of similar efforts is necessary and will be a great boon to spatially-oriented quantitative historical researchers. As the results from this paper show, spatially flexible historical data will allow researchers in multiple disciplines to substantially advance their fields both empirically and theoretically.

References

Anselin, Luc. 2002. Under the Hood: Issues in the Specification and Interpretation of Spatial Regression Models. *Agricultural Economics* 27: 247-67.

Anselin, Luc, and Sergio Rey. 2010. *Perspectives on Spatial Data Analysis*. Berlin: Springer.

Bader, Michael D. M., Jennifer A. Ailshire, Jeffrey D. Morenoff, and James S. House. 2010. Measurement of the Local Food Environment: A Comparison of Existing Data Sources. *American Journal of Epidemiology* 171: 609-17.

Barbujani, G., N. L. Oden, and R. R. Sokal. 1989. Detecting regions of abrupt change in maps of biological variables. *Syst. Zool*. 38: 376-89.

Bradstreet & Dunn. 1980. *Regional Market Indicators*.

Briggs, Xavier De-Souza, Susan J. Popkin, and John Goering. 2010. *Moving to Opportunity: The Story of an American Experiment to Fight Ghetto Poverty*. Oxford University Press.

Bocquet-Appel, J.-P., and L. Jakobi. 1998. Evidence for a spatial diffusion of contraception at the onset of the fertility transition in Victorian Britain. *Population New Methodological Approaches in the Social Sciences* 101: 181-204.

Browning, C. R., D. Wallace, S. Feinberg, and K. A. Cagney. 2006. Neighborhood social processes and disaster-related mortality: The case of the 1995 Chicago heat wave. *American Sociological Review* 71: 665-82.

Chan-Tack, Anjanette M. Forthcoming. The Geography of Retail Inequality – A Spatial Analysis of Race and Class Effects on Supermarkets Access in Chicago, 1970-2000.

Cliff, Andrew D., and John K. Ord. 1981. *Spatial Processes: Models & Applications*. London: Pion Limited.

Cressie, Noel. 1993. *Statistics for Spatial Data*. New York: John Wiley.

Crowder, Kyle, and Scott South. 2008. The Spatial Dynamics of White Flight: The Effects of Local and Extralocal Racial Conditions on Neighborhood Out-Migration. *American Sociological Review* 73 (5): 792-812.

[18] <ww.aag.org/cs/projects_and_programs/historical_gis_clearinghouse/hgis_projects_programs>.

Diggle, P. 2010. Historical Introduction. In *Handbook of Spatial Statistics*, ed. E . Gelfand, Peter J. Diggle, Montserrat Fuentes and Peter Guttorp, 3-17. CRC Press.
Doreian, Patrick. 1980. Linear models with spatially distributed data: Spatial disturbances or spatial effects? *Sociological Methods & Research* 9: 29-60.
Doreian, Patrick. 1981. Estimating linear models with spatially distributed data. *Sociological Methodology* 12: 359-88.
Doreian, Patrick. 1982. Maximum-Likelihood Methods for Linear-Models – Spatial Effect and Spatial Disturbance Terms. *Sociological Methods & Research* 10 (3): 243-69.
Downey, Liam. 2003. Spatial measurement, geography, and urban racial inequality. *Social Forces* 81 (3): 937-52.
Downey, Liam. 2005. The Unintended Significance of Race: Environmental Racial Inequality in Detroit. *Social Forces* 83 (3): 971-1008.
Downey, Liam. 2006. Using Geographic Information systems to Re-conceptualize Spatial Relationships and Ecological Context. *American Journal of Sociology* 112: 567-612.
Entwisle, Barbara. 2007. Putting People into Place. *Demography* 44: 687-703.
Fotheringham, A. Stewart, and Peter A. Rogerson, eds. 2009. *The SAGE Handbook of Spatial Analysis*. Thousand Oaks, CA: SAGE.
GeoBooks. <http://qmrg.org.uk/files/2008/11/38-maup-openshaw.pdf>.
Goodchild, Michael F., and Donald G. Janelle, eds. 2004. *Spatially Integrated Social Science.* New York: Oxford Univ. Press.
Hipp, John P. 2007. Block, Tract, and Levels of Aggregation: Neighborhood structure and crime and disorder as a case in point. *American Sociological Review* 72 (5): 659-80.
Hirsch, Arnold R. 1998. *Making the Second Ghetto: Race and Housing in Chicago 1940-1960.* Chicago: University of Chicago Press.
Hunt, D. Bradford. 2010. *Blueprint for Disaster: The Unraveling of Chicago Public Housing.* Chicago: University of Chicago Press.
Kalleberg, Arne L., Peter V. Marsden, Howard E. Aldrich, and James W. Cassell. 1990. Comparing Organizational Sampling Frames. *Administrative Science Quarterly* 35: 658-88.
Lambert, Dayton M., Jason P. Brown, and Raymond J. G. M. Florax. 2010. A Two-Step Estimator for a Spatial Lag Model of Counts: Theory, Small Sample Performance and an Application. *Regional Science and Urban Economics* 40: 241-52.
Lee, Barrett A., Sean F. Reardon, Glenn Firebaugh, Chad R. Farrell, Stephen A. Matthews, and David O'Sullivan. 2008. Beyond the Census Tract: Patterns and Determinants of Racial Segregation at Multiple Geographic Scales. *American Sociological Review* 73 (5): 766-91.
Lesthaeghe, Ron. 2010. *The Unfolding Story of the Second Demographic Transition.* Report 10-696. University of Michigan, Institute of Social Research.
Massey, D. S., and N. A. Denton. 1993. *American Apartheid: Segregation and the Making of the Underclass*. Cambridge, Mass.: Harvard University Press.
McQuarrie and Marwell. 2009. The Missing Organizational Dimension in Urban Sociology. *City and Community* 8 (3): 247-68.

Morland, Kimberly, Ana V. Diez Roux, and Steve Wing. 2006. Supermarkets, Other Food Stores, and Obesity: The Atherosclerosis Risk in Communities Study. *American Journal of Preventive Medicine* 30: 333-9.

Nosek, Vojtěch, and Pavlína Netrdová. 2014. Measuring Spatial Aspects of Variability. Comparing Spatial Autocorrelations with Regional Decomposition in Internal Unemployment Research. *Historical Social Research* 39 (1): 292-314. doi: 10.12759/hsr.39.2014.2.292-314.

Openshaw, S. 1984. The Modifiable Areal Unit Problem. *CATMOG* Series 38. Norwich.

Openshaw, Stanley, and P. Taylor. 1979. A Million or So Correlation Coefficients: Three Experiments on the Modifiable Area Unit Problem. In *Statistical Applications in the Spatial Sciences*, ed. N. Wrigley, 127-44. London: Pion.

Porter, Michael E. 1995. The Competitive Advantage of the Inner City. *Harvard Business Review* 73: 55-71.

Portes, Alejandro, and Ruben Rumbaut. 1996. *Immigrant America: A Portrait*. Berkeley: University of California Press.

Reardon, Sean F, and David O'Sullivan. 2004. Measures of Spatial Segregation. *Sociological Methodology* 34: 121-62.

Sampson, Robert J. 2012. *The Great American City: Chicago and the Enduring Neighborhood Effect.* Chicago: University of Chicago Press.

Sampson, Robert J., Jeffrey D. Morenoff, and Earls Felton. 1999. Beyond Social Capital: Spatial Dynamics of Collective Efficacy for Children. *American Sociological Review* 64 (5): 633-60.

Sharkey, Patrick. 2012. *Stuck in Place: Urban Neighborhoods and the End of Progress Toward Racial Equality.* Chicago: University of Chicago Press.

Small, Mario Luis, and Monica McDermott. 2006. The presence of organizational resources in poor urban neighborhoods: An analysis of average and contextual effects. *Social Forces* 84: 1697-724.

Small, Mario Luis, and Laura Stark. 2005. Are poor neighborhoods resource deprived? A case study of childcare centers in New York. *Social Science Quarterly* 86: 1013-36.

Small, Mario Luis. 2009. *Unanticipated Gains: Origins of Network Inequality in Everyday Life.* Oxford University Press.

Smirnov, Oleg A. 2010. Modeling spatial discrete choice. *Regional Science and Urban Economics* 40 (5): 292-8.

Stein, Michael. 1999. *Statistical Interpolation of Spatial Data: Some Theory for Kriging.* New York: Springer-Verlag.

Talen E., and L. Anselin. 1998. Assessing Spatial Equity: An Evaluation of Measures of Accessibility to Public Playgrounds. *Environment and Planning* 30 (4): 595-613.

Timberlake, Jeffrey M., and John Iceland. 2007. Change in Racial and Ethnic Residential Inequality in American Cities, 1970-2000. *City and Community* 6 (4): 335-65.

Tolnay, Stewart E., Glenn Deane, and E. M. Beck. 1996. Vicarious violence: Spatial effects on Southern Lynchings, 1890-1919. *American Journal of Sociology* 102 (3): 788-815.

Wang, Fahui. 2010. *Quantitative Methods and Applications in GIS.* New York: Taylor and Francis.

Weber, Alfred. 1929. *Theory of the Location of Industries*, trans. Carl J. Friedrich. Chicago: The University of Chicago Press.
Weeks, J. R. 2004. The Role of Spatial Analysis in Demographic Research. In *Spatially Integrated Social Science*, ed. Michael. F. Goodchild and Donald G. Janelle, 381-99. New York: Oxford University Press.
Wilson, William Julius. 1978. *The Declining Significance of Race: Blacks and Changing American Institutions.* Chicago: University of Chicago Press.
Wilson, William Julius. 1987. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. Chicago: University of Chicago Press.