Policy and Place: A Spatial Data Science Framework for Research and Decision-Making

by

Marynia Aniela Kolak

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved June 2017 by the
Graduate Supervisory Committee:

Luc Anselin, Chair
Sergio Rey
Julia Koschnisky
Ross Maciejewski

ARIZONA STATE UNIVERSITY

August 2017

ABSTRACT

A major challenge in health-related policy and program evaluation research is attributing underlying causal relationships where complicated processes may exist in natural or quasi-experimental settings. Spatial interaction and heterogeneity between units at individual or group levels can violate both components of the Stable-Unit-Treatment-Value-Assumption (SUTVA) that are core to the counterfactual framework, making treatment effects difficult to assess. New approaches are needed in health studies to develop spatially dynamic causal modeling methods to both derive insights from data that are sensitive to spatial differences and dependencies, and also be able to rely on a more robust, dynamic technical infrastructure needed for decision-making. To address this gap with a focus on causal applications theoretically, methodologically and technologically, I (1) develop a theoretical spatial framework (within single-level panel econometric methodology) that extends existing theories and methods of causal inference, which tend to ignore spatial dynamics; (2) demonstrate how this spatial framework can be applied in empirical research; and (3) implement a new spatial infrastructure framework that integrates and manages the required data for health systems evaluation.

The new spatially explicit counterfactual framework considers how spatial effects impact treatment choice, treatment variation, and treatment effects. To illustrate this new methodological framework, I first replicate a classic quasi-experimental study that evaluates the effect of drinking age policy on mortality in the United States from 1970 to 1984, and further extend it with a spatial perspective. In another example, I evaluate food access dynamics in Chicago from 2007 to 2014 by implementing advanced spatial analytics that better account for the complex patterns of food access, and quasi-experimental research design to distill the impact of the Great Recession on

the foodscape. Inference interpretation is sensitive to both research design framing and underlying processes that drive geographically distributed relationships. Finally, I advance a new Spatial Data Science Infrastructure to integrate and manage data in dynamic, open environments for public health systems research and decision-making. I demonstrate an infrastructure prototype in a final case study, developed in collaboration with health department officials and community organizations.

DEDICATION

This dissertation is dedicated to Dante (my son), Churchill (a cat), and all the family, friends, and colleagues who made this experience possible.

ACKNOWLEDGMENTS

This research was funded in part by Award 1R01HS021752-01A1 from the Agency for Healthcare Research and Quality (AHRQ), "Advancing spatial evaluation methods to improve healthcare efficiency and quality." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of AHRQ.

CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

OVERVIEW

1.1 Background

With inequalities increasing across multiple environments and landscapes in the United States, improving policy efficacy and fine-tuning interventions serves to support a more equitable society. Reducing economic and health inequalities is consistently identified as a national and global priority. From long-term trends of rising income disparity (Ryscavage, 2015) to well documented differences in mortality (Deaton and Lubotsky, 2003; Pickett and Wilkinson, 2015; Peltzman, 2009), understanding treatment impacts across populations has become crucial to establishing meaningful policy. While researchers, planners, and some officials have sought to reduce inequalities, increasing levels of micro-level segregation between cities and neighborhoods further complicates the spatial organization of urban landscapes (Massey et al., 2009).

One common new approach to ameliorate this problem is place-based policies that maximize efforts and make interventions efficient. With greater interest in these place-based approaches come greater scrutiny. For example many interventions, while well-intentioned, may further increase inequalities due to a number of complex processes affecting how, where, and to whom treatment is delivered and implemented (White et al., 2009; Lorenc et al., 2013). This can happen when the underlying process driving inequality is poorly understood or missed among complicated landscapes, like highly segregated urban environments with multiple spatial patterns.

At the same time, multiple sectors are eager to use new types of data, from Big to

open, to facilitate better quality and more efficient decision-making. This is especially true in health, where I position my research. A greater interest and scrutiny in place-based approaches takes a central role in increasing urban and health equity (Corak, 2013; Amaro, 2014). A push towards place-based policy is further heightened by calls to incorporate meaningful analytics with new types of Big Data in dynamic decision making. However, issues in existing theoretical frameworks, methods, and technological infrastructures have challenged a fully place-based approach to evaluating spatially dynamic problems.

A core challenge to the determination of intervention efficacy remains in attributing underlying causal relationships where complicated processes may exist. Spatial interaction and heterogeneity between units at individual or group levels can violate both components of the SUTVA assumption that are core to the counterfactual framework, making evaluation effects difficult to assess. As interest in causal inference grows across multiple disciplines, large gaps persist in identifying, understanding, and modeling spatial processes that affect program evaluation. Major methodological innovations are still needed in single-level spatial econometrics analysis to account for issues of selection bias, spatial dependence (including spatial spillovers), and spatial heterogeneity essential to causal inference studies, especially with increased availability of high volume, high variety data. Without an integrated spatial systems framework, spatial effects are not accounted for effectively or consistently.

1.2 Conceptual Framework

A spatial data science framework is implemented to address these challenges. With the dynamic, shifting emergence of data science in the past decade, using new types

of data and methods has become increasingly important for real-time decision making. Much of the new Big data is streaming and social, like the firehose of Twitter, and much of the new open data made available by governments can be linked back to questions posed in the social sciences. With new kinds of data and revolutionary shifts in processing and storage capabilities, cross-sector queries emerge to better understand how people behave. Regardless of planning or disciplinary intention, "data science is happening" (Scott, 2014). While the private sector may capitalize on predicting user behaviors, further possibilities beckon in public sector and academic fields. How can new data and data science approaches identify vulnerable populations (and individuals) for local governments to support? Can we better understand the complex, interconnected relationships between social and environmental dimensions of cities? This momentum is challenged by the need for better training and collaboration across disciplines for approaching these new problems (Cleveland, 2001; Provost and Fawcett, 2013; Schutt and O'Neil, 2013).

Data science has thus emerged as a working concept and framework, though growing differently from different fields. It is commonly viewed from the domain of computer science (Foster et al., 2017), though earliest concepts were defined within statistics (Cleveland, 2001). Defining the field is not a goal of this dissertation, though I take the position that it generally incorporates increasingly scalable quantitative techniques, with attention to increasingly scalable infrastructure capable of processing both data and analytic needs. And perhaps most importantly, it takes an inherently applied approach to resolve complex problems with increasingly creative solutions. Spatial data science, which some also argue can be termed "Geographic Data Science," considers the geography and spatial effects at each juncture of decision-making analysis: theoretical, methodological, and technological. What makes the spatial dimension

of data science powerful is that it serves as the place for integrating research design and methodology, data infrastructure, and decision-making within spatially dynamic challenges.

This dissertation advances a spatial data science framework for causal modeling and decision-making specifically, with a focus on spatial econometric and statistical techniques. The framework builds on the previous work of researchers; in both the causal inference and decision support essays, I extend existing frameworks by making space explicit in a formal way. A spatial framework for distilling causal links for policy and intervention assessment considers a more comprehensive understanding of how spatial effects impact the data generating process. It incorporates a more tuned approach to the nuances of how treatments may act differently in different places, and identify what drives such variability. It also includes the extension and development of new tools and technologies, to both improve assignment and treatment estimates, and integrate a Big Data infrastructure for dynamic decision-making.

1.3 Problem Statement and Research Objectives

We need a more spatially-explicit framework to develop and integrate spatially dynamic causal modeling methods: Not only to derive insights from data that are sensitive to spatial differences and dependencies but also to be able to rely on a more robust, dynamic technical infrastructure to manage and explore both spatial data and analytic processes and outcomes. Dynamically exploring and testing intervention efficacy, across the multivariate components of geographies, would better refine these aspects of social science research, and more effectively develop tailored treatment and interventions in customized policy applications.

The purpose of the three dissertation essays is to address this gap theoretically, methodologically and technologically by (1) developing a theoretical spatial framework (within single-level panel econometric methodology) that extends existing theories and methods of causal inference, which tend to ignore spatial dynamics; (2) demonstrating how this spatial framework can be applied in empirical research; and (3) implementing a new spatial infrastructure framework that integrates and manages the required data for health systems evaluation. The goal it thus to develop, test, and implement a spatial data science framework to systematically describe, evaluate, and extend causal analysis in population-based decision-making when spatial effects are present.

Research objectives are, specifically, to: 1) Investigate methodological gaps in causal research where spatial effects may confound findings by defining different approaches that account for these effects when treatment is biased with interaction between units and comparing relevant methods to extensions of spatial econometric techniques that aim to estimate spatial effects in intervention impact assessments; 2) Develop and extend research methods that take "place" into account when evaluating policy, by developing and comparing a sequence of applied models in different environments using appropriate methodologies to test a) what was the effect of the exposure of a set of units to a program or policy on some outcome, and b) if the assignment of observational units to program or control groups introduces a bias when evaluating the intervention effects; and 3) Apply appropriate methods and decision-making tools to develop the groundwork for a new systems infrastructure that dynamically integrates, manages, and access data relevant to place-based decision-making.

The dissertation has multiple constraints to refine and hone focus. Within these components of a spatial data science framework, I focus on research and infrastructure design challenges in my research, with some considerations of usability. I further

narrow the extent by focusing on evaluation work in health, crossing causal inference methods and decision-support structures used. Methodological work focuses on single-panel econometrics; while there are many other invaluable approaches, they were out of scope for this research. Finally, as will be the case for the systems infrastructure essay, focus is confined to population health applications that are still in relatively early days of infrastructure design.

1.4 Dissertation Significance

This dissertation addresses a crucial research gap posed by not incorporating spatial effects when evaluating interventions and developing place-based policies. If spatial effects are not considered from the theoretical perspective, it may confuse processes being studied, miss important signals, and violate core assumptions. In the counterfactual framework, spatial effects violate the assumption that units and treatments are independent. If spatial effects are not incorporated methodologically, results that impact decision-making can be skewed and/or biased (see Chapter 2 for detailed discussion and references). If spatial effects are not considered in infrastructure design, analysis and/or dynamic decision-making may not be feasible, especially in a Big Data context. Existing systems focus on static data and static analysis, encouraging siloed approaches and closed systems. An integrated system developed for scalability and flexibility would both improve and open research and decision-making.

An integrated spatial counterfactual framework, positioned within spatial data science, addresses this research gap directly. If spatial effects impact a phenomenon being considered, they must be accounted for in each component of analysis and investigation (theoretically, methodologically, technologically). When designing research

for policy and decision-making, identification strategies can be extended with spatial effects in a single-level of analysis using new, integrated methods of counterfactuals and spatial econometrics. Addressing simultaneous spatial processes and distilling spatial phenomenon from a systems framework perspective would resolve many challenges existing in research design and dynamic decision-making. A spatially-minded approach is better suited to evaluating interventions that exhibit spatial trends, as well as for developing place-based policies. This is in contrast to a-spatial approaches that do not account for spatial effects, or approaches that use spatial methods but in a stationary or a-spatial way.

This work serves to further critical work in intervention assessment and evaluation studies, compare and build from the strengths of multidisciplinary frameworks in a scalable environment, extend spatial analysis techniques for more robust population-based social science research, and incorporate evidence-based solutions with effective data architecture and design in a collaborative, decision-driven web environment.

Chapter 2

A SPATIAL PERSPECTIVE ON THE ECONOMETRICS OF PROGRAM EVALUATION

Abstract

As interest in causal inference grows in econometrics, statistics, and related fields, large gaps persist in identifying, understanding, and modeling spatial processes that affect program evaluation and research design. Spatial interaction and heterogeneity between units at individual and group levels can violate both components of the SUTVA assumption that are core to the counterfactual framework, making evaluation effects difficult to assess. I discuss how the following methods have been and may be extended to a spatial framework: fixed effects and differences-in-differences, propensity score and matching, regression discontinuity, and instrumental variables. Methodological innovations are needed in single-level spatial econometric analysis to simultaneously account for selection bias, spatial dependence (including spatial spillovers), and spatial heterogeneity. To address these challenges, I propose a spatially explicit counterfactual framework, within single-level panel econometric methodology. Such a framework considers how spatial effects impact treatment choice, treatment variation, and treatment effects. To illustrate this new methodological framework, I replicate a classic quasi-experimental study about evaluating the efficacy of drinking age policy on mortality, and further extend it with a spatial perspective. A spatially explicit counterfactual framework is shown to add further insight to the evaluation of treatment effects.

2.1 Introduction

Since the evaluation of policies and programs often relies on methods of causal inference, advancing these methods can also improve population outcomes. While causal inference impacts multiple disciplines, recent health care reform in the United States highlights the need to further refine analytical tools that are developed in multiple disciplines. The Affordable Care Act of 2010 (ACA, 2010) supported specific programs tailored to improving population health and integrated, place-based efforts to improve the wellbeing of persons and communities. Place-based interventions and community-specific programs include consideration of social, economic, and physical environments that can affect health and disease (Mueller et al., 2011). A consequence of such place-based emphasis is that future efforts to evaluate targeted policies should properly and jointly account for spatial effects, such as spatial dependence and spatial heterogeneity, coupled with participant selection bias, which together can invalidate the results or conclusions from evaluation studies. While there has been tremendous work done on causal inference in the health sciences (see reviews by Imbens and Rubin (2015); Rothman and Greenland (2005); Greenland (2000); Pearl (2001)), methodological gaps in accounting for such spatial effects remain.

A core challenge to the determination of intervention efficacy remains in attributing underlying causal relationships where complicated processes may exist between individuals and places. Spatial effects of spillover or treatment heterogeneity violate the SUTVA assumption (Rubin, 1974), as is often the case in health, social science, political science, and economics. At the same time, specialized methods in spatial regression and spatial econometrics that account for spatial effects in research design have not been fully connected with program evaluation methodology. Recent reviews

have found gaps of causality in econometrics research and gaps of spatial effects in the evaluation literature. They have consistently called for an improved understanding of underlying processes and assumptions affecting causality overall, including spatial processes (Koschinsky, 2013; Cummins et al., 2007; Baum-Snow and Ferreira, 2014; Pearl, 2009; Greenland, 2000).

As interest in causal inference continues to grow in econometrics, following suit from statistics and other fields, large gaps persist in identifying, understanding, and modeling spatial processes that affect program evaluation and research design. Ample evidence, both theoretical and empirical, indicates that ignoring spatial spillover or heterogeneity effects in the statistical analysis can result in misleading inference (Anselin, 1988b; Anselin and Le Gallo, 2006; Anselin and Lozano-Gracia, 2008; Brueckner, 1998, 2003; Mobley et al., 2004, 2012). Even when many contextual (geospatial) factors are included in a regression model, it is unlikely that all such factors can be accounted for in practice, creating spatial effects as omitted variables. In addition, explicit modeling of simultaneous spatial interaction requires the use of specialized model specifications. From a methodological perspective, spatial spillover effects violate the assumption of independence in statistical analysis and require the application of specialized techniques to produce robust, reliable statistical inference. A new, spatially explicit analytical methodology is required to properly and robustly assess the effectiveness, costs, and benefits of place-based policies.

A spatial perspective of the counterfactual framework considers spatial effects in structure of the research design, influence on assignment, and treatment effect evaluation. This requires a careful accounting for potential selection bias, which arises when the decision to participate in the program is not an exogenous factor. It also requires controlling for violations of the assumption of independence between observations

and outcomes, in combination with spatial effects, i.e. spatial dependence (spillover effects) and spatial heterogeneity (different responses in different contexts). Parallel, multidisciplinary approaches are calling for deeper understandings of processes underlying spatial interaction relationships. Within health research, further development of theoretical and empirical approaches of relational geographies is necessary to refine theoretical models and develop more robust, effective health programs and policy (Cummins et al., 2007). This echoes a bold call for a "paradigmatic shift" needed in traditional statistical analysis to causal analysis of multivariate data. This shift also affects the assumptions that underlie all causal inferences, as well as the languages describing those assumptions, conditional nature of subsequent claims, and methods developed to assess those claims (Pearl, 2009).

To address these challenges, I propose a spatially explicit counterfactual framework, within single-level panel econometric methodology.[1] In this essay, I will review existing approaches, identify areas where the treatment of spatial effects is relevant, and illustrate this with the replication of a classic quasi-experimental example as presented by Angrist and Pischke (2015) and Du Mouchel et al. (1987). To contextualize the discussion, I review disciplinary approaches to causal inference and position the discussion in a Heckman-Rubin blended framework with a spatial perspective. Several common methods of estimating causal effect within this framework are further distilled with attention to how spatial effects may be captured. Methodological innovations that extend a causal inference framework with spatial effects are proposed. To illustrate this new methodological framework, I replicate a classic quasi-experimental study about evaluating the efficacy of drinking age policy on mortality, and further extend

[1]Multilevel methodologies are not included, as they are beyond the scope of this study.

it with a spatial perspective. A spatially explicit counterfactual framework is shown to add further insight to the evaluation of treatment effects.

This paper is structured as follows: Section 2 reviews the fundamental problem of inference and frameworks for causal inference. Section 3 discussed the SUTVA assumption and provides an overview of spatial processes that violate SUTVA. Section 4 summarizes common methods of assessing causality when SUTVA is violated, noting spatial extensions and gaps. Section 5 illustrates an empirical example that incorporates spatial components in causality research in an innovative way, with final conclusions discussed in Section 6.

2.2 The Fundamental Problem of Inference

For unit $i \in [1, ...N]$, let Y_i^{obs} denote the realized or potentially observed outcome, following notation in Imbens and Rubin (2015) with substitution $D_i = B_i$ to indicate treatment [2]:

$$Y_i^{obs} = Y_i(D_i) = \begin{cases} Y_i^{obs}(0) & \text{if } D_i = 0 \\ Y_i^{obs}(1) & \text{if } D_i = 1 \end{cases} \tag{2.1}$$

$Y_i^{obs}(0)$ is equal to the realized outcome without treatment or policy D applied, and $Y_i^{obs}(1)$ is the outcome with treatment D applied. $Y_i(1-D_i)$is equal to the missing potential outcome for each unit. The core challenge of causal inference is that alternate outcomes for a policy, program, or treatment cannot be observed (Holland 1986). No agent can be simultaneously in a control and treatment group, and only one of $Y_i(0)$

[2] I use D as the treatment indicator or dummy variable, and later B to indicate the treatment effect.

and $Y_i(1)$ can be observed. Furthermore, one cannot substantiate causal claims from associations alone, even at the population level, as behind every causal conclusion there must exist some causal assumption that is not testable in observational studies (Pearl, 2001). While identification strategies (many reviewed in this paper) can be successfully implemented to distill credible treatment effect estimates, differences in results across identification strategies can reflect different causal relationships in the data and treatment effect heterogeneity (Baum-Snow and Ferreira, 2014). Thus not only does selection into treatment in a quasi-experimental research design for both observed and unobserved groups complicate model specification, but so does treatment effect heterogeneity. Variations can exist across different regions of space and periods of time within a sample, and likewise affect different populations differently. While simplified models with strong assumptions are necessary to begin to understand a phenomenon, a more complex design is essential to more effectively represent the complex realities that exist.

Causal analysis goes further than a standard statistical analysis of inferring parameters of a distribution from samples drawn of that distribution; it aims to infer aspects of the data generation process. This allows for deduction of likelihood of an event not just under static, but dynamic conditions, including predicting effects of interventions and spontaneous changes, identifying causes of reported events, and assessing responsibility and attribution (Pearl, 2001). Causal inference has not fully reached spatial econometric frameworks, however (Herrera et al., 2013). This is despite exponential growth of interest in other fields and methodological gaps that could benefit from it. For example, tools of causal inference could facilitate tests of dependence on variables in spatial frameworks. In this essay, I argue that a more effective interdisciplinary exchange between these framework paradigms is

needed to address gaps within disciplinary frameworks and strengthen causal inference methodology overall.

Econometric methodology is developed to account for endogeneity that can influence responses. Statistics commonly starts with an analysis of randomization and interpretation of causal statements as comparisons of potential versus observed outcomes, allowing for general heterogeneity in effects of treatment (Imbens and Wooldridge, 2009). While statistical frameworks emerging from randomization techniques focus on causality (Hoover, 2004), the treatment of endogeneity and spatial spillover effects in econometric studies are helpful in addressing SUTVA assumption violations. At the same time, econometric studies may benefit from a shift of focus from treatment effects to treatment interactions, as well as more careful considerations of time series, multilevel modeling, and issues of sampling thriving in statistics (Gelman and Zelizer, 2015). These paradigms consider research questions from different perspectives, serving to on the one hand, isolate approaches and identify methodological gaps; and on the other, deepen understanding of certain techniques and close gaps with interdisciplinary exchange.

2.2.1 Frameworks for Causal Inference

Reviewing dominant frameworks of causal inference provides a point of comparison, and departure, from different approaches. The spatial counterfactual framework later proposed in this essay is positioned within a blended Rubin-Heckman approach, building from the strengths of each perspective. One of the most common paradigmatic approaches to assessing causality incorporates counterfactuals and a potential outcomes framework (see Imbens and Rubin (2015) for in-depth review). This approach builds

from statistical traditions and borrows program evaluation concepts and vocabulary (like "treatment" and "control" group assignments) from some of the health sciences for intervention and policy analysis. It diverges from econometric traditions of more structured, theory-driven models, though some work has emerged to further exchange between these paradigmatic traditions.

The Rubin counterfactual causal inference framework requires potential outcomes with each action-unit pair associated with a potential outcome; the observation of multiple units upholding a stable unit assumption; and an assignment mechanism, also serving as the organizing principle. The assignment mechanism is composed of a potential outcomes and treatment indicator, D_i[3]. Units can be individuals, households, or areas. Causal effect investigations are focused on settings with observations on units exposed or not exposed to some policy or program (i.e. treatment). Evaluation is based on a comparison of units exposed and not exposed (Imbens and Wooldridge, 2009). The causal effect of one action-unit pair relative to another requires comparison of potential outcomes, with any treatment occurring temporally before observation of any potential outcome possible (Imbens and Rubin, 2015). Defining a causal effect does not need more than one unit. However, learning about causal effects requires multiple units, thus requiring certain assumptions to hold true in the comparison of those multiple units.

Following the assumption of ignorability, again using Imbens and Rubin (2015) notation, given covariates X, treatment assignment should be independent of outcomes Y, or:

[3]In the original literature, the treatment indicator is denoted as variable W. To avoid confusion with W as a spatial weights matrix or one of many other meanings, I use D_i to denote the treatment variable here.

$$Y_i(0), Y_i(1) \perp D|X \quad . \tag{2.2}$$

Causal relationships are distilled between a vector of treatment variables D and an outcome y. Following notation from Baum-Snow and Ferreira (2014) a data generating process for outcome Y for each observation i in structural form is as follows, substituting X coefficients $\delta = \beta$ to remain consistent with Anselin's 1988b notation, $T = D$ for treatment variable, and $\beta = B$ for treatment effect:

$$Y_i = D_i B_i + X_i \beta_i + U_i + e_i \tag{2.3}$$

where U_i are unobserved values and e_i is remaining stochasticity in the model. B is the treatment effect, equal to the difference between $Y_i(1)$ and $Y_i(0)$.

Heckman contrasts the treatment effects of a program evaluation approach with economic parameters of a structural approach (see Heckman (2010); Heckman and Vytlacil (2007). A structural approach to causal inference follows econometric traditions, originating from a parametric, explicitly formulated empirical model.Heckman (2010) provides a structural form in translation from program evaluation literature as follows: for ex post outcome Y for observation i,

$$Y_i = \alpha + D_i B_i + \epsilon_i \tag{2.4}$$

where D_i is the dummy variable indicating treatment or program participation, B_i is the treatment effect,[4] and ϵ is the error term. To translate back to a potential outcome framework, Heckman substitutes $\alpha = \mu_0, \epsilon = U_0, Y_0 = \mu_0 + \epsilon$ and $B = (Y_1 - Y_0) = \mu_1 - \mu_0 + U_1 - U_0$. Recent developments consider nonlinear or nonparametric

[4]Here, I substitute individual return to participation or treatment effect β, per Heckman notation, with B to remain consistent in our notation.

identification and estimation for policy evaluation, building on rich economic theoretical traditions and producing estimates that cumulate across studies (Heckman, 2010). Some economics models have built on a "natural experiment" research design and further extend with multiple treatment and control groups, multiple pre- and post-observations, and other design features to increase validity (Meyer, 1995).

A link between a structural and program evaluation approach has been further proposed from the Marschak Maxim dialogue, focusing on combinations of structural parameters rather than identifying the individuals themselves (Heckman, 2010). Sensitivity analyses are also essential to examining how results change with covariate adjustment, with bound analysis serving as a tested example (Little and Rubin, 2000; Horowitz and Manski, 2000). Additional paradigmatic approaches to causal inference include decision-oriented and probabilistic causal inference, and Pearl's innovative framework integrating structural equations and graphical models (Dawid, 2000; Pearl, 2010). An excellent review of both deterministic and probabilistic causal inference frameworks can be found in Mur et al. (2011). A Granger 1988 model of causality in econometrics can be applied to time-series analysis, with specific assumptions on temporality, exogeneity, and independence. It serves in contrast to the counter-factual framework discussed here, focused instead on prediction and forecasting. A recent revival of conjoint analysis in causal inference studies seeks to score and compare different hypothesized outcomes simultaneously (Hainmueller et al., 2014). A statistical or "variable-based" modeling of causal inference can also be contrasted with an agent based model framework, where a generative process is tested using multiple theories, focusing on interaction among units (see Smith and Conrey (2007a) for discussion). Rigorous agent based modeling experiments follow a tradition of inference that test multiple hypotheses with simulations, with empirical extensions increasingly emergent.

2.2.2 Discipline-Specific Perspectives on Causal Inference

Not only are there different frameworks of causal inference, but different applications and interpretations based on disciplinary approach. By taking a multidisciplinary approach to causal inference, as I argue in this essay, integrated disciplinary strengths forge more thoughtful and meaningful results in interdisciplinary problems. Causal inference has been implemented across multiple disciplines: the social and biomedical sciences, epidemiology, labor and economic theory, political science, statistics and econometrics. Depending on the paradigm framework or disciplinary approach, causal inference is an issue of identification strategy, control for confounding variables, or missing data problem. In observational studies of the health sciences, understanding the assumptions underlying how the data were generated is crucial to distilling causal relationships (Pearl, 2001). A common approach is to control for confounding variables, though Pearl has suggested confounding variables serve as causal variables 2001. Rothman and Greenland (2005) suggests that as there is nearly always some genetic and some environmental component causes to every causal mechanism, with such multicausality, most causes are not necessary or sufficient to produce disease. Causal inference in epidemiology may therefore be better viewed as an exercise in measurement of an effect rather than as a criterion-guided process for deciding whether an effect is present or not.

In economics and labor theory, it is often considered an identification strategy. General equilibrium effects that contaminate control groups with influence of treatment are common in urban environments and thus influence labor economic studies (see review by Baum-Snow and Ferreira (2014). Likewise essential to urban economics is the necessity to consider the sources of variation in the treatment variables used

to study causal effects. This is to both consider omitted variables ("the endogeneity problem") as well as consider how representative the population considered is for which the identifying variation exists (Baum-Snow and Ferreira, 2014). Identifying causal effects in the social sciences requires an explicit theoretical model and knowledge about the data-generating process, or outcome process under treatment conditions (see review by Gangl (2010). Aggregation to group level has consistently served as the most common approach to SUTVA-violating data in sociological causal inference. However, interest in connecting structural models and instrumental variables may be growing (Gangl, 2010). A common thread emergent in reviews of causal inference in all of these disciplines has been the calls for increased understanding of the data-generating processes underlying models. Because spatial processes can occur between units and within units, I argue that their understanding and model specifications are likewise essential, even when difficult to distill because of SUTVA assumption violations. By integrating strengths from the geographic sciences, relevant causal inference applications are further clarified.

2.3 A Spatial Framework for Causal Inference

2.3.1 Spatial Challenges to the SUTVA Assumption

In the late 1970s, Rubin introduced the Stable-Unit-Treatment-Value-Assumption (SUTVA) that became core to both econometric and statistical frameworks (Rubin, 1974). It assumes that with some intervention or event, units receiving treatment do not affect units not receiving treatment. The SUTVA principle requires outcomes to be independent of actual treatment assignment at both the individual level and

within the larger population. Core assumptions are that potential outcomes for any unit do not vary with treatments assigned to others. Also, that for each unit, there are no different forms of versions of each treatment level that would lead to a different potential outcome (Imbens and Rubin, 2015).

Requiring no interference and no hidden variations of treatment between units pose challenges in causal inference analysis where there exist social interactions, peer effects, neighborhood effects, spatial fixed effects, information diffusion, norm formation, effects of experimental bias, and other effects in the determination of outcomes (Garfinkel et al., 1992; Sobel, 2006; Shadish et al., 2002; Gangl, 2010). Regularity assumptions must be weakened when analyzing treatments with the presence of social interactions, for example, as exchange between individuals and/or groups can alter treatment effect estimates (Sobel, 2006; Morgan and Winship, 2014; Gangl, 2010). Empirical economic analysis has strived to more precisely define components of social interactions that describe agent behavior and impact models, and characterizing an assumed state of equilibrium for analysis for which agents' actions are mutually consistent.

With individual and group interactions, identification problems may arise as equilibrium outcomes cannot easily distinguish endogenous interactions from contextual interactions. There is also the challenge of differentiating an outcome as aggregated individual behaviors versus actual group behaviors, or the reflection problem (Manski, 1993, 2000). Spatial interaction and heterogeneity between units at individual or group levels can violate both components of the SUTVA assumption. This makes program or policy evaluation effects difficult to assess. Exclusion restrictions, or assumptions from acquired knowledge within an expert domain, are core to addressing the second principle of SUTVA (no hidden variations) and are required for designing

a causal inference study. Failure to incorporate spatial components may further result in inconsistent estimates, biased inference, and incorrect understanding of the causal process (Corrado and Fingleton, 2012).

Strategies to make SUTVA more plausible can include redefining treatment levels to a larger set or coarsening the outcome (Imbens and Rubin, 2015). With aggregation, the data-generative mechanisms may be left unspecified in the analysis and depending on the process, may not be addressed in a standard framework (Gangl, 2010). This common strategy transforms observational data to a more aggregate level where SUTVA can be maintained, and then estimated treatment effects at a macro-level can be observed (Imbens and Rubin, 2015; Moffitt, 2005; Morgan and Winship, 2014; Smith, 2003; Gangl, 2010). Yet, common practices of aggregation and coarsening of outcomes to address SUTVA violations may pose serious challenges to researchers should the causal processes being investigated work at a disaggregate spatial resolution or a more precise distillation of outcomes is desired by policy-makers. This challenge becomes further complicated when working with observational data at varying spatial and temporal resolutions.

Randomized evaluations are standards in providing unbiased causal effects, but even these golden standards are not without drawback as spatial spillover challenges the SUTVA assumption (Baylis et al., 2015). Spatial spillover can underestimate treatment effect estimates in importance, affecting both precision and biasing the estimate. Recent research recommends spatial methods to test assumptions, control for externalities, and identify the mechanisms driving outcomes (Baylis et al., 2015). Without spatially explicit analytics, the measure of fit in a regression analysis can be biased in the presence of spatial autocorrelation, which is a common feature of cross-sectional data (Anselin, 1988a; Griffith, 1987; Anselin and Griffith, 1988).

Yet while spatial methods have been useful for measuring change and variability, they have been less commonly used in identifying causality. In the same way that causal inference research can benefit from a spatial perspective, spatially focused research can benefit from a causal perspective, especially in multidisciplinary applications like health. For example, Browning et al. (2003) implemented a spatial model to distill individual, spatial, temporal variance in self-rated health in Chicago. While results characterized health improvement over a decade along socioeconomic factors, the causality of how and in which way health improvement and socioeconomic status affected each other was not clear. In a similar way, over a decade of food accessibility research from geographic perspectives have not yet substantially moved the discussion past correlation. Calls for taking advantage of natural experiments to evaluate causal pathways between the food environment and health outcomes have been made, and remain (Ver Ploeg, 2010). I argue that taking advantage of natural experiments resulting from policy change and similar quasi-experimental settings will push spatial analysis past associations, into causation, and ultimately benefit both perspectives.

When working with observational data or when randomization is not plausible or cost effective, determining how to account for spillover effects and interactions can prove especially difficult, as understanding of the data-generating process may not always be explicit. The actual generative mechanisms will be left unspecified and to the extent they are built on social interactions or some other SUTVA-violating process, cannot be addressed in a standard statistical framework (Gangl, 2010).

2.3.2 Spatial Effects and Spatial Processes

In most causal inference studies addressing potential SUTVA-violating processes, if spatial effects are mentioned at all, they are usually implemented at a single, unidirectional stage of the modeling process. Spatial patterns may be confirmed and described, but understanding of how spatial effects affect outcomes is not investigated. Spatial effects are defined as spatial dependence and spatial heterogeneity, and can manifest in a variety of spatial patterns (Anselin, 1988a). In Table 1, different spatial concepts are defined according to Anselin (1988a) definitions, with health-relevant examples. Because different data generating process can result in similar spatial patterns, it is essential to make spatial effects explicit to test hypotheses and uncover underlying trends.

Spatial Concept	Definition	Example
Spatial Process	Description of how a spatial pattern is generated.	John Snow's famous example of cholera exposure from contaminated water.
Spatial Dependence	Nearby locations are more likely to share similar attributes than distant locations; special case of cross-sectional dependence.	Similarity in health outcomes due to peer effects (e.g. smoking).
Spatial Autocorrelation	Operationalizes spatial dependence.	
Spatial Lag Model (Spatial Spillover)	Spatial autocorrelation in the dependent variable.	Upgrades in cancer screening in one area related to those of neighbors.
Spatial Error Model	Spatial autocorrelation in error term, i.e. in the unexplained part of the model.	Mismatch of extent of spatial phenomenon and the unit for which data is available.
Spatial Heterogeneity (Spatial Heteroskedasticity)	Relationships between variables differ across regions but due to exogenous factor(s), not interaction.	Clusters of increased lead levels in areas with older homes.

Table 1. Overview of Spatial Effects. Source: Anselin et al. (2008)

I argue that a more meaningful framework for causal inference must consider how spatial effects impact assignment mechanisms and treatment effects, and how estimates can be adjusted to account for those effects. In a spatial counterfactual framework, spatial effects would be made explicit and account for related underlying patterns driven by the data-generating process. Feedback effects and simultaneity of multiple spatial

effects occurring at different levels has generally not been used in causal inference analysis, even when those effects are considered to be present. Interdependence is often explicitly defined as weighted averages or sums of unit outcomes, but in such studies, endogeneity of this spatial lag is not addressed (Franzese Jr and Hays, 2009).

The presence of a spatial lag or spatial spillover effect affects model estimates directly, serving as a substantive spatial process where variables of interest at one location are jointly determined by values at other locations. Furthermore, if the phenomenon studied occurs at a different spatial scale than the geographic area indexing the data, model outcomes are affected. I start with a standard linear regression model:

$$y = X\beta + \epsilon. \tag{2.5}$$

Here y is a vector of observations on the dependent variable; X is a matrix of observations on the explanatory variable; β is a vector of coefficients; and ϵ is vector of error term. With ρ as a spatial lag coefficient and W as a vector of terms correlated with spatial disturbance, a spatial lag model (Anselin (1988b)) can be represented as:

$$y = \rho W_y + X\beta + \epsilon, \tag{2.6}$$

and in reduced form:

$$y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\epsilon. \tag{2.7}$$

The challenge of developing a spatial framework for causal inference remains in incorporating direct, simultaneous interdependence of outcomes across units in a model. Spatial effects may exist on multiple levels, and not accounting for them may lead to omitted variables and biased, inconsistent parameter estimation. In a policy

evaluation setting this impacts interpretations. Muting or over-emphasizing an effect of some intervention across a region of complex spatial processes may occur without accounting for a priori underlying processes affecting regions differentially. Treatment heterogeneity resulting from SUTVA-violating processes may also be ignored in this context. A spatial multiplier effect serves as another example, where if ignored, the non-spatial effect will be exaggerated (Anselin, 2003). The other form of spatial dependence that affects models is spatial error, or spatial interdependence in the unexplained part of the model. I begin with the same simple regression model. If spatial effects are present in the error terms, vector ϵ can be represented by the following:

$$\epsilon = \lambda W_E + \mu \quad \text{can also take on form } \epsilon = (I - \lambda W)^{-1}\mu \tag{2.8}$$

Here, μ is an error term that satisfied an i.i.d. assumption and λ represents a spatial autoregressive coefficient. A regression model with a spatial autoregressive disturbance would then take the reduced form:

$$y = X\beta + (I - \lambda W)^{-1}\mu \tag{2.9}$$

Poor specification of spatial error may result in lack of efficiency (Anselin, 1988b), inconsistency for the standard maximum likelihood estimator in probit models (Fleming 2004), and biased standard errors. If the structure of the error covariance matrix is complex and correlations persist across clusters, accounting for spatial correlations beyond clustering correlations is recommended to avoid biased standard errors (Imbens et al., 2011). Spatial error can be the result of when the geographic unit at which a spatial process occurs does not correspond to the unit used in analysis. Data indexed by a county-level unit is appropriate for wider region policy analysis, though it would

obscure localized neighborhood fluctuations by aggregation, thus introducing error by artifice of data format. Aggregation may serve to reduce SUTVA violations, yet poses challenges if the interest in distilling those local neighborhood fluctuations is central to the analysis.

Spatial heterogeneity takes the form of varying coefficients across data. The heterogeneity is spatial because there is a structure to the cross-sectional unit driven by spatial variables. Discrete spatial variability is defined as a spatial regime Anselin (1988b, 1990). Different methods of modeling continuous spatial variability include geographically weighted regression (Brunsdon et al., 2002), the deterministic spatial expansion model (Casetti, 1997), or a special form of random coefficient variation (as summarized by Anselin (2007)). Standard econometric panel methods using discrete spatial variability techniques can be applied to account for spatial heterogeneity.

2.3.3 An Integrated Spatial Framework for Counterfactuals

I propose a spatially explicit counterfactual framework that incorporates existing standards, and extends causal inference modeling with a spatial perspective. This framework builds on previous literature that calls for inclusion of spatial effects in causal inference, as well as recommendations for integrated approaches. Best practices in causal inference research must consider the source of variation in treatment variables and recognize which effect is being estimated (if any) (Baum-Snow and Ferreira, 2014). It must further incorporate replication over space and time to fully "contextualize variation in treatment effects and its structural and institutional determinants" (Gangl, 2010). This follows calls for development of an integrated counterfactual model that considers the empirical aspects of treatment choice, treatment variability, and resulting

treatment effects (Heckman, 2005). I map aspects of spatial effects to each of these components of a counterfactual framework into a new, spatially explicit counterfactual framework. This framework integrates Heckman principles of structure, Rubin's research design-based counterfactual model, and realities of spatial effects that may influence outcomes, as presented in Table 2. Examples of how spatial effects may impact aspects of the integrated counterfactual framework are provided to connect concepts like "peer effects" to their spatial proxy (spatial dependence). Because each spatial concept may require a different specification, a tuned spatial perspective is vital in learning the best methods of implementation.

How is treatment chosen or assigned?	Treatment is assigned by spatial regime. (SD)
	Proximity to intervention affects likelihood of being treated or not. (SD)
	Places with certain characteristics more likely treated. (SH)
What are potential sources of variation in the treatment variables?	Interaction within or between units may appear as **spatial lag** or spillover. (SD)
	Unit of measurement does not correspond to level at which phenomenon takes place (ie. **spatial error**. (SD)
	Spatial patterns emerge from exogenous factors, showing **spatial heterogeneity**. (SH)
What effects are being estimated (if any)?	Spatial Effects (ie. neighborhood effects, information diffusion)
	Spatial Dependence (SD) (ie. peer effects, social interactions)
	Spatial Heterogeneity (SH) (ie. spatial fixed effects)

Table 2. Spatial effects may impact multiple aspects of counterfactual framework.

A spatial perspective goes beyond the implementation of spatial tools or methods. It considers the inherently spatially and temporally dynamic, interactive nature of the populations being studied. A spatial framework should inform the initial design of the model. It is imperative for the researcher to first consider how they think about space and spatial interactions, and how that affects their research design. A conceptual interrogation of potential spatial effects in the phenomenon being studied is necessary to consider sources of spatial dependence and spatial heterogeneity, and

tools to employ. Is there movement within or between units that can appear as spatial spillover? Does the unit of measurement, if indexed by geographic area, correspond to the level at which the phenomenon takes place? For example, crime may occur and influence neighborhoods at block levels, and so their measurement must occur at that resolution to reduce spatial error.

Next, are there distinct spatial patterns in observational data? When data exhibit spatial autocorrelation, observations from nearby units tend to have similar values. Such spatial effects are at the core of geographic analysis: nearby places are more similar than those further away, and tend to not act as isolated regions (Tobler, 1970). The Moran's I statistic can be tested on OLS residuals from the standard linear regression model to detect the presence of spatial autocorrelation, as suggested by Cliff and Ord 1972; 1973; 1981. This has been further extended as a LISA, or local indicator of spatial association (Anselin, 1995). Tests for spatial dependence can identify the type of spatial effect that fits the data best (Anselin, 1988a). In spatial panel models, a suite of diagnostics are increasingly available to evaluate how different spatial processes may be present in the model (Lee and Yu, 2010; Elhorst, 2003; Anselin et al., 2008; Baltagi et al., 2003, 2013, 2007; Pace and LeSage, 2008).

Finally, the researcher must consider spatial processes affecting assignment and/or treatment, and how are they estimated. Classic problems that may impact accurate evaluation of outcomes are summarized in Table 3.

Identification problem (Fisher, 1966)	Is there a need to distinguish endogenous (spatially influenced) interactions from contextual interactions? Are effects the results of treatment, or results of some unobserved variable, or emerging from an unspecified process?
Reflection problem (Manski, 1993)	Do aggregate behaviors affect the treatment outcome? Does the outcome reflect individuals or emergent group behaviors?
Modifiable areal unit problem (Openshaw, 1984)	If aggregating to ease SUTVA violation, is a new error or spatial effect introduced that would impact outcomes? Does aggregating introduce a structural change that can be confused with the actual effect?

Table 3. Selected problems to consider in model design.

Does aggregating introduce a structural change that can be confused with the actual effect? If either or both conditions are met, furthermore, how can multiple and complex spatial effects be detangled from assignment and/or treatment effects, or how can their simultaneous relationship(s) be specified? For many models, spatial effects at assignment or treatment will make for more straightforward design, as spatial effects affecting both complicates model setup and more importantly, interpretation of results, in a single-level analysis (as opposed to a multilevel approach). Detecting the presence of spatial patterns is only the beginning, serving as descriptive tool rather than predictive or prescriptive. The nature of spatial patterns uncovered must be considered and further specified in a formal model, making the effects explicit. A more in-depth discussion of specific strategies follows in Section 4, incorporating work done on spatial extensions of traditional methods to account for spatial effects.

I argue that care must be taken when considering choice of technique. There is no "one size fits all" spatial specification to account for all spatial effects, and an explicit model requires specific consideration. Rather, a spatial perspective must

inform the researcher when considering which specialized methods should be used to account for research design challenges. This can prevent the misuse of spatially explicit methodology, for example using an incorrect technique to resolve a challenge, or using the "right" tool incorrectly. Considering a more holistic understanding of spatial processes, from the first stages of research design to the final steps of assessing estimates, can support the theoretical and empirical underpinnings of a model. I further investigate this concept using the proposed organizing principle of a spatial counterfactual framework in the empirical illustration of this essay.

2.4 A Review of Causal Inference Methods from a Spatial Perspective

Methodological challenges posed by the identification problem, reflection problem, and modifiable area unit problem impact research resign in causal inference, all violating the SUTVA assumption. An important concern in quasi-experimental settings is the extent to which the assignment of observational units (e.g., individuals or aggregate spatial units such as counties) to program or control groups introduces biases in the quantification of program effects. To properly account for this, an extensive methodological literature in econometrics and statistics has developed. It yields techniques that incorporate quasi-experimental research designs, such as differences-in-differences, instrumental variables, regression discontinuity, and propensity score matching (Abbring and Heckman, 2007; Gangl, 2010; Guo and Fraser, 2010; Heckman, 2010; Imbens and Wooldridge, 2009; Morgan and Winship, 2014; Pearl, 2009; Rubin, 2006; Shadish et al., 2002).

In this section, I review these techniques to account for some of these challenges

in quasi-experimental settings, including existing spatial extensions.[5] I additionally discuss how spatial methodologies could more effectively be included to make the complex nature of spatial effects (and their specification) more explicit. Many of these methods of single-level counterfactual analysis have been and may be extended to a spatial framework of causal inference. However, spatial effects are often not formalized when framed as a research design in these examples, as I argue is essential in this essay. Figure 1 summarizes key studies that focus on spatial extensions in causal inference research, most of which are referenced by the following discussion.

2.4.1 Fixed Effects Models and Difference in Differences

In a simple research design for intervention evaluation, one group is tested before and after some intervention. A fixed effect model can be represented as the following, using the Baum-Snow and Ferreira (2014) taxonomy with previously indicated substitutions, with α_i equal to a fixed effect across observations:

$$Y_{it} = D_{it}B_i + X_{it}\beta + \alpha_i + \epsilon_{it} \tag{2.10}$$

Differences-in-differences (DID) is a type of fixed-effect or panel method of causal inference, generally extending a linear regression. It incorporates a comparison group over the same time period as the treatment group. A DID setup for treatment assessment would be:

[5]Multilevel methodologies are also commonly used for this purpose, however are beyond the scope of this study. Extending multilevel research design would require more than a direct spatial panel econometric application, as proposed here; furthermore, spatial multilevel approaches are not fully developed.

Method	Spatial Extensions	Implementation Examples	Assumptions
Differences in Differences (DID)	Jalil 2014; Delgado and Florax 2015; Conley and Taber 2011; Dubé et al 2014	• Radii-based fixed effects to identify treatment and controls before and after an intervention • Tract-based fixed effects for heterogeneity • Dummy variable marks time-based distance from an intervention	In absence of treatment, average outcomes for all groups expected. Seeking group-specific trends.
Propensity Scores and Matching	Hujer, Rodrigues, and Wolf 2007; Schutte and Donnay 2014	• Probability of being treated linked to estimated degree of spatial correlation • Regional effects added as second-order interaction	Assignment is conditional on covariates, and can be predicted by observations or unobserved variables that do not predict outcomes.
Regression Discontinuity (RD)	Gibbons et al 2014; Black 1999; Holmes 1998; Menon and Giacomelli 2012; Atthias and Wicht 2014; Effer and Lassman 2013; Keele and Titiunik 2015	• Treatment assigned by spatial regime • Spatial weights account for difference between places	Population is otherwise similar on either side of discontinuity. Differences in treatment variables can be identified across regimes.
Instrumental Variables (IV)	Lee 2014; Anselin and Lozano-Gracia 2008; Drukker et al 2010; Diamond 2012; Kelejian, Prucha, and Yuzefovich 2004	• Account for endogenous variables in spatial model • Measure local exogenous factor by interacting cross-sectional differences with regional factor by type	Unmeasured confounding exists in treatment effects. IV chosen have causal effects on treatments but not outcomes.

Figure 1. Identification Strategies with Spatial Implementations.

$$Y_{it} = D_{it}B + X_{it}\beta + \varrho_t + \kappa_i + \epsilon_{it} \tag{2.11}$$

where ϱ_t are period fixed effects and κ_i are individual fixed effects. Early work on this method can be found in Ashenfelter and Card (1984); Card (1990); Card and Krueger (1994); Meyer et al. (1990), with a traditional DID setup developed by Imbens and Wooldridge (2009).

A simple differences-in-differences design observes outcomes for two groups over two time periods. One group is exposed to treatment and the second time period and not the first (or vice versa), and the second group is never exposed to the treatment and serves as a control. The conventional DID design requires that in the absence of treatment, (average) outcomes for treatment and control groups will follow parallel paths over time, requiring strong underlying assumptions. A DID model often incorporates a linear parametric model to derive the DID estimator. However, semi-parametric techniques that allow for relaxed identification assumptions have also been pioneered (see brief review in Abadie (2005)). There are multiple challenges with DID design, as outcomes in treatment and/or control groups may be systematically different for some reason other than the intervention studied. Additionally, resulting standard errors can be inconsistent and common corrections may not perform sufficiently to ameliorate this inconsistency (Bertrand et al., 2002).

There have been a handful of promising, preliminary explorations of spatial effects in causal inference research using DID estimates. Observational data and underlying processes can be complicated in context and generation process and may exhibit heterogeneity of treatment effects across space and time. In one formal specification,

with j indexing spatial units such as census blocks or counties, the DID form[6] is as follows:

$$Y_{ijt} = D_{ijt}B + X_{ijt}\beta + \theta_{jt} + \epsilon_{ijt} \tag{2.12}$$

where θ_{jt} represents spatial fixed effects across cross-sectional observations. Preliminary work incorporating geographic components in fixed effects models has offered new insight and occasional clarification of policy or intervention efficacy, proving promising for future work. Conley and Taber (2011) proposed a simple model that would allow for temporal and spatial dependence and heteroskedasticity, across cross sections, depending on group population. For a base model where data is available at a group j and time t level,

$$Y_{jt} = D_{jt}B + X'_{jt}\beta + \theta_j + \gamma_t + \eta_{jt}, \tag{2.13}$$

where D_{jt} is the treatment assignment, X_{jt} is the vector of regressors with parameter vector β, θ_j is a time-invariant fixed effect for group j, γ_t is a time effect common across all groups but varying across time $t = 1, ..., T$, and η_{jt} is a group and time interaction random effect. In fixed effect modeling, heterogeneity of treatment across geographic space can be found when investigating borders of different policies (Jalil and others, 2014). However, tools to detect the causation for that differentiation are not fully developed. Jalil and others (2014) extended a panel regression model along buffered borders of the boundary of the Atlanta Federal Reserve District to investigate lending policies during the Great Depression. The study found that liquidity intervention by the Atlanta Federal Reserve District reduced incidence of bank suspensions by 34 to

[6]Baum-Snow and Ferreira (2014) taxonomy with previously indicated substitutions.

72 percent, though varying across geographic areas. It also made the results more generalizable to a larger, more diverse area. And it showed that the treatment of liquidity intervention affected different areas differently, even with a net significant change. However, there was no explicit treatment of spatial spillovers, and thus no formal test of its significance.

Allowing for spatial interactions may allow for treatment heterogeneity while assessing intervention impact. For example, Delgado and Florax (2015) applied a spatially structured DID design for data where treatment outcomes of units depended on both unit-specific applied treatment as well as neighboring treatments. They found that direct (unit-specific) and indirect (neighboring unit) treatment effects could be identified in straightforward, spatially explicit models (Delgado and Florax, 2015). Straightforward, spatially explicit models can benefit from a structured DID design where unit-specific treatments influence but differ from neighboring units though such models would require unidirectional effects, ignoring feedback from interactions that may be more realistic. A recent exploration into the development of a spatial differences-in-differences estimator, using a spatial probit model, underscored the need for suitable weight matrices to account for spatial links between observations (Dube et al., 2014).

A spatially sensitive conceptual framing requires considerations of how space affects treatment in a causal inference research design. While some DID models have been extended with spatial effects, as has been reviewed here, most do not consider variations in treatment due to differing spatial processes. I return to these concepts in Section 5, when a spatial conceptual framework is implemented by first considering how space impacts both structural and counterfactual elements of the research design.

2.4.2 Propensity Score and Matching Methods

Propensity score matching is the probability of receiving treatment conditional on covariates, first proposed by Rosenbaum and Rubin (1983). For treatment variable D_i and pre-treatment variables denoted by K-component vector of covariates X_i for unit or agent i, estimating the conditional probability of receiving treatment on the observed covariates is summarized[7] as:

$$Pr(D_i = 1|X_i = x) = E[D_i|X_i = x]. \tag{2.14}$$

This pre-processing method compares outcomes of a group with the same propensity to be treated where some get treated and others do not. Hence the propensity score is the probability of being treated. Propensity scores are used to reduce the dimensionality problem of matching, allowing to condition on a scalar variable rather than a general n-space, serving as a weighting scheme. When pre-treatment is not observed propensity score matching can be used for where time-varying unobservables are different in treatment and control and may influence outcomes (Baum-Snow and Ferreira, 2014). This method assumes no omitted variable bias, and then conditions the definition of the treatment variable and set of variables chosen to control for it.

Propensity score methodology may not serve for more behaviorally complex problems (Heckman and Robb, 1985). Furthermore, it assumes selection in and out of a treatment can be fully predicted by observations or unobservable variables that do not predict the outcome of interest, further limiting certain study designs (Baum-Snow and Ferreira, 2014). For non-experimental settings that consider causal inference, propensity score matching has been successful in reducing bias between treated and

[7]Imbens and Rubin (2015) notation.

comparison units (see review by Dehejia and Wahba (2002)). Nonparametric pre-processing methods can bolster parametric model performance by controlling for potential confounding variables and reducing variance of estimated causal effects (Ho et al., 2007).

The specification of the assignment mechanism is central to the development of a counterfactual model, as is the specification of spatial processes. If a phenomenon across a geographic space is being considered, then, how might spatial effects impact how each unit came to receive the treatment level applied? One approach to accounting for spatial effects in the assignment mechanism has been to control for regional effects with matching (Hujer et al., 2009; Schutte and Donnay, 2014). In the Rubin counterfactual model, regional effects can be added as second-order interaction terms in the second stage of specifying propensity scores, to determine $KL+KQ$ components for the logarithmic odds ratio. Yet even at this level, independence is assumed.

There have been several attempts in incorporating spatial processes with matching techniques more broadly. Hujer et al. (2009) investigated macroeconomic effects of labor market programs in West Germany using an extended matching model that accounted for spatial interaction. Their dynamic panel model controlled for unobserved time-invariant regional effects with an augmented matching function. It estimated the degree of spatial correlation using a system GMM estimator. Here, the assumption of independence of observations across the study area was considered invalid, and the resulting econometric model was specified with spatial dependencies to account for it.

Statistical matching has also been extended with sliding spatiotemporal windows to address the modified areal unit problem and selection bias for micro-level interactions of conflict event studies in a "matched wake analysis" (Schutte and Donnay, 2014). In this model, observations are first sorted by nearest neighbor mapping and then dependent

events are counted. Next, observations are matched to previous events, trends, and geographic information using a coarsened exact matching. Finally treatment effects on the dependent variable are established with a DID design for the matched sample. Computational matching allows for repeated matching and readjustments for all spatial and temporal parameter combinations, matching on observables. However, when spatiotemporal cylinders overlap they violate the SUTVA, leading to biased estimates. Either these overlaps must be removed or numbers of previous events and control events must be matched.

Again there are many promising studies demonstrating that propensity scores can be further extended with a geographic perspective, however formal approaches integrated these concepts remain absent. An accurate approximation of the propensity score by estimated propensity scores is essential, as estimators for treatment effects are sensitive to decisions made in the specification of estimated propensity scores. One extension calculated a spatial propensity score with a spatial logit model (Chagas et al., 2011), which confirms conditional likelihood or spatial dependence. Propensity score matching relaxes spatial effects as the spatial dimension is latent, with spatial controls serving as a precondition for correct identification of the effects of interest. Franzese Jr and Hays (2009) suggests a spatial probit model may further extend this methodology, focusing on estimation-by-simulation methods to estimate spatial effects of assignment rather than in terms of parameter estimates. I argue that a sensitivity analysis incorporating spatial effects could help investigate impacts on assignment selection. It is essential in propensity score matching to develop specification on pre-treatment units, distilling spatial components further.

2.4.3 Instrumental Variables

Instrumental variables (IV) serve as additional "treatments" used to estimate causal effects of an outcome when there is unmeasured confounding. IV estimators are used to recover consistently estimated coefficients on treatment variables when treatments are endogenous, attempting to recover random variation in treatments (Baum-Snow and Ferreira, 2014; Heckman, 1979; Angrist et al., 1996; Angrist and Pischke, 2015). IV estimators recover treatment effect D in the following system, in its basic form:[8]

$$Y_i = D_i B + X_i \beta + \epsilon_i \tag{2.15}$$

$$D_i = Z_i^1 \varsigma_1 + X \varsigma_2 + \omega_i \tag{2.16}$$

where a set of excluded instruments Z_i must have at least one instrument per treatment variable for proper identification. Exogenous variables are denoted as $Z = [Z^1 X]$. If $E(Z_\epsilon) = 0$ and the coefficients on excluded instruments ς_1 are sufficiently different from 0, IV estimators can recover consistent estimates of the treatment effect. While a thorough discussion of IV use is beyond the scope of this study, another common method using IV worth noting is the local average treatment effect (LATE). Set up as the following,[9] for instrumental variable Z_i, treatment variable D_i, and outcome variable Y_i:

[8] Baum-Snow and Ferreira (2014) notation.

[9] Angrist and Pischke (2015) notation with $\rho = \varrho$, $A = X$ and previous substitutions.

$$
\begin{aligned}
\text{First Stage:} \quad & E[D_i|Z_i = 1] - E[D_i|Z_i = 0] \\
\text{or} \quad & D_i = \alpha_1 + \varphi Z_i + \gamma_1 X_i + e_{1i} \\
\\
\text{Reduced Form:} \quad & E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] \\
\text{or} \quad & Y_i = \alpha_0 + \varrho Z_i + \gamma_0 X_i + e_{0i} \\
\\
\text{LATE:} \quad & \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} = \frac{Y_i}{D_i}
\end{aligned}
\tag{2.17}
$$

$$
\begin{aligned}
\text{or} \quad & E[Y_{1i} - Y_{0i}|C_i = 1] \text{ where } C_i \text{ is complier population} \\
\text{and} \quad & E[Y_{1i} - Y_{0i}|D_i = 1] \text{ or treatment effect on treated}
\end{aligned}
$$

Instrumental variables should have casual effects on treatments but not outcomes, and be randomly assigned. The exogenous IV can be used as a covariate if it is relevant for treatment assignment and potential outcomes of the IV factor are independent given the treatment status. More precise model specification and identification analysis with group-level instrumental variables are promising: They may support models where SUTVA violations from underlying processes are present, and group effects are likely. Identifying and specifying structural group interaction effects can be made spurious by unobservables in groups, as easily confused with group interaction effects.

Economics has benefitted from social interaction research that extends more complex individual and group-level behaviors (as decisions made over a discrete set of choices). However, spatial components may further benefit this exchange. In this context, Durlauf and Ioannides (2010) review and frame social interactions work in

econometrics while Manski 1993; 2000 provides an overview and detailed discussion of social interaction methodology.

Instrumental variables impact both assignment and treatment, as they must be randomly assigned and linked with treatment. Instrumental variables can accommodate endogenous variables in spatial models (Anselin and Lozano-Gracia, 2008; Arraiz et al., 2010). If there are sufficient variations in group sizes, endogenous and exogenous interaction effects can be identified with a specified spatial autoregressive model using conditional maximum likelihood and instrumental variables. However, large group sizes weaken identification as estimated converge in distribution (Lee, 2007) Incorporating local and national treatment factors can be used to show geographic sorting of phenomena being studied (Diamond, 2016).

While not implemented in the empirical illustration in this essay, considering instrumental variables is useful when determining potential sources of endogeneity between variables. These associations can impact variations in treatment if not accounted for, confusing a treatment effect with changes in underlying data generating processes not related to the policy or intervention being considered. A spatial counterfactual framework must consider how spatial processes may impact, interact, or proxy such phenomenon.

2.4.4 Regression Discontinuity

Whereas matching methods determine assignment conditional on covariates, regression discontinuity design identifies differences in treatment variables across regimes, using a corresponding running variable. Regression discontinuity design assumes that the population being studied is otherwise similar on either side of the discontinuity,

or at least that any differences can be controlled for (Angrist and Pischke, 2010). While DID design seeks group-specific trends, a regression discontinuity setup seeks a behavior deviating from the norm at the point of discontinuity that would suggest a direct influence from the phenomenon being studied. Excellent reviews of regression discontinuity can be found in Imbens and Lemieux (2008) and Lee and Lemieux (2009), and further contextualized in Baum-Snow and Ferreira (2014). A sharp RD design has no ambiguity about treatment status, and can be characterized as the following, again with Baum-Snow and Ferreira (2014) notation:

$$y_i = \alpha + D_i B_i + X_i \beta + U_i + \epsilon_i \tag{2.18}$$

where $D_i = 1(Z_i \geq Z_0)$, where individuals with instrument $Z_i \geq Z_0$ assigned to the treatment group. If such clarity on treatment assignment is not clear, D_i can be written as the following in a fuzzy RD design:

$$D_i = \theta_0 + \theta_1 G_i + u_i \tag{2.19}$$

where $G_i = 1(Z_i \geq z_0)$. Solving for treatment effect B in this form resembles the LATE definition, as an effect is only recovered for some agents (Baum-Snow and Ferreira, 2014). The first stage must be strong enough to recover θ_1.

Where treatments are assigned by regimes that are spatial in nature, such as administrative boundaries or specified areas, a spatial regression discontinuity design may be appropriate for assignment. Boundary-continuity design is considered a special case of regression discontinuity, using a set of spatial weights to account for observables and unobservables. Spatial connections between places are considered as different than those affecting unobservable variables of interest (see (Gibbons et al., 2014) for a discussion). While not common in the health sciences, spatial regression

discontinuity has been implemented with increasing frequency in evaluating polices across the discontinuity of administrative boundaries in economic and political studies (beginning with Black (1999); Holmes (1998); Giacomelli and Menon (2012).

Recent extensions include Athias et al. (2014) who combine a spatial RDD with fixed effects to avoid omitted variable bias due to unobserved heterogeneity of treatment. Furthermore, Egger and Lassmann (2015) define the forcing variable by spatial unit for more flexibility in an attempt to remove the endogeneity bias of the average treatment effect on outcomes. A detailed overview and formalized definition of geographic regression discontinuity design from a political science perspective can be found in Keele and Titiunik (2014). The authors present challenges unique to a spatial perspective by specifying geographic boundaries as regression discontinuities, thus formalizing spatial effects. I argue that distilling different spatial effects can allow for further, meaningful inference. For example, individual effects of a geographic unit, that may be proxied as a spatially heterogeneous process, may coincide or be absent alongside areal unit group effects. Spatial outcomes may appear similar, but will be driven by different processes. As will be shown in the empirical illustration, additional exploration is required to avoid spatial misspecification when considering using regimes as a discontinuity design.

2.5 Empirical Example: Making a Case for a Spatially Explicit Counterfactual Framework

A thorough review of common identification strategies in causal inference research shows that while spatial extensions are needed and encouraging, they are sparsely found in the literature and lack a formalized, integrated approach. As proposed in this

essay, a new methodological framework with a blended Heckman-Rubin organizing principle and formalized spatial perspective, is needed. This would allow for the joint treatment of spatial dependence in the form of spatial spillovers, spatial correlation in the error term, serial correlation across time periods, heteroskedasticity, extreme spatial heterogeneity in the model coefficients, and endogeneity (e.g., due to selection bias). These innovations are critically important for evaluation research, where different policies (or variants of the same policy) may apply to different regions, which exhibit spatial heterogeneity. In all planned evaluations with a place or space-time component, complications or complexities due to spatial spillovers, selection bias, and unobserved heterogeneity need to be properly accounted for in order to obtain robust and reliable parameter estimates.

I propose that spatial econometric techniques can be extended to causal inference research to jointly deal with spatial dependence (such as spatial spillovers impacting diffusion effects), heteroskedasticity, spatial heterogeneity (extreme geographic variation), and endogeneity (e.g., selectivity). This could involve extrapolating from single cross-sectional data settings to situations where observations are available across both space and over time. Recent results for a single cross-section have allowed the treatment of endogenous variables jointly with a spatial lag and/or spatial error specification (e.g., (Anselin and Lozano-Gracia, 2008)), in combination with heteroskedasticity in the error term (e.g., Kelejian and Prucha (2010)). But they have not also dealt with extreme spatial heterogeneity or panel data. Earlier work by Anselin (1988b, 1990) introduced the concept of spatial regimes, where estimation and specification tests allow for the model parameters to vary across a small number of spatial subsets of the data (e.g. urban-rural). This method can account for extreme spatial heterogeneity, where the different regimes are specified theoretically and a-priori. By making these

differing spatial concepts explicit as potential sources of treatment variation (and/or serving as a component to treatment assignment), their impact on treatment effects is likewise made more clear. I argue that this is increasingly urgent in interdisciplinary problems posed by health and policy evaluation, where understanding outcomes more completely would actually improve health outcomes.

While causal inference and counterfactual frameworks emerged from epidemiology and health science literature, spatial extensions that would facilitate a more effective analysis of place-based interventions and strategies have been largely absent. As both counterfactual models and preliminary spatial extensions have been successful in political science and some labor and economic theory, incorporating techniques to similar methodological frameworks in the health and social sciences is very promising. There are huge strides to be made in the evaluations of public health policies and strategies that consider human-environment relationships and chronic disease impacts. These complicated relations may have one or more spatial effects as a core component affecting outcomes, and so require a more sophisticated empirical analysis. In the following analysis, a quasi-experimental study evaluating the effect of drinking age policy on mortality is reviewed and extended with a spatial perspective.

2.5.1 Capturing Spatial Influence of Drinking Age Policy Effects

The effect of minimum legal drinking age (MLDA) on mortality has been examined substantially in policy research, and serves as such a classic example of applied causal inference (Wagenaar and Toomey, 2002; McCartt et al., 2010; Shults et al., 2001). In this study, I replicate the differences-in-differences fixed model framing posed as a DID textbook example by Angrist and Pischke (2015), as replicated from Du Mouchel

et al. (1987) and discussed in Carpenter and Dobkin (2009) and Norberg et al. (2009). Spatial effects are examined in a exploratory spatial data analysis, and then made explicit in a spatial fixed effects panel model. While state fixed effects and state trends are preserved in traditional analysis, the influence of a policy on nearby states is not. Explicitly accounting for spatial effects may relax the SUTVA assumption, thus still allowing for meaningful interpretation of results.

2.5.1.1 Background

After the end of the prohibition era in 1933, most states implemented a drinking age of 21. In the early 1970s, the voting age was reduced to 18 and following tension over the Vietnam War, drinking age policies were influenced once again. After 1971, several states reduced their drinking age to 18. After 1984, federal policy shifted to pressure states into increasing the drinking age again (by tying it to the receipt of fiscal expenditures for infrastructure development), and by the late 1980s drinking ages returned to 21. Some states, like California, kept their drinking age at 21 the entire time, while others had considerable variation at multiple points (in either direction). This resulting natural experiment of policy patchworks that took place in the United States over nearly two decades proved useful for quasi-experimental researchers.

Following this literature, exposure to lower MLDA has been conclusively connected with increased mortality and additional negative outcomes (Wagenaar and Toomey, 2002; McCartt et al., 2010; Shults et al., 2001). While both mortality rates for all deaths and motor vehicle accident deaths decreased from 1970 to 1984, MVA mortality shows a distinctly different pattern (see Figure 2, which uses data from the analysis).

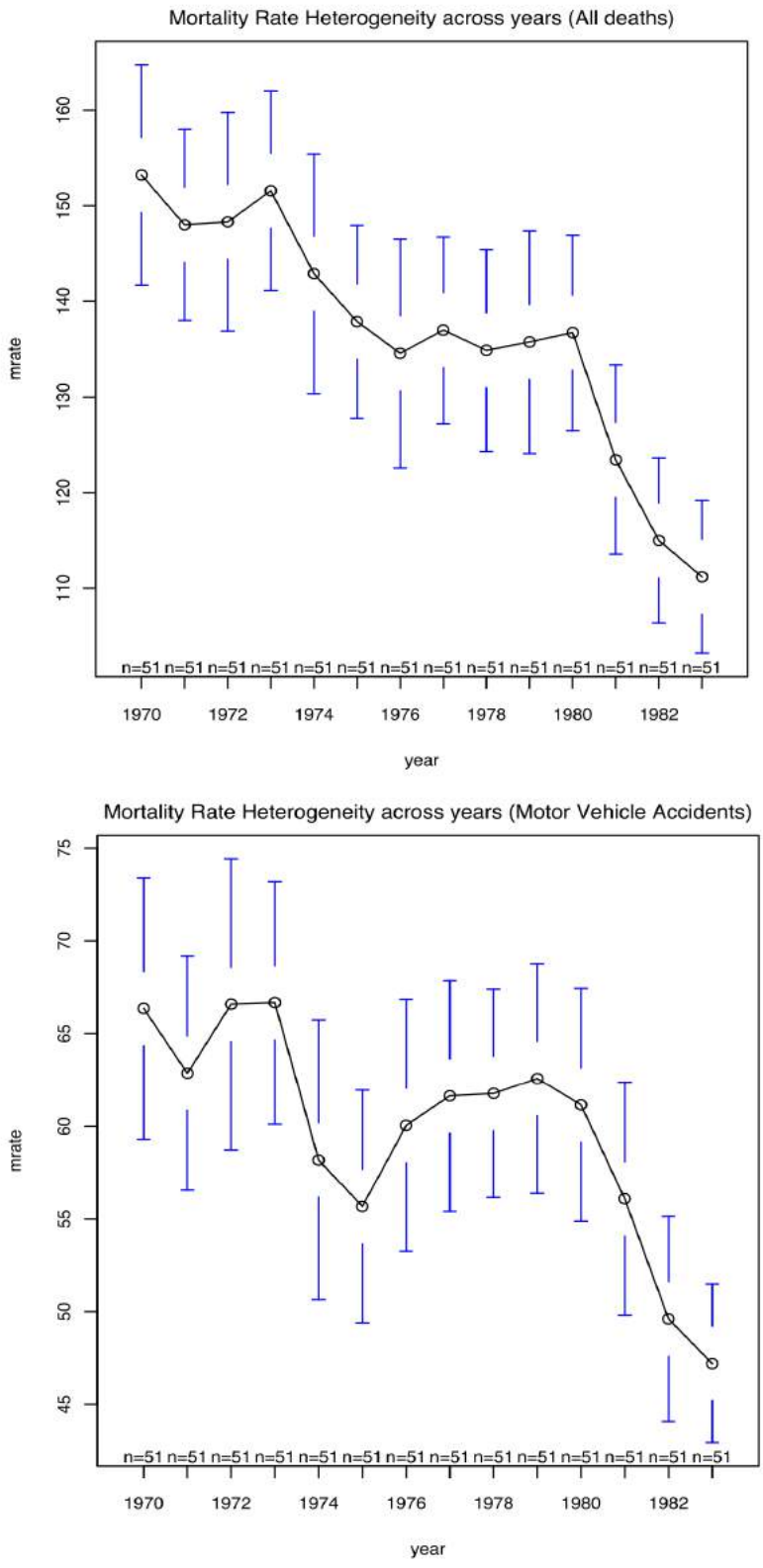


Figure 2. Mortality Rate Heterogeneity, across time.

When controlling for state and birth year fixed effects, lower MLDA is also associated with significantly higher risk of alcohol consumption substance abuse disorder (Norberg et al., 2009). Using a regression discontinuity approach, increased inpatient and emergency department visits have also been significantly associated as an MLDA effect (Callaghan et al., 2013). Much of the literature focuses on the impact of MLDA laws on motor vehicle accidents (MVA), likely underestimating its full impact as it misses alcohol-related conditions requiring hospital settings (Callaghan et al., 2013,?). However, MLDA impact on MVA morality remains a standard policy case, and is thus used in this study to establish a replicable baseline.

In quasi-experimental research investigating MLDA effects, several of the identification strategies discussed earlier in this study are commonly implemented. Regression discontinuity designs were used to evaluate increased MVA mortalities in the year after an individual aged into a legal drinking status, as revisited by Angrist and Pischke (2015). Difference-in-difference or fixed effect model have also been implemented to evaluate the effect legal drinking age has on MVA mortality, as will be discussed here. In early work on this topic by Du Mouchel et al. (1987), a simple spatial regime approach was implemented within the fixed effect model, without success. Spatial effects are revisited in this analysis, though specified in a different way to account for uniquely different spatial processes.

2.5.1.2 Methods

2.5.1.2.1 Data and Definitions

The original data for state, purchase age change, and effective date can be found in Du Mouchel et al. (1987), and the complete dataset with corresponding mortality data used was from the website for Angrist and Pischke (2015). Mortality data for all deaths, motor vehicle accidents, suicide, and internal causes were originally sourced from the US death census for the corresponding time periods, as made available by the National Center for Health Statistics. There are fifty-one "states" (the District of Columbia is included in the original dataset) and fourteen time periods in the panel, ranging from 1971 to 1984. States changed their drinking policies in different years, and with different implementations, making this an unbalanced panel for evaluating treatment effects.

The treatment variable in this analysis refers to the drinking age policy of a state, by year. It is constructed to capture variation due to within-year timing, according to the month when a policy is changed. For example, if Wyoming changed their drinking age to 19 in July 1975, the policy treatment variable $D_{WY,1975}$ is scaled so that young adults over 19 years of age were only able to drink for half that year.

2.5.1.2.2 Original Model Specifications

A multistate regression DID model is implemented, as shown in Angrist and Pischke (2015) with updated notation here:

$$Y_{it} = \alpha + D_{it}B + \kappa_i + \varrho_t + \epsilon_{it} \tag{2.20}$$

where D_{it} represents the treatment variable or legal policy for state i in year t, B as the treatment effect, individual state fixed effect κ_i and fixed time effect ϱ_t. The state and time fixed effect are constructed by multiplying a state or time dummy by a corresponding state or time index. For example, when observations from Wyoming are switched on because the dummy variable for Wyoming is equal to one, all other state dummies are switched off.

To control for state-specific trends like urban versus rural states, changes in physical topography, and differing speed limit policies, a state and year interaction term η_{it} is added. Following Angrist and Pischke (2015) specification with updated notation, this term serves to control for the common trends assumption necessitated in a counterfactual framework.

$$Y_{it} = \alpha + D_{it}B + \kappa_i + \varrho_t + \eta_{it} + \epsilon_{it} \tag{2.21}$$

Both regression estimates and standard errors were replicated, converting STATA code from the original analysis to R. Both models are time-demeaned fixed panel models, and include state and year effects. The HC1 estimator by MacKinnon and White (1985) was necessitated to replicate results more precisely. This heteroskedasticity-consistent (HC) covariance matrix estimator adjusts for degrees of freedom, and is the most common robust standard error estimator used in STATA (Hausman and Palmer, 2012). Parametric models were implemented in R using plm (Croissant and Millo, 2008). Because results were not sensitive to population-weighted implementations in Angrist and Pischke (2015), they were not replicated for this study. Replication results are listed in Table 5.

2.5.1.2.3 Spatial Extension

A spatial counterfactual framework is then implemented to characterize spatial and aspatial impacts on treatment assignment, estimated effects, and variation in treatment variables. I considered how spatial effects can impact each component, and summarized in the Table 4, serving as the organizing principle of the framework.

Prior to spatial model specification, an exploratory spatial data analysis was conducted on regression results to identify potential spatial patterns. Coefficient estimates of MVA mortality for each state in the continental states were mapped, and subjected to spatial testing. Using a queen contiguity spatial matrix, a global Moran's I was conducted to test for spatial autocorrelation. To identify significant spatial clusters and outliers, LISA statistics were calculated. A conditional LISA map showing MVA mortality, conditioned on all mortality, further illustrates how spatial patterns compare across both phenomenon. If all states with higher MVA mortality also have higher mortality overall, the spatial effect may mirror a process not related to the policy being studied. Finally, Lagrange Multiplier tests are conducted on all mortality and MVA mortality fixed panel model regressions to test for spatial dependence.

A spatial fixed effects panel model was then implemented for MVA mortality, with spatial lagged mortality rates indicated as ρW_y, with the following specification:

$$Y_{it} = \alpha + \rho W_Y + D_{it}B + \kappa_i + \varrho_t + \eta_{it} + \epsilon_{it} \tag{2.22}$$

Spatial autocorrelation tests (including the Moran's I, LISA, and conditional LISA) were performed in GeoDa opensource software (Anselin et al., 2006). The spatial panel model was implemented as a "within" model allow-

ing for individual effects, indexed by state and time, using the splm package in R (Millo and Piras, 2012). A complete notebook of results is available at: https://github.com/Makosak/PythonNotebooks/blob/master/MLDA%20Experiment%2C%20Panel%20Data%20Setup.ipynb.

How is treatment chosen or assigned?	Treatment is assigned by state as a policy adoption, and was constructed to account for within-year timing.
What are potential sources of variation in the treatment variables?	Policy is changed according to multiple factors, including increased pressure from residents or groups, and/or increased negative outcomes (ie. motor vehicle accident mortality). Spatial patterns emerge from exogenous factors, showing **spatial heterogeneity**. Variations in speed limits, rural versus more urban driving regimes, and physical geography all serve as such examples. (SH) A state may be influenced by their neighbor's policy implementation; ie. a successful policy nearby may influence state policymakers to adopt (spatial multiplier effect). Thus, interaction within states may appear as **spatial lag**. (SD) While less likely, states may demand a different policy to act in competition with its neighbors. (SD)
What effects are being estimated (if any)?	MLDA effect on mortality (ie. fixed effects) Spatial Heterogeneity (SH) (ie. spatial fixed effects) Spatial Dependence (SD) (ie. spatial multiplier effect)

Table 4. Considering a spatial counterfactual framework to better estimate the effect of drinking age on motor vehicle accident mortality.

2.5.1.3 Results

Panel estimates from Angrist and Pischke (2015), corresponding to equations (1) and (2), are successfully replicated (Table 5). Results report regression estimated of minimum drinking age effects on the death rates (per 100,000) of 18-20 year olds. When accounting for state trends, legal alcohol access added about 6 additional MVA deaths per 100,000 18-20 year olds. Following Angrist and Pischke (2015) formatting, this table shows coefficients on proportion of legal drinkers by state and year from fixed effect models. There was a slight variation in standard error at the hundredth of a decimal point, which was likely due to variations in STATA and R packages. The HC1 estimator by MacKinnon and White (1985) was necessitated to replicate results more precisely. This heteroskedasticity-consistent (HC) covariance matrix estimator adjusts for degrees of freedom, and is the most common robust standard error estimator used in STATA (Hausman and Palmer, 2012).

Dependent Variable	(1)	(2)	(3)
All deaths	10.80	8.47	-
	(4.55)	(4.99)	
Motor vehicle accidents	7.59	6.64	2.54
	(2.47)	(2.60)	(1.68)
State trends	No	Yes	Yes
Spatial Lag	No	No	Yes

Table 5. Drinking age policy effect on motor vehicle accident mortality.
Standard errors are reported in parenthesis.

Tests of spatial dependence for MVA mortality, due to policy effects, at both global and local levels showed significant spatial effects at both scales. When assessing multiple types of mortality from the study period, only MVA deaths have significant

spatial lag dependence at $p = 0.0053$ (Table 6). Additional causes were included as in Angrist and Pischke (2015) to control for lead causes of all deaths. A global Moran's I for only MVA deaths, using coefficients from model (2), shows high positive spatial autocorrelation (0.3625) (Figure 4). There are both significant spatial clusters and spatial outliers, representing a complex spatial pattern overall. When only selecting states that were significant in the analysis, most states are included and account for the majority of spatial autocorrelation (0.3641).

A LISA analysis further distills the patterns (Figure 3). Several states in the West cluster as a group with disproportionally high MVA mortality, and several in the Northeast serve as a low mortality cluster. California, Utah, and Colorado are all spatial outliers of low MVA mortality, compared to relatively higher mortality in surrounding states. New Hampshire is a spatial outlier in the other direction.

Because more MVA deaths may occur in areas with more deaths overall, conditioning on all death rates may offer further insight into clusters or outliers of MVA deaths. To account for this, a conditional LISA map shows MVA deaths due to policy effects conditioned on all deaths (Figure 5). Here, West Virginia emerges as a spatial outlier, with more MVA deaths than surrounding states. South Dakota has slightly less deaths overall than nearby Western states, but with similar levels of higher MVA mortality.

A spatial panel fixed effect model (3) that accounts for state trends (as individual effects) and spatially lagged mortality shows a reduced treatment effect (Table 5). Of the approximately 6 per 100,000 persons affected by the policy overall, a little over 4 per 100,000 were impacted from neighbor state policy influence. This finding is similar in direction, and magnitude of the original, aspatial results, however the original treatment effect is dampened. Following spatial dependence found

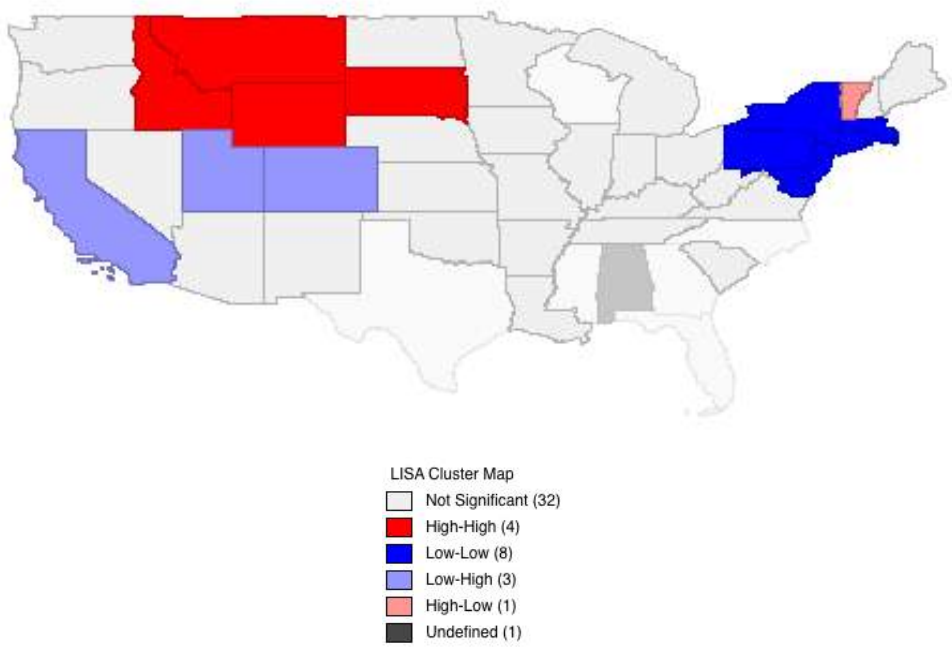


Figure 3. A LISA analysis shows significant spatial clusters and outliers of MVA mortality coefficients, suggesting both spatial dependence and heterogeneity are present. None-significant states from the model are shown in a transparent white color.

in the dataset for MVA deaths in exploratory analysis, confirmation of a spatial multiplier effect of nearby states influencing policy is reasonable and likewise significant.

Mortality (1970-1984)	LM	df	p-value
All deaths	0.0204	1	0.8863
Motor Vehicle Accidents	7.7563	1	0.0053
Suicide	0.5720	1	0.4495
Internal	0.6949	1	0.4045

Table 6. Lagrange Multiplier Test on Spatial Lag Dependence Results.

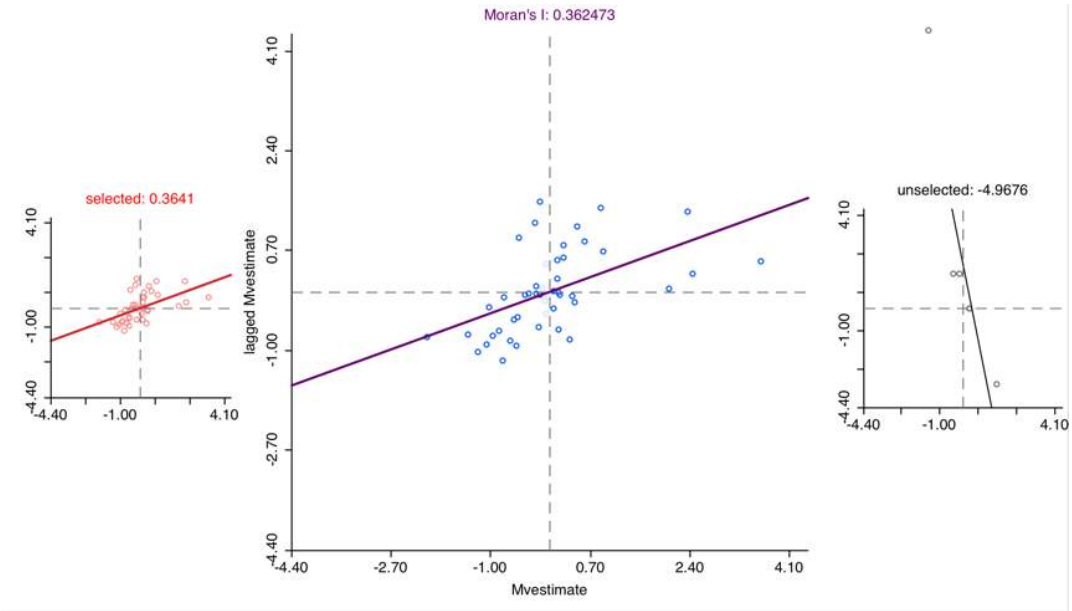


Figure 4. Moran's I shows positive spatial autocorrelation for the entire sample. When selecting only significant state coefficients, the spatial dependence remains.

2.6 Discussion

2.6.1 Overview

In the review on common identification strategies for counterfactual frameworks in quasi-experimental research design, I identified areas where the treatment of spatial effects is relevant. Early applications of spatial processes can be found in differences-in-differences and fixed effects, regression discontinuity, instrumental variable, and matching techniques. However, the complexity and nuance of different spatial concepts underlying different processes remains to be fully implemented in a counterfactual framework. In response, I proposed a spatially explicit counterfactual framework that formally considers how spatial effects can impact each structural component of a

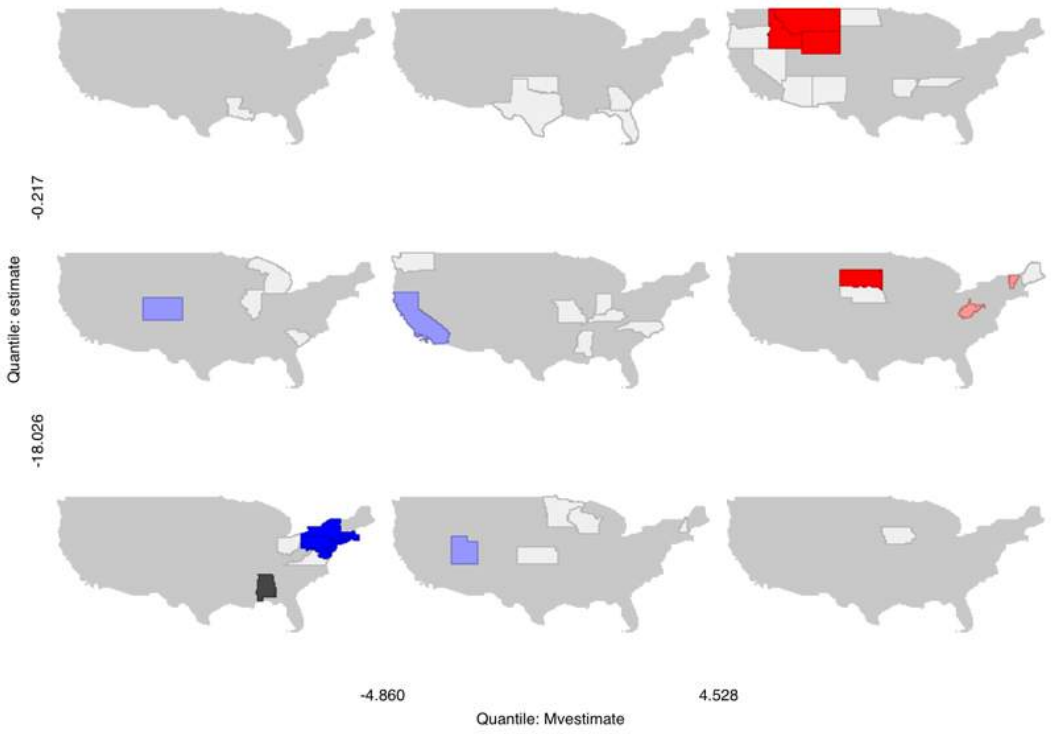


Figure 5. A Conditional LISA map shows significant spatial clusters and outliers of motor vehicle accident deaths (x-axis), conditioned on all deaths (y-axis). LISAs are thus conditioned on two variables (MVA deaths and all deaths per 100,000 persons), rather than only MVA death rates.

causal research design setting (following Heckman organizing principles), and apply those concepts using a Rubin, program-evaluation-specific framework.

Spatial effects can serve as a proxy for many underlying processes in complex problems common to health policy and the social sciences. By specifying these in the correct way, after first effectively diagnosing the spatial pattern observed, the SUTVA assumption of no intra-unit interaction may be relaxed. To demonstrate these concepts in practice, I replicated and extended a standard policy study of measuring drinking age policy effects on mortality. Not only was the treatment effect a highly spatial phenomenon, but its distribution reflected multiple spatial processes.

2.6.2 Case Study Discussion

The results of the MLDA empirical illustration showed that extending policy analysis with a spatial perspective is necessary. If not included, treatment effects can be overestimated. Furthermore, by missing the influence a state may have on its neighbors, the power of a single state switching policies is underestimated as it may not only improve (or worsen) outcomes within the state, but also on nearby ones. While spatial effects provide further insight into treatment effects, their inclusion remains consistent in scale and direction with original results. In other words, the underlying model holds *and* more meaningful interpretation is facilitated, rather than misspecified. The spatial multiplier effect is uncovered by the significance of spatial lag in treatment effects. However, this type of spatial dependence is not the same everywhere, further complicating interpretation. Spatially heterogeneous processes characterize how policy and spatial effects interact in different ways across the country.

The Western states of Idaho, Montana, Wyoming, and South Dakota all lowered their drinking age in the early 1970s, and did not raise to 21 again until the late 1980s. Their MVA mortality rates are exceptionally high, as is the temporal trend of these states working in a similar policy regime. Bordering North Dakota is not included in this cluster, and notably had higher drinking age policies. Nearby California, in further contrast, has a high drinking age (at 21) through the entire period, and low MVA mortality compared to neighboring states. On the other side of the country, eight states reside in a spatial cluster of low mortality rates. While some of these states lowered their drinking ages in the early 1970s, all were raised to 21 (if not already there) by the early 1980s. Because of the strong spatial dependence in both

spatial clusters, and similar policy histories, a clear spatial multiplier effect acted as if policy were "contagious" to nearby states.

The treatment heterogeneity across states are likely due, in part, to spatially heterogeneous processes. From a clustered group perspective; certain regions that behave similarly demonstrate a stronger spatial multiplier effect in policy adoption than others, while other regions do not. By making space explicit and implementing spatial diagnostic techniques to capture these trends, a better understanding as to sources of variation is revealed. In the "high" West cluster, states are large, tend to be more rural, share several common features of physical geography, and likely share similar cultural attributes from parallel historical trends. The "low" cluster in the Northeast has states that tend to be smaller with more urban and well-connected communities, and also share a parallel history. Exogenous processes unique to both of these groups seem to drive MLDA effects in a significant way. At an individual state perspective, additional spatial heterogeneity emerges. In addition to examples already discussed, consider West Virginia. This state has a higher MVA mortality rate than its blue neighbors when conditioned on its overall death rate, and also can be characterized with a different historical, socioeconomic, and physical topography regime. Note, too, that most non-significant states from the fixed model are also spatially proximate, in the South. This spatially heterogeneous process acting both on individual state, and some groups of states, further complicates the distillation underlying process. Spatial lag here serves as a proxy for similarities across neighboring states, though only in some areas of the country.

The original study hypothesized regional variation, and grouped states into twelve regional categories (Du Mouchel et al., 1987). However results varied wildly, with large standard errors, and individual state-level analysis was instead recommended.

Indeed, the spatial effects by regime may miss patterns occurring from neighbor states, as well as unique spatial patterns that emerge from spatially heterogeneous processes. For example, Utah is a spatial outlier of low MVA mortality, as compared to surrounding states. Its unique history and low drinking prevalence, influenced by local cultural patterns, make this a sensible result. Thus grouping outlier Utah in a regional category with Arizona, Nevada, and California (as done in Du Mouchel et al. (1987)) would confuse treatment effects. Furthermore, regions bordering each other are likely to have more influence on each other's policies than those across the country. If spatial effects are manifesting in different ways within region boundaries, in different ways, a multi-state region spatial regime would not be effective (because extreme heteroskedasticity is not present). State-level effects remain plausible, but are further extended by a spatial lag that may more effectively capture treatment heterogeneity influences by neighbor relationships.

2.6.3 Conclusions

Identifying the types of spatial effects presented in a problem is necessary in research design to avoid their misspecification. A suite of existing spatial diagnostics and exploratory spatial data analysis techniques can be implemented to define those spatial concepts, as reviewed in this essay and implemented in the empirical example. By extending a counterfactual analysis directly in a spatial panel econometric model, identified spatial effects (that in turn proxy for underlying processes) are formalized. The spatial counterfactual framework I proposed in this study builds on integrating a robust Heckman-Rubin blended research design where spatial concepts are made

explicit, and has been shown to generate consistent and meaningful results when implemented.

These methodological innovations allow for greater insight in complicated quasi-experimental problems that often challenge SUTVA assumptions. With more place-based programs and interventions driving public health policy across the country, greater understanding is needed to determine how a place-based policy impacts nearby places (for better or worse). Incorporating spatial effects into formal causal inference modeling within this natural experimental framework responds to that call and can not only better model specification, but improve health outcomes with a deepened understanding of the treatment effect.

Chapter 3

URBAN FOODSCAPE DYNAMICS: TRACING FOOD INEQUITY IN CHICAGO FROM 2007-2014

Abstract

The study evaluates food access dynamics by implementing advanced spatial analytics that better account for the complex patterns of food access, and quasi-experimental research design to distill the impact of the Great Recession on the foodscape. It utilizes a validated series of supermarket data in the City of Chicago for 2007, 2011, and 2014. As different processes may drive different trends, isolating unique patterns is essential to testing assumptions that are hypothesized to change the foodscape. An innovative mix of exploratory methods investigates spatial and temporal aspects of supermarket access by building a validated, longitudinal dataset; generating a high-resolution potential food access score for each cross-section; and identifying trends, outliers and significant clusters in both spatial and temporal dimensions. To quantify the effect of the Recession, a sensitivity analysis of quasi-experimental models using different spatial conceptualizations was implemented to explore both consistency and variations in treatment. Chicago neighborhoods with more foreclosure experienced a small but significant worsening in food accessibility after the Great Recession. This is the case even after accounting for variations in income, group effects, and patterns of racial segregation. Persistent trends of inequity across the entire time period study remain, with significantly worse access in segregated black neighborhoods. Inference interpretation is sensitive to both research design framing and underlying processes that drive geographically distributed

relationships. For highly spatial phenomenon like segregation and foreclosure, making space explicit may reduce the magnification of certain results.

3.1 Introduction

Health inequities represent systematic differences in health status between different populations. Socioeconomic, racial, and ethnic disparities in diabetes, hypertension, cancer, and cardiovascular and kidney diseases are just a few examples of the most striking disparities CDC (2005); Davis et al. (1995); Deaton and Lubotsky (2003); Vart et al. (2015). The complex underlying mechanisms of health disparities involve differences in education, employment, health literacy, health insurance, financial status, and access to high-quality medical care, and the medical consequences of stress, bias, and racism Adler and Rehkopf (2008); Pickett and Wilkinson (2015); Phelan et al. (2010). Food accessibility may be an important component in reducing health inequalities, with increasing access to healthy options among socioeconomically disadvantaged populations emergent as a pivotal, interdisciplinary area of research. Systematic differences in the built environment that affect urban neighborhoods' access to healthy foods may also perpetuate health disparities Lake and Townshend (2006); Moore and Diez Roux (2006); Walker et al. (2010); Ball et al. (2009). In neighborhoods with severely restricted access to healthy foods, residents may preferentially consume unhealthy foods associated with increased risks of adverse clinical outcomes due to diabetes, hypertension, atherosclerosis, and kidney disease Gordon-Larsen et al. (2006); Gutiérrez (2015); Wrigley et al. (2003). Additional research suggests there is an association between inequity in neighborhood food environments and diet-related

chronic disease outcomes Gittelsohn and Sharma (2009); Moore et al. (2008), though the causal linkages are not clear.

The term "food desert" is used to describe food access inequity at the neighborhood level, popularized in the last decade with multiple measurement tools published and formalized as policy at multiple scales. In addition to better understanding the relationship between food access and health outcomes, studies also seek to better distill the complexity of food access across different environments and improve methods of measuring accessibility. A call for more robust studies to improve an understanding of the complex relationship between people and food, with a greater focus on reliability and validity of measures of the food environment, is central to food access research McKinnon et al. (2009). Additionally, very little research has looked at changes of food accessibility over time, and few with large-scale, validated datasets.

It is rare for food access studies to go beyond correlation. This may be due in part to the lack of longitudinal data used, though experiments that distill causal pathways are also uncommon. Because a randomized, controlled study targeting food market access may not be feasible, quasi-experiments may be an ideal setting for evaluating different aspects of food access. Calls for taking advantage of natural experiments to evaluate causal pathways between the food environment and health outcomes have been made, and remain Ver Ploeg (2010). Distilling the dynamics of the food retail environment is also essential, and could always benefit from natural experiment research design. We need to understand shifting food access as a result of underlying population change, changes in demographics and/or socioeconomics, policy changes, and/or changes due to national stressors.

The Great Recession of 2008 greatly exacerbated income inequality across the United States Fisher et al. (2015); Danziger et al. (2013). It may have also magnified

unequal access to healthy food by disproportionately reshaping the retail business landscape in vulnerable neighborhoods. With an estimated population of 2.7 million, Chicago is the third largest city in the United States, and one of its most racially and socioeconomically segregated Moore and Diez Roux (2006); Whitman et al. (2012). Chicago serves as a focal point of the food access discussion since 2006, when consultant Mari Gallagher released a privately funded food desert study of Chicago that went on to be front-page news. Heightened awareness of these "food deserts" stimulated local, state, and national programs to incentivize grocery retail development in underserved communities, including the City of Chicago‘s A Recipe for Healthy Places and the Illinois Fresh Food Fund in 2011CDPH (2013); Karpyn et al. (2010). Previous studies reported disparate neighborhood access to healthy food across Chicago Austin et al. (2005); Block and Kouba (2006); Suarez-Balcazar et al. (2006), and cross sectional studies have explored food access relationships Powell et al. (2007); Bower et al. (2014) but none investigated the impact of major historical events like the Great Recession on healthy food access or further distilled how food environments change over time. In addition, Safeway abruptly closed all of its 14 Dominick's supermarkets in Chicago in 2013, further threatening healthy food access in neighborhoods served by the chain Channick (2013).

This study explores a validated series of supermarket data in the City of Chicago for 2007, 2011, and 2014. Over this time period, Chicago experienced and weathered some of the associated fiscal stress of the Great Recession, though inherited fiscal policy challenges persist Hendrick et al. (2010). It is unclear how households in Chicago fared and how shifting economic pressures and/or policies affected the basic human need of access to healthy food. The study evaluates food access dynamics by implementing (1) advanced spatial analysis methods that better account for the complex patterns of food

access over time, and (2) quasi-experimental research design to quantify the impact of the Great Recession on the foodscape. Together, these objectives aim to distill an urgent need to differentiate persistent versus changing relationships in food access, as well as vulnerability to outside shocks. How sensitive is food access to changes in national economic events; and, are certain areas affected more severely than others? Making spatial effects explicit is central to this understanding, though have not been fully implemented in previous research that seeks to detangle causal relationships. Similar or overlapping spatial patterns in food access and sociodemographics may under or overestimate trends if not accounted for.

In Section 2, a potential food access measure is calculated for each cross section, and explored with population characteristics in a spatio-temporal exploratory data analysis. Two cross sections are then isolated, pre- and post-Great Recession, and converted to a spatial panel data setup in Section 3. This period predates policy changes in the supermarket landscape. A sensitivity analysis of Recession effects through different ways of conceptualizing spatial effects in a panel model framework is performed to distill causal relationships between food access, underlying population change, and effects of the Great Recession. Both analyses are distilled in discussion in Section 4.

3.2 Spatio-Temporal Analysis of Food Access Change

Prior to estimating a treatment effect of the Recession, trends in food access must first be identified. By doing so, trends can be compared, validated, and extended within the foodscape literature. Furthermore, change in access can be differentiated from persistent trends that drive the data-generating process. As different processes may

drive different trends, isolating unique patterns is essential to testing assumptions that are hypothesized to change the foodscape. Underlying trends and associated spatial patterns need to be identified, providing some of the spatially informed perspective required in the specification of a spatial counterfactual model.

An innovative mix of methods thus investigates spatial and temporal aspects of supermarket access by (1) building a validated, longitudinal dataset; (2) generating a high-resolution potential food access score for each cross-section; and (3) identifying trends, outliers and significant clusters in both spatial and temporal dimensions.

3.2.1 Methods

3.2.1.1 Data sources and definitions

Using 2010 US Census designations, 791 resident-populated census tracts in the City of Chicago that incorporated a total population of 2.7 million people were investigated. Data on demographic, social, and economic characteristics of the population within the individual census tracts was extracted from the 2012 American Community Survey (ACS) 5-year estimate. This data was used as a baseline rather than tying each time period with corresponding ACS 5-year data, as changes between time periods at such a short interval can be obscured by large margins of error at the tract level Spielman et al. (2014); Folch et al. (2014) Census tracts were used as the spatial unit of observation because much food access research and policy utilizes the census tract as a standard, and following changes to the 2010 Decennial Census, the tract is the finest spatial resolution available for non-decadal socioeconomic data.

Supermarkets are defined as full-service stores that carry a diverse line of groceries

and contain five or more checkout lanes, following industry classification specifics IFM (2009). Supermarkets serve as a proxy for healthy food access, rather than corner stores and local bodegas, because multiple studies show that supermarket access most accurately represents healthy food access in population-based studies Hendrickson et al. (2006); Horowitz et al. (2004). It should be noted, however, that the five checkout lane threshold is relatively low compared to newly-built "big box" supermarkets. The supermarkets included in the database greatly varied in size and included many neighborhood and ethnic groceries. A dataset that detailed the locations of all chain and independent supermarkets in the Chicago area in 2007, 2011, and 2014 was curated using the following complementary public and purchased sources: PolicyMap; InfoUSA; Chicago and Cook County data portals; registry of State of Illinois stores accepting Supplemental Nutrition Assistance Program (SNAP) benefits (food stamps) or Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) coupons; internet research, including Yelp comments and newspaper articles that confirmed the dates that a store had opened or closed; and local supermarket websites. If uncertain about the presence, location, size, or food availability at individual supermarkets, an in-person audit was performed to confirm supermarket status. In order to minimize boundary error that would be introduced by excluding supermarkets that were outside the city limits of Chicago but served adjacent Chicago communities, all supermarkets from the entire surrounding Cook County area were identified.

3.2.1.2 Quantification of Supermarket Food Access

The primary quantitative measure of Chicago residents' access to healthy food was the average distance from residential areas to the nearest supermarket. A food

access measure was generated for each time period to reflect the net effects of the opening and closing of stores. Advanced spatial analysis incorporated validated street networks and land use data, and applied raster coverage and minimum distance techniques using ArcGIS software. Use of street network distance as a proxy for access may better approximate the true distance traveled by residents because it takes the population pattern of each tract into account, while having as its base a much closer approximation of actual distance to the nearest store Block and Kouba (2006); Langford and Higgs (2006); Smoyer-Tomic et al. (2006); Talen (2003). First, the street network of Chicago and surrounding counties was converted from vector (graph representation) to raster (pixelated representation) format to generate a fine-resolution grid map accurate to the nearest ten feet (see Appendix).

Next, the raw street network distances of the closest supermarket was generated to the nearest ten feet; longer distances correspond to less supermarket access. In contrast to approaches that analyze supermarket counts or density per census tract, this approach to generate a fine resolution grid map to calculate raw street network distances to the closest supermarket enabled a better accounting for access to nearby supermarkets that might lie across the border in adjacent census tracts but outside the tracts of individuals' dwellings.

Using the raw values for street network distances to the closest supermarket, the average raw food access index was calculated as the mean street network distance for each individual census tract. In order to minimize error that would be introduced by differential land use across census tracts, streets located in non-residential or industrial land use areas were removed CMAP (2013). A population-adjusted food access index, which standardized the average raw food access index to the total population, was calculated for each census tract.

3.2.1.3 Exploratory Spatial Data Analysis

In order to assess underlying trends of food access dynamics, a range of spatial analytical methods were applied. This includes spatial pattern and outlier detection, spatial cluster analysis, and temporal trend analysis. Identifying consistent patterns in spatial and temporal dimensions is needed to establish a baseline for treatment effect analysis. It facilitates a deeper understanding of hypothesized, underlying trends that are driving generated spatial and temporal patterns. Spatial scales considered were local (at a census tract level), regional (corresponding to data-driven aggregates of tracts), and global (the entire city). A predefined threshold for low or high access was not used. Instead, trends of relative high and low access across different methods of ESDA were examined for their consistency. These methods were implemented in GeoDa opensource software Anselin et al. (2006), and in some cases mapped using QGIS opensource software QGis (2011).

3.2.1.3.1 Spatial Pattern and Outlier Detection

Raw and adjusted food access measures were visualized and explored to uncover trends, outliers, and relationships of high and low access across multiple spatial scales. Census tracts were visualized in color-coded quartiles and outliers of the average raw food access index. Outliers are defined as those tracts with an average raw food access index that was greater than 1.5-times the interquartile range of the overall distribution of the average raw food access index. To account for intrinsic instability of the raw data due to small population bases, I also constructed spatially smoothed maps of the population-adjusted food access index using both traditional spatial smoothing and

empirical-Bayes smoothing Anselin (1995); Rushton and Lolonis (1996); Talbot et al. (2000); Fang et al. (2005).

3.2.1.3.2 Hot/Cold Spot Analysis

To examine spatial clusters of food access, I used LISA maps based on the local Moran I statistic Anselin (1995); Waller and Gotway (2004). These analyses identify hot and cold spots in healthy food access by testing for statistically significant associations of access between each census tract and its neighbors. A spatial weights matrix identifies neighboring tracts if they are contiguous. This was based on the rationale that census tract boundaries do not represent boundaries for grocery shopping, and residents are likely to shop within their own tract or in neighboring ones. An empirical-Bayes-adjusted LISA of the adjusted food access index for each year adjusts for instability across census tracts. Clusters of tracts with either higher or lower rates of supermarket access were identified at a statistical significance level of $p = 0.05$ (based on 999 Monte Carlo permutations, ie. randomly reshuffled locations of supermarket access). These correspond to 'high-high' and 'low-low' LISA statistics. 'High-low' and 'low-high' LISA outliers were not stable at higher significance levels and were excluded because they could also be considered members of hot/cold spot clusters that they bordered.

3.2.1.3.3 Temporal Trends

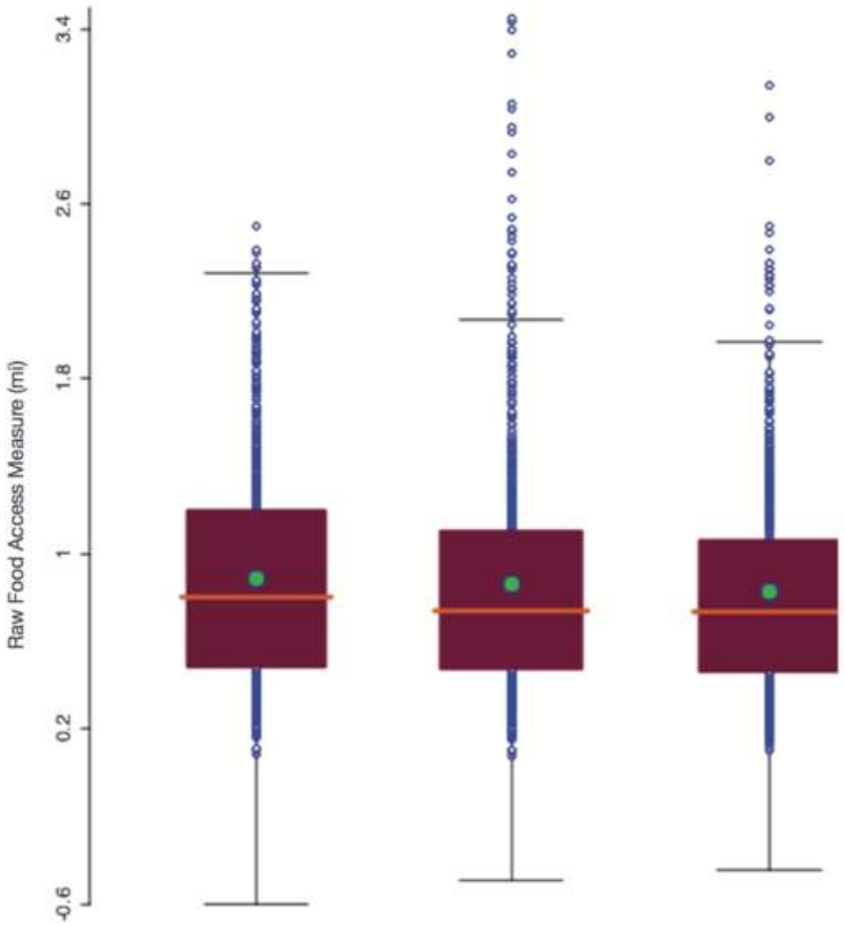
To assess temporal trends in neighborhood food access over time, census tracts were traced across each cross section and coded, using their LISA statistic, as having

persistently high food access, persistently low access, volatile access (improved or worsened access), or no significance (as a hot or cold spot) over time (following work done by Schieb et al. (2013)). For example, a tract must be a significant cold spot for each year to be coded as persistently poor access. This shows relative relationships of high and low access, even if raw cost distance improves globally over time. For increased accuracy, neighbors of significant cluster cores were included in the analysis.

3.2.2 Results

Based on 2012 US Census data, there were 2.7 million residents in the City of Chicago, of which 48.6% were white, 33.7% black, 6.1% Asian and 11.6% other races. Among Chicago residents, 28.4% self-reported Hispanic ethnicity. The distributions of the average raw food access index of the 791 resident-populated census tracts are presented in Supplemental Figure 2 for each year of analysis. Although the total number of full-service supermarkets increased by 20% from 125 in 2007 to 145 in 2011 and 149 in 2014, and the average distance from Chicago residents' dwelling to the nearest supermarket decreased by 8 percent during that period (2007: 0.89 ± 0.49 miles; 2011: 0.86 ± 0.47 miles; 2014: 0.82 ± 0.47 miles), the maximum distance increased from 2.48 miles in 2007 to 3.44 in 2011 and 3.11 in 2014. Furthermore, the number of outlier areas, which correspond to census tracts with extremely long distances to the closest supermarket, increased in the post-recession years of 2011 and 2014 (Figure 2).

Figure 3 presents maps of Chicago that color code quartiles and outliers of the average raw food access index for each census tract. The north side of the city contained most of the high food access census tracts, whereas most of the low food



	2007	2011	2014
Supermarkets (total)	125	145	149
Mean Distance (mi)	0.88	0.86	0.82
Standard Deviation (mi)	0.49	0.54	0.47
Median Distance (mi)	0.80	0.75	0.73
Maximum Distance (mi)	2.48	3.44	3.11
Minimum Distance (mi)	0.09	0.09	0.10

Figure 6. Box plot representation of the distribution of the average raw food access index by each year of analysis.

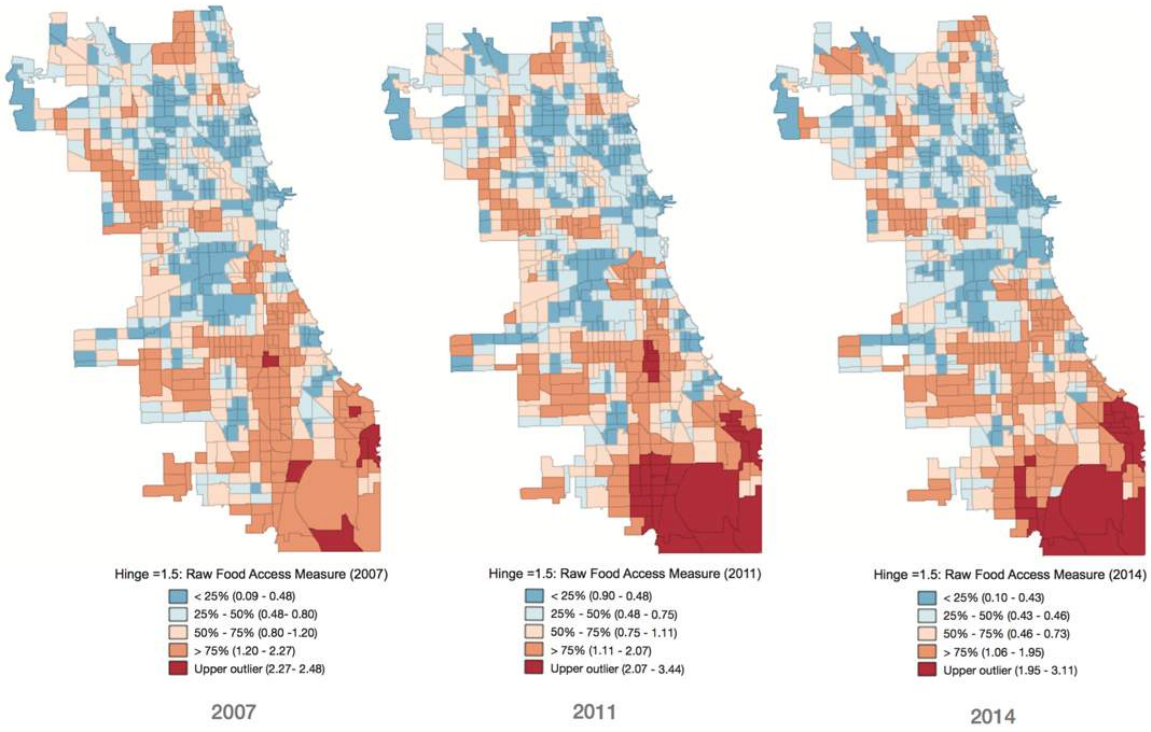


Figure 7. Maps of the City of Chicago that color codes quartiles and outliers of the average raw food access index for each census tract.

access areas, including all of the extremely low access outliers, were located on the south and west sides of the city.

The spatially-smoothed maps of Chicago, with color-coded quartiles of the population-adjusted food access index, demonstrate that parts of the west and south sides of Chicago had the longest distances to the closet supermarket (Figure 4). The LISA maps confirm that throughout 2007, 2011 and 2014, the vast majority of census tracts with significantly low food access (P = 0.05) were located in west and south sides of Chicago, whereas most of the significantly high food access census tracts (P = 0.05) were located in the north side of Chicago (Figure 5).

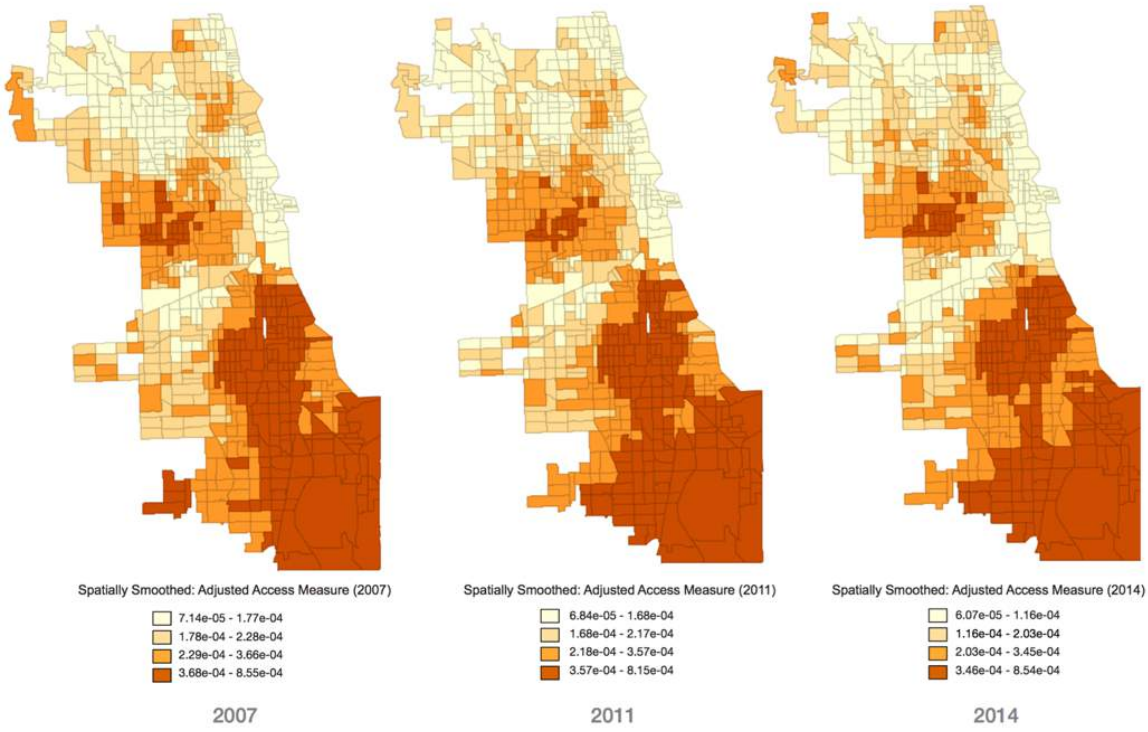


Figure 8. Maps of the City of Chicago that color codes quartiles of the population-adjusted food access using spatial smoothing for regional trends.

The cluster analyses of longitudinal trends in food access across 2007, 2011 and 2014 are summarized in Figure 6. The majority of the census tracts with persistently low or volatile food access were located in the west and south sides of Chicago, whereas most areas with persistently high food access were located in the north side.

Figure 6 presents the demographic characteristics of the City of Chicago overall and according to categories of longitudinal food access defined in Figure 6. Approximately 1.6 million Chicago residents lived in persistently high food access clusters with a mean distance of less than one mile to the nearest full-service supermarket (0.69 ± 0.35 miles in 2007; 0.64 ± 0.30 miles in 2011; 0.62 ± 0.31 miles in 2014). In contrast, more than 0.5 million Chicago residents lived in areas with persistently low food access

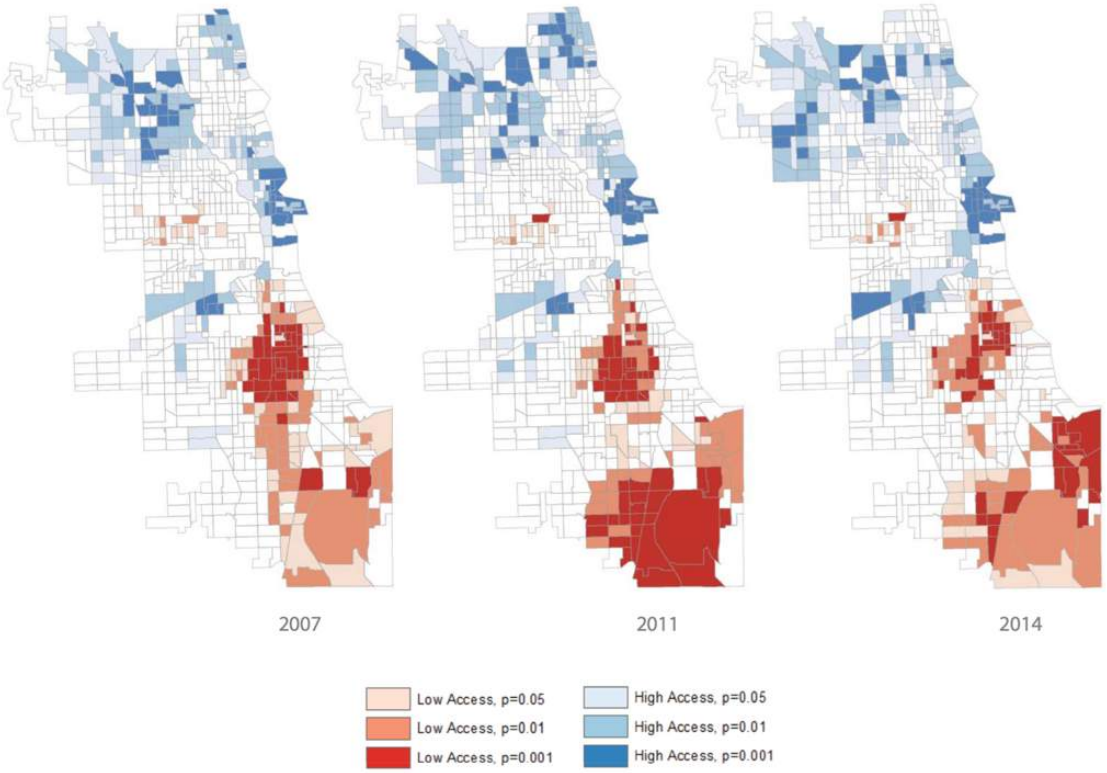


Figure 9. Cluster analyses of longitudinal trends in healthy food access in Chicago between 2007 and 2014.

with a mean distance to the nearest full-service supermarket that was approximately twice the distance in the persistently high food access areas (1.28 ± 0.54 miles in 2007; 1.34 ± 0.68miles in 2011; 1.21 ± 0.54 miles in 2014).

The census tracts with persistently high food access were predominantly white (64.7%) and had the largest proportion of Hispanic residents (37.3%) and the lowest proportion of black residents (11.3%). The high food access census tracts also had the highest median income (58k) and educational achievement (24.5% college graduates), and the lowest rates of unemployment (5.6%), overall poverty (13.5%), children in poverty (21.3%), and residents receiving SNAP benefits. The census tracts with persistently low food access were predominantly black (78.0%), and had the lowest

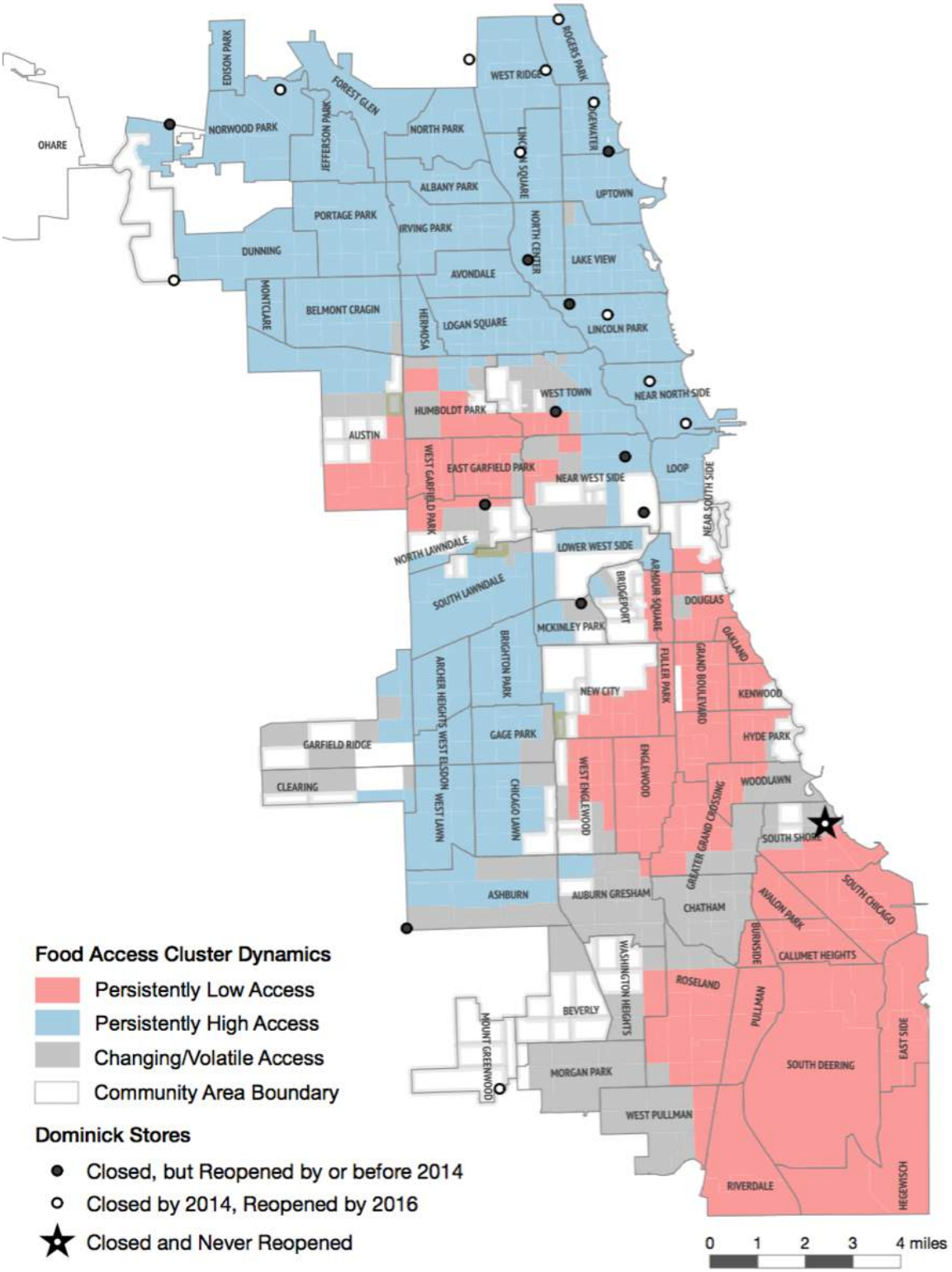


Figure 10. Cluster analyses of longitudinal trends in healthy food access in Chicago between 2007 and 2014.

	All Census Tracts	Persistently High Access	Persistently Low Access	Volatile Access	Persistently No Cluster
Census tracts, n	791	407	203	108	73
Population, n	2,700,519	1,591,800	507,307	341,108	260,304
Mean distance,* miles					
2007	0.89 ± 0.49	0.69 ± 0.35	1.28 ± 0.54	0.92 ± 0.45	0.86 ± 0.43
2011	0.87 ± 0.54	0.64 ± 0.30	1.34 ± 0.68	0.91 ± 0.50	0.78 ± 0.36
2014	0.82 ± 0.47	0.62 ± 0.31	1.21 ± 0.54	0.91 ± 0.45	0.73 ± 0.33
Race-ethnicity					
White, %	48.6%	64.7%	15.8%	25.0%	44.4%
Black, %	33.7%	11.3%	78.0%	67.3%	40.5%
Asian, %	6.1%	8.1%	2.4%	1.8%	7.2%
Other, %	12.9%	17.4%	4.6%	7.0%	9.4%
Hispanic ethnicity, %	28.4%	37.3%	12.5%	15.6%	22.0%
Education					
No high school, %	19.1%	16.0%	24.3%	22.5%	16.7%
High school graduate, %	26.0%	22.6%	30.6%	28.0%	28.7%
College graduate, %	17.5%	24.5%	8.1%	10.8%	14.8%
Socioeconomic data					
Median annual income, $	48,373	58,286	31,430	43,334	47,672
Unemployment, %	6.8%	5.6%	9.0%	8.5%	7.3%
Poverty, % of families	19.6%	13.5%	31.2%	20.9%	19.6%
Children in poverty, %	29.9%	21.3%	45.0%	33.5%	30.2%

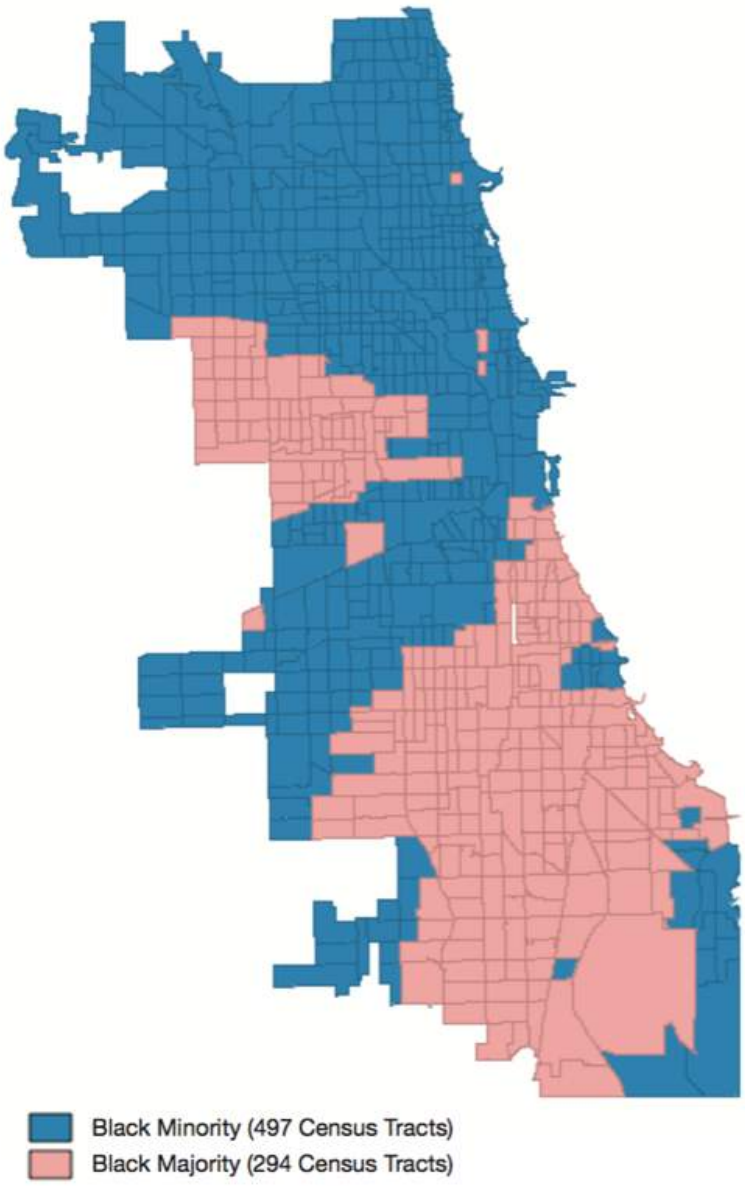
* Distance from Chicago residents' dwellings to the nearest supermarket

Figure 11. Cluster analyses of longitudinal trends in healthy food access in Chicago between 2007 and 2014.

proportion of white residents (15.8%). These census tracts also had substantially lower median income (31k) and educational achievement, and substantially higher rates of unemployment (9.0%), overall poverty (31.2%), children in poverty (45.0%), and residents receiving SNAP benefits. The demographic characteristics of census tracts that experienced volatile food access were most similar to the persistently low food access tracts (Table 1). While healthy food access consistently improved across each time period in Black-minority tracts (0.72 in 2007 to 0.69 in 2011 to 0.65 in 2014), Black-majority tracts initially experienced worsened food access before it subsequently improved (1.17 in 2007 to 1.18 in 2011 to 1.10 in 2014) (Supplemental Figure 3).

During the period of data collection, several Dominick's stores remained closed, reflected in 2014 calculations. Yet all stores residing in high access regions were reopened by 2016. The only Dominick's location in Chicago to remain closed was located in the persistently low food access zone on the south side (Figure 6).

A key finding in the exploratory analysis is that for some residents, food access slightly worsened between 2007 and 2011 before improving in 2014. Furthermore, there was a marked difference between Black majority and non-majority tracts. When using Black majority tracts as a regime indicator, as reported in 2012, the difference is profound. The mean of Black-majority tracts for each year is more than sixty percent greater than Black-minority tracts (see Figure 7). When tracking change in potential food access between 2007 and 2014 with segregated tracts as differing regimes, further interesting findings emerge. As shown in Figure 8, trends across both groups (black majority and minority) follow parallel paths except for a slight worsening in potential food access for black-majority groups in 2011 (see Figure 9). Because this period flanks the Great Recession, a case for extending the study to take advantage of a natural experiment is made.



	2007	2011	2014
Black Majority Tracts	294	294	294
Black Minority Tracts	497	497	497
Mean of Black Majority	1.17	1.18	1.10
S.D. of Black Majority	0.49	0.62	0.48
Mean of Black Minority	0.72	0.69	0.65
S.D. of Black Minority	0.42	0.39	0.38

Figure 12. Map of the City of Chicago that color codes census tracts according to whether or not they include a majority black population (in 2012).

3.3 Distilling the Effects of the Recession on Food Access

The Great Recession impacted the United States with a mélange of increased home foreclosures, increased unemployment, and additional markers of economic decline. The recession period, measured by change of seasonally adjusted real GDP, is from the third quarter of 2008 to the second quarter of 2009 StatExtracts (2015). Because this period took place completely between two cross-sections of food access measured, 2007 and 2011, there is an opportunity to measure its effect. Furthermore, the slight worsening in access for one group in 2011, with otherwise parallel trends, suggests that the impact may be different for differing groups.

To take advantage of this natural experiment, a quasi-experimental research design is implemented to distill the effects of the Recession on food access. These experiments introduce foreclosure risk as a treatment proxying a Recession effect. Because of the strong, consistent spatial patterns made evident by ESDA, a sensitivity analysis of quasi-experimental models using different spatial conceptualizations is implemented to explore both consistency and variations in treatment. By investigating underlying trends that persist over both time periods, and detecting structural breaks or change between periods, key relationships to emerge will provide better insight into foodscape dynamics in Chicago. Furthermore, the effect of the Recession on food access is quantified. Such an economic shock is hypothesized to worsen access, and is likely to impact neighborhoods differently. The following table (1) illustrates the research design setup.

New data is introduced to facilitate this study, including the best census estimates available to proxy a pre- and post-Recession time. The 2000 census is used for pre-Recession estimates, and the 2014 ACS (average of 2009-2014) proxies the post-

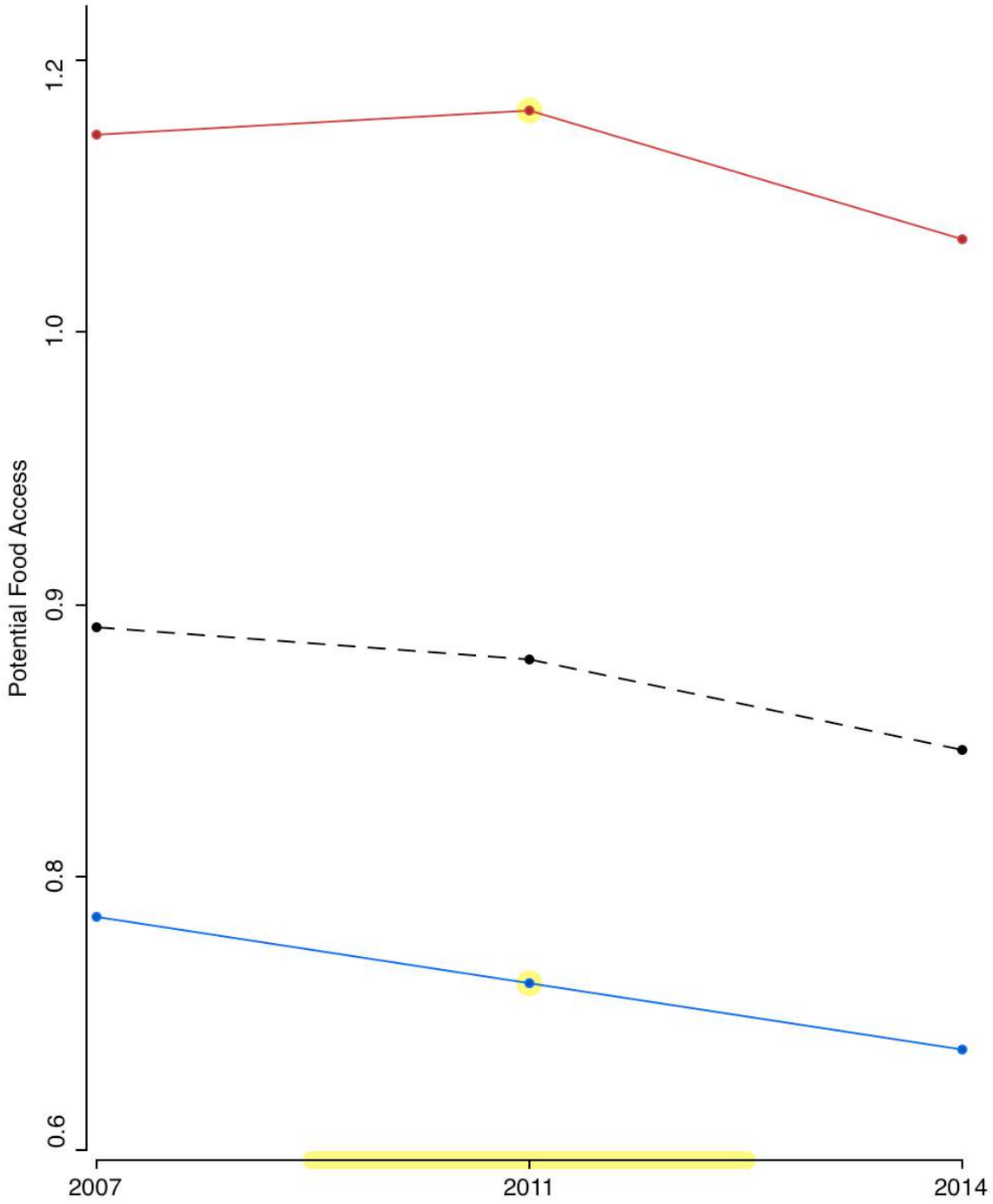


Figure 13. Change in potential food access (in miles) over time. Potential food access is measured as the cost distance to supermarkets, along the road network; longer distance represents poorer access. The red line indicates black-majority tracts; the blue line is black-minority tracks; and the dotted line represented all tracts.

Model	Setup	Spatial Lag	Types of Effects
Simple DID	Aggregate	No	Group, Temporal
Pooled OLS	Balanced Panel	No	Group
Pooled, Lag	Balanced Panel	Yes	Group, Spatial
Fixed Effects	Balanced Panel	No	Group, Temporal
Fixed, Lag	Balanced Panel	Yes	Group, Temporal, Spatial
Random Effects	Balanced Panel	No	Group, Temporal, Individual
Random, Lag	Balanced Panel	Yes	Group, Temporal, Individual, Spatial

Table 1. Research Design Sensitivity Analysis Methods Overview. Group effects are captured by matching estimates to demographic subgroups. Spatial effects are implemented with a spatially lagged food access. Temporal effects for balanced panels are implemented as a time-demeaned within transformation. Individual effects for tract and temporal dimensions are characterized as random.

Recession period. No other form of census data was available for all variables required that did not mix pre- and post-Recession years in estimates or averages. Because of this limitation, continuous estimates are converted to categorical ones to measure demographic group effects.

3.3.1 Methods

3.3.1.1 Variable Definitions

Y_i is the adjusted food access measure for census tract $i \in [1, ...N]$, defined as log of the population-adjusted mean cost distance. The mean cost distance calculation was discussed in Section 2: in summary, it is the cost distance C to supermarkets along the road network, calculated at 10-ft pixel resolution p, and averaged to the the census tract.[10] The log was taken to transform the population adjusted cost distance measure to a normal distribution for parametric analysis.

[10]Pixels not included in the road network had a value of zero.

$$Y_i = \log \frac{\frac{\sum_{p=1}^{N_i} C_p}{N_i}}{P_i} \tag{3.1}$$

To proxy Recession effects, foreclosure estimates for 2008 were used from the U.S. Department of Housing and Urban Development (HUD). The tract-level foreclosure rate estimate was calculated by HUD from the Federal Reserves Home Mortgage Disclosure Act Data on high cost loans, Office of Federal Housing Enterprise Oversight Data on falling home prices, and the Bureau of Labor Statistics data on place and county unemployment rates. The measure was validated at a statewide level by the Mortgage Bankers Association National Delinquency Survey; the foreclosure estimated was predicted at 75 percent. Furthermore, not all census tracts in the study area are available; several tracts with lower populations, and likely lower foreclosures, are omitted. This, however, was reasonable when compared with tracts that also had lower populations that impact lower food access, serving as a new baseline. While imperfect data, this was the only dataset available to provide the finest-resolution foreclosure data for analysis. As such, an excess risk estimate from the data is calculated to serve as a dummy variable for the Recession effect, rather than the continuous variable provided.

Excess risk of foreclosure was calculated as a relative risk measure for the entire study area region. Following Anselin et al. (2006), as adapted to this application, consider foreclosure estimate R_i as events of interest and population P in areal units $i \in [1, ...N]$. The reference risk estimate is then:

$$\tilde{\pi} = \frac{\sum_{i=1}^{N} R_i}{\sum_{i=1}^{N} P_i} \tag{3.2}$$

which is different from the sum of observed events per unit. Risk here is the weighted average of region N specific rates, each weighted by their share in the whole population:

$$\tilde{\pi} = \sum_{i=1}^{N} \tilde{\pi}_i * \frac{P_i}{\sum_{i=1}^{N} P_i} \quad \text{where } \tilde{\pi}_i = R_i/P_i \tag{3.3}$$

and only when each unit has the same population will the region N specific rates equal the study region's rate Anselin et al. (2006). Using the standardized incidence ratio with population, an expected incidence value can then be calculated for each areal unit. Higher risk ratio values correspond to populations experiencing elevated risk for foreclosure. An elevated risk greater than two times the relative risk for all tracts was set as the threshold for this analysis.

$$D_i = 1 \quad \text{if } \tilde{\pi}_i > 2\tilde{\pi} \tag{3.4}$$

Thus, for census tract $i \in [1, ...N]$, let Y_i^{obs} denote the realized or potentially observed outcome. Excess risk of foreclosure is indicated as a treatment D_i.

$$Y_i^{obs} = Y_i(D_i) = \begin{cases} Y_i^{obs}(0) & \text{if } D_i = 0 \\ Y_i^{obs}(1) & \text{if } D_i = 1 \end{cases} \tag{3.5}$$

3.3.1.2 Quasi-Experimental Research Design

For the next series of analyses, the effect of the Recession on potential supermarket access is made explicit by the introduction of a treatment variable in a counterfactual framework. A dummy variable that represents excess foreclosure risk, proxying the effect of the Recession, follows the convention that $D_i = 1$ if the tract has over two

times excess risk for foreclosure. This treatment is set to zero for all tracts in the first time period, and changes in the second; treatment here is considered harmful, with the effects of the Recession assumed to be negative. Thus, treatment is assigned according to place with certain characteristics more likely to be treated, serving as a case of spatial heterogeneity. Potential sources of variation in the treatment variable are mainly that the unit of measurement (here, a tract) may not correspond to the level at which the phenomenon takes place, and that additional patterns may emerge from exogenous factors. Both of these potential sources can be made explicit as cases of spatial dependence in the form of spatial error, and spatial heterogeneity.

The SUTVA principle of a counterfactual framework requires outcomes to be independent of actual treatment assignment at both the individual level and within the larger population. Core assumptions are that potential outcomes for any unit do not vary with treatments assigned to others. Also, that for each unit, there are no different forms of versions of each treatment level that would lead to a different potential outcome Imbens and Rubin (2015). Spatial interaction and heterogeneity between units at individual or group levels can violate both components of the SUTVA assumption. The following experiments serve as different quasi-experimental framings to relax the SUTVA assumption and consider the appropriate exclusion restrictions that impact hidden variation in outcomes.

3.3.1.2.1 Simple DID Analysis

A common approach to make SUTVA plausible is to redefine treatment levels at a coarser level, averaging out the potential SUTVA-violating variation. A difference-in-difference (DID) quasi-experimental design with matching, at an aggregated level, is

first implemented to track group-specific trends. A simple differences-in-differences design observes outcomes for two groups over two time periods. One group is exposed to treatment (here, foreclosure risk) in the second time period and not the first, and the second group is never exposed to the treatment and serves as a control. The conventional DID design requires that in the absence of treatment, (average) outcomes for treatment and control groups will follow parallel paths over time, requiring strong underlying assumptions.

One approach to accounting for spatial effects in the assignment mechanism has been to control for regional effects with matching Hujer et al. (2009); Schutte and Donnay (2014). Because of the strong correlation and spatial effects of segregation and food accessibility in Chicago, demonstrated in Section 2, treatment assignment was further segmented in a matching pre-processing step. Black-majority ($X_{B,i} = 1$) and Black-minority ($X_{B,i} = 0$) census tracts were added as conditional criteria. By comparing segregated tracts with and without treatment, the research design can further distill the links being studied.

The counterfactual thus has four cases. The first two cases are controls, and the last two are treatments:

$$Y_i^{obs} = Y_i(D_i) = \begin{cases} Y_i^{obs}(0)|X_{B,i} = 0 & \text{if } D_i = 0 \\ Y_i^{obs}(0)|X_{B,i} = 1 & \text{if } D_i = 0 \\ Y_i^{obs}(1)|X_{B,i} = 0 & \text{if } D_i = 1 \\ Y_i^{obs}(1)|X_{B,i} = 1 & \text{if } D_i = 1 \end{cases} \tag{3.6}$$

An average for each case reflects the aggregated, observed food measure according to treatment and control group. DID and matching assumptions hold that absent treatment, groups will otherwise remain similar. The group matching results must confirm this assumption.

3.3.1.2.2 Parametric DID Analysis, With and Without Spatial Effects

Next, a DID fixed-effect panel model is implemented at a finer resolution, extending the linear regression in equation 6 with a treatment variable D_{it} and the measured effect of the Recession, γ, the variable of interest. To make explicit the variation in non-Black majority tracks, White-majority and Hispanic-majority tracts are represented. Furthermore, majority status is preserved for each time period.

First, a pooled OLS is implemented as a baseline, ignoring the panel structure of the data.

$$Y_{it} = \alpha + \gamma D_{it} + X_{INC,it}\beta_1 + X_{B,it}\beta_2 + X_{W,it}\beta_3 + X_{H,it}\beta_4 + \epsilon \tag{3.7}$$

Spatial lag ρW_y is implemented based on specification tests. This corresponds to expectations of the dependent variable having significant positive spatial autocorrelation, as part of the construction of the access measure. Because potential supermarket access crosses tracts, interaction between tracts is further expected. Note that this may challenge the set-up because of the strong spatial patterns in segregation.

$$Y_{it} = \alpha + \rho W_y + \gamma D_{it} + X_{INC,it}\beta_1 + X_{B,it}\beta_2 + X_{W,it}\beta_3 + X_{H,it}\beta_4 + \epsilon \tag{3.8}$$

Pooled OLS estimators are biased and inconsistent because individual tract effects are omitted and likely correlated with other regressors. An index of time and space, measured as individual tract members, is next implemented in a formal panel econometrics model.

A fixed effect DID model is as follows:

$$Y_{it} = \alpha + \gamma D_{it} + X_{INC,it}\beta_1 + X_{B,it}\beta_2 + X_{W,it}\beta_3 + X_{H,it}\beta_4 + \varrho_t + \kappa_i + \epsilon \tag{3.9}$$

where ϱ_t are period fixed effects and κ_i are individual, census tract-level fixed effects. A t-test of coefficients for this fixed effect model was performed to determine structural breaks over time. With significant change present, a time-demeaned fixed (or within) panel model was implemented. The addition of a tract and year interaction term (as in Conley and Taber (2011)) did not significantly change estimates, and because time and individual effects were specified using fixed and random model estimates, is not included in this analysis.

Finally, the variation of an individual tract across time, in both treated and un-treated groups, is made explicit in a random effects model. In a random effects model, the individual effect is characterized as random, drawn not from a population of individuals but of decisions Baltagi (1995). In this implementation it is run from two regressions; the first run from the fixed effect or "within" model above, and the second a "between" model running a regression of averages across time. In this case, it follows the same formula as (13) but time and tract-level effects are random. This model is also implemented in the plm package of R Croissant et al. (2008).

While the influence of individual, tract-level effects may seem necessary in differentiating treatment heterogeneity, their inclusion could pose challenges. A critical assumption in random effects models is that unit-specific effects and explanatory variables are uncorrelated. If this assumption is not held, it leads to inconsistent estimates in both non-spatial and spatial random panel models Mutl and Pfaffermayr (2011). A spatial Hausman test is performed to test this assumption and determine if the fixed or random model is more efficient. From a causal perspective, unit-level variations correlated with explanatory variables may violate critical assumptions underlying the experimental setting.

The experiments made explicit above take into account slightly different assump-

tions, and likewise test for sensitivity in results according to those considerations. Multiple panel econometric diagnostics are taken at each step to evaluate heteroskedasticity, normality of errors, serial correlation concerns, and spatial autocorrelation/dependence. These were implemented in R using plm and splm packages Croissant et al. (2008); Millo et al. (2012).

3.3.2 Results

Results from each experiment are presented in the following tables and figures. Maps in Figures 9 through 14 show distribution of racial and ethnic makeup for Chicago, illustrating the dynamics of demographic movement pre- and post-Recession. Black-majority tracts remained relatively stable, though white-majority tracts expand in Western regions of Chicago. Multiple areas on the west side have new, overlapping White and Hispanic majority status post-Recession.

Figure 15 shows an excess risk of foreclosure map, using the population in 2010 as a base variable. Areas with over 2 times relative risk were selected as treatment variables proxying the effects of the Great Recession. Missing tracts were not reported for 2010 tract boundaries, and therefore not included in the analysis.

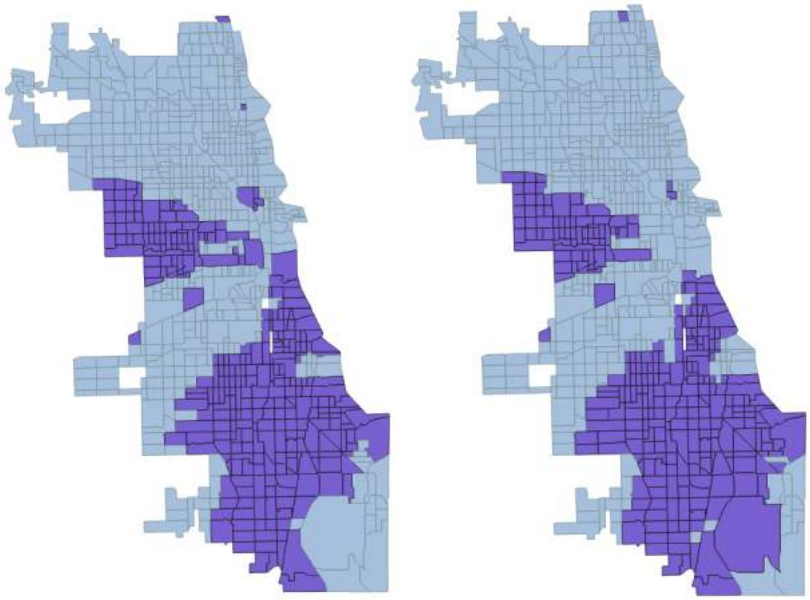


Figure 14. Black Majority tracts pre- (left) and post-(right) Recession

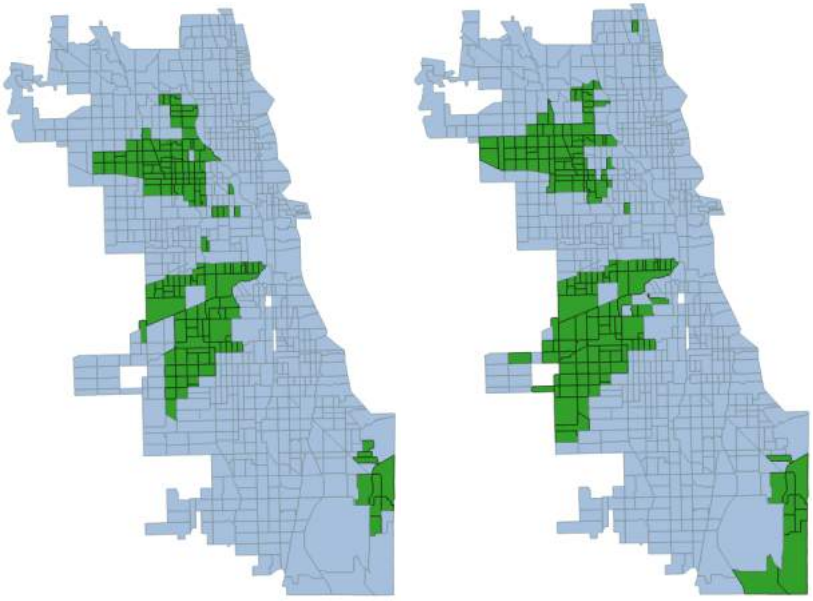


Figure 15. Hispanic Majority tracts pre- (left) and post-(right) Recession

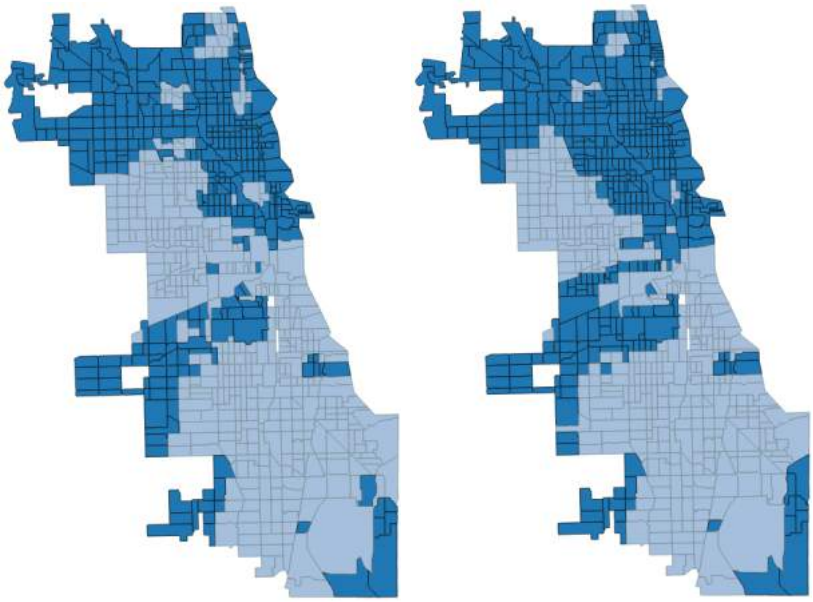


Figure 16. White Majority tracts pre- (left) and post-(right) Recession

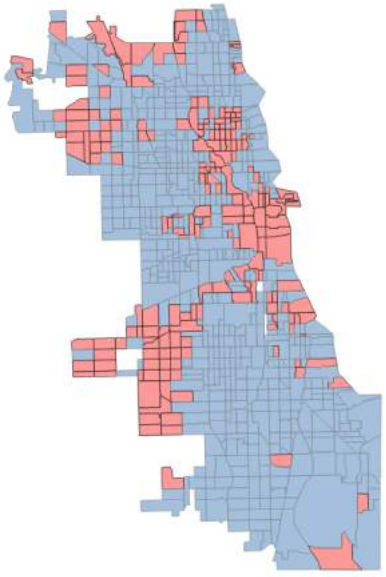


Figure 17. Tracts that gained population after the Recession. All but one tract not highlighted, shown here in light blue, lost some population.

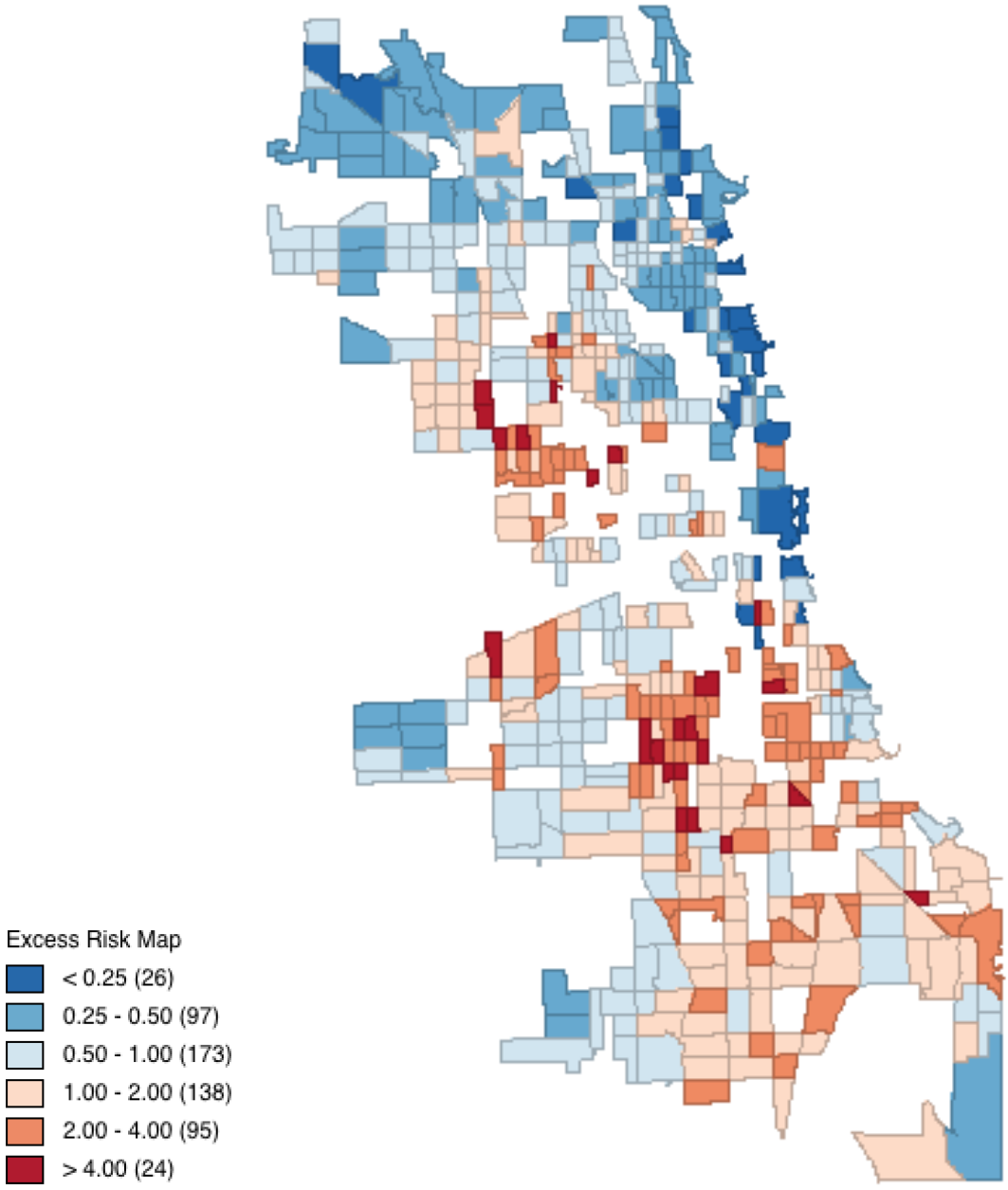


Figure 18. Excess risk of foreclosure in 2008.

3.3.2.1 Quasi-Experimental Design

3.3.2.1.1 Aggregate DID Analysis

Tracts in treatment areas (with over two times excess risk for foreclosure) had significantly worse food access than those in control groups, however there was no significant change in either direction after the Recession (see simple DID analysis results in Table 2). An OLS model group effects estimated a Recession effect of 0.018, but was not significant (see Appendix). Tracts with stable black majorities in both time periods had the highest distance to travel for supermarkets in all categories. However, black majority tracts with excess risk for foreclosure had no significant change after the Recession. Black minority groups that were not at risk, however, did have a significant change post-Recession; food access improved by about 0.05 miles. This shift moved average distance to the nearest supermarket from 0.72 miles pre-Recession to 0.67 miles post, making it the most accessible group overall. Weighted results, adjusted according to proportion of tracts that were in treatment and control groups, follow similar trends, and show that this group had a magnitude higher order of effect in improved access (in Appendix).

With this analysis, the mechanics of supermarket service area expansion and shrinkage are still not clear. Furthermore, how did tracts in non-stable demographic regions fare – did tracts that became white-majority have a different potential food access outcome than tracts that became black-majority? When accounting for demographic change, interesting patterns of inequity emerge. Only seven tracts became black-majority areas in post-Recession Chicago. Three of those (42.9%) had both worsening food access and losing population. These tracts worsened, on average, al-

Category	N tracts	Baseline	Endline	Change
Treatment (All)	200	1.12 ±0.55	1.13 ±0.63	0.01 (0.06)
Control (All)	590	0.81 ±0.45	0.78 ±0.47	-0.03 (0.03)
Difference		0.31***	0.35***	0.04
		(0.05)	(0.04)	
Treatment (B=1)	145	1.24 ±0.51	1.22 ±0.59	-0.02 (0.07)
Control (B=1)	135	1.10 ±0.47	1.13 ±0.62	0.03 (0.07)
Difference		0.14**	0.09	0.05
		(0.06)	(0.07)	
Treatment (B=0)	48	0.72 ±0.44	0.73 ±0.50	0.01 (0.10)
Control (B=0)	437	0.72 ±0.40	0.67 ±0.36	-0.05* (0.03)
Difference		0.00	0.06	0.06
		(0.06)	(0.06)	

Table 2. Aggregate DID Results using raw cost distance measures (in miles).

most a quarter mile, all of which had more than two miles cost distance to the nearest supermarket. At the same time, eighty-four census tracts became white-majority tracts, and over 76 of those lost population (90.4%). For 18% of tracts that became white and had a loss in population, food access actually improved. Only two tracts in this category had worsened access, and those only worsened by less than a tenth of a mile. Thus some supermarkets were opened in areas of population loss, but generally only in areas that were or became white majority tracts.

After net change for each group is accounted for, there is an unexplained benefit for black minority tracts in low-risk areas post-Recession. Supermarket access increased overall, even in tracts losing residents. In this aggregate DID analysis, areas at greater risk for foreclosure did not have a consistent change in access across demographic groups, though black majority areas persistently had worse access overall.

3.3.2.1.2 Parametric Analysis

A panel analysis further distills tract-level variation of potential food access and foreclosure risk, providing deeper insight than an aggregate analysis could. A pooled OLS that does not take temporal variation into account shows highly significant associations with black majority, Hispanic majority, and tracts with excess risk of foreclosure (Table 3). Areas with higher income, and Hispanic majority tracts, are more likely to have better access. Areas with excess risk of foreclosure, and black majority tracts, are more likely to have worse access. When spatially lagged food access is added, associations remain similar in direction, but with reduced magnitude.

Variable	Estimate
Income	-0.008
	(0.017)
Black Maj.	0.064
	(0.060)
Hispanic Maj.	-0.034**
	(0.013
White Maj.	0.015
	(0.014)
Foreclosure	0.069***
	(0.013)

Table 3. Change in Variables Over Time. Results from a t-test of coefficients in the fixed model without spatial effects. *p= 0.05, ** p=0.01, ***p=0.001.

A time-demeaned fixed effect model similar associations in directions. Foreclosure effect shifts from 0.303 to 0.065 in a spatial lag model; the effect of black majority tracts follows a similar directional pattern, though is no longer significant. A random effects model that privileges individual effects confirms previously validated correlations and associated relationships for black and Hispanic majority tracts, though income is no

longer significant. In all models, foreclosure and black-majority status is significant, but with a reduced effect when accounting for spatial effects. In pooled and random spatial panel models, the effect of foreclosure is slightly higher than black-majority status, though this is not consistent when not accounting for spatial effects.

Variable	Pooled	Pool, Lag	Fixed	Fixed, Lag	Random	Random, Lag
Constant	-3.221***	-0.815***	-	-	-3.562***	-1.033***
	(0.144)	(0.118)		-	(0.088)	(0.069)
Income	-0.106***	-0.062**	-0.093**	-0.041**	-0.029	-0.009
	(0.030)	(0.023)	(0.031)	(0.013)	(0.018)	(0.014)
Black Maj.	0.213***	0.055*	0.215***	0.023'	0.213***	0.075***
	(0.029)	(0.022)	(0.029)	(0.012)	(0.022)	(0.017)
Hispanic Maj.	-0.108***	-0.061***	-0.105***	-0.034***	-0.061***	-0.032*
	(0.023)	(0.018)	(0.0230)	(0.010)	(0.018)	(0.014)
White Maj.	0.008	0.023	0.012	0.002	-0.010	-0.002
	(0.026)	(0.019)	(0.025)	(0.011)	(0.015)	(0.012)
Foreclosure	0.277***	0.201***	0.303***	0.065***	0.104***	0.053***
	(0.024)	(0.019)	(0.026)	(0.004)	(0.014)	(0.009)
Spatial Lag	-	0.705***	-	0.741***	-	0.709***
	-	(0.022)	-	(0.021)	-	(0.022)

Table 4. Regression estimates for panel data analysis with tract and year interaction term to control for common trends assumption. Coefficients and standard errors are reported here, with 'p=0.10, *p= 0.05, ** p=0.01, ***p=0.001. Full results available in Appendix.

All non-spatial models (pooled, within, and random) had significant spatial lag dependence ($p < 2.2e-16$), justifying spatial panel model implementation. A t-test of coefficients in a non-spatial fixed panel model shows significant change in Hispanic majority tracts (Table 3). Foreclosures are expected to be significant because they were constructed as dummy variables for each time period. Over time, areas with a Hispanic-majority are more likely to have better food access (ie. lower cost distance).

Model	Treatment Effect	Significance
Aggregated DID	0.018	*none*
	(0.058)	
Pooled OLS	0.277	$p = 0.001$
	(0.024)	
Pooled, Lag	0.201	$p = 0.001$
	(0.019)	
Fixed Effects	0.303	$p = 0.001$
	(0.026)	
Fixed, Lag	0.065	$p = 0.001$
	(0.004)	
Random Effects	0.104	$p = 0.001$
	(0.014)	
Random, Lag	0.053	$p = 0.001$
	(0.009)	

Table 5. Summary Table: Quasi-Experimental Results Overview. Note that for the Aggregated DID analysis, the treated group did not show a change, however the Black-minority control group had a significant reduction in distance to supermarkets.

A pooled regression is thus rejected in favor of a time-demeaned within transformation. For the spatial Hausman test comparing fixed and random effects, the null hypothesis (of uncorrelated individual effects and explanatory variables) is rejected. The spatial fixed panel model is thus considered to produce the most efficient and consistent estimates.

A summary table (Table 5) shows the results treatment effect variation according to model specification. The treatment effect is significant in all tract-level panel model designs, and is sensitive to model specification with and without spatial effects. The spatial fixed panel model estimates a similar effect to a non-spatial fixed model with heteroskcedasticity-consistent (HC) standard errors (as estimated and reported in Table 2). Though in this comparison, both income and spatial lag are additionally significant in the spatial model, as compared to the aspatial model.

3.4 Discussion

3.4.1 Overview

When considering the entire time period available (from 2007 to 2014), access to healthy food seemingly improved in Chicago, based on the increase in its total number of full-service supermarkets and the overall decrease in residents' mean distance to their nearest supermarket. However, these citywide summary statistics obscure wide disparities in healthy food access between persistently high access areas throughout Chicago's north side and persistently low or volatile food access areas across much of its west and south sides. Given the high degree of residential segregation in Chicago, the local food inequity disproportionately burdens some racial minorities. Most notably, African Americans make up approximately one third of Chicago's population, but almost 80% of the residents of persistently low or volatile food access areas, which are also home to overwhelmingly high rates of family and childhood poverty. Furthermore, food access for black and socioeconomically disadvantages residents worsened before slightly improving, with more census tract outliers with extremely low access in 2011 and 2014. While this ESDA finding suggested perhaps worsening inequity among populations, further analysis of the Recession effect imply a more complex underlying process.

Chicago neighborhoods with more foreclosure experienced a small but significant worsening in food accessibility after the Great Recession. This is the case even after accounting for variations in income, group effects, and patterns of racial segregation. The Recession effect is an estimated 0.065 in increased cost distance to supermarkets (with an adjusted score) when using a spatial fixed panel model that accounts for

group, temporal, and spatial effects. Tract-level individual effects are correlated with other explanatory variables in this dataset, which may confuse relationships if both resulting correlation and spatial patterns are not accounted for. The spatial fixed panel model provides the best estimate for measurement because it does not assume random individual effects, and likewise takes into account the highly spatially autocorrelated behavior of food accessibility.

3.4.2 Spatial effects

When these specifications are made, Black majority areas no longer predict low food accessibility. Instead, areas of lower income and more foreclosures (that in turn had higher unemployment and predatory, high risk loans) worsen supermarket access in a statistically significant way. Making space explicit is necessary for not only revealing this phenomenon, but for meaningful interpretation. The spatial pattern of food access is similar to the spatial pattern of foreclosure and Black segregation, as both foreclosure effect and racial effect get smaller when space is accounted for. Because the racial effect goes away when accounting for correlated explanatory variables and space, it is clear that it is not racial makeup but socioeconomic factors that ultimately drive this disparity. While there is a strong spatial cluster of segregated groups in Chicago, therefore, they result from an exogenous process. This finding leads to new insight that would have been missed in a strictly exploratory setting, as similar spatial patterns can emerge from different phenomenon. This finding also underscores the need to better understand the role segregation may have in perpetuating environments that contribute to health disparities Moore and Diez Roux (2006); Blanchard and Lyson (2002); Kaufman (1999); Morland and Filomena (2007); Morland et al. (2002);

Powell et al. (2006); Zenk and Powell (2008); Landrine and Corral (2009); CDPH (2012).

Interestingly, Hispanic majority tracts are consistently correlated with better food access. The inclusion of independent (and often locally owned) supermarkets in the longitudinal dataset, prevalent in several of these Hispanic neighborhoods, may have provided a more complete picture of access missed in previous research that tends to use chain stores. The pattern could be resulting from expanding white-majority tracts across many of these communities, often overlapping Hispanic-majority areas. The aggregate analysis showed increase in food access for some tracts that became white-majority areas, though this finding was not significant in a more explicit panel setting. This spatially heterogenous pattern could alternatively suggest that the underlying supermarket process differs here than other areas because of some additional, exogenous impacts. Because many of these areas have high proportions of immigrants as compared to the rest of the city, this increased healthy food availability may be associated with the immigrant health paradox. This well-documented phenomenon is an association between recent immigrant and Hispanic communities and better health outcomes, despite low socioeconomic status Marks et al. (2014). Future work should further distill these associations.

A surprising finding was that the addition of several markets in White-majority areas was not significant in shifting foodscape dynamics. This may seem counterintuitive, without a spatial perspective; how could the addition of so many supermarkets in socioeconomically privileged, white-majority areas *not* dramatically change the foodscape? The location of supermarkets, not the net change of total stores, is what drives spatial equity of resources. New markets were added in already high clusters of market access, as confirmed by the exploratory analysis. Consider the necessity of

making spatial effects explicit, too, in a quasi-experimental setting. A global analysis of Recession effects showed a significant change for black-minority regions not at risk of foreclosure, with better food access post-Recession; however, a tract-level analysis removes this association. Spatial Effects are essential to distilling the complexity of a changing foodscape. Even if there is a net positive change in supermarket access, with more additions over time, the location of those gained and lost is what shifts the inequity of global access. Without accounting for spatial distribution of supermarket access, model outcomes magnify segregation effects and confuse treatment effects. Spatial effects present can also complicate results if relationships are not specified correctly.

3.4.3 Methodological Innovations

A mixture of methodological innovations were implemented in this study, including the construction of a fine-resolution access measure, use of a validated and longitudinal dataset for food access in major urban environment, research design implemented a sensitivity analysis, and the extension of a counterfactual framework to account for spatial effects.

Traditional methods to quantify urban food access rely on techniques that can introduce substantial error Langford and Higgs (2006); Apparicio et al. (2008). For example, 1-mile distance thresholds that are commonly used in standard buffer analyses may misclassify food access when census tract sizes or built environment landscapes vary widely Smoyer-Tomic et al. (2006). Container techniques that quantify total supermarkets per unit of urban space without considering adjacent and nearby areas incorporate arbitrary administrative boundaries, such as census tract borders, that do

not influence consumers Sadler et al. (2011). Insufficiently validated supermarket and road network data and lack of accounting for variation in population density present additional sources of error Jaskiewicz (2010).

Methodological strengths of this study include field validation of supermarket data, exclusion of non-residential and industrial land use areas from the analyses, and robust spatial analyses that map the closest supermarkets to all residential street addresses, regardless of whether they lie beyond an arbitrary census tract border. By excluding corner stores and local bodegas that are found disproportionately in socioeconomically disadvantaged neighborhoods and primarily market convenience items, sugary and alcoholic beverages, and highly-processed snack foods, this methodology provides a starker portrayal of food inequity than if these stores were considered healthy food sources. Furthemore, unlike most prior studies that evaluated food access at single time points, longitudinal trends are examined and the impact of a major economic event on food access was estimated.

As shown in Summary Table 11, results were sensitive to research design setup. After reviewing the complexity of results, this sensitivity study conclusively shows that inequity in the foodscape is driven by persistently segregated geographies, despite milder influence due to demographic change and the Recession. Using a combination of ESDA, non-experimental, and quasi-experimental analysis was likewise essential to understanding how food access was changing. ESDA provided a robust summary of key relationships and trends that persisted, and emerged, over the time period. Quasi-experimental analyses were sensitive to research design framing, and provided unique insights with and without spatial effects.

In a traditional Rubin counterfactual framework model, only the aggregated DID analysis would meet strict assumptions necessitating no interaction between units.

Because food access measures demands interaction between units by construction, tract-level variation would violate the SUTVA assumption. A challenge of this aggregated implementation is that it may miss variation at a scale finer than the level of analysis. Such analysis not only underestimates the effect of the Recession on food access, but also misses important relationships in other explanatory variables. By making space explicit and testable in a counterfactual framework, the SUTVA assumption can be relaxed. This also permits analysis of treatment heterogeneity at a finer-resolution scale. Spatial panel models that accounted for time and individual effects estimated a treatment effect of similar direction and magnitude to the aggregated measure. As such, spatial effects made explicit at a finer-scale resolution may follow average trends estimated globally, but with greater detail and consistency. (Whereas aspatial models at a finer resolution miss both spatial patterns and violate core assumptions.) Further work is needed to test the implementation of spatial effects in counterfactual frameworks.

3.4.4 Study Limitations

In addition to its strengths, this study has several limitations. By limiting the analyses to Chicago, it's not possible to evaluate if similar trends are evident in other US cities. By calculating food access as the distance from residential areas to the nearest supermarket, the approach failed to account for the effects of commuting behaviors, for example, residents shopping near their places of employment rather than their homes. Although unlikely to be a common occurrence, the approach to validating supermarket status at different discrete time points would have missed stores that opened and closed between the data collection points. The study further considers

'potential' food access rather than actual access, which would necessitate consumer behavior and market data. Future studies could use mixed measures that incorporate both mapping of potential access and qualitative and quantitative measurements of food buying behavior to provide an even richer assessment of food access landscapes. Further research is needed to link food access data with regional health administration and claims data to investigate whether residence in persistently low or worsening food access areas is associated with worsening health outcomes that are plausibly related to diet, and specifically how highly segregated populations are further impacted by these disparities Landrine and Corral (2009).

Additional limitations underlie the analysis to estimate the effect of the Recession on food access. While the foreclosure data used was the best available, it was only available for one year, and had missing data for several tracts (likely corresponding to tracts with low populations). While converting the continuous score to an excess risk dummy variable reduces some of the resulting error, a complete and matching set of continuous data for both years would have provided the most complete approximation. Lack of high quality census data available at precise cross-sectional years, without mixing data averages, adds to the challenge. Finally, there was was a lack of spatial variation in the dependent variable, even after log transformation, that challenged methodology. However, even despite all of these challenges, underlying consistency in model estimates for strong trends remain promising.

3.4.5 Conclusion

This study remains innovative in distilling the complexity of foodscape inequity over time in Chicago, quantifying the effect of the Recession on local food access, and

pushing food access research past descriptive summary. The findings conclusively recommend a shift in focus from refining measures of access to the underlying processes that drive black segregation (which in turn drive low food access). At a minimum, tract-level food access policies must incorporate values of nearby tracts to avoid misguided attention to "food desert islands." Instead, policies should be geared towards shifting resources towards segregated neighborhoods, without forcing a demographic change (ie. gentrification). Spatially lagged food access may serve as a proxy for this effect.

This study underscores the need for additional, rigorous research in not only extending foodscape studies to causal analysis frameworks, but also the need for making spatial effects explicit. Inference interpretation is sensitive to both research design framing and underlying processes that drive geographically distributed relationships. For highly spatial phenomenon like segregation and foreclosure, making space explicit may reduce the magnification of certain results. This shifts the interpretation, then, from singular variables of analysis to the actual trend driving the wider geographic pattern (that thus influence multiple variables). By making these assumptions explicit in a single-level econometric analysis, these components can be distilled at the scale of interest. New innovations are needed to refine methods for problems with high spatial autocorrelation and temporal correlation.

Chapter 4

TOWARDS A SPATIAL DATA SCIENCE INFRASTRUCTURE IN PUBLIC HEALTH INFORMATICS

Abstract

Data integration of disparate, heterogeneous data sources is necessary for advancing policy and planning that improve population health. The need for developing a new systems framework capable of integrating different types of data to promote equitable interventions follows shifts towards a ecosocial view of health; moved towards open data practices and increased availability of new types of data; and calls to move from siloed to shared datascapes that promote collaboration. To address these challenges, I propose a new Spatial Data Science (SDS) Infrastructure for integrating, accessing, and managing spatial and non-spatial data in dynamic, open environments for public health systems research and decision-making. While several studies have incorporated components of a spatial systems infrastructure, in this essay I argue that a more complete, formal model is required to effectively address data integration and exploration in health informatics. A SDS infrastructure is thus a spatial infrastructure, but additionally must be dynamic, reproducible, adaptive, and participatory. Space is made explicit as a place of integration for heterogeneous data that includes population health outcomes, socio-economic variables, built environment indicators, and other social determinants of health. I demonstrate data integration and client-facing application components of an SDS infrastructure prototype in a Chicago case study, developed in collaboration with health department officials and community organizations. Health systems infrastructures can further

support community health improvement frameworks by facilitating shared data and decision support implementations across health partners.

4.1 Introduction

4.1.1 Justification and Context

Policy-driven disciplines require not only effective theoretical frameworks and methodology, but also technologic infrastructures that allow for intervention evaluation and development. New types of data are transforming research and decision-making, though organizations struggle with fully using, accessing, and sharing that data. A spatial perspective serves to resolve not only integration challenges, but also provides methods of storing, accessing, analyzing, and using heterogenous data for research and decision-making.

The need for developing new frameworks capable of integrating different types of data to promote equitable interventions is the results of multiple developments. Within a health framework, these can be characterized as (a) a shift in policy priorities to better understand, model, and quantify place-based relationships between people and local environments; (b) a shift towards open data practices and increased availability of new types of data, characterized as "Big" across multiple dimensions, in multiple health sectors; and (c) a shift in priorities from siloed to shared and integrated data platforms, and associated challenges in data systems infrastructures that results.

To address these challenges, I propose a new Spatial Data Science (SDS) Infrastructure for integrating, accessing, and managing spatial and non-spatial data in dynamic, open environments for public health systems research and decision-making.

While several studies have incorporated components of a spatial systems infrastructure, in this essay I argue that a more complete, formal model is required to effectively address multivariate data integration and exploration in health informatics. A SDS infrastructure is thus a spatial infrastructure, but additionally must be dynamic, reproducible, adaptive, and participatory.

4.1.1.1 Call for Better Understanding of Place-Based Relationships

Place-based approaches have taken a central role in work to increase urban and health equity (Corak, 2013; Amaro, 2014). Increasing levels of micro-level segregation between cities and neighborhoods, incorporating complex trends of racial- ethnic and rising class segregation, further complicates the spatial organization of urban landscapes (Massey et al., 2009). In the new eco-social view of health, understandings of health outcomes necessitate insight into the complex relationships between populations and the environments they inhabit (Levins and Lopez, 1999; Krieger, 2003). Multiple components impact the health of individuals and populations, building from genetic predispositions to powerful social determinants of health, built environment factors, natural environment access, local public health department and health worker initiatives, cultural factors, economically supported or preferred activities, and more. This follows the increasing focus on the importance of social determinants of health, defined as "conditions in the environments in which people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality-of-life outcomes and risks" (People et al., 2000). In a parallel trend, a political ecology approach that considers disease and health as a part of socioecological system has become dominant in contemporary health and medical geography (Mayer, 1996; King,

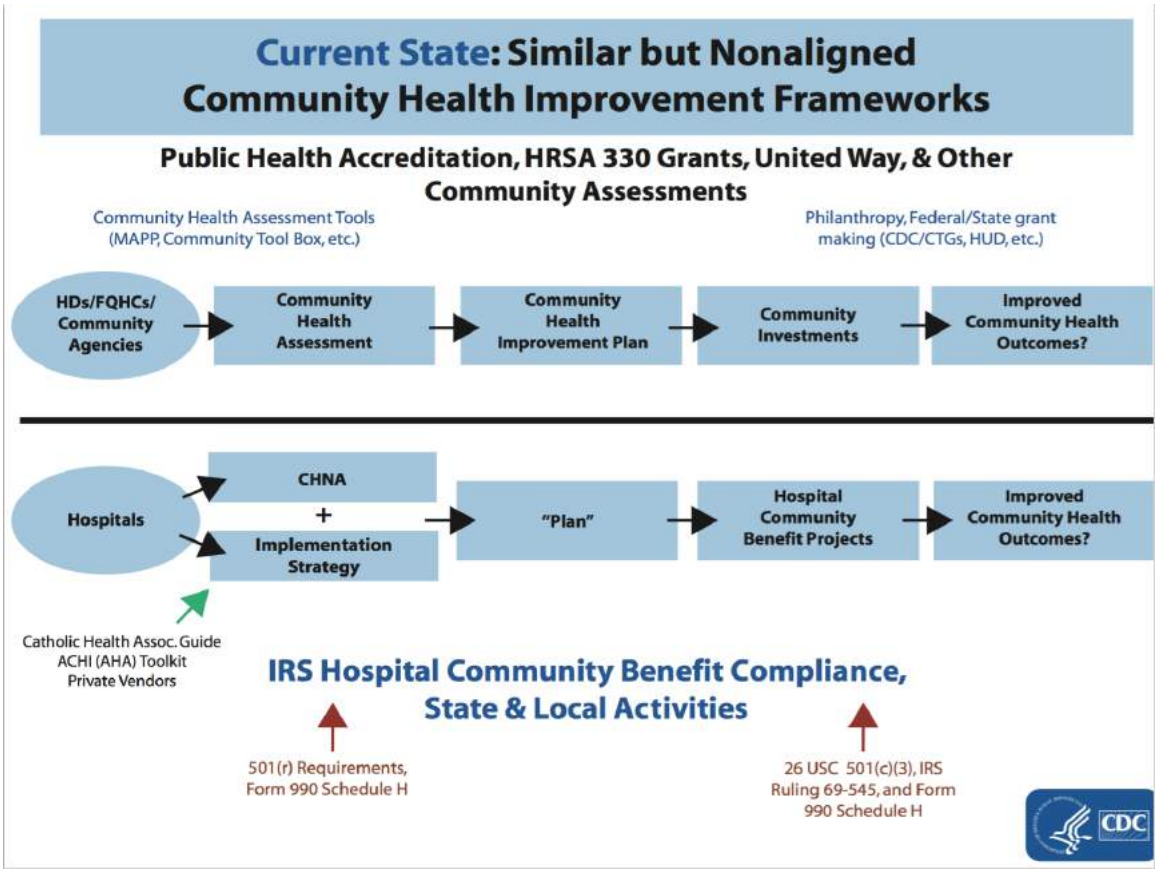


Figure 19. Current State of Community Health Improvement Framework. CDC

2010; Krieger, 2011; Chitewere et al., 2017). Social, economic, and political factors interact to shape local structures that impact individual and group health outcomes. Dramatic policy changes in the United States health system in the past decade reflect these developments, and include a move to address increasing health spending, disparities in outcomes, and new legal framework requiring tax-exempt hospitals to conduct formal community health assessments (Rosenbaum, 2016). These policy moves further prioritize population and neighborhood-level health outcomes, necessitating both a more complex understanding of health environments and technological ability to evaluate effectively.

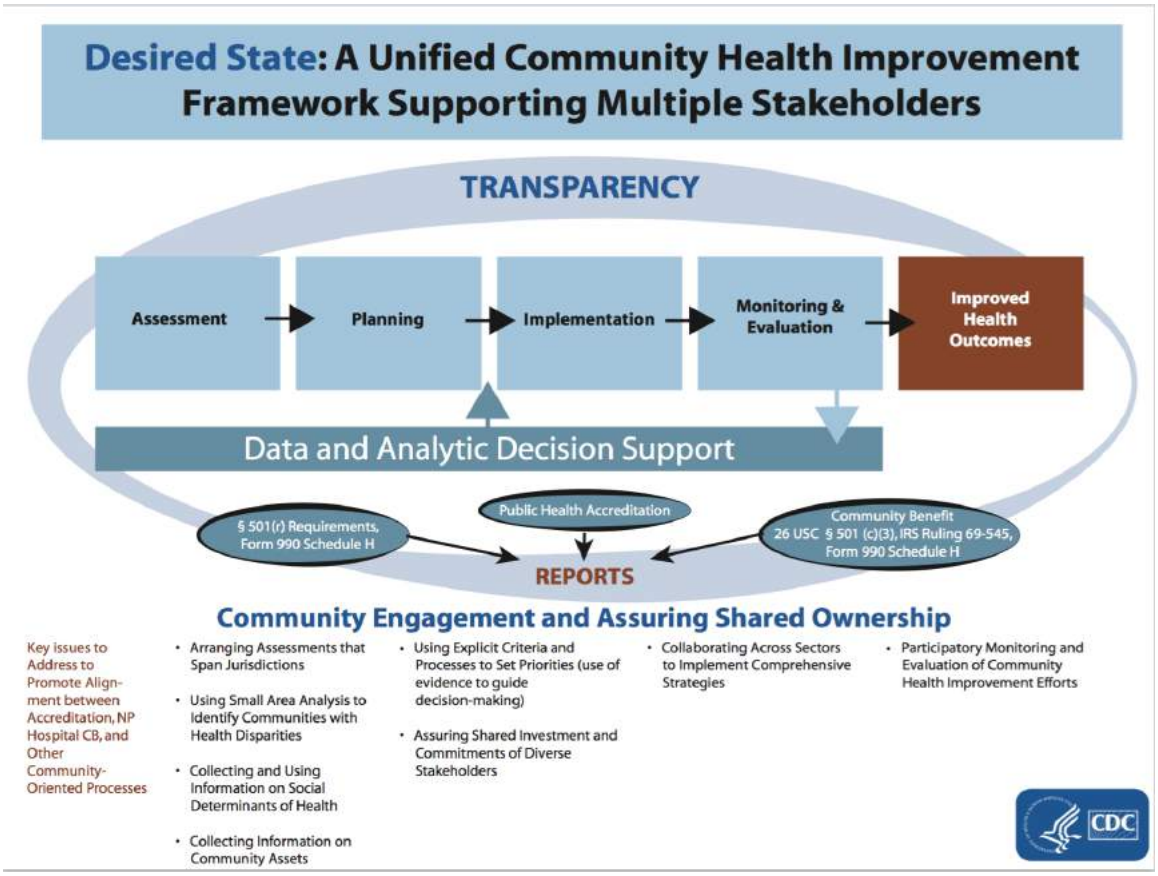


Figure 20. Desired State of Community Health Improvement Framework. CDC

While stakeholders develop health interventions geared towards improved health outcomes in an eco-social perspective, existing frameworks of community health remain siloed, rather than a desired state of shared ownership and collaboration (CDC, 2015). The Center for Disease Control and Prevention (CDC) demonstrates the difference between the current and desired frameworks of community health improvement (figures 1 and 2 above). Calls for increasingly data and analytic decision supports have been made to facilitate assessment, monitoring, and intervention evaluation. The great promise and challenge of the new era of data in healthcare and public health informatics is thus integrating and managing traditional data with new types of data, necessitating new types of cross-collaborative technological infrastructures.

4.1.1.2 New Era of Bigger Data Access and Availability

Progress in establishing cross-sector partnerships and open source data movements has, furthermore, made data more accessible in previously unimaginable ways. The emergence of data portals to allow easy access, via download or application program interface (API), emerged in the past decade in a move towards open data across multiple sectors (Mayer-Schönberger and Cukier, 2013). By making available data on governmental portals, researchers and developers would be able to expand the knowledge base and gain new insights. Open data platforms such as Socrata could be accessed by interested citizens, researchers, or application developers. The importance of collaborations and partnerships in sharing such data, previously siloed, is not to be underestimated, as it was unprecedented. While not all levels of government shared their data in similar proportions, cities that shared more found interesting projects. In the City of Chicago, for example, a coalition of public health officials and computer scientists combined restaurant inspections with social media data to prioritize inspections according to voiced complaints on Twitter. This prioritized system increased the number of uncovered violations at restaurants (as compared to the previous system), allowing for more effective use of resources. Such examples are one of many exercises emerging when multiple types of data are incorporated by multi-disciplinary teams in new and unexpected ways. While governmental and multi-sector data becomes increasingly available on data portals, even newer, more robust repository frameworks now expand points of accessibility and data sharing.

4.1.1.3 Technological Infrastructure Challenges

In multidisciplinary fields like health and the social sciences, integrating multiple types of data serves as a prerequisite to evaluate complex relationships of populations, neighborhoods, and the built environment. However, providing an overview of these relationships, as abstracted in a decision support or data infrastructure environment, serves as a burden to many systems. While surveillance of diseases and injuries is a routine component of public health in the United States, incorporation of related built and public health environment features at a comprehensive neighborhood-level scale is rare. Measurement tools for aspects of the environment have been developed and tested in research settings (i.e. parks, workplaces, walkability, access to healthy foods), but have not been used routinely to gather data to inform decision support systems (Dannenberg and Wendel, 2011). Furthermore, in a new environment of Big and bigger types of data, updating systems with new types or versions of data has become increasingly important. Within new data environments of Smart City dashboards or platforms, integrating multiple datasets must be paired with user-friendly methods of access and exploration. Data may be sourced from multiple users with various forms of expertise, some known by the end users, and others pulled from cloud platforms (sourced from organizations previously unknown to end users). In an environment of multidisciplinary stakeholders, data and analytics could be better integrated in a flexible technological framework to allow for continuous assessment of population-level programs and outcomes.

4.1.2 Introducing a Spatial Data Science Infrastructure

To preserve and investigate complicated relationships between entities, spatially-enabled "bigger" data infrastructures are needed, though not yet fully implemented, in social science and health for research and decision-making systems. A heterogeneous, flexible framework centered around place can provide the means of integration and dynamic infrastructure crucial to decision-making. A technological framework that supports streaming data updates, dynamic integration, and ultimately on-the-fly-analysis could transform how research and science are implemented. Inverse infrastructures (ie. user-driven, decentralized infrastructures) would furthermore support participatory methods of contributing and editing data, a need central in increasingly distributed systems. To address these challenges, I argue that a new Spatial Data Science (SDS) Infrastructure is required for integrating, accessing, and managing spatial and non-spatial data in dynamic, open environments for research and decision-making. A SDS infrastructure is thus a spatial infrastructure, but additionally must be dynamic, reproducible, adaptive, and participatory.

In Section 2, I review existing challenges to data integration in the health sciences, positioned within a contemporary "Big" and "bigger" datascape. To resolve these issues, I argue that space should serve as a place of integration for heterogenous data systems in public health informatics. In Section 3, I propose a new framework that integrates spatial, dynamic, reproducible, adaptive, and user-centric characteristics, effectively linking rich and varied traditions of existing data infrastructure frameworks in a new, formalized way. To illustrate this concept, I implement a SDS infrastructure for public health data management in Section 4. Section 5 concludes, reviewing core characteristics of the infrastructure, remaining challenges, and future work.

4.2 Facing Data Challenges with a Spatial Perspective

4.2.1 Data Integration Challenges in the Health Sciences

Integrating different types of data from different sources responds to multiple calls made in healthcare, including goals to better understand relationships between neighborhoods and built environments on health outcomes (Pastor and Morello-Frosch, 2014); better identify strategies of measuring health disparities (Bilheimer and Klein, 2010); and reducing waste and inefficiency in clinical operations, research and development, public health, evidence-based medicine, genomic analytics, pre-adjudication fraud analysis, device/remote monitoring, and patient profile analytics (Raghupathi and Raghupathi, 2014).

Challenges in integrating different kinds of data for varying purposes in healthcare have been well documented, and include the difficulties of working with different types of information, different ways of storing data, sharing data across organizations, managing data updates, visualization, data access, varying budgets for technology across organizations, domain and technological knowledge mismatches across organizations, data integration, data standars, managing large amounts of data, meeting privacy and security standards, fiscal limitations interoperability, common terminology, and data confidentiality, sample size, missing data, and measurements errors challenge strategies of measuring health disparities (Johnson et al., 2014; Richardson et al., 2013; Shah et al., 2014; Bilheimer and Klein, 2010).

These challenges can be categorized according to characteristics of "Big" data. Big data has been characterized as data with at least one of the great "V"s : volume, variety, velocity, and veracity (Laney, 2001), though has not been formalized with a

rigorous definition (Mayer-Schönberger and Cukier, 2013). The term also incorporates new types of data previously unavailable, like the "smart cities" movement and the sensors that drive it; the "open data" movement with increasing access to governmental and administrative data; and, the "volunteered geography movement" incorporating crowd-sourcing techniques to curate relevant unstructured geographic data (Anselin, 2015). Much of this data has spatial components, like associated addresses, zip codes, states, regions, and so forth, though may only be initially available in a non-spatial data format (like a CSV).

There have been multiple public health decision support systems and visualization tools published and implemented that seek to connect complex, varied data into a single database for further analysis. Examples of public health decision support systems incorporating GIS include the Community Health Needs Assessment Toolkit (Community Commons) developed by the non-profit organization IP3 (the Institute for People, Place and Possibility) (2014), the North Carolina Health Data Explorer developed by East Carolina University (2012), the Pennsylvania Cancer Atlas developed by Penn State (MacEachren et al., 2008), the Dartmouth Health Atlas (Wennberg, 1996), EpiVue developed by Washington State University (Yi et al., 2008; Fuller, 2011), and the Common Data Warehouse Project by Florida State University (Berndt et al., 2003).

Most of these decision support tools rely on GIS visualization of public health summary data, often aggregated at no finer resolution than zip code, across familiar base maps or boundaries. Charts, graphs, and descriptive text accompany maps to provide enriched information. The finest resolution tool available was the Community Commons CHNA Toolkit (Catlin et al., 2014), which allowed for customized reports of more detailed information, though not all data was available at the finest resolution

(ie. block level). Some tools allow for the user to input their own data as a layer (like EpiVue). In many cases, technology used at the time of implementation became outdated, improved, or otherwise changed only a few years later.

Conceptual models of incorporating big data and big data analytics in healthcare have emerged, though the field is still in early stages, and generally approached from a non-spatial perspective (Raghupathi and Raghupathi, 2014). Yet there remains an increasingly urgent need to create "distributed, interoperable spatial data infrastructures to integrate health research data across and within disparate health research programs," as powerful means for generating hypotheses, detecting spatial patterns, and responding to health threats (Richardson et al., 2013). In this essay, I argue that formalizing space as an organizing principle for data systems not only resolves several integration issues, but can also be further extended when made open and user-oriented.

4.2.2 The Need for Spatial Perspective as a Place of Integration

Following calls to better understand how relationships between neighborhoods and built environments impact health outcomes (Pastor and Morello-Frosch, 2014), the need for integrating multiple types of data from different sources becomes central. While integrating data according to spatial relationships is both prevalent and central to Earth and environmental sciences, it has not fully formalized in health. Indeed, the article boasting the "spatial turn in health science" by Richardson et al. was only published in 2013. Research incorporating spatial concepts in health informatics tends to simplify elements to geocoded locations or overlay comparisons, rather than underlying structures that make space explicit. Indeed, integrating spatial techniques

as a means of data management and integration has not been discussed explicitly in health literature (Bilheimer and Klein, 2010). Strategies to resolve these new data challenges in population health domains are evolving and varied.

In this essay, I argue that health informatics can be transformed by using space as a place of integration for heterogeneous data that includes population health outcomes, socio-economic variables, built environment indicators, and other social determinants of health. Representations of space offers integrating principles from a conceptual and technological perspective; how space is made explicit in a systems infrastructure also affects how the infrastructure is implemented.

Space, and in some cases time, serve as a key between data. For example, zip code areas will have multiple types of data associated: census demographic, socioeconomic characteristics, survey data, population statistics, corresponding tiles of satellite data that could show both land change and environmental organizations within, environmental sensors, civilian data from cell phones and personal devices, movement data as persons and vehicles traverse in and out of the area, and built environment components, from locations of sidewalks to characteristics of parks and buildings within and nearby the zip code. Most of this data can be represented at different scales, from a macro perspective (regions, counties, states) to meso-scale (neighborhoods, buildings, streets), to micro scale (individual or family characteristics). While each dataset can be investigated on its own, especially at an individual scale, insight is often still desired an appropriately aggregate population-level scale. Furthermore, insight can be desired to explore the dynamics of relationships between characteristics. Often group effects are different from individual ones, as is well documented and explored in complex systems research (Mitchell and Newman, 2001; Wolfram, 1985; Smith and Conrey, 2007b). This phenomenon is also captured in multiple other concepts, like the

ecological fallacy (Robinson, 2009). To distill these concepts, identifying confounding variables and analyzing hypothesized relationships and investigating the actual scale of the phenomenon manifestation is necessary. To accomplish these tasks, combined data is a prerequisite.

In social science relevant fields like policy and public health, getting the data necessary for more effective insight and decision-making remains a core challenge. In the move towards evidence-based decision-making, better and more comprehensive data is needed. This is increasingly important as calls for data-intensive quasi-experimental research design approach public health, for example, or as data science collaborations with health officials drive applied mechanisms for policy development.

4.3 Components of a Spatial Data Science Infrastructure

I argue that in a new spatial systems architecture for health informatics, space must be made explicit as an underlying principle of integration, organization, and analysis. It must be relevant to emerging technological needs that require work with heterogeneous, bigger, and distributed data sources. In this section I review multiple models of infrastructures that are relevant to this concept including basic spatial database abilities, an open science framework, service oriented architecture, complex adaptive systems, and user-driven abilities. I advance a new categorization, building on these traditions, to characterize a modern Spatial Data Science Infrastructure that is thus Spatial, Reproducible, Dynamic, Adaptive, and User-Centric. As such, the infrastructure must also support basic exploratory capabilities and consider the scale(s) of analysis.

4.3.1 Basic Spatial Infrastructures

In its most simplified form, enabling a spatial infrastructure begins with concepts of a spatial database and spatial thinking. A spatial database must, at a minimum, be (a) a database system, (b) offer spatial data types in data model and query language, and (c) support spatial data types in implementation, like spatial indexing and algorithms for spatial joins (Güting, 1994). A spatial database is thus a database extended, enabled, and optimized for spatial data. Spatial data incorporates information and shapes of geographic features and relationships between them, generally stored in a standardized topology (ie. point, line, polygon, multipolygon, etc). A data is not considered spatial until it has been enabled as such. For example, a table with latitude and longitude fields is not "spatial" until a vertex has been generated for each row, perhaps generated from the recorded latitude and longitude, and additional spatial metadata and geometry recorded for that dataset. Because a spatial database is also a database, it can incorporate non-spatial data (ie. a table with attributes) as well as spatial data (ie. projected coordinates that include attributes of their locations). In other words, a spatial database is flexible enough to accommodate all types of data, enabling geometries when appropriate. Attributes and fields between datasets can be linked by attribute or by location in a spatial database. Data can likewise be searched by attribute or location. Spatial database systems vary in their capabilities, with more developed systems like POSTGIS able to accommodate a multitude of complex spatial (and non-spatial) operations. While querying spatial views allows for quick and effective visualization abilities, a spatial database primarily serves as a storage tool, analysis tool, and organization tool (Obe and Hsu, 2015). By allowing spatial

relationships to drive storage, analysis, and organization, the spatial database serves as a technological manifestation of "thinking spatially."

A Spatial Data Infrastructure (SDI) builds from these concepts, and at a minimum serves as a framework for spatial data, metadata, users, and tools that work together to use spatial data effectively. Formally, though simplistically, a SDI connects people and data with appropriate standards, policy, and an access network (Rajabifard and Williamson, 2001). SDI has no universally accepted, standardized definition because spatial data infrastructure is a multi-faceted concept of different perspectives (Grus et al., 2007). Developments and directions has been documented in a thriving and rich literature (see reviews in Borba et al. (2015); Mirto et al. (2016); ?? (Gru); Cooper et al. (2011); Coetzee and Wolff-Piggott (2015); Hendriks et al. (2012); Mansourian et al. (2015)). A more thorough description of modern SDI will follow (see Section 3.5).

While SDI works effectively and continues to mature for spatial data integration, organization, visualization, and analysis, it can also serve hybrid fields because of its abilities to link non-spatial data (or rather, spatially-enable "non-spatial" data when necessary), as will be implemented in the case study (see Section 4). When considered a framework connecting distributed systems and services, a SDI can also be flexible enough to be reproducible, dynamic, adaptive, and collaborative, as will be discussed in following sections. When able to accommodate each of these characteristics, a Spatial Data Science infrastructure is possible. A Spatial Data Science infrastructure must thus be flexible and incorporate principles from multiple fields and perspectives. It thrives as a collection of technologies, collaborations, and conceptual framing that serves to unify different types of data for new and old types of analysis and insight.

4.3.2 Open Science Data Frameworks

As Big and "bigger" data becomes more available, a scientific perspective demands that it be held to the same standards of traditional, "small" data (Shah et al., 2015). A new open science data framework for research and decision-making must thus allow for citable, reusable data. Following Shah et al. (2015), such a data framework should at minimum provide (1) extensible storage options and APIs for access, (2) allow users to subset the data with persistent links and author attributions, and (3) provide data curation tools to allow data and metadata to be updated. From a wider perspective, this can be positioned within the emerging paradigm of explicitly open science, or one linked to open data, open modeling, open software, open collaboration, and open publication (Rey, 2014). Rey (2014) defines these pillars of an open science within the applications of regional science, a multidisciplinary field that considers spatial dimensions of social science and human-environment dynamics, thus making an extension to public health informatics even more feasible.

While data portals allow for more dynamic data access via API, data curation at the point of access is still desired to produce a more transparent and replicable data framework environment. The DataVerse project, led by Harvard Data Science teams, serves as example software with a flexible architecture that allows for storage, control, sharing, versioning, and data citation that can allow for data and multiple levels of scale and privacy be shared across institutions. Data can be searched according to research project, topic, or keyword, and then downloaded or accessed via API according to security level. Spatial data can be uncovered by keyword search query, or found within curated datasets, like the "Data and Code for Spatial Analysis for the Social Sciences Dataverse" (King, 2007). Such powerful repositories are built on more

flexible data architectures allowing for more heterogeneous, multivariate data formats, with some data visualization capabilities available. Plenar.io is another example, serving as a spatio-temporal open data repository and place of data visualization and exploration (Catlett et al., 2014). An existing dataset available on the platform can be subset according to time and/or spatial query within the user interface, and then extracted via API. The technology is built using a spatial data architecture, querying a simple but powerful PostGresSQL/PostGIS data warehouse, to make data extraction more effective and accessible to non-spatial-data developers or researchers.

A scalable, flexible infrastructure could also allow for increasingly automated and open science, allowing for greater transparency and validation of results (Rey, 2014; Soranno et al., 2015; Shah et al., 2015). However, leveraging the spatial data capabilities remains to be fully implemented in existing infrastructures common in health, social science, and public sector platforms. Searching by location, rather than keyword, is technologically possible but can be computationally intensive (depending on architectures implemented), and generally not available on major data portals and national data frameworks. Carto, an open source spatial database and geovisuzalization tool, serves as a notable exception and private sector leader in scaling spatial data infrastructure (Zastrow, 2015). Furthermore, the increased availability of datasets has not reduced the challenges of data integration, management, and interoperability, but rather exposed them in fields eager to incorporate heterogeneous data for new analysis and insight.

Open-Source Software and Open Specifications allow for the development of these solutions with low-cost, simplicity, compatibility, and interoperability (Anderson and Moreno-Sanchez, 2003). The emergence of PostGresSQL-PostGIS, for example, had been highlighted as a robust and scalable (object relational) database management

system (Vitolo et al., 2015). Open-Source software is also considered an opportunity for public health institutions with low resources to access modern data analysis and visualization tools for minimal cost (Yi et al., 2008). While the Google Map Engine and Google Fusion Tables are not Open-Source, their low cost is also relevant for spatial data science systems in appropriate settings (Gonzalez et al., 2010; Hu and Dai, 2013).

4.3.3 Service-Oriented and Grid Architecture

Siloed system infrastructures of manually downloaded and manually updated data poses many challenges to new demands on bigger data infrastructures. Data must be updated on a more regular basis because of increasingly available datasets. The volume of new types of data can be too massive for downloading onto single desktops, and/or the processing and computational analysis of the data can prove too strenuous for traditional systems. To address new challenges, integrating other types of data may be identified as an optimal solution, however multiple integration challenges at an infrastructure, tool, and platform level increase the struggle for siloed systems. Different organizations possess different expertise, and may not be familiar with certain types of data that are new to them. And even when individual datasets are not large or qualify as "big" volume, the integration of dozens of additional datasets quickly accumulate to larger masses.

Moving from siloed systems to distributed networks necessitates new types of architectures. Service-oriented architectures (SOA) and grid architecture (or "cyber-infrastructure") are technology agnostic and have been successfully used to integrate data across distributed, interoperable infrastructures. SOA is a set of components

that can be invoked, generally as communication protocols over a network, and whose interface descriptions can be published and discovered (Consortium, Consortium). This can incorporate the polices, practices, and frameworks that enable such application functionality, including the following principles: technology neutral, standardized, consumable, reusable, abstracted, published, formal, and relevant (Sprott and Wilkes, 2004). Consuming data through API services within a data infrastructure framework serves as an example use of SOA. SOA is increasingly used to access data available as web services, serving as a standard in much web development. However, it is underutilized in multiple fields including public health and decision-making, specifically when considering the ability for leveraging SOA to consume and integrate multiple different types of data from different sources. The underlying challenge of sharing data across distributed systems was initially called the "grid problem," with the goal of creating a more "flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions, and resources" (Foster et al., 2001). Grid architecture was proposed as a possible solution, incorporating protocols, services, APIs, and software development kits like "Globus" (Ananthakrishnan et al., 2015; Foster and Kesselman, 1999).

4.3.4 Infrastructure as a Complex Adaptive System

Spatial data infrastructures have been increasingly approached as complex adaptive systems (Grus et al., 2010; Brous et al., 2014). This concept is especially crucial when considering a dynamic and de-centralized approach to spatial data infrastructure. Data infrastructure networks are made of individuals and groups that perform activities and interact, either in nearby physical spaces or virtually, across distributed nodes

across the web. Features of a CAS infrastructure (from Grus et al. (2010)) are (1) Components of stable and simple building blocks, (2) Complexity, with system behavior emerging from the (simultaneous) interaction of its simple components, (3) Sensitivity to initial conditions, (4) Openness, meaning the system can be impacted by external influences, (5) Unpredictability, and (6) Scale independence, or perhaps here viewed as scalability.

Coordinating mechanisms serve to manage relationships infrastructure network agents, and include principles of: (a) Self Organization, where infrastructure systems can develop more complexity over time as the result of external and internal factors, (b) Feedback, where a feedback loop mechanisms use the infrastructure output to adjust inputs and processes, (c) Planning, driven by anticipatory coordination, and (d) Allocation of Resources, where required activities are divided into subtasks to be carried out by different specialist agents (as summarized by Brous et al. (2014)). They are furthermore adaptable and subject to the changes of dynamic input-output relations. The Allocation of Resources concept defines data infrastructures as CASs that "take the form of a patchwork of functional components working together" (Brous et al., 2014), a concept that drives much of contemporary development. Rather than singular, standalone desktop environments, modern infrastructures pull from multiple web services, code libraries, and open source constructs woven together to produce an efficient system. Specialist agents could be contractors or human experts, but they could also be viewed as different specialized technological tools (and the humans that develop them) pieced together in an open source framework, such as: data services (ie. SODA API of Socrata), analytical services (ie. Google routing API, PySal spatial analysis library, MapZen), visualization services (ie. Leaflet JS library, Carto JS library), hosting services (ie. Github.io, Heroku, MapServer), data processing and

management (ie. Hadoop, Spark, POSTGIS). Spatial concepts must be incorporated at the ground-level of architectural organization, as data management and workflow processes would be difficult to adapt to spatial data at a later stage. Because a spatial data infrastructure is also a data infrastructure, in the same way that a spatial database system is a database system, it can adapt to non-spatial settings with ease, though the reverse is not true. This reflects the "sensitivity" feature of spatial data infrastructure as a CAS, where initial conditions must be set to "spatial" and coordinating mechanisms of "planning" should incorporate spatial thinking.

Some of the greatest remaining challenges in accessing and utilizing SDI involve difficulties of collaboration between SDI developers and the user community, and coordination between different SDI (Castelein et al., 2013). For example, intended users outside the SDI community may not be able to effectively access the high quality geographic data from SDI systems. Because of a lack of technological expertise in the area, they have difficulty combining the data and services from different sources. This phenomenon is well documented in multidisciplinary fields like the health sciences, as summarized in Section 2. Additional barriers to SDI collaboration include siloed approaches where organizations do not want to give up autonomy in their field of expertise, as well as technical and interoperability problems stemming from highly developed siloed SDI systems that have difficulty communicating with each other (Castelein et al., 2013). Again, these challenges of collaboration can be viewed as behaviors, coordinating mechanisms, and features to be improved in a Complex Adaptive System Data Infrastructure.

Research on SDI initiatives traces a move from techno-framework (or technology centric) to socio-technical actor networks, emerging from continuous processes of negotiations and alignments between the actors, or agents, of the system (Grus et al.,

2007). Rather than focusing narrowly on the data or technology alone, considering the human dynamics of developing such an infrastructure is essential to its success. Rather than viewing this dynamic process as a frustrating limitation, it can instead be framed as a core, desirable characteristic for a framework that ensures increasingly more powerful results. For example, a positive feedback loop using aspects of Volunteered Geographic Information could be incorporated to test and improve data quality. Participatory methods could be incorporated into the design stage to inform the various user needs and goals that developers or architects may overlook. Extract, Transform and Load (ETL) workflows could be automated and coded into pipelines for reproducible and validated outputs. By viewing a SDI as a Complex Adaptive System, it becomes both transdisciplinary and user-intensive.

4.3.5 Decentralized Spatial Infrastructures

At the beginning of the 21st century, traditional GIS systems were considered no longer appropriate for modern, distributed, heterogeneous network environments because of their closed architecture and inflexible infrastructure (Tsou and Buttenfield, 2002). However, they are still commonly used to house spatial data in multiple health sectors and disciplines as one of many isolated data system silos. Spatial system infrastructures have begun to move from closed, desktop systems to open, distributed systems that are flexible enough to accommodate dynamic interaction by users. The most recent developments reflect a radical change in infrastructure architecture, moving towards increasingly inverse systems (Coetzee and Wolff-Piggott, 2015). SDIs have thus developed from technology-centric to user-centric models.

Borba et al. (2015) considers three, sometimes overlapping, generations: the First SDI generation (1990-1999), a data-centric model; The Second SDI generation (2000-2006), the process-oriented model; and the Third SDI generation (2007-), the User-Centric model. Whereas previous SDI initiatives emphasized public and private sectors and the spatial community, new SDI systems have expanded to public spheres of distributed power and seek to include spatial and non-spatial communities. Whereas the first generation was oriented to the data, driven by data "keys," the second generation shifted to a focus on domain variables according to process being studied. Third generation systems are oriented to user requisites, with different domains and purposes. Borba et al. (2015) suggests three core principles characterizing the most recent SDI systems: Openness initiatives, a Culture of Participation, and "Inverse Spatial Injection," which can be characterized as an inverse or de-centralized data infrastructure.

Inverse infrastructures serve in great contrast to Hughesian large-scale technical systems that have dominated data infrastructures with top-down approaches (Egyedi and Mehos, 2012). These new, emerging inverse infrastructures can exist alongside or in place of traditional systems, tend to develop independently, and are user-driven (Vree, 2003; Egyedi and Mehos, 2012; Egyedi et al., 2007; Coetzee and Wolff-Piggott, 2015). The "virtual organizations" of Foster's Grid increasingly serve as agents that impact the evolution of a data infrastructure, impacted by new types of data getting produced by new types of producers (from geo-tagged FitBit sensor readings to OpenStreetMap volunteers), new types of governmental data getting released or withheld according to feedback or administration, or new types of disciplinary collaborations necessitating new types of data integration. I argue that if an inverse infrastructure is desired to

collect and share data across organizations, then user-centric design must be likewise be integrated from the start.

Notably, a modern SDI does not necessitate a standalone Geographic Information System (GIS) that stores the data and allows for analysis and visualization. In what is termed as the "Post-GIS Era", GIS moves from an application technology to a piece of ubiquitous computing that possesses a "bit of geospatial in everything" (Ed Parsons, as quoted by Harvey (2013)). In this context, GIS has become part of the infrastructure itself, with spatial data available in a multitude of forms and places, though unevenly distributed. A SDI framework may not even require a static or standalone database, for example, if the data used can be distributed, transformed, and analyzed across computational pipelines as an organized framework. The database may be replaced with automated ETL workflows that serve to manage and distribute the data, however the spatial perspective remains core as an underlying organizing principle.

4.3.6 Tying it Together: Principles of a Spatial Data Science Infrastructure.

In a review of infrastructure traditions from a spatial systems perspective, I demonstrated that while multiple approaches exist for integrating different types of heterogeneous, increasingly bigger data, there has been no formalization of critical components required for a modern SDI in health informatics. Here, I propose a spatial data science infrastructure that incorporates multiple contemporary data framework perspectives as categorical components (see Figure 3). Space is made explicit as an organizing principle for data integration, further driving technological requirements. Such a framework is thus **spatial**, meaning it supports both spatial and non-spatial

Data Framework Perspective	Qualifications	Key Characteristics that meet SDSS
Spatial Database	(a) is a database system (ie. supports non-spatial data) (b) offers spatial data types in data model & query language (c) support spatial data types in implementation	• Spatial data extensions • Spatial operations • Query-able • Retains metadata • Standards-based
Open Science Data Framework	(a) extensible storage options and APIs for access (b) allows users to subset the data with persistent links and author attributions (c) provides data curation tools to allow data and metadata to be updated	• Query-able • Reusable Design • Retains metadata • Scalable • Open • Agile/On-Demand • User-driven
Service-Oriented Grid Architecture	(a) coordinates resources that are not subject to centralized control (b) uses standard, open, general-purpose protocols and interfaces (c) delivers nontrivial qualities of service	• Query-able • Unevenly distributed • Virtual organizations • Scalable • Secure • Open (Sometimes) • Agile/On-Demand • Adaptable • User-driven • Standards-based
Complex Adaptive System	(a) Components of stable and simple building blocks (b) Complexity, with system behavior emerging from the (simultaneous) interaction of its simple components (c) Sensitivity to initial conditions, (d) Openness; can be impacted by external influences (e) Unpredictability (f) Scale independence, or perhaps here viewed as scalability.	• Virtual organizations • Reusable Design • Scalable • Adaptable • Open • Agile/On-Demand
Third-Generation Plus Spatial Data Infrastructure	(a) Openness initiatives (b) Culture of Participation (c) Inverse or de-centralized data infrastructure	• Spatial data extensions • Spatial operations • User-driven • Unevenly distributed • Inverse Architectures • Virtual organizations • Open

Figure 21. Spatial Data Science Infrastructure - Characteristics

data formats, operations, and query languages. Metadata of both spatial and non-spatial are retained in storage, or if the data is never permanently stored, the metadata can be retained along the ETL workflow. This spatial ability serves as the simple building block of the system.

An SDS infrastructure is inherently **open**, with data that can be subsetted and queried without reducing quality. It incorporates reusable design for replication and testing required in science. An SDS system is **agile**, or has the capability of on-demand processing at key locations of the workflow (ie. data extraction, data transformation, data access). It supports service-oriented architecture and/or cloud computing systems that integrate different data and analytic needs on-demand, following standardized protocols. These services can be varied, including data sourcing services, data cleaning services, analytic services, or visualization services. Components of the SDS infrastructure are likely unevenly distributed, and may change in load demands over time. A SDS infrastructure is **adaptive**, building from a simple building block to increasing complexity over time. It supports updating or expanding different components over time, or according to the needs of the users. In this manner, the infrastructure is scalable and technology-agnostic. Finally, an SDS infrastructure is **user-driven**. It supports a decentralized infrastructure that is user-driven, rather than technology or data-driven. Goals of the infrastructure can be determined through participatory infrastructure design, and other methods geared towards adjusting the system according to business, client, or user needs. As an inverse architecture, it must be capable of connecting different agents or so-called virtual organizations to a front (client-facing) and/or back (server-side) end, depending on design need.

A SDS infrastructure must consider all of these components, formalizing relevant data systems traditions in a new framework organized along spatial relationships. To

do so successfully, it must likewise consider the spatial scale(s) of the infrastructure and how to effectively adapt to new datastreams.

4.3.6.1 Defining the Spatial Scale of an Infrastructure

In this essay, I argue that a new SDS infrastructure can effectively incorporate multiple, heterogenous types of information that include both social determinants of health and built environment characteristics. To allow for greatest flexibility, the finest resolution available for the majority of socioeconomic and built environment data is used as a basic areal unit of aggregation. This can, for example, be the census tract or zip code level of the data, rather than a county or state-level. This finer resolution allows for more meaningful analysis of phenomena occurring at smaller scales. By using the smallest available unit as a starting point, phenomena occurring at larger spatial scales would be made clear in an exploratory spatial data analysis.

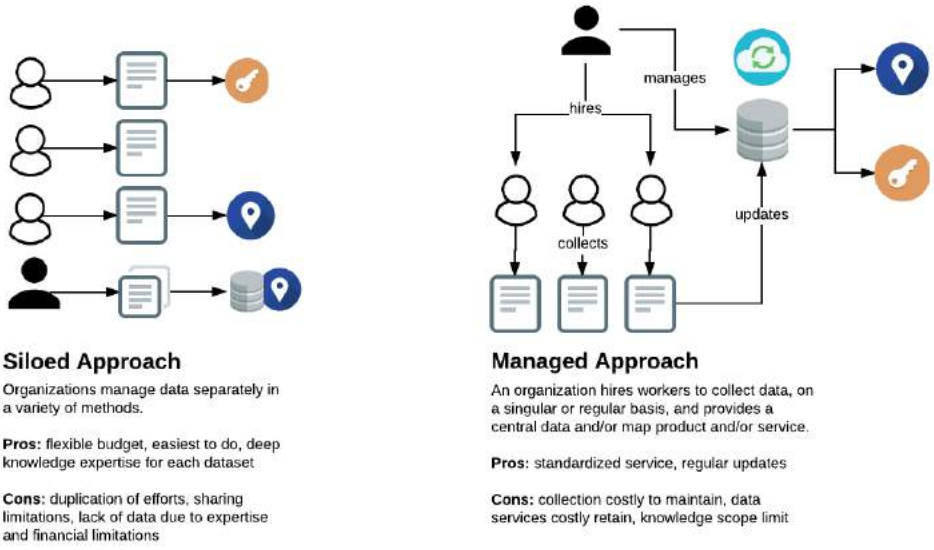
When connecting data on socioeconomic and built environment features with patient records, the areal unit chosen must likewise preserve privacy and confidentiality. Covered personal health information (PHI) can include finer resolution geographic zip codes if there are fewer than 20,000 persons residing therein, with additional restraints for certain zip codes and rare illness in the United Stated. The data linkage can thus work in either direction. Aggregated PHI at the zip code level can be linked to a warehouse of social determinants and built environments at the zip code level. Or, the zip code level SDS infrastructure could be connected to a protected, offline electronic data warehouse to retain PHI data resolution and confidentiality. By using the appropriate geographic unit as the point of data linkage, the systems infrastructure can be geared according to the final needs of the users.

4.3.6.2 From Siloed to Shared Systems

By enforcing a participatory design methodology, the ultimate objectives of an SDS infrastructure are customized to meet the final needs of the end users. By allowing for a complex systems approach within the infrastructure, the system can change over time to accommodate new needs as they are uncovered. A desired balance achieves an inherently useful but flexible system, where the systems design may adapt to new types of datastreams.

This framework can be used to improve community asset mapping in health, for example, by translating a centralized, top-down approach to a distributed, collective model. Community asset mapping requires an updated inventory of resources available to a community, from food pantries to cultural centers. This information is then compared with underlying socioeconomic needs, health outcomes, and other health indicators to assess what interventions may be necessitated to improve community outcomes. This inventory of resources serves as a core component of a community health improvement model, as reflected in the first phase of the CDC diagram discussed in the introduction. In the figure below, different approaches to gathering community assets are considered as a siloed, managed, or shared infrastructure.

The dominant method of community asset collection today is siloed data systems, where each group constructs and maintains their data. Organizations may record data in spreadsheets, word documents, databases, and/or spatial database systems. Some may geocode locations and convert their data to a map, though there may be a high cost associated in required staff expertise and/or software required. While these challenges prevent several groups from maximizing their use of the data, the knowledge of data maintained tends to be high. A small group of community workers



Siloed Approach

Organizations manage data separately in a variety of methods.

Pros: flexible budget, easiest to do, deep knowledge expertise for each dataset

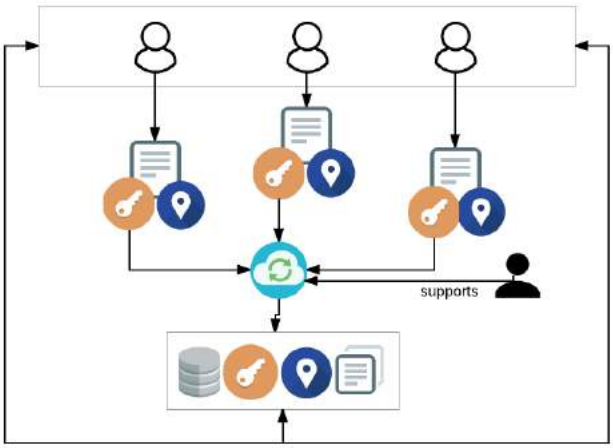
Cons: duplication of efforts, sharing limitations, lack of data due to expertise and financial limitations

Managed Approach

An organization hires workers to collect data, on a singular or regular basis, and provides a central data and/or map product and/or service.

Pros: standardized service, regular updates

Cons: collection costly to maintain, data services costly retain, knowledge scope limit



Shared Approach

Organizations manage assets with API enabled systems, and contribute to collective. Automated integration tools and technical organizations support merge processes. Final dataset available in multiple forms for use by all organizations.

Pros: flexible budget, flexible implementation, deep knowledge expertise for each dataset, free data and map products, free data and map services, standardized services, can start simple and expand over time

Cons: more difficult to implement initially because of technical expertise and collaborations required, budget range for implementation

Figure 22. Three approaches to community asset mapping as a system infrastructure.

may know all the food pantry locations and their updated information on a monthly basis, for example, but not have the technical budget to digitize that information. One contemporary approach to this problem has been the emergence of proprietary datasets that establish a baseline of technological and knowledge-base standards. Workers are hired to collect data, who may or may not be familiar with the community and are rarely content experts, who in turn update the database. The data is then sold to large organizations (like hospitals), and occasionally made available in non-profit settings or as limited views on public systems.

A third approach, as is argued here, is an inverted structure that builds from a simple, shared spatial database. In this model, each organization contributes their data with free, user-friendly methods that can accommodate both non-profit and corporate settings alike. For example, a group can update an online form, create an online spreadsheet or Fusion Table if they use Google services, or upload their existing spreadsheet to a shared cloud drive. A mostly automated processes then scrapes the information together, updates a simple and flexible data model, and converts the resource data to spatial formatting. The resulting combined data stream, with spatial and and non-spatial formats, can then be shared with the entire group. In the next section, this example is implemented as a case study component.

4.4 Case Study: A Spatial Data Science Data Infrastructure for Asset Management in Health Informatics

4.4.1 Overview

Siloed approaches in community health frameworks continue to challenge multiple stakeholders, as it can result in overlapping and redundant work, lack of communication and/or increased competition between groups, and both fragmented and incomplete datasets for all groups. Data and analytic decision supports could better support the process of improving community health outcomes by supporting the assessment and evaluation stages (CDC, 2015). Improved access to data and higher quality, more nuanced analysis could also transform the planning stage.

To address these challenges, I developed a Spatial Data Science Decision Support System (SDS-DSS) prototype as a cross-sector collaboration with the Chicago Department of Public Health and multiple Chicago community organizations. The goal of this framework was to develop a user-friendly, low cost system that would better model and evaluate community health outcomes for different types stakeholders. To achieve this, the decision support system would need to gather, integrate, and use information on community assets, health indicators, and social determinants of health. The system must also enable participatory monitoring and evaluation of community health improvement efforts, to not only encourage shared investment and multi-sector collaboration, but also improve the data quality required for the decision support over time.

These objectives were designated to meet key issues addressed by the CDC to promote alignment between accreditation, hospital, and other community-oriented

processes (see Figure 2). They also provide a framework to inform a Healthy Chicago 2.0 strategy to "analyze geographic access to health and human services and address gaps in care." Healthy Chicago 2.0 is a place-based policy initiative by the City of Chicago and the Chicago Department of Health designed to reduce inequities and improve the health and vitality of its residents. This strategy addresses the first action area of the policy initiative – to increase capacity and availability of health and human services, and maximize impact of existing resources. By integrating existing community assets information with social determinants of health and built environment characteristics, policy makers could better target interventions for improved health outcomes.

I implemented this project with the City of Chicago Department of Public Health (CDPH) as a Public Service Intern and Volunteer, working with officials at Innovations, Planning, and Epidemiology groups. In a parallel effort, I worked with multiple community organizations from the West Humboldt Park neighborhood via a community collective known as the West Humboldt Park Healthy Community Initiative. In particular, I collaborated with representatives from the Northwestern Hospital Community Extension Office, La Casa Norte, Logan Square Neighborhood Association, Northwest Food Partners Network, and the local Salvation Army. CDPH was interested in generating a web application with publicly available health indicators and socioeconomic parameters to facilitate their planning and resource allocation process, and ultimately share the generated dataset through an open API. The West Humboldt Park community collective was interested in pooling their individual resource lists into a larger one, and making it easily accessible in an open environment. In the final decision support, a SDS infrastructure allows for both. A robust back-end integrates,

processes, and queries the data—making multiple front-ends possible to serve different stakeholder needs.

4.4.2 Solution Framework

I implemented the SDS infrastructure through three methodological phases: (1) System Architecture Conceptualization, (2) Data Integration and Warehousing Workflows, and (3) Client-Facing Web Applications. During system conceptualization, I conducted an inventory of needs with end-users to establish a baseline of standards required, data desired, and application prototypes needed. In the data integration phase, I established standardized data workflows to convert heterogeneous data to the desired architecture.

The decision support system infrastructure prototype ultimately serves two core groups of stakeholders: health department officials and community organizations. To accommodate health department officials, I designed a simple web application that allows for data access and exploration (Healthy Access, Healthy Regions Explorer - HARE). This data includes selected health indicators, socioeconomic variables, and available resource data. For community organizations I collaborated with in the West Humboldt Park neighborhood of Chicago, I designed a simple web application to access and query resource data generated from the project (West Humboldt Park Resource Map). While each stakeholder has a different front end application, the data comes from the same data warehouse.

A model of how the back- and front-ends connect is shown in the following figure (5). In this case study, a SDS prototype focuses on the data integration/processing,

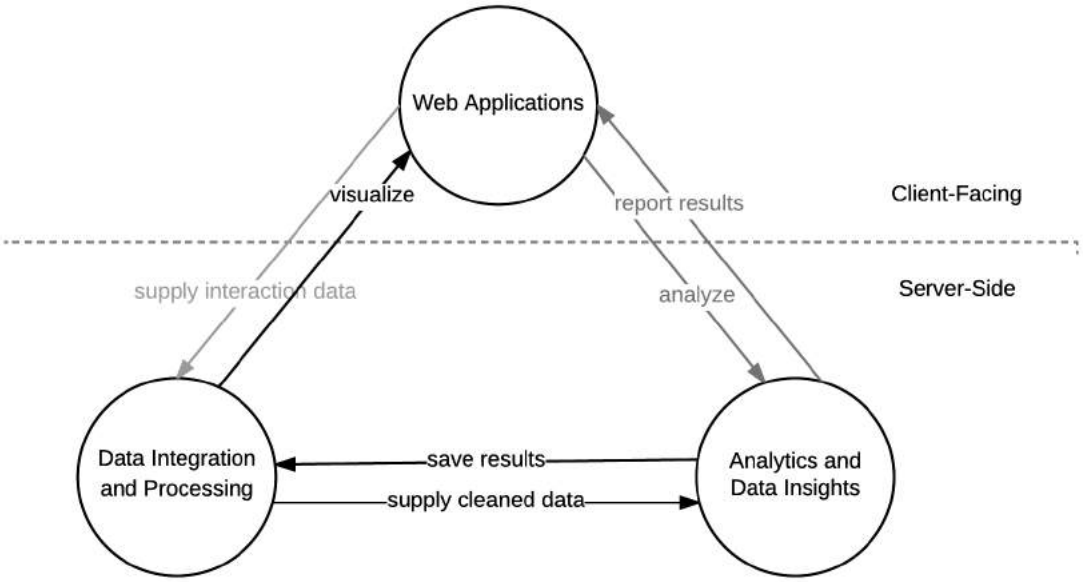


Figure 23. Core components of a SDS system. This study primarily focuses on aspects of data integration and processing, with initial links to simple front-end applications.

with initial extensions to web application components. A more complete SDS system would also facilitate data analytics, likewise exchanging between front- and back-ends.

4.4.2.1 Abstracting the Public Health Environment

With stakeholder participation, I first determined which aspects of the public health environment were to be abstracted in a data warehouse. First, I reviewed existing public health data models to evaluate what data should be included in a standardized approach. Then, data to include were refined and pared down according to stakeholder needs. To identify essential data the framework required, I worked with partners at the health department and select community organizations in parallel

projects, following an aligned but previously siloed approach to monitoring community health outcomes.

4.4.2.1.1 Generating a Baseline

Multiple components affect the health of individuals and populations, building from genetic predispositions to powerful social determinants of health, built environment factors, natural environment access, local public health department and health worker initiatives, cultural factors, economically supported or preferred activities, and more. There is no standardized public health data model; this may be partly due to technical and interoperability challenges, theoretical model challenges, lack of public health data available, and difficulty in communication across varying fields of research. Health data models focused on a clinical environment commonly connecting patients, facilities, and records, but do not relate social determinants of health. Local government data models focused on infrastructure, transportation, and facilities, but did not include local public health department components. A more unified and conceptualized (urban) public health environment data model demands more research and investigation. In a big-picture theoretical model of health, multiple factors influence outcomes. The *Science of Eliminating Health Disparities* symposium (2012) identified four components of the environmental context in public health: social, policy, physical/natural, and built environment. A political ecology model from health geography further refines the "eco-social" concept integrating social, economic, and political components of a location that influence local health (Krieger, 2011). In a systems infrastructure, these concepts must be abstracted accordingly.

For example, the urban design of a city at a neighborhood level, such as the

placement and design of buildings and parking lots, affects the walkability and bikeability of a neighborhood. While the relationship between urban design and travel is inconsistent and not completely understood (Frank and Engelke, 2001), including transportation modes is useful in a public health data model. In some areas that are more friendly to pedestrians because of built form, the threat of localized high crime (another public health risk) keeps residents indoors, thus negating the potential healthy effect (Dannenberg and Wendel, 2011). Relationships between built environment and physical activity may be complex, but evidence on the correlation have been deemed sufficient to suggest policy changes (Milner and Milner, 2016). Personal and subjective factors influencing the choice to walk or bicycle include cost, distance, safety, and circumstances, and environmental factors include weather, topography, and infrastructure features (Frank et al., 2003; Frank and Engelke, 2001). Transportation systems, land use, and health-promotive environments also influence these systems. Walking primarily takes place in public areas (streets and public facilities), and is supported by safe, attractive, and comfortable environments. In a public health data infrastructure, sidewalks, public facilities, and crime statistics may serve as a seed of this representation.

4.4.2.1.2 Defining the Data Inventory

A basic data model was constructed to initialize the public health environment, integrating available health indicators and outcomes, social determinants of health, and built environment characteristics. To determine which variables and statistics to incorporate in the next version, interviews with different stakeholders were held to uncover common and specific needs. These conversations were scenario-based,

attempting to uncover the process of data usage and inquiry for each group. Each approach reflects the interests of the stakeholders involved.

Of the forty-two recommended health metrics for community health assessment posed by the for Disease Control and Prevention (2015), only eleven were available below county level (housing, marital status, language spoken at home, foreign born, employment status, poverty level, age, sex, race/ethnicity, income, and educational attainment). All of these variables are available from the U.S. Census at a tract-level, at minimum, and most can be easily extracted from already cleaned data available on the the Social Vulnerability Index website (for Disease Control et al., 2014; Flanagan et al., 2011).[11] These variables are integrated into the data warehouse, however additional data was still needed to meet the goals of the stakeholders and requirements of a eco-social health data model.

The following overview includes data topics integrated from multiple sources, made available at the tract-level, across built environment, social determinants of health, health outcomes, and resource data categories. **Built environment** indicators included population density, food access indicators, brownfields sites, high performing schools, no vehicle households, proportions of car commuters, proportions of public transit users, commute time averages, walk scores, perceived safety (Naik et al., 2014), property crime and violent crime rates.[12] **Social determinants of health**

[11]The Social Vulnerability Index is constructed from demographic and socioeconomic factors to approximate social vulnerability at the tract-level, and is used by emergency planners for disaster mitigation efforts.

[12]Property and Violent crime was coded from police records made available on the Chicago Data Portal. Following stakeholder specification, I calculated the total number of crimes reported for the year, geocoded and aggregated at the census tract, per census tract population. Property crime was tagged as: burglary, larceny, motor vehicle theft, and arson. Violent crime was coded if the incident was tagged as: homocide 1st and 2nd degree, criminal sexual assault, robbery, aggravate assault, and aggravated battery.

included race and ethnicity, limited English speaking populations, disability status, elderly populations, young children, children and youth, institutionalized populations, crowded housing, multi-unity housing, educational attainment, per capita income, unemployment status, persons in poverty, children in poverty, uninsured populations, foreclosure risk, the economic hardship index (Shih et al., 2013), the childhood opportunity index (Acevedo-Garcia et al., 2014), and the health literacy index (Pleasant, 2013). **Health outcomes** data from the health department, made available for open sharing, includes premature mortality rate and years of life lost.

Resource data includes service locations linked to one of the following categories: (1) emergency needs and social services, (2) medical providers and health services, (3) wellness and healthy living, (4) education and job resources, and (5) behavioral support and counseling. These categories were generated and updated through community meetings, rather than top-down taxonomies. Resource data was made available through pooled data sharing of the community organizations involved, as well as the health department. It is not considered complete, as this pilot continues to go through further iterations of data updates as new organizations contribute.

Finally, to represent different political boundaries that may impact results, fiat boundaries are imported at different scales, including: census tract, zip code, ward, and Chicago community area. When appropriate, data is made available at different aggregations. Because it is the finest resolution available for most of the data, the census tract serves as the spatial scale of interest for this system. Data was pulled for 2015 or 2014, unless otherwise noted, for this initial prototype; while the system remains flexible, increasing temporal resolution was out of scope for this study. Data was sourced from the Chicago Department of Public Health, Chicago Data Portal, Medicaid Data Portal, IDOT (Illinois Department of Transportation) Data Portal,

Cook County Data Portal, GTFS feed (General Transit Feed Service), the Center for Disease Prevention and Control, Decennial Census and American Community Survey 5-year estimates for 2014 via Census.gov, and the West Humboldt Park Pilot project.

4.4.2.2 Server-Side Infrastructure

I implemented a SDS infrastructure to integrate multivariate heterogenous data representing dimensions of public health in Chicago with multiple stakeholders (each requiring unique needs). By integrating the data in a dynamic (where available), documented, and replicable way, the process remains open and adaptable to changes in data or stakeholder needs over time. Because of the iterative process of data integration using service-oriented architecture (SOA) and participatory methods by stakeholders, the process furthermore ensures increased quality (and potentially complexity) over time.

The SDS infrastructure begins with the simple building block of place: all data is linked through its geographic relationships in a public health environment. Data is retained at its original spatial resolution, be it a location recorded as a point (ie. service address, latitude/longitude of street image), line (ie. sidewalk, street network), or polygon (ie. tracts, community areas, building footprints, park shapes). Statistics and reported calculations such as health outcomes, health indicators, and scores/indices are generally linked to some geographic area (ie. zip code, census tract) as its reporting coverage. These datasets were thus "spatially enabled" in this infrastructure according to those geographic relationships. The resulting data model serves as an enhanced snowflake model, with geographic indicator attributes (ie. FIPS code, zip code) and/or geographic locations (according to spatial index and

spatial metadata) joining the multivariate, heterogenous data. Complex data models of source data are thus retained, as only attributes related to location are pulled with either integrating and/or developing data views.

4.4.2.2.1 Data Model

A snowflake schema ties together data tables in a multidimensional database. A centralized "fact table" connects multiple dimensions. In this case, spatial features representing a geographic area connects multiple dimensions of data. The spatial feature may be census tracts, with hundreds of attributes pulled and linked from other spatial and non-spatial data. Multiple spatial features (or "snowflakes") exist, corresponding to different spatial resolutions. ETL workflows and SQL queries relate these dimensions; several of these workflows are integrated as services to further automate the processes. While these data linkages could be done on-the-fly, the joined tables are precomputed to improve computational efficiency because storage is less expensive than on-the-fly computational processing. The solution is scalable, which becomes increasingly important as more data dimensions are merged with central spatial features.

Some data is available as a data service, harvestable through a standard web service such as REST or SOAP using SOA. Other data can be extracted through websites in excel, CSV, or other data format as a download. While not a standardized web service, the digitized download enables dynamic processing when connected with ETL workflows. While more data is becoming available as web services, transforming the datascape, much data still requires manual manipulation. In this project, because of the diversity of multiple stakeholders, there is great variety in how data was sourced

and stored. Some was only available as offline shared excel files or PDF documents, whereas other datasets were easily queried and extracted from standardized data portals. I privileged updating data sourcing processes according to need and goals of the project overall, rather than technological capability. This reflects the user-driven objectives and participatory approach of this framework.

4.4.2.2.2 Shared Systems Approach for Resource Data

Resource data from the West Humboldt Park Pilot is updated using a shared approach, following the inverse infrastructure concept discussed in Section 3. I worked with several community organization representatives to determine how each group collected and maintained their data. Then, data collection for each group was migrated online using a Google Fusion Table. I incorporated a basic data standard that retained flexibility, and included core data essentials that were meaningful to the group (ie. site name, description, and address, source). This data standard was refined with community input, and may still further be updated.

Data is thus shared as a service, harvestable through the online format each organization made available. I trained groups to share their Google Fusion Table by finding their unique table id code. Data is then geocoded, merged, and/or updated, according to data standards established. The source for each row is retained, but added to the collective. Initially, updates were performed manually to ensure compliance. Then, machine learning techniques were implemented to de-duplicate the structured data using Dedupe.io, an opensource engine. The updated, shared datastream is then made available on the public website as both product and service. A schematic of this process follows.

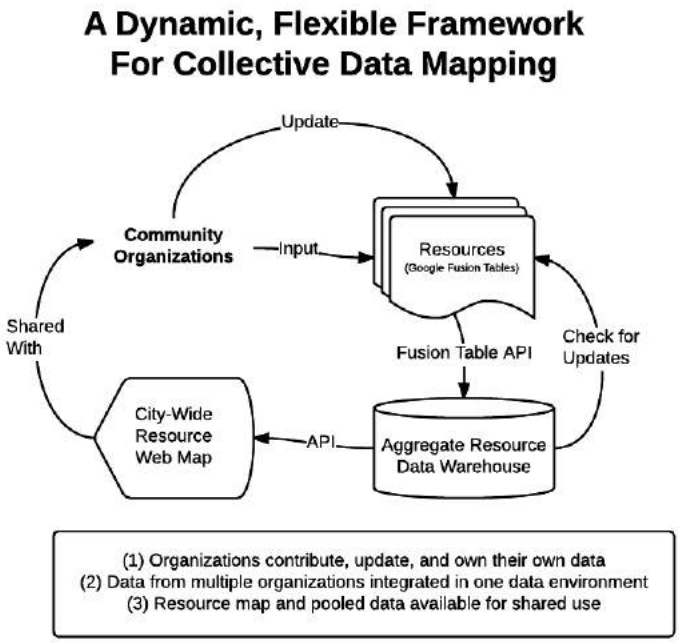


Figure 24. Collective Data Mapping Infrastructure

4.4.2.2.3 Overall Data Integration Approach

ETL workflows operationalize data processing and updates on the back side of a data warehouse architecture, automating transformations in computational pipelines to data extraction, transformation, and warehouse loading (Vassiliadis and Simitsis, 2009). To integrate data in the warehouse, I incorporate ETL workflows to extract heterogeneous from multiple data formats, transformed according to established quality rules and data standards, and then loaded into the data warehouse. The basic methodology involves the design of the overall warehouse architecture, or planning and implementation of a centralized spatial data warehouse and relevant ETL workflows to integrate and pre-process data from multiples sources and formats. This involved

establishing proper data models and planning towards spatial and non-spatial queries to support analytical functions.

Data was saved in its initial form, and processed through an ETL workflow to the final database for storage in cleaned, re-projected form. The Feature Manipulation Engine (FME by Safe Software) for ETL processing and integration of data, and PostGres SQL and POSTGIS were used to store data.[13] ETL (Extract, Transform, and Load) workflows followed QA/QI standards I established for the system architecture (see Appendix). In pgAdmin, the PostGresSQL database is viewed, constraints added, and SQL queries tested. A schematic of this integration process follows.

4.4.2.3 Client-Facing Applications

Two lightweight front-end web application prototypes were developed to allow user interaction with generated results from different components of the warehouse. Data integration of select health indicators, socioeconomic characteristics, and social determinants of health is featured in the "Healthy Access and Regions Explorer" (HARE) web application (http://makosak.github.io/chihealthaccess). The "Healthy Regions" web application allows users to interact with and explore multiple choropleth maps and associated data. Maps are generated from multiple attributes that characterize select social determinants of health, with data table updated live according to where the user points to on the map.

[13]While FME is as a proprietary technology, it remains useful in the opensource community as an ETL workhorse. Its workflows can be shared, allowing for replicability. I additionally diagram most central workflows used. While I have already begun to convert each transformation into an opensource script in python and/or POSTGIS, completing each component was beyond the scope of this study.

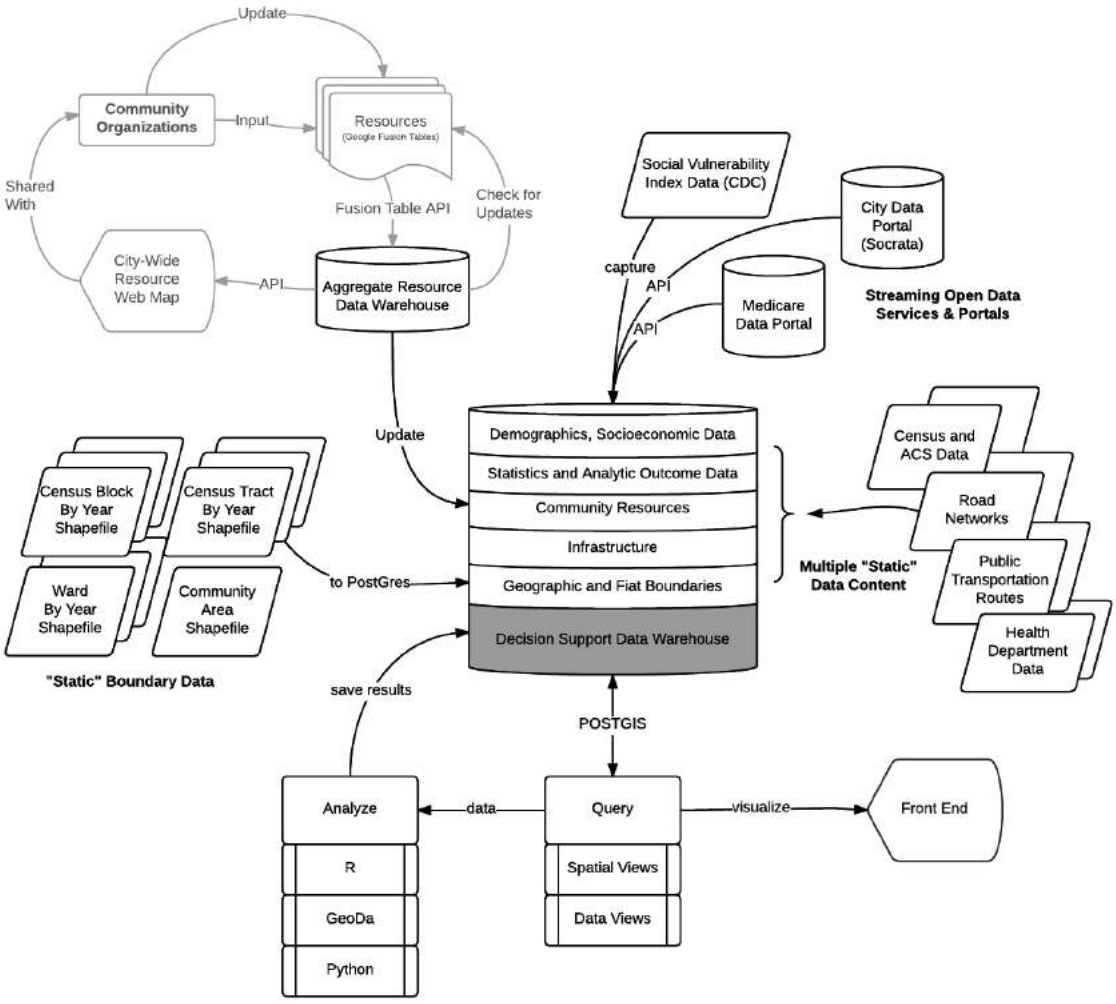


Figure 25. System Architecture Overview

The "West Humboldt Park Resource Map" (http://makosak.github.io/HumboldtResources) serves as both a data integration and dynamic asset mapping application. The "West Humboldt Park Resource Map" web application allows users to query resource data with simple buffer analysis, with immediate results made available for interaction and exploration. In both cases, data can be downloaded on the site in multiple formats.

I developed these applications with open software, hosted on Github. The "Healthy

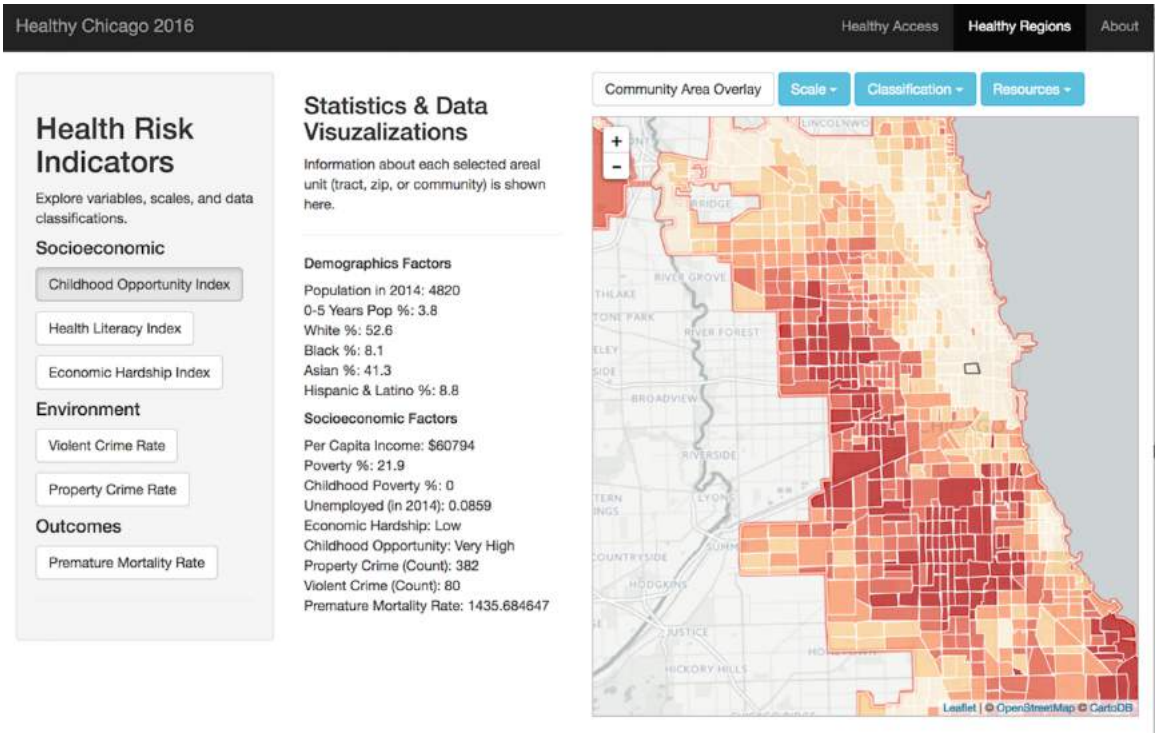


Figure 26. Screenshot of HARE web application

Regions" app primarily uses Leaflet and D3 javascript visualization libraries, with a HTML/CSS, Bootstrap, javascript, and jquery framework. The "West Humboldt Park Resource Map" serves as a customized version of the Derek Eder web map template, and uses Google Maps API, Google Fusion Table API, and also a a HTML/CSS, Bootstrap, javascript, and jquery framework.

4.5 Discussion

In the Chicago systems infrastructure case study, I illustrate a SDS system prototype that integrates data relevant to population health for various stakeholders. While front-end applications reflect simplified queries, as driven by stakeholder needs, they may likewise obscure more complicated systems beneath. The Chicago case

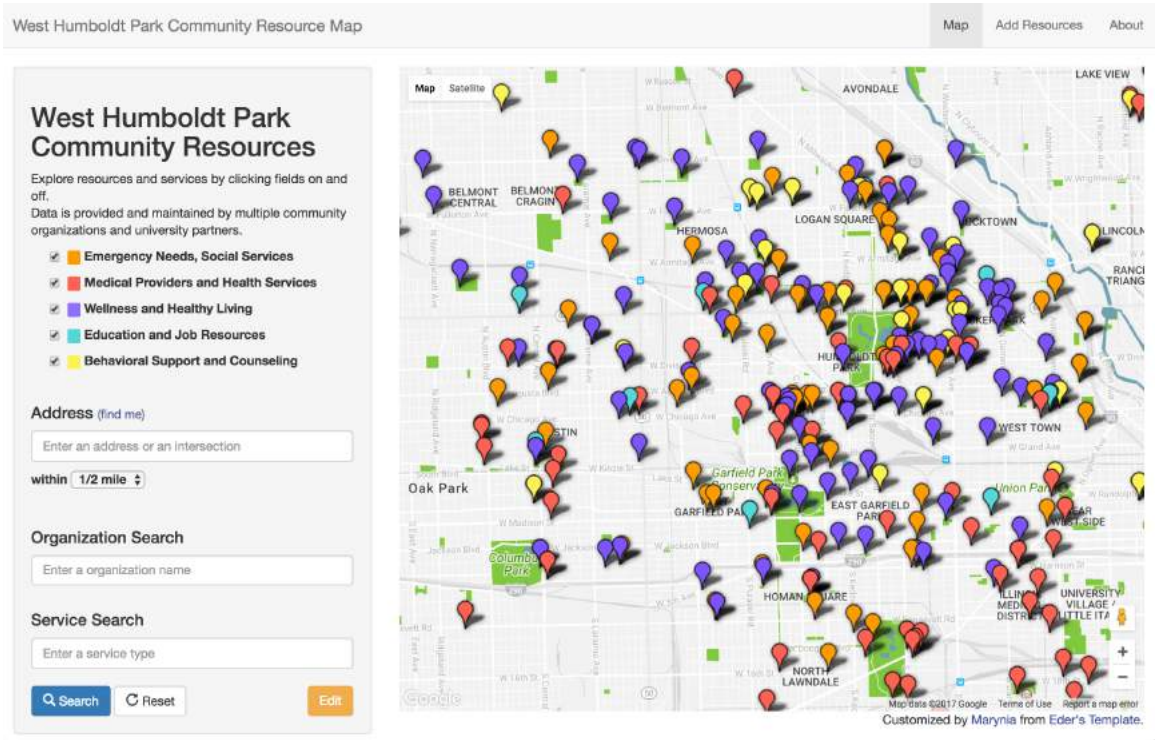


Figure 27. Screenshot of West Humboldt Park Resource Map

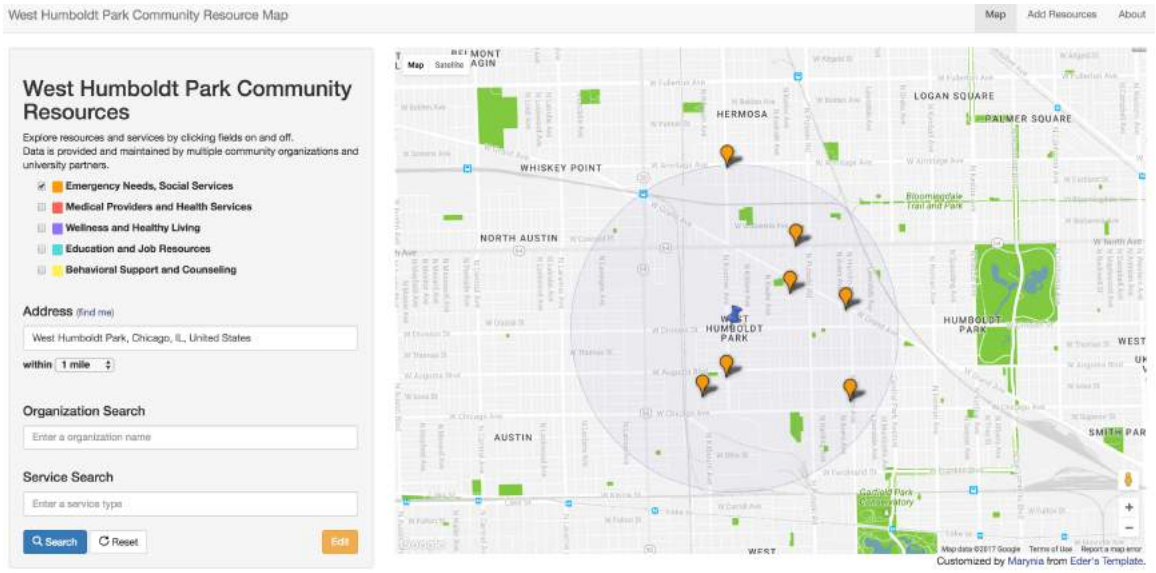


Figure 28. Screenshot of West Humboldt Park Resource Map with Query Selection

study is a successful SDS system prototype because it meets the core specifications. It uses spatial relationships as a means of organizing information, and takes advantage of spatial warehouse architecture. By implementing ETL workflows to integrate and process data, it is likewise dynamic and adaptive, accommodating data updates according to stakeholder needs. By making workflow diagrams, code, and integrated data streams public, and standards explicit, the system is also reproducible. It is likewise user-oriented; the Chicago system incorporated participatory design with multiple users from system conceptualization through final web applications reviews. The West Humboldt Pilot also adds an inverse infrastructure feature, as organizations are able to contribute data on the back-end.

While the infrastructure was developed to meet SDS requirements, it must also be considered within a wider public health informatics framework. To evaluate the public health SDS infrastructure developed, the following dimensions were additionally considered (see Table 1) from recommendations of a CDC working group: system usefulness, flexibility, system acceptability, portability, and system costs (Buehler, 2004). While these characteristics were designed to evaluate public health surveillance systems with a focus on infectious disease breakouts, the model was adapted to a contemporary public health community health approach with a focus on chronic disease surveillance. The system is assumed moderately flexible, acceptable, and generally portable because of assumed challenges in organizations retooling for a spatial explicit approach. The cost will also be related to spatial expertise and coding abilities, even with opensource and free software. Still, the system usefulness remains high, addressing a well documented need and also facilitating updated stakeholder needs according in each cycle/iteration.

In this essay, I proposed a new Spatial Data Science Infrastructure that formalizes

Category	*Evaluation*
System Usefulness	**High.** Addresses well-documented need
Flexibility	**Moderate.** Very flexible in system set-up, geodatabase architecture needs, and workflows implemented. Can be linked to additional data sets in low or high security settings. However, warehousing approach with spatial linking requires spatial data handling abilities
System acceptability	**Moderate.** Transferring to a spatial database as an underlying data infrastructure strategy may be difficult when not familiar or well understood.
Portability	**Moderate.** Low barrier to software acquisition. Higher barrier to staff expertise required; training and/or hiring may be required.
System Cost	**Low to Moderate.** Depending on staff expertise, may vary. Most related software is free or low cost, though staff cost may be higher for development. Computational pre-processing on server and storage cost will also vary depending on needs.

Table 6. Data Integration Component Evaluation

desired characteristics of a modern infrastructure essential to health informatics. In an ecosocial view of health, multiple factors from social, environmental, and economic societal structures can contribute to community health outcomes locally. As such, a new infrastructure is required to successfully abstract, integrate, and access the associated data representations. By making space explicit as an organizing rule for data relationships, a flexible but powerful architecture is made possible. A SDS infrastructure is a spatial infrastructure that is also dynamic, reproducible, adaptive, and participatory. By allowing for these additional characteristics, health systems infrastructures can further support community health improvement frameworks by facilitating shared data and decision support implementations across health partners.

Data integration of disparate, heterogeneous data sources is necessary for advancing policy and planning relevant to public health. As argued in this essay, it can be

accomplished when using space as a place of integration. This integration facilitates new insight as it allows new kinds of users to be able to access multivariate data, from community organizations to health officials. This can also made data more accessible across organization as mutually shared knowledge, but only if the conceptual design remains open and participatory.

REFERENCES

(2009). Supermarket facts. *Institute FM*.

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies 72*(1), 1–19.

Abbring, J. H. and J. J. Heckman (2007). Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. *Handbook of econometrics 6*, 5145–5303.

Acevedo-Garcia, D., N. McArdle, E. F. Hardy, U. I. Crisan, B. Romano, D. Norris, M. Baek, and J. Reece (2014). The child opportunity index: improving collaboration between community development and public health. *Health Affairs 33*(11), 1948–1957.

Adler, N. E. and D. H. Rehkopf (2008). Us disparities in health: descriptions, causes, and mechanisms. *Annu. Rev. Public Health 29*, 235–252.

Amaro, H. (2014). The action is upstream: place-based approaches for achieving population health and health equity. *American journal of public health 104*(6), 964.

Ananthakrishnan, R., K. Chard, I. Foster, and S. Tuecke (2015). Globus platform-as-a-service for collaborative science applications. *Concurrency and Computation: Practice and Experience 27*(2), 290–305.

Anderson, G. and R. Moreno-Sanchez (2003). Building web-based spatial information solutions around open specifications and open source software. *Transactions in GIS 7*(4), 447–466.

Angrist, J. and J.-S. Pischke (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. Technical report, National Bureau of Economic Research.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association 91*(434), 444–455.

Angrist, J. D. and J.-S. Pischke (2015). The path from cause to effect: Mastering 'metrics. Technical report, Centre for Economic Performance, LSE.

Anselin, L. (1988a). Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical analysis 20*(1), 1–17.

Anselin, L. (1988b). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht.

Anselin, L. (1990). Spatial dependence and spatial structural instability in applied regression analysis. *Journal of Regional Science 30*(2), 185–207.

Anselin, L. (1995). Local indicators of spatial association?lisa. *Geographical analysis 27*(2), 93–115.

Anselin, L. (2003). Spatial externalities, spatial multipliers, and spatial econometrics. *International regional science review 26*(2), 153–166.

Anselin, L. (2007). Spatial econometrics in RSUE: Retrospect and prospect. *Regional Science and Urban Economics 37*(4), 450–456.

Anselin, L. (2015). Spatial data science for an enhanced understanding of urban dynamics.

Anselin, L. and D. A. Griffith (1988). Do spatial effecfs really matter in regression analysis? *Papers in Regional Science 65*(1), 11–34.

Anselin, L. and J. Le Gallo (2006). Interpolation of air quality measures in hedonic house price models: spatial aspects. *Spatial Economic Analysis 1*(1), 31–52.

Anselin, L., J. Le Gallo, and H. Jayet (2008). Spatial panel econometrics. In *The econometrics of panel data*, pp. 625–660. Springer.

Anselin, L., N. Lozano, and J. Koschinsky (2006). Rate transformations and smoothing. *Urbana 51*, 61801.

Anselin, L. and N. Lozano-Gracia (2008). Errors in variables and spatial effects in hedonic house price models of ambient air quality. *Empirical economics 34*(1), 5–34.

Anselin, L., I. Syabri, and Y. Kho (2006). Geoda: an introduction to spatial data analysis. *Geographical analysis 38*(1), 5–22.

Apparicio, P., M. Abdelmajid, M. Riva, and R. Shearmur (2008). Comparing alternative approaches to measuring the geographical accessibility of urban health services: Distance types and aggregation-error issues. *International journal of health geographics 7*(1), 7.

Arraiz, I., D. M. Drukker, H. H. Kelejian, and I. R. Prucha (2010). A spatial cliff-ord-type model with heteroskedastic innovations: small and large sample results. *Journal of Regional Science 50*(2), 592–614.

Ashenfelter, O. and D. Card (1984). Using the longitudinal structure of earnings to estimate the effect of training programs. Technical report, National Bureau of Economic Research.

Athias, L., P. Wicht, et al. (2014). Cultural biases in public service delivery: Evidence from a regression discontinuity approach. *MPRA Paper 60639*.

Austin, S. B., S. J. Melly, B. N. Sanchez, A. Patel, S. Buka, and S. L. Gortmaker (2005). Clustering of fast-food restaurants around schools: a novel application of spatial statistics to the study of food environments. *American Journal of Public Health 95*(9), 1575–1581.

Ball, K., A. Timperio, and D. Crawford (2009). Neighbourhood socioeconomic inequalities in food access and affordability. *Health & place 15*(2), 578–585.

Baltagi, B. (1995). Econometric analysis of panel data.

Baltagi, B. H., P. Egger, and M. Pfaffermayr (2013). A generalized spatial panel data model with random effects. *Econometric Reviews 32*(5-6), 650–685.

Baltagi, B. H., S. H. Song, B. C. Jung, and W. Koh (2007). Testing for serial correlation, spatial autocorrelation and random effects using panel data. *Journal of Econometrics 140*(1), 5–51.

Baltagi, B. H., S. H. Song, and W. Koh (2003). Testing panel data regression models with spatial error correlation. *Journal of econometrics 117*(1), 123–150.

Baum-Snow, N. and F. Ferreira (2014). Causal inference in urban and regional economics. Technical report, National Bureau of Economic Research.

Baylis, K., A. Ham, and others (2015). How important is spatial correlation in randomized controlled trials? In *2015 AAEA & WAEA Joint Annual Meeting, July 26-28, San Francisco, California*. Agricultural and Applied Economics Association & Western Agricultural Economics Association.

Berndt, D. J., A. R. Hevner, and J. Studnicki (2003). The catch data warehouse: support for community health care decision-making. *Decision support systems 35*(3), 367–384.

Bertrand, M., E. Duflo, and S. Mullainathan (2002). How much should we trust differences-in-differences estimates? Technical report, National Bureau of Economic Research.

Bilheimer, L. T. and R. J. Klein (2010). Data and measurement issues in the analysis of health disparities. *Health services research 45*(5p2), 1489–1507.

Black, S. E. (1999). Do better schools matter? Parental valuation of elementary education. *Quarterly journal of economics*, 577–599.

Blanchard, T. and T. Lyson (2002). Access to low cost groceries in nonmetropolitan counties: Large retailers and the creation of food deserts. In *Measuring Rural Diversity Conference Proceedings, November*, pp. 21–22.

Block, D. and J. Kouba (2006). A comparison of the availability and affordability of a market basket in two communities in the chicago area. *Public health nutrition 9*(07), 837–845.

Borba, R. L., J. C. Strauch, J. M. Souza, and D. J. Coleman (2015). Architectural and technological aspects for the next generation of sdi.

Bower, K. M., R. J. Thorpe, C. Rohde, and D. J. Gaskin (2014). The intersection of neighborhood racial segregation, poverty, and urbanicity and its impact on food store availability in the united states. *Preventive medicine 58*, 33–39.

Brous, P., I. Overtoom, P. Herder, A. Versluis, and M. Janssen (2014). Data infrastructures for asset management viewed as complex adaptive systems. *Procedia Computer Science 36*, 124–130.

Browning, C. R., K. A. Cagney, and M. Wen (2003). Explaining variation in health status across space and time: implications for racial and ethnic disparities in self-rated health. *Social science and medicine 57*(7), 1221–1235.

Brueckner, J. K. (1998). Testing for strategic interaction among local governments: The case of growth controls. *Journal of urban economics 44*(3), 438–467.

Brueckner, J. K. (2003). Strategic interaction among governments: An overview of empirical studies. *International regional science review 26*(2), 175–188.

Brunsdon, C., A. Fotheringham, and M. Charlton (2002). Geographically weighted summary statistics?a framework for localised exploratory data analysis. *Computers, Environment and Urban Systems 26*(6), 501–524.

Buehler, J. W. (2004). Review of the 2003 national syndromic surveillance conference?lessons learned and questions to be answered. *Morbidity and Mortality Weekly Report*, 18–22.

Callaghan, R. C., M. Sanches, and J. M. Gatley (2013). Impacts of the minimum legal drinking age legislation on in-patient morbidity in canada, 1997–2007: a regression-discontinuity approach. *Addiction 108*(9), 1590–1600.

Callaghan, R. C., M. Sanches, J. M. Gatley, and J. K. Cunningham (2013). Effects of the minimum legal drinking age on alcohol-related health service use in hospital settings in ontario: a regression–discontinuity approach. *American journal of public health 103*(12), 2284–2291.

Card, D. (1990). The impact of the mariel boatlift on the miami labor market. *Industrial & Labor Relations Review 43*(2), 245–257.

Card, D. and A. Krueger (1994). The economic return to school quality: A partial survey. Technical report.

Carpenter, C. and C. Dobkin (2009). The effect of alcohol consumption on mortality: regression discontinuity evidence from the minimum drinking age. *American Economic Journal: Applied Economics 1*(1), 164–182.

Casetti, E. (1997). The expansion method, mathematical modeling, and spatial econometrics. *International Regional Science Review 20*(1-2), 9–33.

Castelein, W. T., A. K. Bregt, and L. Grus (2013). The role of collaboration in spatial data infrastructures. *URISA Journal 25*(2), 31–40.

Catlett, C., T. Malik, B. Goldstein, J. Giuffrida, Y. Shao, A. Panella, D. Eder, E. van Zanten, R. Mitchum, S. Thaler, et al. (2014). Plenario: An open data discovery and exploration platform for urban science. *IEEE Data Eng. Bull. 37*(4), 27–42.

Catlin, B., K. Barnett, and P. Stange (2014). Community health needs assessment (chna) toolkit.

CDC (2005). Racial/ethnic disparities in prevalence, treatment, and control of hypertension–united states, 1999-2002.

CDC (2015). Community health improvement navigator.

CDPH (2012). Policy development as a tool for increasing access to healthy foods.

CDPH (2013). A recipe for healthy places: Addressing the intersection of food and obesity in chicago.

Chagas, A. L. S., R. Toneto, and C. R. Azzoni (2011). A spatial propensity score matching evaluation of the social impacts of sugarcane growing on municipalities in Brazil. *International Regional Science Review*, 0160017611400069.

Channick, R. (2013). Final closing time for dominick's on saturday. *Chicago Tribune*.

Chitewere, T., J. K. Shim, J. C. Barker, and I. H. Yen (2017). How neighborhoods influence health: Lessons to be learned from the application of political ecology. *Health & Place 45*, 117–123.

Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International statistical review 69*(1), 21–26.

Cliff, A. and K. Ord (1972). Testing for spatial autocorrelation among regression residuals. *Geographical analysis 4*(3), 267–284.

Cliff, A. D. and J. K. Ord (1973). *Spatial autocorrelation*, Volume 5. Pion London.

Cliff, A. D. and J. K. Ord (1981). *Spatial processes: models & applications*, Volume 44. Pion London.

CMAP (2013). Inventory tp-b.

Coetzee, S. and B. Wolff-Piggott (2015). A review of sdi literature: Searching for signs of inverse infrastructures. In *Cartography-Maps Connecting the World*, pp. 113–127. Springer.

Conley, T. G. and C. R. Taber (2011). Inference with ?difference in differences? with a small number of policy changes. *The Review of Economics and Statistics 93*(1), 113–125.

Consortium, W. W. W.

Cooper, A. K., P. Rapant, J. Hjelmager, D. Laurent, A. Iwaniak, S. Coetzee, H. Moellering, and U. Düren (2011). Extending the formal model of a spatial data infrastructure to include volunteered geographical information.

Corak, M. (2013). Income inequality, equality of opportunity, and intergenerational mobility. *The Journal of Economic Perspectives 27*(3), 79–102.

Corrado, L. and B. Fingleton (2012). Where is the economics in spatial econometrics? *Journal of Regional Science 52*(2), 210–239.

Croissant, Y. and G. Millo (2008). Panel data econometrics in r: The plm package. *Journal of Statistical Software 27*(2), 1–43.

Croissant, Y., G. Millo, et al. (2008). Panel data econometrics in r: The plm package. *Journal of Statistical Software 27*(2), 1–43.

Cummins, S., S. Curtis, A. V. Diez-Roux, and S. Macintyre (2007). Understanding and representing ?place? in health research: a relational approach. *Social science & medicine 65*(9), 1825–1838.

Dannenberg, A. L. and A. M. Wendel (2011). *Measuring, assessing, and certifying healthy places*. Springer.

Danziger, S., F. T. Pfeffer, S. Danziger, and R. F. Schoeni (2013). Wealth disparities before and after the great recession. *The ANNALS of the American Academy of Political and Social Science 650*(1), 98–123.

Davis, S. K., M. A. Winkleby, and J. W. Farquhar (1995). Increasing disparity in knowledge of cardiovascular disease risk factors and risk-reduction strategies by socioeconomic status: implications for policymakers. *American journal of preventive medicine*.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association 95*(450), 407–424.

Deaton, A. and D. Lubotsky (2003). Mortality, inequality and race in american cities and states. *Social science & medicine 56*(6), 1139–1153.

Dehejia, R. H. and S. Wahba (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics 84*(1), 151–161.

Delgado, M. S. and R. J. Florax (2015). Difference-in-differences techniques for spatial data: Local autocorrelation and spatial interaction. *Economics Letters 137*, 123–126.

Diamond, R. (2016). The determinants and welfare implications of us workers' diverging location choices by skill: 1980–2000. *The American Economic Review 106*(3), 479–524.

Du Mouchel, W., A. F. Williams, and P. Zador (1987). Raising the alcohol purchase age: its effects on fatal motor vehicle crashes in twenty-six states. *The Journal of Legal Studies 16*(1), 249–266.

Dube, J., D. Legros, M. Theriault, and F. Des Rosiers (2014). A spatial Difference-in-Differences estimator to evaluate the effect of change in public mass transit systems on house prices. *Transportation Research Part B: Methodological 64*, 24–40.

Durlauf, S. N. and Y. M. Ioannides (2010). Social interactions. *Annu. Rev. Econ. 2*(1), 451–478.

Egger, P. H. and A. Lassmann (2015). The causal impact of common native language on international trade: Evidence from a spatial regression discontinuity design. *The Economic Journal 125*(584), 699–745.

Egyedi, T. M. and D. C. Mehos (2012). *Inverse Infrastructures: Disrupting networks from below*. Edward Elgar Publishing.

Egyedi, T. M., J. L. Vrancken, and J. Ubacht (2007). Inverse infrastructures: Coordination in self-organizing systems. In *Standardization and Innovation in Information Technology, 2007. SIIT 2007. 5th International Conference on*, pp. 23–36. IEEE.

Elhorst, J. P. (2003). Specification and estimation of spatial panel data models. *International regional science review 26*(3), 244–268.

Fang, S., G. Z. Gertner, Z. Sun, and A. A. Anderson (2005). The impact of interactions in spatial simulation of the dynamics of urban sprawl. *Landscape and urban planning 73*(4), 294–306.

Fisher, F. M. (1966). *The identification problem in econometrics*. McGraw-Hill.

Fisher, J., D. S. Johnson, and T. M. Smeeding (2015). Inequality of income and consumption in the us: measuring the trends in inequality from 1984 to 2011 for the same individuals. *Review of Income and Wealth 61*(4), 630–650.

Flanagan, B. E., E. W. Gregory, E. J. Hallisey, J. L. Heitgerd, and B. Lewis (2011). A social vulnerability index for disaster management. *Journal of Homeland Security and Emergency Management 8*(1).

Folch, D. C., D. Arribas-Bel, J. Koschinsky, and S. E. Spielman (2014). Uncertain uncertainty: Spatial variation in the quality of american community survey estimates.

for Disease Control, C. and Prevention (2015). Community health assessment for population health improvement: Resource of most frequently recommended health outcomes and determinants. Technical report.

for Disease Control, C., A. f. T. S. Prevention, A. Disease Registry, Geospatial Research, and S. Program (2014). Social vulnerability index.

Foster, I., R. Ghani, R. S. Jarmin, F. Kreuter, and J. Lane (2017). *Big Data and Social Science–A Practical Guide to Methods and Tools*.

Foster, I. and C. Kesselman (1999). The globus toolkit. *The grid: blueprint for a new computing infrastructure*, 259–278.

Foster, I., C. Kesselman, and S. Tuecke (2001). The anatomy of the grid: Enabling scalable virtual organizations. *International journal of high performance computing applications 15*(3), 200–222.

Frank, L., P. Engelke, and T. Schmid (2003). *Health and community design: The impact of the built environment on physical activity*. Island Press.

Frank, L. D. and P. O. Engelke (2001). The built environment and human activity patterns: exploring the impacts of urban form on public health. *Journal of Planning Literature 16*(2), 202–218.

Franzese Jr, R. J. and J. C. Hays (2009). Empirical modeling of spatial interdependence in time-series cross-sections. In *Methoden der vergleichenden Politik-und Sozialwissenschaft*, pp. 233–261. Springer.

Fuller, S. (2011). From intervention informatics to prevention informatics. *Bulletin of the American Society for Information Science and Technology 38*(1), 36–41.

Gangl, M. (2010). Causal inference in sociological research. *Annual Review of Sociology 36*, 21–47.

Garfinkel, I., C. F. Manski, and C. Michalopoulos (1992). 7 Micro Experiments and Macro Effects. *Evaluating welfare and training programs*, 253.

Gelman, A. and A. Zelizer (2015). Evidence on the deleterious impact of sustained use of polynomial regression on causal inference. *Research & Politics 2*(1), 2053168015569830.

Giacomelli, S. and C. Menon (2012). Firm size and judicial efficiency in italy: evidence from the neighbour's tribunal.

Gibbons, S., H. G. Overman, and E. Patacchini (2014). Spatial methods.

Gittelsohn, J. and S. Sharma (2009). Physical, consumer, and social aspects of measuring the food environment among diverse low-income populations. *American journal of preventive medicine 36*(4), S161–S165.

Gonzalez, H., A. Y. Halevy, C. S. Jensen, A. Langen, J. Madhavan, R. Shapley, W. Shen, and J. Goldberg-Kidon (2010). Google fusion tables: web-centered data management and collaboration. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pp. 1061–1066. ACM.

Gordon-Larsen, P., M. C. Nelson, P. Page, and B. M. Popkin (2006). Inequality in the built environment underlies key health disparities in physical activity and obesity. *Pediatrics 117*(2), 417–424.

Granger, C. W. (1988). Some recent development in a concept of causality. *Journal of econometrics 39*(1), 199–211.

Greenland, S. (2000). Causal analysis in the health sciences. *Journal of the American Statistical Association 95*(449), 286–289.

Griffith, D. A. (1987). Spatial autocorrelation. *A Primer (Washington, DC, Association of American Geographers)*.

Grus, L., J. Crompvoets, and A. Bregt (2010). Spatial data infrastructures as complex adaptive systems. *International Journal of Geographical Information Science 24*(3), 439–463.

Grus, L., J. Crompvoets, and A. K. Bregt (2007). Multi-view sdi assessment framework. *International Journal of Spatial Data Infrastructures Research 2*, 33–53.

Guo, S. and M. Fraser (2010). *Propensity score analysis: Statistical methods and analysis*. Thousand Oaks, CA: Sage.

Gutiérrez, O. M. (2015). Contextual poverty, nutrition, and chronic kidney disease. *Advances in chronic kidney disease 22*(1), 31–38.

Güting, R. H. (1994). An introduction to spatial database systems. *The VLDB Journal?The International Journal on Very Large Data Bases 3*(4), 357–399.

Hainmueller, J., D. J. Hopkins, and T. Yamamoto (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis 22*(1), 1–30.

Harvey, F. (2013). A new age of discovery: The post-gis era.

Hausman, J. and C. Palmer (2012). Heteroskedasticity-robust inference in finite samples. *Economics Letters 116*(2), 232–235.

Heckman, J. J. (1979). *Statistical models for discrete panel data*. Department of Economics and Graduate School of Business, University of Chicago.

Heckman, J. J. (2005). The scientific model of causality. *Sociological methodology 35*(1), 1–97.

Heckman, J. J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. Technical report, National Bureau of Economic Research.

Heckman, J. J. and R. Robb (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of econometrics 30*(1), 239–267.

Heckman, J. J. and E. J. Vytlacil (2007). Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. *Handbook of econometrics 6*, 4779–4874.

Hendrick, R., M. Luby, and J. M. Terzakis (2010). The great recession's impact on the city of chicago. *Municipal Finance Journal*.

Hendrickson, D., C. Smith, and N. Eikenberry (2006). Fruit and vegetable access in four low-income food deserts communities in minnesota. *Agriculture and Human Values 23*(3), 371–383.

Hendriks, P. H., E. Dessers, and G. Van Hootegem (2012). Reconsidering the definition of a spatial data infrastructure. *International journal of geographical information science 26*(8), 1479–1494.

Herrera, M., M. Ruiz, and J. Mur (2013). Detecting dependence between spatial processes. *Spatial Economic Analysis 8*(4), 469–497.

Ho, D. E., K. Imai, G. King, and E. A. Stuart (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis 15*(3), 199–236.

Holmes, T. J. (1998). The effect of state policies on the location of manufacturing: Evidence from state borders. *Journal of Political Economy 106*(4), 667–705.

Hoover, K. D. (2004). Lost causes. *Journal of the History of Economic Thought 26*(2), 149–164.

Horowitz, C. R., K. A. Colson, P. L. Hebert, and K. Lancaster (2004). Barriers to buying healthy foods for people with diabetes: evidence of environmental disparities. *American Journal of Public Health 94*(9), 1549–1554.

Horowitz, J. L. and C. F. Manski (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American statistical Association 95*(449), 77–84.

Hu, S. and T. Dai (2013). Online map application development using google maps api, sql database, and asp .net. *International Journal of Information and Communication Technology Research 3*(3).

Hujer, R., P. J. Rodrigues, and K. Wolf (2009). Estimating the macroeconomic effects of active labour market policies using spatial econometric methods. *International Journal of Manpower 30*(7), 648–671.

Imbens, G., T. Barrios, R. Diamond, and M. Kolesar (2011). Clustering, Spatial Correlations and Randomization Inference.

Imbens, G. W. and T. Lemieux (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics 142*(2), 615–635.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Imbens, G. W. and J. M. Wooldridge (2009). Recent developments in the econometrics of program evaluation. In *Journal of Economic Literature*. Citeseer.

Jalil, A. J. and others (2014). Monetary intervention really did mitigate banking panics during the Great Depression: Evidence along the Atlanta Federal Reserve District border. *The Journal of Economic History 74*(01), 259–273.

Jaskiewicz, L. J. (2010). *Potential spatial access to supermarkets: Does the measure matter?* University of Illinois at Chicago, Health Sciences Center.

Johnson, G. S., G. S. Birkhead, R. Block, S. Kelley, J. Coates, R. J. Campbell, and B. Fowler (2014). Public health informatics in high population states: New york and ohio. In *Public Health Informatics and Information Systems*, pp. 531–554. Springer.

Karpyn, A., M. Manon, S. Treuhaft, T. Giang, C. Harries, and K. McCoubrey (2010). Policy solutions to the ?grocery gap? *Health Affairs 29*(3), 473–480.

Kaufman, P. R. (1999). Rural poor have less access to supermarkets, large grocery stores. *Rural Development Perspectives 13*, 19–26.

Keele, L. J. and R. Titiunik (2014). Geographic boundaries as regression discontinuities. *Political Analysis*, mpu014.

Kelejian, H. H. and I. R. Prucha (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics 157*(1), 53–67.

King, B. (2010). Political ecologies of health. *Progress in Human Geography 34*(1), 38–55.

King, G. (2007). An introduction to the dataverse network as an infrastructure for data sharing.

Koschinsky, J. (2013). The case for spatial analysis in evaluation to reduce health inequities. *Evaluation and program planning 36*(1), 172–176.

Krieger, N. (2003). Does racism harm health? did child abuse exist before 1962? on explicit questions, critical science, and current controversies: an ecosocial perspective. *American journal of public health 93*(2), 194–199.

Krieger, N. (2011). *Epidemiology and the people's health: theory and context*. Oxford University Press.

Lake, A. and T. Townshend (2006). Obesogenic environments: exploring the built and food environments. *The Journal of the Royal society for the Promotion of Health 126*(6), 262–267.

Landrine, H. and I. Corral (2009). Separate and unequal: residential segregation and black health disparities. *Ethnicity & disease 19*(2), 179.

Laney, D. (2001). 3d data management: Controlling data volume, velocity. Technical report, and variety. Technical report, META Group.

Langford, M. and G. Higgs (2006). Measuring potential access to primary healthcare services: the influence of alternative spatial representations of population. *The Professional Geographer 58*(3), 294–306.

Lee, D. S. and T. Lemieux (2009). Regression discontinuity designs in economics. Technical report, National Bureau of Economic Research.

Lee, L.-f. (2007). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics 140*(2), 333–374.

Lee, L.-f. and J. Yu (2010). Some recent developments in spatial panel data models. *Regional Science and Urban Economics 40*(5), 255–271.

Levins, R. and C. Lopez (1999). Toward an ecosocial view of health. *International Journal of Health Services 29*(2), 261–293.

Little, R. J. and D. B. Rubin (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual review of public health 21*(1), 121–145.

Lorenc, T., M. Petticrew, V. Welch, and P. Tugwell (2013). What types of interventions generate inequalities? evidence from systematic reviews. *Journal of epidemiology and community health 67*(2), 190–193.

MacEachren, A. M., S. Crawford, M. Akella, and G. Lengerich (2008). Design and implementation of a model, web-based, gis-enabled cancer atlas. *The Cartographic Journal 45*(4), 246–260.

MacKinnon, J. G. and H. White (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics 29*(3), 305–325.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies 60*(3), 531–542.

Manski, C. F. (2000). Economic analysis of social interactions. Technical report, National bureau of economic research.

Mansourian, A., A. Lubida, P. Pilesjö, E. Abdolmajidi, and M. Lassi (2015). Sdi planning using the system dynamics technique within a community of practice: lessons learnt from tanzania. *Geo-spatial Information Science 18*(2-3), 97–110.

Marks, A. K., K. Ejesi, and C. García Coll (2014). Understanding the us immigrant paradox in childhood and adolescence. *Child Development Perspectives 8*(2), 59–64.

Massey, D. S., J. Rothwell, and T. Domina (2009). The changing bases of segregation in the united states. *The Annals of the American Academy of Political and Social Science 626*(1), 74–90.

Mayer, J. D. (1996). The political ecology of disease as one new focus for medical geography. *Progress in Human Geography 20*(4), 441–456.

Mayer-Schönberger, V. and K. Cukier (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McCartt, A. T., L. A. Hellinga, and B. B. Kirley (2010). The effects of minimum legal drinking age 21 laws on alcohol-related driving in the united states. *Journal of Safety Research 41*(2), 173–181.

McKinnon, R. A., J. Reedy, M. A. Morrissette, L. A. Lytle, and A. L. Yaroch (2009). Measures of the food environment: a compilation of the literature, 1990–2007. *American journal of preventive medicine 36*(4), S124–S133.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of business & economic statistics 13*(2), 151–161.

Meyer, B. D., W. K. Viscusi, and D. L. Durbin (1990). Workers' compensation and injury duration: Evidence from a natural experiment. Technical report, National Bureau of Economic Research.

Millo, G. and G. Piras (2012). splm: Spatial panel data models in r. *Journal of Statistical Software 47*(1), 1–38.

Millo, G., G. Piras, et al. (2012). splm: Spatial panel data models in r. *Journal of Statistical Software 47*(1), 1–38.

Milner, C. and J. Milner (2016). Impact of policy on physical activity participation and where we need to go. *Annual Review of Gerontology and Geriatrics 36*(1), 1–32.

Mirto, M., S. Fiore, L. Conte, L. V. Bruno, and G. Aloisio (2016). A spatial data analysis infrastructure for environmental health research. In *High Performance Computing & Simulation (HPCS), 2016 International Conference on*, pp. 435–442. IEEE.

Mitchell, M. and M. Newman (2001). Complex systems theory and evolution.

Mobley, L. R., H. Frech III, and L. Anselin (2004). Spatial Interaction, Hospital Pricing and Hospital Antitrust. *Urbana 51*, 61801.

Mobley, L. R., L. Watson, and G. G. Brown (2012). Geographic disparities in late-stage cancer diagnosis: Multilevel factors and spatial interactions. *Health and Place 5*(18), 978–990.

Moffitt, R. (2005). Remarks on the analysis of causal relationships in population research. *Demography 42*(1), 91–108.

Moore, L. V. and A. V. Diez Roux (2006). Associations of neighborhood characteristics with the location and type of food stores. *American journal of public health 96*(2), 325–331.

Moore, L. V., A. V. D. Roux, J. A. Nettleton, and D. R. Jacobs (2008). Associations of the local food environment with diet quality?a comparison of assessments based on surveys and geographic information systems the multi-ethnic study of atherosclerosis. *American journal of epidemiology 167*(8), 917–924.

Morgan, S. L. and C. Winship (2014). *Counterfactuals and causal inference*. Cambridge University Press.

Morland, K. and S. Filomena (2007). Disparities in the availability of fruits and vegetables between racially segregated urban neighbourhoods. *Public health nutrition 10*(12), 1481–1489.

Morland, K., S. Wing, A. D. Roux, and C. Poole (2002). Neighborhood characteristics associated with the location of food stores and food service places. *American journal of preventive medicine 22*(1), 23–29.

Mueller, K., C. MacKinney, M. Gutierrez, and J. Richgels (2011). Place based policies and public health: The road to healthy rural people and places.

Mur, J., M. Herrera, M. Ruiz, and others (2011). Selecting the W Matrix: Parametric vs Non Parametric Approaches. In *ERSA conference papers*. European Regional Science Association.

Mutl, J. and M. Pfaffermayr (2011). The hausman test in a cliff and ord panel model. *The Econometrics Journal 14*(1), 48–76.

Naik, N., J. Philipoom, R. Raskar, and C. Hidalgo (2014). Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 779–785.

Norberg, K. E., L. J. Bierut, and R. A. Grucza (2009). Long-term effects of minimum drinking age laws on past-year alcohol and drug use disorders. *Alcoholism: Clinical and Experimental Research 33*(12), 2180–2190.

Obe, R. O. and L. S. Hsu (2015). *PostGIS in action*. Manning Publications Co.

Openshaw, S. (1984). Ecological fallacies and the analysis of areal census data. *Environment and Planning A 16*(1), 17–31.

Pace, R. K. and J. P. LeSage (2008). A spatial hausman test. *Economics Letters 101*(3), 282–284.

Pastor, M. and R. Morello-Frosch (2014). Integrating public health and community development to tackle neighborhood distress and promote well-being. *Health Affairs 33*(11), 1890–1896.

Pearl, J. (2001). Causal inference in the health sciences: a conceptual introduction. *Health services and outcomes research methodology 2*(3-4), 189–220.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys 3*, 96–146.

Pearl, J. (2010). An introduction to causal inference. *The international journal of biostatistics 6*(2).

Peltzman, S. (2009). Mortality inequality. *The Journal of Economic Perspectives 23*(4), 175–190.

People, H., U. D. of Health, and H. Services (2000). Healthy people 2010.

Phelan, J. C., B. G. Link, and P. Tehranifar (2010). Social conditions as fundamental causes of health inequalities theory, evidence, and policy implications. *Journal of health and social behavior 51*(1 suppl), S28–S40.

Pickett, K. E. and R. G. Wilkinson (2015). Income inequality and health: a causal review. *Social Science & Medicine 128*, 316–326.

Pleasant, A. (2013). Health literacy around the world: Part 1. health literacy efforts outside of the united states. *Health Literacy*.

Powell, L. M., M. C. Auld, F. J. Chaloupka, P. M. O?Malley, and L. D. Johnston (2006). Access to fast food and food prices: relationship with fruit and vegetable consumption and overweight among adolescents. In *The economics of obesity*, pp. 23–48. Emerald Group Publishing Limited.

Powell, L. M., S. Slater, D. Mirtcheva, Y. Bao, and F. J. Chaloupka (2007). Food store availability and neighborhood characteristics in the united states. *Preventive medicine 44*(3), 189–195.

Provost, F. and T. Fawcett (2013). Data science and its relationship to big data and data-driven decision making. *Big Data 1*(1), 51–59.

QGis, D. (2011). Quantum gis geographic information system. *Open Source Geospatial Foundation Project 45*.

Raghupathi, W. and V. Raghupathi (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems 2*(1), 3.

Rajabifard, A. and I. P. Williamson (2001). Spatial data infrastructures: concept, sdi hierarchy and future directions.

Rey, S. J. (2014). Open regional science. *The Annals of Regional Science 52*(3), 825–837.

Richardson, D. B., N. D. Volkow, M.-P. Kwan, R. M. Kaplan, M. F. Goodchild, and R. T. Croyle (2013). Spatial turn in health research. *Science (New York, NY) 339*(6126), 1390.

Robinson, W. S. (2009). Ecological correlations and the behavior of individuals. *International journal of epidemiology 38*(2), 337–341.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Rosenbaum, S. (2016). Hospital community benefit spending: Leaning in on the social determinantsof health. *The Milbank Quarterly 94*(2), 251–254.

Rothman, K. and S. Greenland (2005). Causation and causal inference in epidemiology. *American journal of public health 95*(S1), S144–S150.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688.

Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge University Press.

Rushton, G. and P. Lolonis (1996). Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine 15*(7-9), 717–726.

Ryscavage, P. (2015). *Income inequality in America: An analysis of trends*. Routledge.

Sadler, R. C., J. A. Gilliland, and G. Arku (2011). An application of the edge effect in measuring accessibility to multiple food retailer types in southwestern ontario, canada. *International Journal of Health Geographics 10*(1), 34.

Schieb, L. J., L. R. Mobley, M. George, and M. Casper (2013). Tracking stroke hospitalization clusters over time and associations with county-level socioeconomic and healthcare characteristics. *Stroke 44*(1), 146–152.

Schutt, R. and C. O'Neil (2013). *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc.".

Schutte, S. and K. Donnay (2014). Matched wake analysis: finding causal relationships in spatiotemporal event data. *Political Geography 41*, 1–10.

Scott, S. L. (2014). Carpe datum! comment on 'data science: An action plan for expanding the technical areas of the field of statistics'. *Statistical Analysis and Data Mining 7*(6), 418–419.

Shadish, W. R., T. D. Cook, and D. T. Campbell (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning.

Shah, D. V., J. N. Cappella, W. R. Neuman, M. Crosas, G. King, J. Honaker, and L. Sweeney (2015). Automating open science for big data. *The ANNALS of the American Academy of Political and Social Science 659*(1), 260–273.

Shah, S. N., E. T. Russo, T. R. Earl, and T. Kuo (2014). Peer reviewed: Measuring and monitoring progress toward health equity: Local challenges for public health. *Preventing chronic disease 11*.

Shih, M., K. Dumke, M. Goran, and P. Simon (2013). The association between community-level economic hardship and childhood obesity prevalence in los angeles. *Pediatric obesity 8*(6), 411–417.

Shults, R. A., R. W. Elder, D. A. Sleet, J. L. Nichols, M. O. Alao, V. G. Carande-Kulis, S. Zaza, D. M. Sosin, R. S. Thompson, and T. F. on Community Preventive Services (2001). Reviews of evidence regarding interventions to reduce alcohol-impaired driving. *American journal of preventive medicine 21*(4), 66–88.

Smith, E. R. and F. R. Conrey (2007a). Agent-based modeling: A new approach for theory building in social psychology. *Personality and social psychology review 11*(1), 87–104.

Smith, E. R. and F. R. Conrey (2007b). Agent-based modeling: A new approach for theory building in social psychology. *Personality and social psychology review 11*(1), 87–104.

Smith, H. L. (2003). Some thoughts on causation as it relates to demography and population studies. *Population and Development Review*, 459–469.

Smoyer-Tomic, K. E., J. C. Spence, and C. Amrhein (2006). Food deserts in the prairies? supermarket accessibility and neighborhood need in edmonton, canada. *The Professional Geographer 58*(3), 307–326.

Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association 101*(476), 1398–1407.

Soranno, P. A., E. G. Bissell, K. S. Cheruvelil, S. T. Christel, S. M. Collins, C. E. Fergus, C. T. Filstrup, J.-F. Lapierre, N. R. Lottig, S. K. Oliver, et al. (2015). Building a multi-scaled geospatial temporal ecology database from disparate data sources: fostering open science and data reuse. *GigaScience 4*(1), 28.

Spielman, S. E., D. Folch, and N. Nagle (2014). Patterns and causes of uncertainty in the american community survey. *Applied Geography 46*, 147–157.

Sprott, D. and L. Wilkes (2004). Understanding service-oriented architecture. *The Architecture Journal 1*(1), 10–17.

StatExtracts, O. (2015). Quarterly national accounts: Quarterly growth rates of real gdp, change over previous quarter.

Suarez-Balcazar, Y., M. Hellwig, J. Kouba, L. Redmond, L. Martinez, D. Block, C. Kohrman, and W. Peterman (2006). The making of an interdisciplinary partnership: the case of the chicago food system collaborative. *American journal of community psychology 38*(1-2), 113–123.

Talbot, T. O., M. Kulldorff, S. P. Forand, and V. B. Haley (2000). Evaluation of spatial filters to create smoothed maps of health data. *Statistics in medicine 19*(1718), 2399–2408.

Talen, E. (2003). Neighborhoods as service providers: a methodology for evaluating pedestrian access. *Environment and Planning B: Planning and Design 30*(2), 181–200.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, 234–240.

Tsou, M.-H. and B. P. Buttenfield (2002). A dynamic architecture for distributing geographic information services. *Transactions in GIS 6*(4), 355–381.

Vart, P., R. T. Gansevoort, M. M. Joosten, U. Bültmann, and S. A. Reijneveld (2015). Socioeconomic disparities in chronic kidney disease: a systematic review and meta-analysis. *American journal of preventive medicine 48*(5), 580–592.

Vassiliadis, P. and A. Simitsis (2009). Extraction, transformation, and loading. In *Encyclopedia of Database Systems*, pp. 1095–1101. Springer.

Ver Ploeg, M. (2010). *Access to affordable and nutritious food: measuring and understanding food deserts and their consequences: report to Congress*. DIANE Publishing.

Vitolo, C., Y. Elkhatib, D. Reusser, C. J. Macleod, and W. Buytaert (2015). Web technologies for environmental big data. *Environmental Modelling & Software 63*, 185–198.

Vree, W. G. (2003). Internet en rijkswaterstaat: een ict-infrastructuur langs water en wegen.

Wagenaar, A. C. and T. L. Toomey (2002). Effects of minimum drinking age laws: review and analyses of the literature from 1960 to 2000. *Journal of Studies on Alcohol, Supplement* (14), 206–225.

Walker, R. E., C. R. Keane, and J. G. Burke (2010). Disparities and access to healthy food in the united states: A review of food deserts literature. *Health & place 16*(5), 876–884.

Waller, L. A. and C. A. Gotway (2004). *Applied spatial statistics for public health data*, Volume 368. John Wiley & Sons.

Wennberg, J. E. (1996). *The Dartmouth Atlas of Health Care in the United States (incl. Diskette)*. American Hospital Association.

White, M., J. Adams, and P. Heywood (2009). How and why do interventions that increase health overall widen inequalities within populations. *Social inequality and public health*, 65–82.

Whitman, S., J. Orsi, and M. Hurlbert (2012). The racial disparity in breast cancer mortality in the 25 largest cities in the united states. *Cancer epidemiology 36*(2), e147–e151.

Wolfram, S. (1985). Complex systems theory. *Princeton: The Institute for Advanced Study*.

Wrigley, N., D. Warm, and B. Margetts (2003). Deprivation, diet, and food-retail access: Findings from the leeds ?food deserts' study. *Environment and Planning A 35*(1), 151–188.

Yi, Q., R. E. Hoskins, E. A. Hillringhouse, S. S. Sorensen, M. W. Oberle, S. S. Fuller, and J. C. Wallace (2008). Integrating open-source technologies to build low-cost information systems for improved access to public health data. *International Journal of Health Geographics 7*(1), 29.

Zastrow, M. (2015). Science on the map. *Nature 519*(7541), 119.

Zenk, S. N. and L. M. Powell (2008). Us secondary schools and food outlets. *Health & place 14*(2), 336–346.

APPENDIX A

URBAN FOODSCAPE DYNAMICS: TRACING FOOD INEQUITY IN CHICAGO FROM 2007-2014

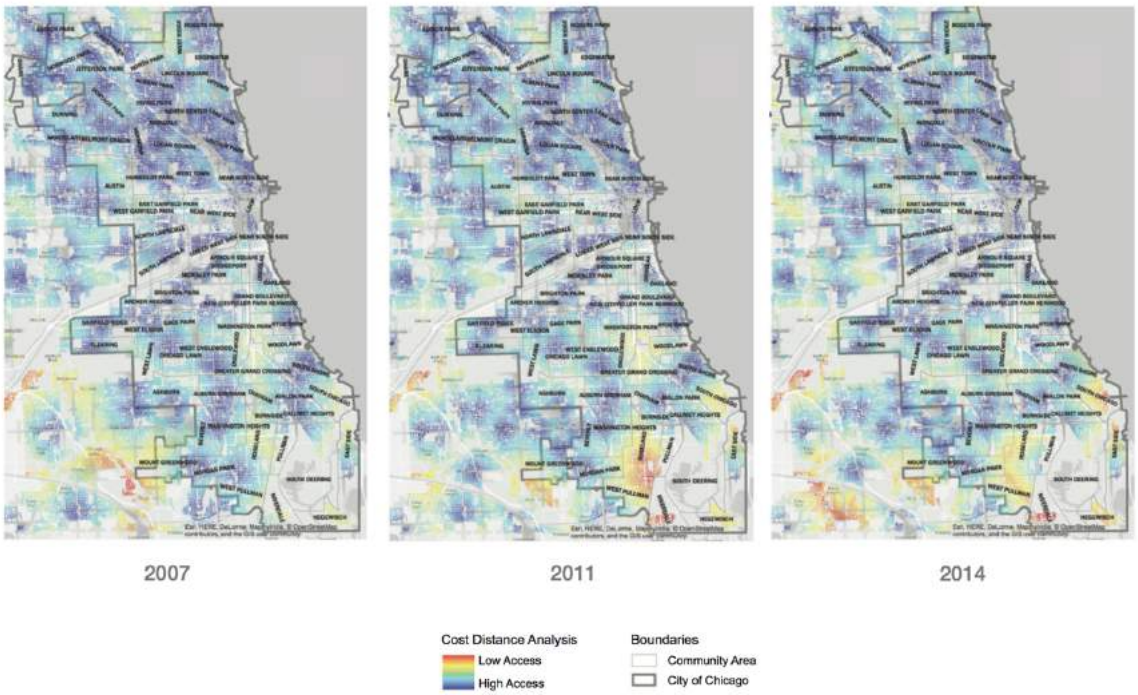


Figure 29. Cost distance calculations on residential and mixed-use street networks for each year of analysis.

A.1 ESDA Supplementary Figures

A.2 Quasi-Experimental Analysis

Category	Proportion	P*Baseline	P*Endline	Change
Treatment (B=1)	0.73	0.93	0.92	-0.01
Control (B=1)	0.23	0.19	0.18	-0.01
Treatment (B=0)	0.27	0.19	0.20	+0.01
Control (B=0)	0.77	0.55	0.52	-0.03

Table 7. Weighted Aggregate DID Results

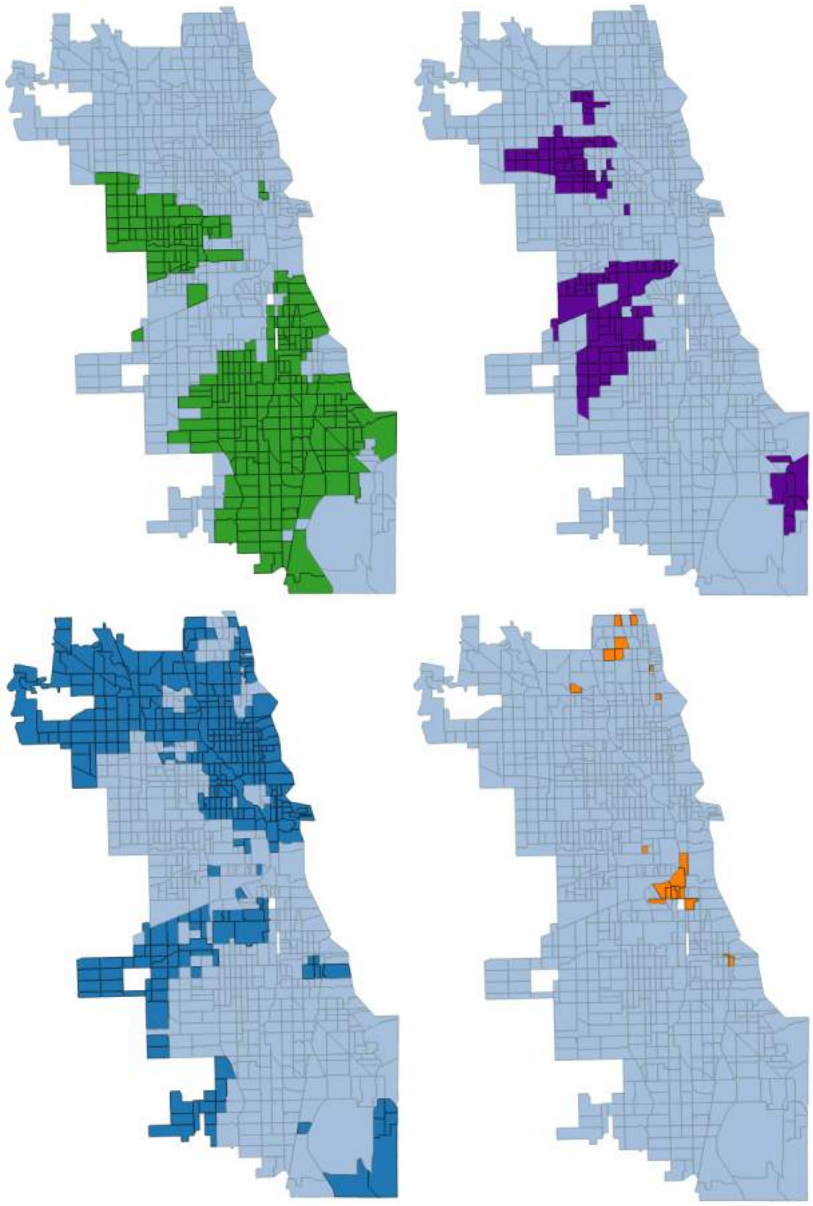


Figure 30. Tracts with stable majorities pre-and post-Recession, by race and ethnicity, from left to right: (a) Black, (b) Hispanic, (c) White, and (d) Diverse (with no racial or ethnic majority)

Category	Change (miles)
Difference in Treated, Black Majority Tracts: $E(Y_{i,t=0}\|B_i = 1, D_i = 1) - E(Y_{i,t=1}\|B_i = 1, D_i = 1)$	-0.02
Difference in Treated, Black Minority Tracts: $E(Y_{i,t=0}\|B_i = 0, D_i = 1) - E(Y_{i,t=1}\|B_i = 0, D_i = 1)$	$+0.01$
Difference in Non-Treated, Black Majority Tracts: $E(Y_{i,t=0}\|B_i = 1, D_i = 0) - E(Y_{i,t=1}\|B_i = 1, D_i = 0)$	$+0.03$
Difference in Non-Treated, Black Minority Tracts: $E(Y_{i,t=0}\|B_i = 0, D_i = 0) - E(Y_{i,t=1}\|B_i = 0, D_i = 0)$	$-0.05*$
Difference in Treated and Control, Black Majority $(E(Y_{i,t=0}\|B_i = 1, D_i = 1) - E(Y_{i,t=0}\|B_i = 1, D_i = 0))$ $-(E(Y_{i,t=1}\|B_i = 1, D_i = 1) - E(Y_{i,t=1}\|B_i = 1, D_i = 0))$	$+0.03$
Difference inTreated and Control, Black Minority $(E(Y_{i,t=0}\|B_i = 0, D_i = 1) - E(Y_{i,t=1}\|B_i = 0, D_i = 0))$ $-(E(Y_{i,t=1}\|B_i = 0, D_i = 1) - E(Y_{i,t=1}\|B_i = 0, D_i = 0))$	$+0.06$

Table 8. Simpe DID Analysis Results with counterfactual specfication. If difference is significant following t-test, indicated as such following previous conventions. Note that the difference between Treatment and Control groups was already shown to be significant in Section 3.1

A.3 Full Model Results

```
>>06/12/2017 10:15:29 PM
REGRESSION (DIFF-IN-DIFF, COMPARE REGIMES AND TIME PERIOD)
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data Set            :  FA-Finalff-only
Dependent Variable  :  Group (time 0,time 1)
Number of Observations: 1530
Mean dependent var  :     0.87213  Number of Variables   :    4
S.D. dependent var  :    0.510561  Degrees of Freedom    : 1526

R-squared           :    0.069975  F-statistic           :     38.2719
Adjusted R-squared  :    0.068146  Prob(F-statistic)     :  7.6139e-24
Sum squared residual:     370.921  Log likelihood        :    -1086.95
Sigma-square        :    0.243068  Akaike info criterion :     2181.89
S.E. of regression  :    0.493019  Schwarz criterion     :     2203.22
Sigma-square ML     :    0.242432
S.E of regression ML:    0.492374

------------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------------
          CONSTANT       0.805987      0.0206142        39.0987      0.00000
             SPACE       0.300735       0.041041        7.32768      0.00000
    Ttime 0_time 1     -0.0241385      0.0291528      -0.827997      0.40780
          INTERACT      0.0185545      0.0580407       0.319681      0.74931
------------------------------------------------------------------------------
```

Figure 31. Aggregage DID OLS

```
In [158]: pooledNSE<- spml(FA~  LINC + BlkD + WhtD + HispD + FRisk,data=index1,listw=wr,model="pooling", lag=FALSE,spatial.error="none")
          summary(pooledNSE)

          ML panel with iid errors

          Call:
          spreml(formula = formula, data = data, index = index, w = listw2mat(listw),
              w2 = listw2mat(listw2), lag = lag, errors = errors, cl = cl)

          Residuals:
             Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
          -0.9490 -0.2120  0.0133  0.0000  0.2100  1.0400

          Coefficients:
                       Estimate Std. Error  t-value  Pr(>|t|)
          (Intercept) -3.2212841  0.1444849 -22.2949 < 2.2e-16 ***
          LINC        -0.1064681  0.0301794  -3.5278  0.000419 ***
          BlkD         0.2125689  0.0286360   7.4231 1.144e-13 ***
          WhtD         0.0077438  0.0245971   0.3148  0.752894
          HispD       -0.1083539  0.0229072  -4.7301 2.244e-06 ***
          FRisk        0.2767047  0.0241404  11.4623 < 2.2e-16 ***
          ---
          Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 32. Pooled OLS

```
In [41]: pooledSE<- spml(FA~  LINC + BlkD + WhtD + HispD + FRisk,data=index1,listw=wr,model="pooling", lag=TRUE,spatial.error="none")
summary(pooledSE)

ML panel with spatial lag and iid errors

Call:
spreml(formula = formula, data = data, index = index, w = listw2mat(listw),
    w2 = listw2mat(listw2), lag = lag, errors = errors, cl = cl)

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -3.60   -2.80   -2.55   -2.56   -2.33   -1.42

Spatial autoregressive coefficient:
       Estimate Std. Error t-value  Pr(>|t|)
lambda 0.705347   0.021827  32.316 < 2.2e-16 ***

Coefficients:
             Estimate Std. Error t-value  Pr(>|t|)
(Intercept) -0.814972   0.111762 -7.2920 3.054e-13 ***
LINC        -0.062081   0.023344 -2.6593 0.0078294 **
BlkD         0.054858   0.022151  2.4766 0.0132643 *
WhtD         0.023283   0.019026  1.2237 0.2210601
HispD       -0.061198   0.017719 -3.4538 0.0005528 ***
FRisk        0.200747   0.018673 10.7506 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 33. Pooled with Spatial Lag

```
In [166]: fixedNSEtime<-plm(fm,data=index1,model="within",effect="time")
summary(fixedNSEtime)

Oneway (time) effect Within Model

Call:
plm(formula = fm, data = index1, effect = "time", model = "within")

Balanced Panel: n=790, T=2, N=1580

Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
-0.9330 -0.2130  0.0161  0.2070  1.0500

Coefficients :
        Estimate Std. Error t-value  Pr(>|t|)
LINC  -0.092700   0.030711 -3.0184  0.002582 **
BlkD   0.214833   0.028660  7.4958 1.095e-13 ***
WhtD   0.012294   0.024676  0.4982  0.618385
HispD -0.104852   0.022960 -4.5668 5.338e-06 ***
FRisk  0.303063   0.026451 11.4574 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    198.73
Residual Sum of Squares: 145.15
R-Squared:      0.26965
Adj. R-Squared: 0.26845
F-statistic: 116.15 on 5 and 1573 DF, p-value: < 2.22e-16
```

Figure 34. Fixed Effects Model

```
In [167]: fixedSEtime<- spml(fm,data=test,listw=wr,model="within",effect="time",
                      lag=TRUE,spatial.error="none")
summary(fixedSEtime)
```

```
Spatial panel fixed effects lag model


Call:
spml(formula = fm, data = test, listw = wr, model = "within",
    effect = "time", lag = TRUE, spatial.error = "none")

Residuals:
   Min. 1st Qu.  Median 3rd Qu.    Max.
-0.3700 -0.0723 -0.0160  0.0439  0.9490

Spatial autoregressive coefficient:
       Estimate Std. Error t-value  Pr(>|t|)
lambda 0.740504   0.020746  35.694 < 2.2e-16 ***

Coefficients:
       Estimate Std. Error t-value  Pr(>|t|)
LINC  -0.0414729  0.0129513 -3.2022 0.0013637 **
BlkD   0.0226755  0.0124638  1.8193 0.0688635 .
WhtD   0.0021237  0.0105432  0.2014 0.8403663
HispD -0.0336087  0.0098468 -3.4131 0.0006422 ***
FRisk  0.0645053  0.0106053  6.0824 1.184e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 35. Fixed Effects with Spatial Lag

```
In [44]: randomNSE<- spml(fm,data=index1,listw=wr,model="random",effect="individual",
                    lag=FALSE,spatial.error="none", method="ec2sls")
summary(randomNSE)
```

```
ML panel with , random effects

Call:
spreml(formula = formula, data = data, index = index, w = listw2mat(listw),
    w2 = listw2mat(listw2), lag = lag, errors = errors, cl = cl,
    method = "ec2sls")

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
-0.98900 -0.22300  0.00891  0.21200  1.21000

Error variance parameters:
    Estimate Std. Error t-value  Pr(>|t|)
phi  5.21234    0.41937  12.429 < 2.2e-16 ***

Coefficients:
             Estimate Std. Error  t-value  Pr(>|t|)
(Intercept) -3.562451   0.088434 -40.2839 < 2.2e-16 ***
LINC        -0.028988   0.018293  -1.5847 0.1130330
BlkD         0.213073   0.022487   9.4753 < 2.2e-16 ***
WhtD        -0.010386   0.015406  -0.6741 0.5002172
HispD       -0.061295   0.017620  -3.4787 0.0005039 ***
FRisk        0.103519   0.012199   8.4860 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 36. Random Effects Model

```
In [130]: randomLagML<- spml(fm,data=index1,listw=wr,model="random",
                    lag=TRUE,spatial.error="none", method="ec2sls")
summary(randomLagML)
```

```
ML panel with spatial lag, random effects

Call:
spreml(formula = formula, data = data, index = index, w = listw2mat(listw),
    w2 = listw2mat(listw2), lag = lag, errors = errors, cl = cl,
    method = "ec2sls")

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -3.63   -2.83   -2.58   -2.58   -2.34   -1.32

Error variance parameters:
    Estimate Std. Error t-value  Pr(>|t|)
phi  5.06629    0.40316  12.566 < 2.2e-16 ***

Spatial autoregressive coefficient:
       Estimate Std. Error t-value  Pr(>|t|)
lambda 0.708809   0.022136  32.021 < 2.2e-16 ***

Coefficients:
             Estimate Std. Error  t-value  Pr(>|t|)
(Intercept) -1.0334549  0.0690211 -14.9730 < 2.2e-16 ***
LINC        -0.0089207  0.0142797  -0.6247   0.53216
BlkD         0.0746132  0.0174819   4.2680 1.972e-05 ***
WhtD        -0.0015918  0.0120211  -0.1324   0.89466
HispD       -0.0320473  0.0137269  -2.3346   0.01956 *
FRisk        0.0534015  0.0095307   5.6031 2.106e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 37. Random Effects with Spatial Lag

```
In [99]: test3<-sphtest(fm,data=index1, listw = wr, spatial.model = "lag",model="within")
test3
```

```
        Hausman test for spatial models

data:  x
chisq = 41.711, df = 5, p-value = 6.74e-08
alternative hypothesis: one model is inconsistent
```

Figure 38. Spatial Hausmann Test

APPENDIX B

TOWARDS A SPATIAL DATA SCIENCE INFRASTRUCTURE IN PUBLIC HEALTH INFORMATICS

B.1 Data Dictionary

The following data was identified as necessary for all of these approaches, and thus incorporated in the final data system. The finest spatial resolution available is indicated, as well as the data source:

1. Social Determinants of Health
    - I Demographic Characteristics - from ACS 2014 5-year average
        - i. Population (tract level)
        - ii. Age (tract level)
        - iii. Sex (tract level)
        - iv. Race and Ethnicity (tract level)
    - II Socioeconomic Characteristics - from ACS 2014 5-year average
        - i. Median Income Level (tract level)
        - ii. Female Single Head of Household (tract level)
        - iii. High School graduate (tract level)
    - III Crime Statistics
        - i. Property Crime (tract level) - Chicago Data Portal API
        - ii. Violent Crime (tract level) - Chicago Data Portal API
    - IV Childhood Opportunity Index (tract level) - CDPH
    - V Economic Hardship Index (tract level) - CDPH
    - VI Health Literacy Index (tract level) - UNC at Chapel Hill
    - VII Social Vulnerability Index (tract level) - US
2. Built Environment Characteristics
    - I Population Density (tract-level)
    - II Street Network (line) - MapZen Metro Extract, OpenStreetMap
    - III Environmental Features (line, polygon) - MapZen Metro Extract, OpenStreetMap
    - IV Transit Indices
        - i. Walk Score (point) - Walk Score API
        - ii. Bike Score (point) - Walk Score API
        - iii. Public Transit Score (point) - Walk Score API
    - V Perceived Safety (point) - MIT Street Score Project
    - VI Food Security Index (tract-level) - USDA
3. Community Assets
    - I Medical Service Providers

i. Hospitals (point) - CDPH
ii. Community Clinics (FQHC) (point) - CDPH
iii. School-based Clinics (point) - CDPH

II Emergency and Social Service Providers (point) - WH-Pilot
III Well-Being resources (point) - WH-Pilot
IV Education and Job-Training Resources - WH-Pilot
V Food Resources

i. Farmers Markets (point) - Chicago Data Portal API, WH-Pilot
ii. Produce Carts (point) - Chicago Data Portal API
iii. Grocery stores (point) - Chicago Data Portal API, WH-Pilot
iv. Food Pantries and Hot Meals (point) - WH-Pilot

4. Health Indicators and Outcomes

I Public Health Statistics

i. Premature Mortality Rates (tract level) - CDPH
ii. Select PH Indicators (community area) - Chicago Data Portal API
iii. Blood Level Screening (community area) - Chicago Data Portal API
iv. Insurance Coverage (tract level) - ACS 2015 5-year average
v. Chronic Diseases Indicators (community area) - Chicago Data Portal API
vi. Infectious Diseases Indicators (community area) - Chicago Data Portal API
vii. Reproductive Health Indicators (community area) - Chicago Data Portal API

B.1.1 QA and QI rules and protocols

The following rules were established as a basic template for the QA/QI rules of a data import. The goal of the QA/QI guide for each dataset is to look for automated steps in cleaning up the data. Each step in the process is indicated as such: for example, adding a new address field that would equal the concatenated string of other fields would be a step; deleting an empty column would be a step; adding a new field that would equal a timestamp would be a step.

Rules and Protocols. For each dataset, consider the following:

1. Note ID fields (document if primary/foreign key/unsure).
2. Search for Nulls in ID fields, and indicate if other columns should be searched for nulls, spaces, etc. This is only relevant if the value would be incorrect with a null or space.
3. Note geometry, if there is one: point, line, or polygon. If there is a facility location, address must be concatenated into one line. A new field may need to be created; please note if needed. Additionally, a vertex will need to be created from an address or lat/long; that's another step.
4. Search for empty columns, or not useful columns (some subjectivity with the latter). ID columns that should be deleted when importing.
5. Every table should have a timestamp field added.
6. PostGresSQL is a lowercase database system, so most datasets field names will have to be converted to lower case as well. Field values don't need to, unless you see something fishy.
7. Default to keeping data, if you're not sure about something.
8. Include a "Project" step in the process map. This signifies that it will be converted into the correct geographic projection system of the data warehouse.
9. Documentation: Notes for each dataset are recorded in a text editor, but then recorded as a simple process map.