BRIDGING THE DATA-THEORY DIVIDE USING REGIONALIZATION AND UNCERTAINTY FOR NEIGHBORHOOD IDENTIFICATION

by

REBECCA MASON DAVIES

B.A., Columbia University, 2010

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirement for the degree of
Master of Arts
Department of Geography
2017
This thesis entitled:
Bridging the Data-Theory Divide Using Regionalization and Uncertainty for Neighborhood Identification
written by Rebecca Mason Davies
has been approved for the Department of Geography

__________________________________________
Seth Spielman

__________________________________________
Fernando Riosmena

__________________________________________
Carson Farmer

Date ____________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.
Davies, Rebecca Mason (M.A., Geography)

Bridging the Data-Theory Divide Using Regionalization and Uncertainty for Neighborhood Identification

Thesis directed by Associate Professor Seth Spielman

Census tracts are often equated with neighborhoods to study small-scale social phenomena. The association of tracts and neighborhoods suffers from two major problems: data quality and theoretical implications. Tract-level data in the American Community Survey (ACS) contains high uncertainty measured as sampling error. Neighborhoods are inherently vague, formed from a mix of complex social, political, economic, historical, and environmental factors. This analysis attempts to reconcile statistical and geographic uncertainty accompanying the association of tracts and neighborhoods by applying Spielman & Folch’s data-driven regionalization algorithm to identify neighborhoods in the Denver metro area using ACS data. The results of the algorithm are unstable, undermining efforts to compare data-driven regions to neighborhoods as defined in theory or practice. The challenge of reconciling data and theory as demonstrated in this experiment suggests that rather than equating tracts and neighborhoods, tracts should be interpreted as zones that can be aggregated in flexible combinations.
ACKNOWLEDGEMENTS

This thesis is submitted with sincere gratitude to:

Seth Spielman, for the marvelous opportunity to pursue a master’s degree and write this thesis, and for continuously guiding me towards more nuanced understanding of the material.

David Folch, for introducing me to his and Seth’s research, and for his patience and support as I learned to navigate graduate school.

My thesis committee for past instruction, thoughtful feedback, and much good advice along the way (some of which I heeded and some of which I did not, but now see its value!).

The Geography Department and my classmates for creating a positive and fruitful learning and research environment. In particular, I will remember Professor Babs Buttenfield’s warm welcome to the Geography Department in my first semester, Professor Emily Yeh’s insightful feedback in Research Design, and Joe Tuccillo for providing a relatable example of how to succeed in graduate school.

Professors Sudhir Venkatesh and David Smiley of Columbia University, for inspiring and encouraging my early interest in people and places, and for introducing me to the research process. Also, thanks to Senator Richard Devlin for the opportunity to better understand the challenge of translating ideas into real improvements in people’s lives.

My family, Brendan, and Teeni for all other forms of support including occasional needed distractions.
TABLES

Table
1. 2011-2015 American Community Survey 5-Year Estimates for Race (variable B02001) in Census Tract 1.02, Denver County, Colorado .................................................................6
2. General Scenario Variables ..................................................................................................47
3. Tabulated Sample Regionalization Results ...........................................................................50
4. Source-Target Pair Frequency for Sample Regionalization Results ........................................51
5. Nodes and Edges Meeting Frequency Thresholds for General Scenario .........................67
6. OLS Correlation Statistics for Frequency, Degree, and Entropy ........................................77
7. Degree Summary Statistics for Variable Scenarios ...............................................................85
8. Count and Percent of Tract Pairs Meeting Frequency Threshold for All Scenarios .............86
FIGURES

1. Census Geographies...........................................................................................................20
3. Sample Regionalization Results Overlaid on Census Tracts for Denver, Colorado Metropolitan Area ........................................................................................................35
4. Denver-Aurora-Lakewood Metropolitan Statistical Area Counties .........................41
5. Denver County Statistical Neighborhoods ........................................................................42
6. Sample Regionalization Results .....................................................................................50
7. Network Graph for Sample Regionalization Results ..................................................53
8. Network Graph Illustration of Perfect Region Stability ...............................................55
9. Adjacency Matrix for Sample Regionalization Results ...............................................56
10. Sample of Six Regionalization Results ...................................................................62
11. Network Graph for General Scenario Regionalization Output ................................63
12. General Scenario Tract Degree Distribution ................................................................64
13. a. Network Graph Highlighting Tract 08005080200 ..................................................65
    b. Maximum Size Regionalization Trial Result Containing Tract 08005080200 ..........65
    c. Minimum Size Regionalization Trial Result Containing Tract 08005080200 .........65
14. a. Network Graph Highlighting Tract 08031004700 ..................................................66
    b. Maximum Size Regionalization Trial Result Containing Tract 08031004700 ..........66
    c. Minimum Size Regionalization Trial Result Containing Tract 08031004700 ..........66
15. Distribution of Edge Frequency Among Tract Pairs in the Generalization Scenario ....67
16. Network Graph for Tract Pairs that Occur in 90 or More Regionalization Trials .........69
17. Census Tracts in Denver MSA that Participate in High Frequency Pairs .................70
18. Network Graph for Tract Pairs that Occur in 60 or More Regionalization Trials ........71
19. Census Tracts in Denver MSA that Participate in Moderate Frequency Pairs ..........72
20. a. Cluster B Map View ..................................................................................................73
    b. Cluster B Network Graph View ................................................................................73
21. a. Cluster C Map View ..................................................................................................73
    b. Cluster C Network Graph View ................................................................................73
22. Boxplot of Entropy for Census Tracts in General Scenario ........................................74
23. Entropy Map of Denver MSA, General Scenario .......................................................75
24. Local Moran's I Cluster for General Scenario Entropy Values ....................................76
25. Regression Plots for Degree, Frequency, and Entropy Measures for General Scenario
    a. Degree vs. Entropy ......................................................................................................78
    b. Maximum Frequency vs. Entropy ..............................................................................79
    c. Mean Frequency vs. Entropy ....................................................................................79
    d. Degree vs. Maximum Frequency ............................................................................79
    e. Degree vs. Mean Frequency ....................................................................................79
26. a. Aerial View of Cluster A ..........................................................................................80
    b. Entropy Map of Cluster A ........................................................................................80
27. a. Aerial View of Cluster B ......................................................................................................................... 82
   b. Entropy Map of Cluster B ......................................................................................................................... 82
28. a. Aerial View of Cluster C .......................................................................................................................... 83
   b. Entropy Map of Cluster C .......................................................................................................................... 83
29. a. Network Graph for Tract Pairs that Occur in 90 or More Regionalization Trials,
      Housing Scenario ........................................................................................................................................... 86
   b. Network Graph for Tract Pairs that Occur in 90 or More Regionalization Trials,
      Poverty Scenario .......................................................................................................................................... 86
   c. Network Graph for Tract Pairs that Occur in 90 or More Regionalization Trials,
      Transportation Scenario ........................................................................................................................... 87
30. Entropy Distribution of Tracts in All Scenarios ............................................................................................. 88
31. Entropy Map of Denver MSA, Transportation Scenario .................................................................................. 88
INTRODUCTION

The following study undertakes the challenge of reconciling data and theory on neighborhoods by applying a regionalization algorithm as a data-driven method for identifying neighborhoods. The analysis focuses on the Denver metropolitan region with variables derived from U.S. Census Bureau’s 2011-2015 American Community Survey Five-Year Estimates. By contrasting repeat results from the algorithm with neighborhood identities based on theoretical and practical evidence, this thesis investigates the potential benefits and pitfalls of using purely data-based or theory-based approaches to defining neighborhoods.

Data-Driven and Theoretical Approaches to Defining Neighborhoods

Politicians, business managers, and average citizens celebrate “data-driven” approaches to decision-making (Patterson, 2016; McAfee & Brynjolfsson, 2012; Wolf, 2010). In common usage, the term “data-driven” often refers to implementation of new technologies and methods to glean insights from large data sets. In academic research, the meaning of the term “data-driven” and its relationship to the scientific process is more complex. Semantics suggest that in some sense the distinction between ‘data-driven’ and ‘theoretical’ approaches to research is false. Science relies on data, and scientific contributions inform the development of theory. In practice the terms express broader ideas about the scientific process and how knowledge is cultivated. Data-driven pursuits accumulate relevant information and make inductive conclusions from the evidence. Theory-driven research approaches a subject with a theoretical framework from which to interpret the accumulated data. Picardi and Masick (2014) distinguish between a “theory-driven hypothesis,” which is “crafted through utilizing existing theory to propose a relationship or effect on the variables of interest,” and a “data-driven hypothesis,” which is “based on
previous findings from studies that have been conducted” (p. 21). Others contrast “hypothesis-driven” and “data-driven” or “hypothesis-neutral” research, implying hypotheses become irrelevant when data points the way forward (Mazzocchi, 2015).

Even in these simplified dichotomies, the boundary between data-driven and theoretical research is blurry. Deeming any project as ‘data-driven’ requires judging what constitutes relevant data and how to weight them – decisions that do not occur in isolation of established theoretical knowledge. Neighborhoods, as persistent and valuable but conceptually vague units of social study, capture and symbolize the tension between data-driven and theoretical approaches. Neighborhoods have a long record in sociology, public health, urban planning, and related fields as units and subjects of analysis. Questions regarding how individuals move, interact, thrive or fail to thrive often concern their local environs, a space that has no firm boundaries to match those of cities, zip codes, and school districts. Neighborhoods are constructed internally by residents and externally by outsiders, leading to conflicting perceptions of where one neighborhood starts and the next begins, and who belongs to a neighborhood.

When neighborhoods are the unit of interest and boundaries are required to quantify and organize demographic data, researchers often look to the census tracts to approximate neighborhoods. Census tracts are convenient: population estimates for census tracts are published annually, comparable across survey years, maintained diligently by the Census Bureau, and align with other jurisdictional boundaries. Yet as substitutes for neighborhoods they are inadequate. Census tracts are neither clearly data-driven nor theory-driven, whereas neighborhoods reside at the intersection of both.
The Importance of Census Tracts

As one of the most substantial survey undertakings in American society, the U.S. Census benefits from a long history of dedicated thought concerning the meaning of geographical boundaries and their implications for how census data represents the U.S. population. In the first edition of a series of Census Tract Manuals published between 1934 and 1956, Howard Whipple Green and Leon E. Truesdell (1934) describe the historical evolution of census tracts: “Dr. Walter Laidlaw, working with population statistics in New York, became convinced that in order to study neighborhoods it was absolutely necessary to have population data for local areas smaller than boroughs or wards, and to establish these areas so that they would remain unchanged from census to census” (p. 1). Interest in using census data to study neighborhoods arose early in the federal census’ trajectory towards becoming the large, detailed, scientific, and costly undertaking that it is today.

Green and Truesdell proceed to describe how tract delineations spread to other major cities – 16 by the 1930 census – and were crafted by local organizing committees in partnership with the Census Bureau. Census tract uses were many and diverse, for “city departments and welfare organizations and by commercial concerns, such as newspaper, utility companies, and selling organizations” (p. 4). The authors list groups who used census tract data at the time, ranging from hospitals, to Boy Scouts, and chambers of commerce.

Census tracts remain important today to understand the composition, welfare, and behaviors of the U.S. population at small scales. However, the environment for data production, maintenance, and application has changed drastically since the early 20th century. Amidst exponentially more powerful tools, advanced statistical knowledge, and new sources of information about people and places, the U.S. Census remains critical to statistical knowledge
about Americans and population change in the U.S.

In *The American Census: A Social History*, Margo Anderson (2015) contends that despite recent technological change and new competing sources of information, “the census is a public institution that grounds the official statistical infrastructure of the nation” (p. 254). U.S. census data drives the allocation of over 400 billion dollars in federal assistance and guides additional resource allocation in state and local government (Reamer, 2010, p. 5; American Community Survey Office, 2014). Anderson (2015) explains how the U.S. census has evolved with the rest of the data landscape through collaborations with data users and producers to advance methods that improve the census amidst limited resources and declining survey response rates. The rise of the American Community Survey in place of the Decennial Census to collect detailed demographic data represents one of the most significant changes to the U.S. Census Bureau’s practices, and provides the impetus for this research project.

**From the Decennial Census to the American Community Survey**

After 2000, the American Community Survey (ACS) supplanted the Decennial Census as the U.S. Census Bureau’s primary method for obtaining detailed population statistics. With declining response rates and budget constraints, the ACS presented an opportunity to shift the emphasis from the decadal census to a more frequent and responsive survey process. The change benefitted from parallel advances in computing and statistical methods across the public and private sectors, and reflected the Census Bureau’s efforts to engage with census data users who provided expertise to help enable the transition and pressure to push it forward (Anderson, 2015).

The Decennial Census remains in a shortened form as the primary means of fulfilling the government’s constitutional mandate to enumerate the U.S. population for purposes of
apportioning representatives (U.S. Const. art. I, § 2). In addition to counting the number of people in each household, the 2010 short form Decennial Census asked questions about age, sex, race, and ethnicity. The long form version that asked more in-depth questions no longer exists. Instead, the ACS is now the primary source for a much broader set of demographic variables that plays a large role in research, policymaking, and collective understanding of the characteristics of the US population. Demographic statistics like income, education, and employment all require accessing ACS data.

In 2000, the last year that the Census included a long form survey, approximately 5 out of 6 households received a short form version of the Decennial Census questionnaire and the remaining 1 out of 6, or approximately 17%, received the more detailed long form (U.S. Census Bureau, 2009b, p. 99). The ACS now samples approximately 300,000 people every month, or cumulatively about 1% of the total population every year. The monthly survey data is then aggregated to create one-, three-, and five-year sample statistics published annually (Torrieri et al., 2014, p. 32).

The resulting data is less precise than the decennial census, but much more timely. Whereas the Decennial Census occurs every 10 years and is a true census of all residents (in its short form), the ACS gathers information on samples of the population continuously. By combining estimates across time, the ACS statistics represent a moving window rather than a point-in-time assessment. Whereas the short form Decennial Census provided baseline knowledge about the entire population from which the long form survey gained precision, with the ACS there is no baseline census data about the underlying population distribution from which to make better inferences regarding survey weights (Spielman, Folch & Nagle, 2014).

At large scales such as states, counties, and cities, aggregated samples overcome the
imprecision accompanying ACS statistics. At small scales, such as census tracts and block
groups, the tradeoffs of the ACS’ timeliness become apparent. For instance, Census Tract 1.02 in
Denver County, Colorado provides a reliable estimate of the population who identifies as “White
Alone,” producing a CV of 5.0% (Table 1). However, the White Alone population represents
88.9% of the census tract. For all other race categories, high CV values render the statistics
useless. Although an estimated 30 people identify as American Indian or Alaska Native Alone,
with a margin of error of 47 the actual value could range from 0 to 77, resulting in a CV of
94.9%. Errors of this magnitude can nullify research findings used in decision-making, leading to
proposals for solutions that will alleviate the impacts of uncertainty upon census tract estimates.

<table>
<thead>
<tr>
<th>Race</th>
<th>Estimate</th>
<th>Margin of Error</th>
<th>90% Confidence Interval</th>
<th>CV (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>White Alone</td>
<td>2,847</td>
<td>233</td>
<td>2,614 – 3,080</td>
<td>5.0</td>
</tr>
<tr>
<td>Black or African American Alone</td>
<td>38</td>
<td>32</td>
<td>6 – 70</td>
<td>51.0</td>
</tr>
<tr>
<td>American Indian and Alaska Native Alone</td>
<td>30</td>
<td>47</td>
<td>0 – 77</td>
<td>94.9</td>
</tr>
<tr>
<td>Asian Alone</td>
<td>7</td>
<td>10</td>
<td>0 - 17</td>
<td>86.6</td>
</tr>
<tr>
<td>Native Hawaiian and Other Pacific Islander Alone</td>
<td>0</td>
<td>11</td>
<td>0 - 11</td>
<td>NULL</td>
</tr>
<tr>
<td>Some Other Race Alone</td>
<td>212</td>
<td>121</td>
<td>91 - 333</td>
<td>34.6</td>
</tr>
<tr>
<td>Two or More Races</td>
<td>69</td>
<td>53</td>
<td>16 - 122</td>
<td>46.6</td>
</tr>
</tbody>
</table>

Table 1: 2011 - 2015 American Community Survey 5-Year Estimates for Race (variable B02001) in Census Tract 1.02, Denver County, Colorado

REVIEW OF THE LITERATURE

Uncertainty

Uncertainty in spatial data has garnered attention across disciplines and data types yet
continues to present challenges in analysis and communication. The attention devoted to
uncertainty stems from multiple converging sources: the risks of decision-making based on data
with high uncertainty (Devillers et al., 2010), obstacles to communicating uncertainty (MacEachren et al., 2005), growth of digitally sourced information containing high uncertainty such as volunteered geographic information (Grira, Bedard, & Roche, 2010), and the potential to glean additional knowledge from uncertainty. Regarding the latter, Anselin (2006) notes, “The importance of the statistical insights lies in the quantification of the uncertainty associated with various estimates and in exploiting the spatial characteristics of this uncertainty in the decision process” (p. S6). Uncertainty is both a challenge and an important form of knowledge about the subject of interest.

Whether uncertainty is framed as a tool or liability, it is unavoidable. Acknowledging the inevitability of uncertainty in spatial data, Zhang and Goodchild advocate for “a re-oriented view of GIS-related errors” in which they are “conceived as an integral part of human knowledge and understanding concerning geographical reality, so that information is provided and used with well-informed assessments of uncertainty” (2002, p. 6). In the re-oriented view, uncertainty is not metadata, it is a necessary complement to the estimates. Estimates and uncertainty combined offer the maximum statistical insight achievable with the data.

Researchers’ definitions of uncertainty and error vary, resulting in Devillers et al.’s (2010) call for “a generally agreed-upon ontology” (p. 392) that formalizes those and other terms associated with data quality. Zhang and Goodchild (2002) distinguish uncertainty from error by arguing that error occurs when data is inaccurate because of a knowable truth, whereas uncertainty reflects the gap between known values and unknown truth. Evans (2012) describes uncertainty as the product of accumulated error: “It is [often] the case that we know that there is some error in our understanding, and this leads to assumptions in our models and uncertainty about our model results that need to be communicated to users of the results” (p. 309). Evans
connects error and uncertainty in a sequential process. Errors originating early in the analysis merge and mutate to form uncertainty.

Taxonomies of uncertainty and error similarly conceptualize uncertainty as a large umbrella that encapsulates concrete types of error within it. Fisher (2005) separates uncertainty into two types depending on the object of interest: uncertainty surrounding well-defined objects and uncertainty surrounding poorly defined objects. For a well-defined object uncertainty is the product of error. For a poorly defined object, error, vagueness, and ambiguity collectively form uncertainty. In Fisher’s framework errors refer to specific ways in which inaccuracies are introduced. Fisher identifies seven kinds of error: measurement error; improper class assignment; class generalization, in which objects are improperly grouped together; spatial generalization, when object details are generalized before digitization; data entry error; temporal error resulting from differences between when data is collected and when it is used; and processing errors that arise in computation. Vagueness occurs when concepts are unclearly defined, and ambiguity results from divergent perceptions of the same subject.

Offering a simpler taxonomy, Griffith, Wong, and Chun (2015) distinguish four types of error contributing to uncertainty in spatial data: sampling error, measurement error, specification error, and location error. Sampling error results from differences between statistics derived from a sample population versus a full census. Measurement error stems from differences between an instrument reading and the actual value. Dissonance between a model and what it attempts to represent results in specification error. Finally, translating real locations to a coordinate plane introduces georeferencing error. Data aggregation compounds these errors.

Differences between taxonomies of uncertainty and error result in part from the nature of the data and domain under study. Landscape data derived from remote sensing, environmental
monitors, and field observations incur errors different from those confronting population data such as census or survey information. For instance, landscape data produces uncertainty around object identification when boundaries are unclear, either due to image resolution or gradual shifts between landscape types such as forest to meadow. A different type of boundary uncertainty arises in demographic data with imposed areal units like census data, in which boundaries may be well-defined but suffer from conceptual arbitrariness when applied to domains for which the boundaries were not designed (Fisher, 2005, p. 193; Griffith, Wong, and Chun, 2016, p. 7). Questions of power and ownership in geopolitical domains encounter uncertainty due to what Fisher labels “discord” (p. 198), exemplified by disputes over territorial boundaries. The data’s characteristics determine its associated types of uncertainty and error. The focus of this analysis is pre-collected ACS data from the U.S. Census Bureau, which has implications for both the types of uncertainty of interest and potential resolutions. The ACS contains statistical uncertainty associated with attribute data and boundary uncertainty associated with the imposed geography of census tracts.

Manski (2014) provides a useful comparison and critique of uncertainty in reported federal statistical data, arguing for better clarity surrounding reported and unreported uncertainty. He commends the Census Bureau’s Current Population Report for publishing sampling error with its statistics (as does the ACS) but notes that the Current Population Survey does not report nonsampling error (p. 633). Nonsampling error encompasses all sources of error outside of sampling error in survey data and is more challenging to quantify and resolve (U.S. Census Bureau, 2009a, Appendix 3 and 6; Manski, 2014, p. 634). Nonsampling error arises from participants’ failure to respond to the survey, inaccurate answers, and bias in participant selection (Freedman, 2004). Manski also takes issue with sources of error disguised by statistical
corrections applied to the data before it is released. For example, the U.S. Census Bureau applies imputation to address missing answers in census responses. Imputation can be accomplished using a variety of statistical methods, but generally imputation employs known information about a population to derive estimates for missing count or attribute data (Cohn, 2011).

Most federal statistical data users have no way to distinguish imputed values from measured responses, to judge the accuracy of imputed values for themselves. In this way, invisible uncertainty is woven throughout the data. Manski (2014) suggests using intervals instead of point estimates as one solution, or providing the unaltered data to users. (Of note, some data users can access unaltered data at federal research centers after passing through an extensive screening process.) However, providing unaltered data risks further confusion from data users who may already lack understanding or willingness to incorporate error (p. 650-651).

Manski admits that nonsampling errors are difficult to quantify, and consumers have a low tolerance for uncertainty in data (p. 634). Nonetheless, he believes that simply noting nonsampling error is insufficient:

Agencies already make judgments when they extrapolate trends to construct advance GDP estimates, when they use weights and imputations to deal with survey nonresponse, and when they seasonally adjust a multitude of statistics. They implicitly judge nonsampling error to be inconsequential when they choose not to quantify it. They should reject this flawed implicit judgment and use their professional expertise to develop useful measures of nonsampling error. (p. 650)

Manski’s argument highlights the complex political ramifications of how data uncertainty is handled and communicated. One of the most obvious effects of data uncertainty upon tangible outcomes involves apportionment of political representation. In 2001, Utah lost a lawsuit arguing
that the 2000 Decennial Census included federal employees living abroad but failed to include Utah residents living abroad, specifically Mormon missionaries. The resulting population count shifted a congressional seat from Utah to North Carolina (Kirkland, 2001). In Utah v. Evans II, a subsequent lawsuit that reached a hearing before the U.S. Supreme Court, Utah argued against the Census Bureau’s “hot deck” method of imputation, claiming it represented an unlawful use of statistical sampling to determine congressional representation. Hot deck imputation estimates missing responses for households based on the attributes of other households with similar known characteristics to reduce nonsampling error from nonresponse. The Census Bureau used hot deck imputation in the 2000 U.S. Census to estimate household size when other survey and follow-up methods failed to establish an accurate number of people in a household, rather than assume no additional residents lived at the household. The Supreme Court ruled in the Census Bureau’s favor, asserting that imputation is not equivalent to sampling and prior precedent supports the Census Bureau’s use of imputation to obtain accurate results (Utah et al. v. Evans, Secretary of Commerce, et al., 2002).

Although the ramifications of uncertainty rarely involve the Supreme Court, ignoring uncertainty sets a dangerous precedent for future statistical analyses and risks accumulating erroneous research findings based on highly error-prone data. Increased attention to uncertainty and spatial data quality over the past 25 years has helped shed light on the importance of comprehensive approaches to data analysis that incorporate uncertainty. Furthermore, the parallel rise in computing power and explosion of population data generated through digital sources requires continued diligence to ensure enthusiasm for new data sources and methods does not compromise the quality of analytical findings. Federal census data retains an important place in this landscape, benefitting from recent conceptual and technological advances while maintaining
standards for the construction and provision of comprehensive, publicly-accessible data about the U.S. population.

Uncertainty in the American Community Survey

As the prevailing source of public data on the U.S. population, the ACS approaches uncertainty from a survey data perspective in which uncertainty is the result of errors categorized as sampling or nonsampling errors. Sampling error results from the sample size and population heterogeneity (Spielman, Folch, & Nagle, 2014, p. 3). It is difficult to represent a diverse population with a small sample. Expanding the sample size is the most obvious remedy to sampling error, however larger samples are costly and can augment nonsampling error (Freedman, 2004). In practice, limited funding prohibits the Census Bureau from surveying more people (Powers, Beede & Telles, 2015). Barring access to additional resources to expand the sample, sampling error can at least be estimated and reported using statistical techniques.

One of the primary benefits of switching from the Decennial Census to the ACS was the publication of sampling error with ACS estimates. The ACS provides sampling error measured as the Margin of Error (MOE) for a 90% confidence interval, a range of numerical certainty within which the actual estimate is expected to fall. The ACS also provides instructions to compute alternate measures from the MOE such as the coefficient of variation (CV), which provides a measure of sampling error relative to the magnitude of the estimate (U.S. Census Bureau, 2009a, Appendix 3). As described in Table I, a large MOE can translate to a small CV if the population estimate is large, and vice versa.

Spielman, Folch, & Nagle (2014) elaborate upon sources of nonsampling error in the ACS. The ACS lacks the population benchmarks from a full enumeration as provided in the
Decennial Census, thereby potentially skewing weights applied to survey results. The Master Address File used to identify housing units can contain inaccuracies that result in coverage errors. Recent years have brought declining response rates, which is a growing problem among social science surveys in general (Tourangeau & Plewes, 2013).

The ACS categorizes sources of nonsampling error as coverage error, unit nonresponse, item nonresponse, response error, and processing error, and publishes measures for the first four types. Coverage error occurs when a potential survey respondent is improperly included or excluded from the survey sample. Unit nonresponse results when no survey response is received from a household or a member of the household, the survey ‘units.’ Item nonresponse occurs when the Census Bureau receives a survey response lacking answers to one or multiple questions, referred to as survey ‘items.’ Response errors result from responses that are inaccurate or recorded incorrectly. Errors that arise in compiling responses become processing errors (U.S. Census Bureau, 2009a, Appendix 6).

By providing some measures of uncertainty the ACS meets Manski halfway, conveying sampling error clearly, offering access to nonsampling error for motivated users, but leaving some nonsampling error unquantified (U.S. Census Bureau, 2009a, p. A-25). (See Torrieri et al., 2014 for extensive details on ACS methodology to further elucidate how survey values are manipulated for data quality and privacy purposes.) Uncertainty values provided with ACS data enable census data users to judge data quality, adapt their research to accommodate uncertainty, and qualify their findings. Error reporting in the ACS also opens new avenues of research to understand the nature of uncertainty in ACS data, explore its ramifications, and devise solutions. The ready availability of sampling error for every estimate as expressed by the MOE lends itself well to exploration.
Folch et al. (2014) demonstrate the potential for sampling error to reveal spatial demographic patterns that reflect upon the surveyed population and the survey process. Folch et al. assess sampling error associated with median household income across the continental U.S. in the 2006-2010 ACS Five-Year Estimates for census tracts. They find that MOE varies in relation to geography, racial diversity, income, and tract boundary stability at statistically significant levels. They also find greater sampling error in tracts near city centers and with lower income levels, and different patterns of uncertainty between northern and southern states. Sampling error in ACS data varies non-randomly across space and demographic characteristics. It is patterned in relation to the underlying estimates.

Sociodemographic causes can potentially explain the nonrandom distribution of sampling error, though such explanations require deeper investigation into the perceptions and experiences of survey respondents. Bazuin and Fraser (2013) undertake such an exploration in a case study of census tract 160 in Nashville, TN, an area populated by African-American homeowners, recent Caucasian arrivals who buy or rent, and low-income African-Americans in low-quality rental housing. The authors find that survey respondents’ willingness to participate in the survey is affected by their distrust of government and the census, language and literacy challenges, and privacy concerns (p. 295). Furthermore, surveyors’ perceptions such as whether a house is vacant or occupied affect survey outcomes. For census tract 160, Bazuin & Fraser find that cumulative errors result in a much lower estimated poverty rate than the actual rate, potentially affecting the allocation of resources to that part of the city. Real consequences ensue from high ACS uncertainty.
Regionalization

Given resource constraints preventing expanding the ACS sample size and pursuing more follow-up for survey participants who fail to respond, researchers within and outside of the Census Bureau have focused on advancing statistical and geographic methods that help alleviate high uncertainty. Spielman, Folch, and Nagle (2014) discuss potential remedies for small scale census geographies, including data aggregation and incorporating additional data sources. In data aggregation, census units are combined to create larger samples, thereby reducing the sampling error in the estimates. Units can be combined based on adjacency (for contiguous units), shared attributes (can be non-contiguous units), or both. Although aggregation reduces geographic detail, it can also recapture the utility of estimates that previously contained unacceptable levels of error. Spielman, Folch and Nagle also explain how alternative geospatial data sources could be employed to create population controls. Such sources might include data from other federal agencies that could be compared to ACS estimates to improve weighting of survey responses. In either strategy, the data collection process remains unchanged.

The present analysis focuses on data aggregation, and specifically regionalization, as a solution to sampling error in ACS data. Regionalization is a form of data aggregation in which adjacent units are combined to create new, larger regions by optimizing one or more parameters of the underlying spatial units. Put simply, Montello (2003) describes regionalization as “the creation or identification of regions” or “spatial categories” (p. 174) in which spatial objects are grouped based on their identification within the same category for one or more variables. Regionalization helps describe places and organize them in an intuitive fashion (p. 173-174): “Humans think in discrete pieces of truth or reality, even if that way of thinking is in fact false or distorted” (p. 176). Constructing regions can add interpretive value even when the process of
defining a region confronts ambiguous boundaries across time, space, and other measures (p. 181).

Historically, descriptive regions have provided a means of distinguishing here from there, despite transitions between regions that are gradual and vague (Creswell, 2013, Ch. 4). From a computational and statistical perspective, even vaguely-defined regions must assume concrete areas that can be outlined and measured using geographic coordinates. Users can download a shapefile outlining the U.S. Forest Service’s ecoregions, but observers will recognize that transitions in the ecological landscape are typically gradual rather than sharp (Forest Service, n.d.). Transitions in the built environment may also be gradual, like the shift from a central downtown district to lower density suburbs, or very clear as occurs at the edge of a suburban housing development bordering countryside. Similarly, population characteristics may change gradually across a city with individual outliers sprinkled throughout, or they might shift rapidly due to housing developments or other socioeconomic patterns.

Regionalization is most pertinent to geographic data like neighborhoods that take the cartographic form of areal units. Point data or other discrete objects are ill-suited for regionalization, lacking a method to account for the space between them. In clustering methods, such as geodemographic classification, non-adjacent areal units can be grouped together based on shared attributes. A geodemographic group might exist as a collection of noncontiguous units associated by another quality, for instance, land cover type, which is also demonstrated by the U.S. Forest Service’s ecoregions (Forest Service, n.d.). Regionalization requires unit adjacency, amplifying the importance of boundaries.

Boundaries play a critical and contentious role in regionalization as the point of distinction between places. By drawing a visible line, boundaries make evident how contrived
regions fail to represent a place or population. Although debate may arise over the boundaries of a physical object – what is the exact point where the forest ends and the meadow begins? – objects that have no physical definition, what Smith and Mark (1998) call “geopolitical objects like nations and neighborhoods” (p. 6), provoke the greatest debate. Such geopolitical objects “exist only as the hybrid spatial products of human cognition and action” (p. 6), giving arbitrary lines sociopolitical meaning.

Regions can also be described as zones, and regionalization as a form of zone design. Alvanides, Openshaw, and Macgill (2001) explain zone design as “the systematic aggregation of areal units into zones subject to contiguity constraints” (p. 141). Like regionalization, zone design employs a parameter of interest to group adjacent zones into larger regions that better suit the research objective. Regarding zones as flexible instead of fixed, results can be evaluated in relation to zone design, and zones can be modified. As Openshaw (1978) states, “the performance of a model, the values assigned to its undetermined parameters, and the choice of zoning system are all interdependent” (p. 782). Envisioning regionalization as zone design, and zones as dynamic rather than static components of the analytical process, regions/zones assume meaning beyond serving as spatial containers for data.

A spatial analyst might employ any variety of methods to delineate regions for diverse purposes across disparate domains. Regionalization as a form of iterative zone design exploits the interdependence between the input units of study and the output results of the regionalization process. Through repeat trials, regionalization can produce multiple potential zoning solutions that in turn highlight patterns which, under a single, fixed zone design, might otherwise remain obscure. Regionalization can be a process of identifying a pattern of value and forming geographic units to match it.
In the case of uncertainty in ACS data, regionalization can reduce statistical uncertainty to acceptable levels. A simple regionalization strategy would combine tracts until population sample sizes grow to a level containing acceptable sampling error. While accomplishing the main objective of reducing MOE values, such a simple strategy overlooks the characteristics of the tracts and the potential for areal units with very different demographic characteristics to land within the same output region. An advanced solution combines units while considering both the sampling error and the value of the estimates of interest. In the advanced regionalization process, output regions reduce statistical uncertainty while maintaining demographic patterns in the input data.

*Ecological Fallacies and the Modifiable Areal Unit Problem*

Regionalization applied to ACS data can reduce statistical uncertainty associated with attribute data but cannot remedy the geographic uncertainty accompanying fixed census tract boundaries. Regions formed through tract aggregation will retain the boundaries associated with census tracts. Only with individual-level survey responses could regionalization address any mismatch between census tract boundaries and the social phenomena of interest. Even then, the risk of losing detail through aggregating smaller units into larger units occurs at every scale, including when individual survey responses are first aggregated into areal units. In the aggregation process, vital details may be lost and homogeneity assumed where heterogeneity exists.

Openshaw (1984) articulates these challenges as “ecological fallacy problems” (p. 30) in which analytical results are misaligned with the unit of study. Data analyzed at the individual or household level can produce results that differ widely from the results of data aggregated into
areal units. Openshaw demonstrates the consequences of ecological fallacy problems through analyzing historical census data from the UK and Italy. His analysis reveals “averaging effects” (p. 29) that result from combining individual data points into summary statistics for areal units. Outliers among the individual data points weight heavily on the average, leading to “selective amplification or filtering of the averaged-out data” (p. 29). In turn, correlations evident at the individual scale disappear or appear disproportionately strong at the scale of areal units.

Subsequent researchers have argued that although conflicting results produced from individual and group data can lead to divergent analytical outcomes, the problems arise not from the disparities between the results but from the misattribution of results based on group data to individuals or vice versa (Schwartz, 1994, p. 823; Susser, 1994).

Intertwined closely with ecological fallacy problems is the Modifiable Areal Unit Problem (MAUP), which considers the effects of the scale and shape of areal units upon spatial data analysis (Openshaw, 1996). The shape of an areal unit alters which spatial data points are included within it, which in turn affects the statistics tabulated for the unit. The scale of the unit likewise affects the number of data points contained within the unit, affecting mean estimates. Different zone design schemes can result in shape and scale effects that lead to variable spatial patterns and outcomes. The danger is that “areal objects are not natural units…but are variously subjectively defined zonations of two-dimensional map space that may or may not have any intrinsic meaning” (p. 64), yet they have strong statistical implications and therefore their definitions are important. The implicit assumptions entangled in the shape and scale of the areal units may not be apparent to consumers of analytical results.

Regionalization’s interactions with the dual problems of the ecological fallacy and the MAUP are complex. While reducing uncertainty, regionalization requires attention to the
potential for ecological fallacies, scale, and zone effects as the units of study change. Where ten input units previously existed, after regionalization the new output may contain only three regions ranging in size from one to six input units. Regionalization relies on the geography of the input data, thereby incorporating its idiosyncrasies and susceptibility to the MAUP. Yet informed regionalization also can improve upon the input geographies by creating new units based on specific variables of interest and offer a tool to create multiple zone designs that enable comparative analysis of analytical outcomes. Understanding the geography of the input data is vital to recognize the implications of using regionalization as a tool to reduce uncertainty in ACS data.

**Census Geographies**

Applying regionalization to ACS data to reduce uncertainty requires understanding the nature of ACS geographies, which are based on the Decennial Census (U.S. Census Bureau, 2016a, October 13). The Census Bureau, like statistical offices throughout federal agencies, employs a mix of legal and statistical geographies (Figure 1). Legal units are based on political representation and government administration boundaries. Statistical units such as census tracts and block groups were devised for statistical purposes and have no direct relationship to jurisdictional boundaries, although they align with administrative or political units to form hierarchical spatial zones. U.S. Census Bureau
geographies range from the nation to blocks, the smallest unit of measure. The Census Bureau only publicly releases data at the block group level and above. Intermediary units span the two extremes. Alternative spatial units such as Zip Codes, School Districts, and Congressional Districts occupy places between the extremes in size but are not directly within the core hierarchy (U.S. Census Bureau, 2016b, October 13).

Small-scale geographies such as census tracts are vital for statistical data collection purposes and highly desired by researchers seeking to understand neighborhood-level social dynamics, risks, and outcomes (Anderson, 2015). Without them, the Census Bureau could not share detailed population data due to privacy constraints. Individual data points are not publishable. Census tracts conform to county and state boundaries, contain between 1200 and 8000 people, and roughly follow physical landscape features (U.S. Census Bureau, 2010a). Originally, tracts were designed to be demographically homogenous, but the Census Bureau prioritizes the continuity of census boundaries for historical comparison regardless of internal heterogeneity (U.S. Census Bureau, 1958; U.S. Census Bureau, 1994).

Tracts provide a level of demographic detail lacking in larger jurisdictions like counties, which due to their geographic size and population can obscure small-scale patterns. For instance, Denver County in Colorado contains 649,654 people in 144 census tracts covering a land area of 154.63 square miles (U.S. Census Bureau, 2016a). Denver County residents are diverse in income, education, race, ethnicity, and family structure, and this diversity manifests in small-scale spatial patterns across the county. Viewing Denver County as a collection of census tracts reveals that African American residents are concentrated in the northeastern part of the county (Figure 2a) and the foreign-born population is concentrated in the southwest region (Figure 2b). The fine-scale distribution of Denver County residents paints a nuanced demographic portrait,
illustrating why tract-level statistics are so compelling. Figures 2a and 2b do not include sampling error, which could significantly alter the evident spatial patterns, particularly for demographic groups that represent small fractions of the overall tract populations.

**Figure 2a**: African American Population as Percent of Total Population: Denver County, Colorado, 2011-2015 ACS Estimates

**Figure 2b**: Foreign-Born Population as Percent of Total Population: Denver County, Colorado, 2011-2015 ACS Estimates
Census tracts also contain sufficiently large sample sizes to maintain survey respondents’ privacy. Individual household data would be ideal from a research and policymaking perspective, providing the greatest accuracy and insight, however the Census Bureau guarantees respondents’ privacy in accordance with Title 13 of the United States Code and implements extensive measures to ensure it fulfills that promise (U.S. Census Bureau, n.d.a). Block group data is also valuable as the finest geographical unit published by the Census Bureau, however block group data suffers from even higher uncertainty and some estimates are suppressed for privacy when the Census Bureau determines their publication could risk revealing information about respondents (U.S. Census Bureau, 2016b). Possessing utility and familiarity, tracts inform a wide range of public policy decisions at various levels of government, as well as public and private research (Edmonston & Schultz, 1995, Appendix D-F).

Despite the value of census tracts and other statistical geographies for data collection and publication, users of ACS data must take caution when using spatial units designed for purposes independent of their study. Bearing in mind the ecological fallacy and MAUP, census data users can and should question the implications of census tract boundaries, both how they impact quantitative results and how they affect measurement and interpretation of demographic data. Griffith, Wong and Chun (2015) highlight a salient example with the interpretation of census units or other areal unit boundaries as real neighborhoods:

Many studies adopt a simplistic and pragmatic approach to select existing areal units, such as census tracts or block groups, as proxies for neighborhoods. Thus, individuals geocoded to a unit, regardless of their socioeconomic and demographic differences, and their dispersions of locations within the unit, are assumed to have the same neighborhood demarcated by an artificial unit boundary. (p. 6)
Interpreting census tracts as neighborhoods amplifies the statistical challenges of areal data by layering a sociological interpretation onto tract or block group boundaries that may not be appropriate. Rather than accept or dismiss this practice, this analysis takes advantage of regionalization’s potential to create meaningful units through data-driven aggregation of tracts with like characteristics. Exploring whether the new regions resemble neighborhoods requires understanding what a neighborhood is and how it is constructed.

**Defining Neighborhoods**

Neighborhoods are defined in diverse and conflicting ways as products of the physical landscape, culture, history, and demographics. The precedence of any definition depends upon the reason for defining neighborhoods, whether for political aims, historical insight, navigation, commercial pursuits, or other purposes. The method used to delineate a neighborhood or any geographic space can affect its definition, as can human perception of the space as a recognizable place. Motivations, methods, and perception combined complicate any effort to construct a coherent representation of a neighborhood or collection of neighborhoods.

The term ‘neighborhood’ can have significance absent geography or rely on spatial relationships. ‘Neighborhood’ could imply people who are related in some way, by a feeling of kinship or other identification synonymous with community (Chaskin, 1997). More often it connotes some form of proximity defined in vague or explicit geographical terms, but also may be defined temporally, or along another vector. For computational purposes, a neighborhood might be specified concretely as all points or objects within a designated distance of a central point. Yet as a sociological construct that tries to capture individual and collective perceptions, the term neighborhood requires more than the measure of people within a given area.
A long history of efforts to quantify and describe neighborhoods exists, yet despite advances in theory, tools and techniques the process of defining neighborhoods remains as much art as science. In the 1920s and 1930s, the Chicago School attempted to codify a structured approach to understanding communities, a superclass to neighborhoods, while acknowledging that even within that structure definitions diverge and overlap (Park, Burgess & McKenzie, 1925). Asking, “Can neighborhood work have a scientific basis?” (p. 144-154), Burgess provides three lenses through which to understand communities: ecological, cultural, and political. The ecological perspective sees communities as the products of geography, resources and competition. The cultural lens views communities through the customs of its residents and the interplay between social mores and individuals’ movements. Finally, the political lens approaches communities through the power structures through which people relate and achieve, or fail to achieve, projects for the common good. Although their applied examples are outdated, Park, Burgess, and McKenzie’s questions regarding the meaning of community boundaries and explanations of the forces that define them remain relevant in today’s context of increasing technological resources and sociological insight.

Alternative theoretical approaches to neighborhoods graduate from views of the neighborhood as a spatially and socially isolated unit to interpreting it as one of many overlapping and hierarchical spatial and social units that serve different purposes. Chaskin’s (1997) literature review situates the neighborhood within a larger context of “nested” (p. 535) spatial units that gain legitimacy through political representation, external forces, and internal relationships within and between groups. The neighborhood is nested further still when considering individuals’ perspectives on their own neighborhood or other neighborhoods, perspectives which vary in relation to the individuals’ attributes and situation (p. 537). Chaskin
provides a layered view in which “the delineation of boundaries is a negotiated process; it is a product of individual cognition, collective perceptions, and organized attempts to codify boundaries to serve political or instrumental aims” (p. 539).

Neighborhoods are also physically nested and negotiated within the urban landscape as viewed through the people who inhabit them. Lynch (1960) surveys residents and visitors in a selection of diverse cities to understand how they perceive and interact with the physical urban landscape. He segments the landscape into five features: paths, edges, districts, nodes, and landmarks. Each feature can reinforce partitions within the city or conversely attract interaction and exchange. An edge in the form of a highway can separate a neighborhood from its surroundings: “The isolation of the North End in Boston by the Central Artery was glaring, in the eyes of residents and non-residents alike” (p. 64). In contrast, a busy thoroughfare can be an edge that unites: “Charles Street carries heavy traffic but also contains the local service store and special activities associated with the Hill. It pulls the residents together by attracting them to itself” (p. 65). Districts most closely resemble neighborhoods in Lynch’s nested structure, occupying physical space while possessing the intangible quality of “character” (p. 41). Collectively, Lynch’s interviews reveal the aggregate impacts of physical form upon the actual and perceived structure of places:

Rather than a single comprehensive image for the entire environment, there seemed to be sets of images, which more or less overlapped and interrelated. They were typically arranged in a series of levels, roughly by the scale of area involved, so that the observer moved as necessary from an image at street level to levels of a neighborhood, a city, or a metropolitan region. (p. 86)

Physical characteristics unique to every place, from the nuances of a street corner to patterns of
development throughout a metropolitan region, can create real distinctions between how people perceive and interact with cities and their parts.

Instead of undermining efforts to define neighborhoods, the variegated and subjective qualities of neighborhoods encourage the pursuit of a simplified neighborhood representation. Suttles (1972) distinguishes “folk models” of communities born from popular perception, from “social science images” created through the concerted efforts of researchers. Both, he argues, are important for their role in making sense of the intricacies of urban space: “These simplified images serve us well by reducing the complexity of the urban landscape to a range of discrete and contrastively defined ecological units despite the general continuity, gray areas, and constant changes in any section of the city” (p. 4). People, planners, and institutions use these simplified images to build relationships, plan future development, and organize in defense or support of place-based priorities.

Motivations for Defining Neighborhoods

Neighborhoods have utility internally and externally for political representation, government administration, navigation, commercial purposes and research on neighborhood-level population outcomes. None of these motivations acts alone. Suttles articulates the development of a neighborhood identity as the result of push and pull forces that involve both “adversaries and advocates” (p. 50) so that one neighborhood is distinguished relative to another. “Community identification, then, can be conceived of as a broad dialogue that gravitates toward collective representation which have credence to both residents and nonresidents alike” (p. 52). The community identity results from collective actions, although the motivations for those actions differ.
Community organizing efforts are deeply intertwined with neighborhoods throughout history. Robert Fisher cites examples of neighborhood-based organizing in the U.S. beginning with the social settlement movement in the late 19th century, moving into political activism in the 1930s, and segregationist homeowner associations in the 1920s and 1950s. “[Neighborhood organizing] is a means of democratic participation, a means of extra-political activity, a way to build community, obtain resources, and achieve collective goals” (Fisher, 1996). Similar community organizing forces are at work today constructing neighborhood identities. In many cities lacking any formal neighborhood designations, residents can represent themselves through forming neighborhood associations. Modern social networking resources like Nextdoor reinforce neighborhood boundaries by enabling people to connect virtually to their neighbors (Nextdoor, 2017).

Political and government organizations use neighborhoods to build and distribute resources. Barack Obama’s presidential election team centered their local organizing strategy around ‘neighborhood leadership teams’ that were responsible for empowering volunteers and mobilizing voters within a defined geographic area (Ganz, 2009). The U.S. Dept. of Housing and Urban Development allocates funds for housing revitalization to “address struggling neighborhoods with distressed public or HUD-assisted housing” implemented via their parent cities (U.S. Dept. of Housing & Urban Development, n.d.). Cities as diverse as New Orleans, Boston, Detroit, Rochester, Fort Worth, Missoula, St. Louis, and Palm Springs fund offices devoted to engaging with neighborhoods. Neighborhoods help political and government groups organize relationships with diverse constituents into a manageable structure.

Businesses assume commercial interests in neighborhoods for marketing and business siting. Marketers use geodemographic classification systems such as PRIZM to identify
prominent demographic patterns in a localized area (Claritas PRIZM, 2017). Entering a zip code into the PRIZM database yields a portrait of common geodemographic segments in that zip code as determined by PRIZM’s internal analytical process and data. Population segments have titles such as “Upper Crust,” “Connected Bohemians,” and “Heartlanders,” each describing a group with a specific mix of demographic qualities. Increasingly sophisticated geospatial tools and data enable businesses to target digital advertisements to residents of localized areas, and support user searches for businesses based on neighborhood designations (Hall, 2011; Austin, 2013).

The physical arrangement of neighborhoods helps reduce a vast urban landscape into smaller portions that ease navigation. City marketing associations like Choose Chicago, Visit San Antonio, and the Nashville Convention & Visitors Corp. promote neighborhood-based tourism (Choose Chicago, n.d.; Nashville Convention & Visitors Corp., 2017; Visit San Antonio, 2017). Real estate advertisements promote housing based on neighborhood desirability, which is then reflected in prices. For diverse public and private interests, neighborhood designations impose spatial logic at a scale useful to achieve their aims. The practical utility of neighborhoods, whether clearly defined or vague, harkens to Lynch’s vision of an urban landscape composed of recognizable features and Montello’s explanation of the cognitive value of regions.

**Neighborhood-based Research**

Research interests in examining neighborhood-level population processes merit special attention. Much later in the 20th century after Park, Burgess, and McKenzie put forth their framework for neighborhoods, interest in neighborhoods burgeoned in social science as a path to understand how local, contextual factors influence outcomes for people in those neighborhoods.
Examples of such studies abound in health, education, crime, employment, youth welfare, and other arenas, often falling under the vast umbrella of ‘neighborhood effects’ research (Dietz, 2002, p. 556-663; van Ham et al., 2012).

Rapid growth in neighborhood-based studies despite skepticism of the validity of study findings has led to retrospection and introspection regarding neighborhood-based research choices. Dietz reviews a range of neighborhood-based studies conducted between 1982 and 2001 and finds neighborhood effects to be relatively small and highly-dependent on study-specific factors. Theoretical approach, data source, neighborhood definition, time frame, and statistical techniques all affect study outcomes, making comparison difficult. Are neighborhood effects at work, or are the results simply the product of “spatial clustering of socioeconomic phenomena” (p. 569)? Other studies addressing localized influences upon population outcomes refer to this distinction as one of ‘context,’ factors endemic to a place, or ‘composition,’ factors related to people clustered in a place (Roux, 2004).

With broad recognition of the importance of nuance in neighborhood effects research, more recent studies have explored how study design affects research on neighborhood-based study outcomes. Writing in the American Journal of Preventative Medicine, Weiss et al. (2007) comment on the common use of predefined statistical units to delineate neighborhoods for urban health studies: “Predefined boundaries are easily identified, replicable, and obviously allow for the use of secondary source data. Their disadvantage rests in possible discrepancies with contemporary settlement patterns and resident perceptions of neighborhood boundaries” (p. S155). The authors proceed to demonstrate a process for defining neighborhoods for a future study. Deeming resident perceptions too resource-intensive and contested for practical use, Weiss et al. establish neighborhood boundaries using census block groups as the core units of
analysis in conjunction with land use data and in-person observations to gauge residential density, physical boundaries such as busy transit routes and large institutions, internal resident homogeneity and external heterogeneity. Acknowledging the subjectivities of the method, they argue it nonetheless represents a “relatively efficient methodology that allows for consideration of a range of factors commonly used in neighborhood definition” (p. S158).

Whether attempting to draw lines on a map or construct a narrative, defining neighborhoods becomes more challenging when the effort requires reconciling quantitative population data with perception and theory. To the extent that like attributes of people in close proximity reinforce the notion of neighborhood, and disparities between people reinforce boundaries, quantitative measures are valuable. Despite declines in formal survey response rates, rapid growth in information about people via digital technologies coupled with advances in computing power have expanded resources to pursue quantitative approaches. However, a purely data-driven approach has the potential to contradict longstanding beliefs and interpretations of a neighborhood’s boundaries informed by popular perception or academic theory. Furthermore, the influence of physical structure and local history upon a neighborhood’s boundaries might overshadow the influence of demographic homogeneity, and thus theory and history provide needed lenses for identifying neighborhoods. Arguably the flexible nature of neighborhood definitions benefits from incorporating diverse data types, as in the strategy of Weiss et al., so that the limits and subjectivities of each balance or validate the other.

**Spielman & Folch’s Regionalization Algorithm**

An expanding landscape of demographic information must confront a legacy of critical thought around how people organize themselves across the physical landscape and how they are
counted. Tension arises between demand for fine-grained detail and the importance of data quality. The rise of the ACS mirrors these trends, representing a new, novel way of constructing the primary source of public information about the U.S. population possessing clear limitations. The challenges brought by these collisions afford opportunities to evaluate the intersection of population data, uncertainty, regionalization, and the role of neighborhoods in theory and practice.

Predefined statistical census units in use today benefit from neither a data-driven nor theoretical approach. As described in the introduction, they have a historical basis in the local landscape as perceived by local tract organizing committees working in partnership with the Census Bureau at the time of their creation. Prioritization of boundary consistency above demographic homogeneity over time means there are no guarantees that current tract boundaries represent contemporary neighborhood delineations, even if they confirmed to popular perceptions of neighborhood boundaries at the time of their creation. Census geographies are not completely arbitrary since they adhere to the parameters outlined earlier, yet neither are they necessarily natural units that hew to Park, Burgess and McKenzie’s cultural, political, or ecological boundaries. They have meaning only as statistical units outside of any identified research or planning pursuit.

The following analysis integrates these larger themes in a succinct exploration of one regionalization method to address uncertainty in the ACS. Spielman & Folch (2015)’s regionalization algorithm reduces uncertainty in the ACS by combining census tracts or block groups to create larger regions with reduced sampling error, as measured by the MOE. Sampling error is reduced because the sample sizes for estimates become larger through data aggregation. The algorithm converts sampling error from MOE into CV values. Since CV is a relative
measure of sampling error, it enables comparison and combination of variables such as *median household income (in dollars)* and *commute time (in minutes)* that otherwise are measured on different scales.

As the regionalization algorithm reduces uncertainty it also seeks to create regions that are demographically similar, reducing the potential for loss of heterogeneity that occurs when data units are merged. This aspect of the algorithm makes it a compelling tool to identify conceptually meaningful units of study and provides the impetus for using the algorithm to explore neighborhood identification. By constructing regions based on the similarity of ACS estimates the algorithm attempts to maintain spatial patterns in the input data. The regions should then reflect those underlying patterns, capturing demographic trends in the input estimates while improving data quality. If the patterns in the underlying data are sufficiently strong as measured using existing census tract boundaries, the algorithm would be expected to identify the same regions in any given trial. If the patterns in the underlying data are weak, either because of population heterogeneity, the effects of uncertainty, or the influence of census tract boundaries upon statistical outcomes, the regionalization algorithm could fail to identify meaningful regions in the input data.

Consequentially, the following analysis is premised on two key assumptions: that meaningful units of study including neighborhoods can be assessed using demographic data, and that sufficiently strong patterns exist in the input tract data to produce region formation patterns in the output. Neighborhood definitions that assert the importance of the physical landscape, history, and other forms of social, cultural, and political cohesion for neighborhood formation challenge the first assumption. However, the same assumption underlies any application that uses census tracts and census data to represent neighborhoods. Demographic variables from the ACS
can capture some though not all of these factors, acknowledging some aspects of neighborhoods will be excluded but that ACS data can provide sufficient information to explore the algorithm’s potential for meaningful region formation. Thus, this research hypothesizes that the output regions produced by Spielman & Folch’s regionalization algorithm to minimize variance in input tract data represent socially and theoretically meaningful units of analysis.

Users can choose one or more single variables to incorporate into the algorithm or create computed metrics from census variables in the form of ratios and proportions. Importantly, users also set the limit for acceptable sampling error (CV) in the output regions. For data derived from ACS estimates with low levels of uncertainty and few input variables, or for relaxed algorithm parameters, the output regions may resemble the inputs. The input units may already meet or nearly meet the given constraints. Data with high uncertainty, more variables, and strict constraints will require combining more regions to find a suitable solution, resulting in larger output regions (see Spielman & Folch, 2015 for further explanation).

The algorithm is designed for tracts and block groups, the census geographies containing the smallest sample sizes and thus highest sampling error. Combining tracts and block groups produces new geographies that are larger than tracts but smaller than counties, thereby retaining some of the advantages of smaller census units without the problems of high sampling error. Figure 3 illustrates a sample outcome of the algorithm with output regions overlaid on input tracts for the Denver metropolitan area.
The output of one regionalization trial depends on the shape and formation of the input regions, their attribute data including uncertainty, as well as a stochastic element of the regionalization algorithm that selects a random region from which to begin formulating possible solutions. Ten trials of the algorithm applied to the same input data with the parameters held constant but a different random starting seed number applied to each could yield ten different results. Alternatively, the starting seed value can be fixed or a tract ID specified to initialize the algorithm, both of which would remove the randomness in repeat trial outcomes. Exploiting the variability in the output from applying random starting seed values presents an opportunity to employ Spielman & Folch’s regionalization algorithm as a tool for iterative zone design. Regionalization based upon sampling error, proximity, and attribute similarity offers more reliable ACS estimates. Could these new regions also possess meaning in a theoretical or sociological sense? That is, can a thoughtful, data-driven approach to regionalization based on ACS estimates identify places with real-life significance as captured in some of the many definitions of a neighborhood? Seeing potential for Spielman & Folch’s regionalization
algorithm to tackle both challenges of reducing uncertainty and identifying meaningful neighborhood units, I pose the following questions for analysis:

- Can Spielman & Folch’s regionalization algorithm serve as a zone design tool beyond reducing uncertainty?
- Do regions produced by the regionalization algorithm resemble neighborhoods as identified in theory, popular media, or government designations?
- If patterns of uncertainty reflect characteristics of the respondents and where they live, do regionalization patterns also reveal information about places and the people who live there?

DATA

The following analysis attempts to answer these questions by applying Spielman and Folch’s regionalization algorithm to the Denver-Aurora-Lakewood, CO Metropolitan Statistical Area (Denver MSA) using 2011-2015 ACS 5-Year Estimates (U.S. Census Bureau, 2016a). The 2011-2015 ACS 5-Year Estimates are the most recent ACS estimates available containing tract-level data. Choosing to focus on one metropolitan region affords deeper exploration of regionalization outcomes in relation to the local landscape. Also, limiting the analysis to one MSA enables more in-depth comparison of regionalization outcomes using different combinations of input variables. Additional data in the form of references to common definitions of neighborhoods in the Denver MSA provide important context for interpreting the utility of regionalization outcomes.
2011-2015 American Community Survey Five-Year Estimates

According to ACS sample size information, in the five-year period from 2011-2015, 17.44 million housing unit addresses were selected to receive a survey resulting in 11 million housing units that were successfully surveyed (U.S. Census Bureau, n.d.b). Surveys were distributed on a rolling, monthly basis with approximately equal numbers of participants surveyed each month (Torrieri et al., 2014). Participation is mandatory, although the Census Bureau has not exacted punishment for lack of participation since the 1970 census (Selby, 2014). In the first step of the survey process, the Census Bureau mailed selected participants a request to complete the survey online, followed by another mailing containing the paper version as an alternative (Torrieri et al., 2014). If the participant failed to respond, the Census Bureau conducted follow-up via phone and finally an in-person visit.

The 2015 survey asked over 75 questions, some specific to the household and others directed at each person in the household (U.S. Census Bureau, n.d.c). The questions included multiple choice and open-ended questions. Topics covered included housing, income, sex, age, race, household size and structure, employment, education, transportation, marital status, and disability, among others.

The ACS’ continuous sampling methodology and annual data releases enable improvements in survey methods and techniques to be implemented frequently (Powers, Beede & Telles, 2015, p. 12). Continuous sampling can also create inconsistency across years. The survey received by participants in 2015 was slightly modified from the survey used in 2011. For example, the 2015 survey asks three questions about computer and internet access that were not asked in 2011 (U.S. Census Bureau, 2015; U.S. Census Bureau, 2011). Most questions are very consistent between years.
Continuous sampling also means that the estimates contain temporal inconsistency. Someone interviewed in July, 2011 might give very different responses if surveyed instead in July, 2015. An interview response provided in July, 2011 contributed to multi-year estimates beginning with the 2007-2011 Five-Year Estimates and will persist in ACS data through the 2011-2015 Five-Year Estimates. One survey thus contributes to five-year estimates spanning nine years (i.e. 2007-2015). For a July, 2011, survey, the span covers a period just before the Great Recession through the ensuing economic recovery, a period demonstrating how national trends like changes in employment and housing can have very different effects year-to-year. On a smaller scale, cities and neighborhoods can also change quickly during periods of rapid growth or decline in population, land value, or another high-impact metric. There is no good solution to the temporal vagueness in the data. 1-year and 3-year estimates are available but only for geographies at the county level or higher. The 5-year estimates should be interpreted as a moving window, and the Census Bureau warns against comparing datasets with overlapping time frames.

**Denver Metropolitan Region**

Metropolitan statistical regions (MSAs) are defined by the Office of Management and Budget (OMB) within the U.S. Department of Commerce as an official statistical geography for census reporting purposes (U.S. Census Bureau, 2016, April 27). Each MSA consists of a group of counties representing a metropolitan region and is revised periodically to account for population growth and decline. MSAs are useful units for understanding regional trends and are frequently employed to compare data between cities, such as employment statistics. MSAs were most recently updated in OMB Bulletin No. 15-01 released in July 2015.

The Denver-Aurora-Lakewood, CO, MSA, depicted in Figure 4, is the largest
metropolitan area in Colorado. It encompasses ten counties home to approximately 2.7 million people, representing nearly half of Colorado’s total population of 5.5 million (U.S. Census Bureau, 2016a). Denver is the most populous city and county in the region with 683,906 residents as of July 2015, followed closely by adjacent Arapahoe County to the east (pop. 630,564), Jefferson County to the west (pop. 565,230), and Adams County (pop. 490,829) to the northwest (State Demography Office, 2017). The Denver MSA shares borders with the Colorado Springs, Boulder, and Greeley MSAs, creating a vast swath of contiguous urbanized areas blanketing the front range of the Rocky Mountains. The Regional Plan Association frames these MSAs within a larger Front Range megaregion stretching from Cheyenne, WY to Albuquerque, NM, one of 11 megaregions they identify in the US (Regional Plan Association, 2016).

The Denver MSA covers 8,346 square miles, spreading across the plains in all directions and into the foothills of the Rocky Mountains to the west (Denver-Aurora-Lakewood, CO Metro Area; Colorado, 2013). The region is characterized by largely suburban development patterns, although numerous cranes in the city center skyline attest to recent investment in high density housing near the urban core, accompanying renewed emphasis on public transit (Goetz, 2013). New development projects throughout the region include condominium and apartment buildings, traditional single-family houses, and all forms between, such as large planned developments that represent a mash-up of the historical company town and smart cities (Grabar, 2016).

The Regional Plan Association forecasts ongoing population growth in the region, a trend that the Denver MSA exemplifies. In 2010, Denver ranked 21st in population among all MSAs, with 2.55 million residents. By 2015 Denver contained an estimated 2.81 million people and moved into 19th position, eclipsing St. Louis and Baltimore. In that same period, the net population change rose from 10,868 people in 2010 to 58,474 in 2015, placing the Denver MSA

Recent growth fits a pattern throughout Denver’s history of periodic rapid expansion (Ballast, 1995). From the city’s early establishment as an outpost of the gold rush in the mid-19th century, through its development as a center for transportation and energy in the early 20th century, Denver weathered economic depressions and overcame resource constraints – namely water – to become a hub of government employment and international investment in the latter half of the 20th century.

Like much of the U.S., Denver experienced substantial post-WWII population growth that led to increasing investment in business and housing outside the Denver core, aided by road and infrastructure investments (Ballast, 1995). Between 1950-1957, Denver’s population grew 39% while the suburbs exploded, ranging from 152% growth in Littleton to 334% growth in Arvada. By 1957, Denver was the 3rd fastest growing metro area after Houston and San Diego. A pattern of growth in the suburbs continued through the late 20th century, despite moderate population loss in the City of Denver between 1970 and 1980. The history of the Denver-Aurora-Lakewood MSA followed suit, growing from four counties in 1950 to today’s present ten county extent by 2003 (U.S. Census Bureau, 2016, April 27).
Figure 4: Denver-Aurora-Lakewood Metropolitan Statistical Area Counties (Baselayer: Bing Maps Road, 2010)
Denver MSA Neighborhoods

Neighborhood designations vary greatly throughout the Denver MSA depending on the history, landscape, and organization of the region’s constituent cities. The City of Denver, which is equivalent to Denver County, designates 70 official statistical neighborhoods, which were established in 1970 by the Denver Dept. of Community Planning and Development (Figure 5). The neighborhoods are composed of census tracts, so they align with ACS data. Neighborhood names were informed by popular knowledge and the history of the area (City and County of Denver, 2016).

Figure 5: Denver County Statistical Neighborhoods (City and County of Denver, 2016)

The City of Denver also allows communities to register as Registered Neighborhood Organizations (RNOs), “groups of residents and property owners that represent significant geographic areas within the city and that usually do not charge membership fees” (City and
County of Denver, 2017). The community chooses its RNO boundaries, which tend to be much smaller than census tracts though they vary widely in size. Sometimes an RNO is coincident with a housing development, such as a condominium association, though an RNO is not a Homeowners Association. Currently 194 RNOs are registered in Denver. Although there is no broad mission for all RNOs, each sets its own priorities. The Highland United Neighbors, Inc. RNO states its mission is “to facilitate consistent and responsible communication among Highland neighbors and the community at-large, to improve the quality of life for its residents, organizations and businesses and to provide advocacy and promotion for our community” (Highland United Neighbors Inc., 2016). The RNO organizes residents and businesses around social events, volunteering, and advocacy. Denver’s statistical neighborhoods better represent an externally imposed neighborhood, with value for political and commercial recognition from city government and other organizations. RNOs, in contrast, reflect internally-organized neighborhoods as perceived and valued by the people living and working within them.

As the central city in the Denver MSA, Denver’s neighborhoods are widely adopted and advertised by tourist and media groups. Visit Denver is a nonprofit trade association that serves as Denver’s tourism organization, funded by private organizations, public taxes, and business membership fees (Visit Denver, n.d.). Visit Denver provides neighborhood guides for some Denver neighborhoods (e.g. Downtown Denver, Cherry Creek, Five Points) as well as outlying suburban areas (e.g. Olde Town Arvada, Golden, Littleton). The guides identify neighborhoods on a map centering on the main commercial district but do not demarcate boundaries. Denver.com, another city guide, identifies a wider swath of neighborhoods and suburbs with a larger geographic extent including small towns in the fringes of the metro area (Boulevards, 2017).
Neighborhood designations exist beyond Denver County but are fewer and less well-defined. The cities of Aurora and Littleton both designate neighborhoods on a map (City of Aurora, n.d.a; Littleton, CO 2017). Aurora’s neighborhoods align closely with census tracts while Littleton’s do not. The City of Aurora partitions neighborhoods into three groups that correspond with Neighborhood Liaisons that the city employs (City of Aurora, n.d.b). The City of Golden identifies nine neighborhoods in a series of ‘Neighborhood Plans’ that attempt to incorporate “community values” into city planning processes (City of Golden, 2016). The Neighborhood Plans function as a tool to incorporate local input into city planning. Other cities such as Thornton allow communities to register neighborhood associations but do not specify existing neighborhoods or associations (City of Thornton, 2017).

Master-planned communities on the edge of urban and rural parts of the Denver metro region present a unique case for neighborhood identification. Developers define the boundaries of master-planned communities, but whether clear neighborhood identities exist within them is unclear. Sterling Ranch is a sizeable master-planned community rising in Douglas County (Sterling Ranch Development Company, 2017). Plans for Sterling Ranch contain nine designated villages, imposing neighborhood-like structure onto suburban communities yet to be built. The Pinery, also in Douglas County, bills itself as a neighborhood but is defined and regulated as an HOA (The Pinery HOA, 2015). Imposed boundaries drawn by developers and HOAs work in tandem and contrast with the internal, organic evolution of neighborhoods modeled by neighborhood associations and RNOs. The evidence across the Denver MSA suggests Suttles’ ‘push and pull’ dynamics of neighborhood identification at work (Suttles, 1972).

Rural parts of the MSA present a separate challenge for neighborhood identification. Elbert County to the southeast and Park County to the southwest are predominantly rural in
character. Elbert County’s largest city by population is Elizabeth, containing fewer than 1,500 residents. Seven census tracts capture all of Elbert County’s residents (U.S. Census Bureau, 2016a). Given the rural character of these places, are there neighborhoods to be found beyond town centers? Could the entire county be considered a neighborhood? These theoretical questions lie beyond the scope of the present analysis but provide additional evidence for use in understanding the utility of regionalization output.

**Variable Selection**

As noted previously, Spielman & Folch’s regionalization algorithm accepts a user-specified selection of variables. ACS variables contain differing levels of error and the spatial distribution of those errors across the Denver region can also vary, so the variables chosen can impact the regions produced. If the regions produced by the algorithm have meaning beyond reducing error as hypothesized, then the choice of variables to include is an important parameter.

Like the identification of neighborhood boundaries, complex factors concern variable selection. Using a set of variables related to housing will produce regions that with reduced sampling error for housing metrics but not necessarily for another topic of interest such as transportation. Variables with especially high sampling error and heterogeneity across the study area might exercise a heavy influence on the results if they exhibit spatial clustering patterns. Anytime they are included they could steer the output toward certain solutions. Conversely, high variable randomness could mean no spatial patterns exist in the input data, leading to no consistent patterns of region formation. The opposite circumstance, extreme homogeneity in inputs data, could also negate the algorithm’s ability to capture spatial patterns in region output, again leading to no consistent patterns of region formation. Spielman & Folch (2015) also found
that the number of variables chosen affects the size of output regions (p. 17). Finding a solution that reduces error sufficiently for two variables will likely not suffice for ten variables.

For a narrow topic, such as an investigation of poverty among seniors, variable choice might present a simple task. This analysis seeks to understand the potential to identify neighborhoods via the regionalization algorithm, begging a more difficult question, what set of variables describes a neighborhood? As outlined earlier, neighborhoods are complex, vague, and defined through internal and external forces. Conversely, they can be simplistic and utilitarian, as when they are equated with census tracts. The regionalization approach seeks a middle ground, which will be reflected in variable selection.

Spielman & Folch (2015) test four sets of variables to demonstrate the potential of the regionalization algorithm. The sets center on four themes or “scenarios” (p. 7), housing, poverty, transportation, and a general scenario. The four scenarios contain a combination of variables computed from two or more ACS variables as well as a population count variable (see Appendix for ACS variables used in each scenario). The housing scenario, for instance, consists of seven computed variables and an eighth population count variable, all derived from 12 ACS variables. The computed variables include ratios, which compare two distinct ACS variables such as average rooms per housing unit, and proportions, which calculate the fraction of a sub-variable within a larger group, such as percent of housing units occupied.

Spielman & Folch designed the four scenarios to demonstrate prospective applications of the algorithm across domains. This analysis will utilize the same variable sets out of utility and convenience. The selections represent a combination of metrics addressing specific themes and have been tested in the algorithm (using the 2008-2012 ACS Five-Year Estimates instead of the 2011-2015 estimates). By applying different combinations of input variables to the
regionalization for each MSA, the effects of variable selection and their accompanying sampling error upon regionalization outcomes may become evident.

This analysis will test each scenario but focus primarily on the general scenario. The general scenario contains variables addressing diverse themes that provide a broad information base to define a neighborhood such as age, education, residential mobility, race and ethnicity, housing, family structure, and income (Table 2). One could craft numerous other variable combinations for consideration, just as neighborhood boundary lines can be drawn any variety of ways. Promising results might demonstrate value in tweaking the input variable set, and investing time in aligning variables with specific research pursuits. As a preliminary investigation, using the four scenarios will help illustrate the extent to which the variable combination employed impacts outcomes.

<table>
<thead>
<tr>
<th>General Scenario Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average rooms per housing unit</td>
</tr>
<tr>
<td>Average household income</td>
</tr>
<tr>
<td>Average persons per household</td>
</tr>
<tr>
<td>Percent of housing units occupied</td>
</tr>
<tr>
<td>Percent of households with married-couple family</td>
</tr>
<tr>
<td>Percent of population age 25+ with bachelor's degree or higher</td>
</tr>
<tr>
<td>Percent of population that lived in the same house 1 year ago</td>
</tr>
<tr>
<td>Percent of population non-Hispanic white</td>
</tr>
<tr>
<td>Percent of population non-Hispanic black</td>
</tr>
<tr>
<td>Percent of population Hispanic</td>
</tr>
<tr>
<td>Percent of population under 18</td>
</tr>
<tr>
<td>Percent of population over 65</td>
</tr>
<tr>
<td>Total population</td>
</tr>
</tbody>
</table>

Table 2: General Scenario Variables

METHODOLOGY

Adopting an exploratory approach, the methodology for this analysis applies quantitative methods followed by a qualitative discussion to evaluate the potential for Spielman and Folch’s
regionalization algorithm to produce statistical units with analytical meaning beyond reducing sampling error. Quantitative measures based on network graph theory are employed to identify stable regions produced through repeat trials of Spielman and Folch’s regionalization algorithm. Findings from the network graph analysis are then used to compare stable regions produced by the algorithm to Denver MSA neighborhoods, as understood through various demographic, landscape, historical, and other factors.

**Running the Algorithm**

Spielman & Folch’s regionalization algorithm relies on user-defined inputs including census geographies, ACS variable data, and an acceptable CV threshold. The algorithm also depends on a random seed value set at the start of each trial, which determines the tract from which the algorithm begins to formulate solution regions. Changing the seed value results in a new random starting point, potentially leading to a different set of solution regions. Exploiting the variation in regionalization results, this analysis runs 100 trials of the regionalization algorithm with different starting seed values for each trial ranging from 789 - 888 (the value of the starting seed is not important beyond its role in aiding randomization).

A U.S. Census shapefile containing 621 tracts in the 2015 delineation of the Denver MSA provides the input geography (U.S. Census Bureau, 2017b, March 7). ACS variables were obtained from the Census API and used to compute the variables in the general, poverty, housing, and transportation scenarios (U. S. Census Bureau, 2017a, March 7). The acceptable CV threshold is set at 0.12, consistent with Spielman & Folch’s implementation of the algorithm in consultation with other sources (see Spielman & Folch, 2015, p. 3 for additional detail). Another user-defined parameter is the acceptable size of a small estimate to exclude from the CV
constraint. The purpose of this parameter is to prevent disproportionate influence from variables calculated as proportions that have very low estimates compared to the total population, which inflates the impact of the CV constraint. The parameter is set to 0.05, meaning that for variables calculated as proportions in which the estimates are less than 5% of the population then the CV threshold of 0.12 will not apply (see Spielman & Folch, 2015, p. 9 for further explanation of the effects of including CV values for very low estimates). Other user-defined parameter options available are not applied in this analysis (see Spielman & Folch, 2014, for code, complete parameter descriptions, and examples).

In total, 400 trials of the algorithm are run, 100 for each scenario applying the same range of starting seed values. The result of running the algorithm as described is a set of 100 regionalization outcomes for each scenario, including a map of output regions for each trial, and estimates and CV values calculated for the new regions. (For analysis of how input tract estimates and CV values compare to output region estimates and CV values, see Spielman & Folch, 2015, p. 10-14.) The cumulative results of each batch of 100 trials are aggregated to form a larger data set that enables region stability assessment.

**Measuring Region Stability**

Quantifying differences in regionalization outcomes presents an opportunity and challenge. Each regionalization solution can produce unique groupings of input tracts into output regions. The current analysis uses those repeat results from 100 regionalization trials as the basis for identifying stable regions that might inform or correspond with real demographic and development patterns. The challenge lies in constructing a metric to measure the stability of those groups when they may constantly change. To demonstrate the challenge, a simplified
example follows.

In the example below, tracts A through I are the input units to the regionalization algorithm. For simplicity, they are spatially organized in a gridded square (Figure 6). After inputting the tracts with their associated estimates and sampling error into the regionalization algorithm, the input tracts are grouped into output regions for each trial as depicted in Figure 6 and quantified in Table 3.

![Figure 6: Sample Regionalization Results](image)

**Table 3: Tabulated Sample Regionalization Results**

<table>
<thead>
<tr>
<th>Trial</th>
<th>Input tracts</th>
<th>Output region</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A, B, D, E</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>C, F</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>G, H, I</td>
<td>III</td>
</tr>
<tr>
<td>2</td>
<td>B, C, F</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>A, D, E, G, H, I</td>
<td>II</td>
</tr>
<tr>
<td>3</td>
<td>E, F, H, I</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>G, D</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>A, B, C</td>
<td>III</td>
</tr>
<tr>
<td>4</td>
<td>D, E, F</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>A, B, C</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>G, H, I</td>
<td>III</td>
</tr>
<tr>
<td>5</td>
<td>A, D, G</td>
<td>I</td>
</tr>
<tr>
<td></td>
<td>B, E</td>
<td>II</td>
</tr>
<tr>
<td></td>
<td>C, F</td>
<td>III</td>
</tr>
<tr>
<td></td>
<td>H, I</td>
<td>IV</td>
</tr>
</tbody>
</table>

Tracts change membership between output regions across trials. In the example in Figure 6, Tract A occurs in output region I (Trials 1 and 5), II (Trials 2 and 4), or III (Trial 3). The number of output regions in each solution also changes, ranging from two regions in Trial 2 to four regions in Trial 5. The locations of the regions are not fixed. Region I represents the northwest corner of the map in Trial 1, the northeast corner in Trial 2, and the southeast corner in Trial 3. Inconsistency between tracts and regions means that a region-based identification
approach would be difficult or impossible to implement. Instead, a metric that measures group stability must focus on the relationships between tracts as they move in and out of regions across trials.

Instead of assessing relationships between regions and tracts, stability can be evaluated as a product of each tract’s relationship to other tracts that are grouped into the same region. Using the prior example, tract A occupies the same region as tracts B, C, D, E, G, H, and I when all trial results are combined. Tract A never falls in the same group as tract F, meaning tract A has seven unique partners out of eight possible tract partners. As evidenced in this example, tracts need not be adjacent to share a group. Some tracts will act as bridges between tracts in the output region. For instance, tract A pairs with C, I, G, and H at least once during five trials despite not sharing a boundary with them.

<table>
<thead>
<tr>
<th>Source Tract</th>
<th>Target Tract</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 4: Source-Target Pair Frequency for Sample Regionalization Results**

Tract A shares an output region with some of those unique partners more often than others, as shown in Table 4. Tract A is grouped with tracts B and D in three trials but pairs with tracts H and I in only one trial. Repeating the same calculation for each partner of tract A produces the third column in Table 4 describing the number of times tract A, the source tract, pairs with each target tract, labeled here as the ‘frequency.’ Pair frequency is a proxy for shared group membership, and thus pair stability across trials. The maximum potential frequency is the
number of trials run, which is 5. High frequency means tracts share a region in regionalization output often. A frequency value of 0 means the two tracts never share a region. A low frequency of 1 means the tracts seldom share a region.

Repeating frequency calculations for all source tracts and their target pairs based on their shared membership in output regions creates a body of data to explore and assess region stability. The data consists of an expanded version of Table 4 in which all tracts 1 through \( n \), contain a list of partners 1 through \( n \), each with a frequency value of 1 or greater, excluding tract pairs with a frequency of 0, i.e. those that do not share a region in any trials. The data contains duplication because tract A’s pairing with tract B mirrors tract B’s pairing with tract A. The relationship has no directionality and thus each pair is the same regardless of whether tract A or B is considered the source.

A Network Graph Approach

The resulting data set lends itself well to a network graph approach to analyze regionalization output. In a network graph representation of the output, a ‘node’ represents each tract and an ‘edge’ (the connection between nodes) indicates shared membership in a group, referred to above as a source-target pair (Table 4). A network graph representing input tracts and output regions for 100 regionalization trials for the Denver MSA would contain one node for each of the 614 tracts in the MSA and an edge between any pairs of nodes that share membership in a group in a least one trial. As in the first example, tracts need not share a boundary to form a relationship if they share membership in a group, hence tracts that are not adjacent on a map can be represented as connected nodes in a network graph.

Methods available to evaluate network graphs derive from various domains and
applications. Network graphs are used to evaluate physical pathways like transportation networks, migration trajectories, and biological systems, as well as abstract networks such as social networks and economic transactions (Barthélemy, 2011; Kolaczyk, 2009). Spatial phenomena represented as network graphs can abstract from a coordinate interpretation of space but retain spatial information in the network design: “the topological aspects of the network are…correlated to spatial aspects such as the location of the nodes and the length of edges” (Barthélemy, 2011, p. 2). The network graph’s abstraction from a strict representation of space supports an altered interpretation of space in which relationships between potentially distant places gain precedence relative to objects that may be near in coordinate space but have no meaningful relationship.

In a network graph representation of regionalization results, the frequency of connections between tracts measures the closeness of the relationship across trials, and thus their stability as a pair. Euclidian distance is not significant for regionalization outcomes if neighboring tracts never share a region, although adjacency is required for cumulative region construction. Tract A may be more likely to pair with tracts B and D than any other tract because they share borders, but if the attributes (estimates and sampling error) of tracts A and D are not complementary as interpreted by the regionalization algorithm, tract A might have a higher likelihood of sharing a region with tract E via the pathway through tract B. Adjacency does not guarantee a significant relationship.

Figure 7: Network Graph for Sample Regionalization Results
Network graph utility extends to its visual representation of spatial relationships via region membership. The example depicted in Figure 6 is represented as a network graph in Figure 7. Map-based depictions of the Denver MSA suffer from the distortion imposed by size differences between measured units, a common problem with choropleth maps (Heer, Bostock & Ogievetsky, 2010). Census tracts are roughly based on population size, so rural tracts consume greater land area than urban tracts to accommodate comparable population counts. Visually comparing spatial patterns between small tracts in central Denver and vast tracts in rural Elbert County is difficult. In the network graph, each node occupies the same amount of space so relationships and stability can be evaluated with less visual distortion.

**Metrics**

The value of the network graph lies in its ability to depict and describe a complex network of relationships between tracts and region membership. Extreme complexity stemming from numerous nodes and connections with little visible order can obscure underlying trends, requiring additional metrics to describe and analyze the graph. To understand the organization of census tracts into groups as a product of regionalization, this analysis applies three metrics that offer related but nuanced views on the relationships between tracts and regions. Understanding the relationship between pairs contributes to insight about regions when all pairs are combined, measured, and visualized.

The three measures assessed, frequency, degree, and entropy, aim to isolate and characterize clusters of tracts to determine if they represent stable regions. Understanding region stability from tract-level measures requires determining, who are the tract’s partners? How often does the tract form those partnerships? Is the tract located with other tracts that demonstrate
similar pair formation patterns? A tract with a modest number of consistent partner tracts suggests membership in a stable region. If that tract is surrounded by other tracts with similar tract-pair relationship patterns, a stable region is more likely. Conversely, a tract with many sporadic pairings located in an area in which other tracts demonstrate random patterns of shared group membership will be less likely to participate in a stable region.

Degree, frequency, and entropy help answer these questions and identify patterns of stability for each tract in the region. Degree quantifies a tract’s partners, frequency measures how often the tract shares a region with those partners, and entropy provides a metric for localized dispersion of tract relationships. Frequency is the primary measure used to identify stable regions, but the combination of all three metrics provides a more complete picture of regionalization outcomes.

Figure 8 depicts a perfectly stable region. All nodes (tracts) connect to all other nodes, and these partnerships occur 100 times in 100 trials. One measure of stability, the Gamma Index is useful for evaluating isolated regions like that in Figure 8. The Gamma Index ($\gamma$) compares existing edges in a network graph ($E$) to all potential edges ($E_{max}$).

$$\gamma = \frac{E}{E_{max}}$$

$E_{max}$ is calculated based on the number of nodes (N), as $N(N-1)/2$. For the graph in Figure 8, ten edges exist out of ten possible edges ($5(4)/2 = 10$), leading to a gamma index of 1.0. Perfect region stability like that in Figure 8 is unlikely, but using degree, frequency, and entropy to identify groups that come close to approximating the relationships in Figure 8 throughout regionalization outcomes would provide substantial evidence for the
persistence and importance of meaningful output regions.

Degree

Table 3 can be reconfigured as an adjacency matrix (Figure 9) in which paired tracts receive a value of 1 and unpaired tracts receive a value of 0 (Barthélemy, 2011, p. 5). Summing the values in each column or row results in a tract’s degree. A node’s degree is the number of unique nodes with which it shares an edge. In regionalization results, a tract’s degree reveals how many other tracts it shares a group with throughout 100 trials. The maximum degree possible is 613, the total number of census tracts containing residential population in the region less the source tract. A degree of 613 is an unlikely result since that would mean at least one outcome of the algorithm is a single all-encompassing region. The minimum degree possible is zero, which might occur if a tract meets the algorithm’s parameters without forming a pair.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 9: Adjacency Matrix for Sample Regionalization Results

Degree provides context for understanding stability. A node with a low degree might pair with a few tracts repeatedly. A node with a high degree might imply instability because it shares group membership with many different tracts across the 100 trials. However, degree is an
indirect indicator of stability. If a tract shares an edge with another tract only once in 100 trials, that pairing is weighted as heavily as a connection with another tract that occurs in 99 out of 100 trials. Degree does not account for the frequency of connections. Conversely, a tract with a moderately high degree might participate in one very large region rarely, raising its degree by the number of tracts in that rare region, but participate in a consistent, small region repeatedly. Or a tract may have a moderately high degree because it participates in a consistently occurring large region. A tract’s degree has implications for potential region stability and region size, which requires untangling the interaction of both. The degree of one tract in isolation provides limited information, but degree values for all tracts amount to collective insights. Examining the distribution of degree values within and across scenarios should prove especially useful for comparing network structures.

*Frequency*

Pair frequency helps contextualize tract degree. The example described in Table 3 and Figure 6 described how to assess the frequency of tract pairs, a metric known in network graph parlance as the ‘edge weight.’ Frequency captures how often two tracts share membership in the same output region. An edge weight of 10 indicates two tracts are part of the same region in 10 trials. Since this analysis runs 100 regionalization trials for each variable set, the maximum potential edge weight for any pair of tracts is 100, representing perfect stability. An edge weight of 1 would indicate the opposite extreme for connected tracts, two tracts that co-habit a region but only very rarely, once in 100 trials. Perfect stability as embodied by a high frequency pairing suggests two conclusions: first, that the level of sampling error in the input tracts necessitated combining at least those two tracts and possibly more depending upon the size of the group.
Second, it suggests that those tracts possess ACS estimates that are more alike relative to other possible tract combinations.

Measuring frequency between pairs translates into insights about regions when all pairs are combined, measured, and visualized. A stable region that manifests in most trials must consist of tracts that form pairs in most trials. A perfectly stable region would consist of a set of tracts that form pairs only with each other and no other tracts in the region 100 times in 100 trials. Since frequency is a proxy for assessing stable regions, distinguishing between more and less stable pairs is important. Edge weight thresholds help evaluate the distribution of high, moderate, and low frequency pairs among all pairs that occur.

This analysis applies edge weight thresholds of 90 and 60. Two tracts that form pairs in 90 or more out of 100 trials demonstrate highly stable pairings. They rarely belong to separate regions. Two tracts that form pairs in 60 or more out of 100 trials demonstrate moderately stable pairings. They share regions in more than half of the trials, but often form alternate pairings. Edge weight thresholds other than 60 and 90 might merit exploration depending on the outcomes of initial data exploration, but as benchmarks these two levels can help depict overall trends in pair formation and isolate parts of the network most likely to contain stable regions for further analysis.

**Entropy**

In the context of regionalization outcomes, entropy characterizes the spread of a tract’s edge weights among all pairs it forms. While degree is agnostic to the frequency of pairings, entropy normalizes pair frequency relative to the tract’s degree. Very different measures of entropy have been employed in diverse applications. Cabral et al. (2013) summarize the
evolution of entropy from its origins in physical science into social science and spatial statistics, with a focus on its use in urban systems starting in the 1970s. The form of entropy applied here was developed by Henry Theil in 1972 as a measure of “spatial dispersion” (p. 5225) and is calculated as follows:

\[ H_n = \frac{\sum_i^n p_i \log(\frac{1}{p_i})}{\log(n)} \]

\( H_n \) is a measure of relative entropy, comparing a tract’s total entropy (the numerator) to its maximum potential entropy (the denominator). Total entropy for a tract is calculated by summing the entropy values for each pair involving the target tract. For each pair, entropy is calculated as the probability that the source tract pairs with target tract \( i \) \( (p_i) \) multiplied by the log of \( 1/p_i \). Maximum entropy for each tract is calculated by finding the log of the tract’s degree \( (n) \).

The result of calculating relative entropy is a decimal value ranging from 0 to 1. Relative entropy of 0 means that the source tract pairs heavily with a few target tracts relative to all tracts it pairs with at least once in 100 trials. Relative entropy of 1 signifies that the source tract pairs with all partner tracts across trials with equal probability, and thus there is high entropy in the system. High entropy indicates no clear pattern of edge weights among partner tracts that would separate random associations from consistent associations. Low entropy indicates concentrated connections. High entropy implies instability in a tract that shifts pairings throughout trials, and low entropy implies localized partner consistency. As with degree, examining relative entropy across the study region will likely help identify stable regions better than focusing on individual relative entropy values for tracts.

After calculating and mapping entropy values for the study area, this analysis will
calculate Moran’s I to assess global spatial autocorrelation of entropy values. If clustering is indicated, Local Moran’s I will be calculated and mapped for the study area to identify clusters of low or high entropy values, which would highlight potential areas of region stability.

**Correlation**

Frequency, degree, and entropy are related in this conceptual framing of stable regions. Stable regions are expected to consist predominantly of tracts with high frequency pairings, low degree, and low entropy. A modest mathematical relationship exists between them as well. Relative entropy incorporates frequency in the numerator in the probability of pair formation, and degree in the denominator. Degree is the unique set of all pairs captured in the frequency measure. However, the mathematical relationships are indirect, encouraging an effort to statistically quantify the relationship between these three measures among all tracts.

Statistical correlation analysis using ordinary least squares (OLS) linear regression will be applied to quantify how tightly or loosely bound degree, frequency, and entropy are for each tract. This analysis will conduct OLS regression for each pair of metrics (frequency and degree, degree and entropy, frequency and entropy). Since degree and entropy are calculated for each tract and frequency is calculated for each pair of tracts, each tract will be assigned a frequency value equivalent to the maximum edge weight among all pairs involving it. The correlations will be repeated using mean edge weight for comparison. Demonstrated correlations between the three variables would strengthen their value as indicators and provide evidence for the regionalization algorithm’s ability to identify stable regions, and ultimately neighborhoods.
From Data to Neighborhoods

Neither frequency, degree, nor entropy definitively identifies a stable region, but collectively they can help identify relative stability. Divergence in any one of these factors does not preclude stability, however some combination of indicators is necessary to provide sufficient evidence that a region merits further investigation. If stability exists, the next step of this exploratory analysis will compare the data-driven evidence of stable regions to Denver neighborhoods identified in practical applications as described above (see Data section), incorporating discussion of how the stable regions may or may not also coincidence with the theoretical underpinnings of neighborhood identification (see Literature Review section).

The final step of the analysis will consist of comparing results for the general scenario with summary results for the transportation, housing, and poverty scenarios. Subject-specific regionalization results could contrast or reinforce the results of the general scenario, and illustrate the effects of variable selection on regionalization outcomes. The comparison will rely on the same three metrics of frequency, degree, and entropy, evaluating their similarities and differences in the aggregate and among areas of especially high or low stability.

RESULTS

Degree, frequency, and entropy results for the general scenario are presented first. Stable regions identified using these metrics are then analyzed for their potential significance as meaningful units of analysis. Subsequently, summarized results for the same indicators applied to the housing, poverty, and transportation scenarios are provided to examine how variable selection impacts regionalization outcomes. The overall picture that emerges is one of instability in regionalization outcomes, hindering comparison of data-driven regionalization solutions and
neighborhoods as defined in theory and practice.

**General Scenario**

Running 100 trials of the regionalization algorithm for the general scenario variable set applied to the Denver MSA’s 621 census tracts produces solutions containing between 139 and 150 regions, with an average of 142 regions in each solution. Five tracts are excluded by the algorithm because they contain a population count of zero, meaning they lack residents, leaving 614 input tracts. The regionalization algorithm reduces the number of input zones (614) by approximately 75% to create output zones (142). The resulting regions range in size from 1 to 22 tracts but trend towards the lower side of the range with an average of 4 tracts per region. Figure 10 illustrates regionalization results for six trials for a small segment of the Denver MSA. As a representative sample, these six outcomes demonstrate modest consistency among output regions, along with consistent region boundary variation. No two outcomes are exactly alike in this sample.

![Figure 10: Sample of Six Regionalization Results](image)
Figure 11 depicts the aggregate results of all trials as a network graph. The graph appears to be highly interconnected with a high density of links obscuring any visible patterns, but the graph’s gamma index of 0.06 indicates it has low interconnectivity relative to all potential edges between nodes. For the full network graph, a low gamma index is to be expected if there is some consistency in region formation that limits very distant nodes from sharing membership in a very large region. Independently, the network graph visualization provides little information to evaluate tract-pair stability or identify stable regions within the results, requiring interpreting the graph through tract/node-based indicators.

**Figure 11:** Network Graph for General Scenario Regionalization Output
Degree

Of the 614 input tracts, the minimum degree exhibited by any tract is 7 and the maximum is 88, meaning that a single tract combines at most with 14% (88/613) of all tracts in the region across 100 trials. Figure 12 illustrates a high concentration of tracts in the degree range of 18-50, with a moderate tail extending towards higher values. Overall, the histogram does not depict a concentration of low degree values that would suggest the formation of small, stable regions. Moderate and large degree values could reflect high variability or the formation of large stable regions. Considering that the average region size is 4 tracts, the formation of large stable regions is unlikely for most high degree tracts. The mean degree for all tracts is 38, which although far from the theoretical maximum degree of 613 is still much higher than the maximum degree of a tract in region containing 4 tracts, which would be 3.

The discrepancy between mean degree and mean region size implies large variations in region composition across trials for small regions. A tract might participate in small regions in each trial, but the composition of those regions differs each trial. Figures 13 and 14 illustrate the implications of degree for two tracts, one with a high degree and one with a low degree. Figure 13a depicts the network connections for a tract with a high degree, paired with 46 other tracts. Examining the tract’s region membership across 100 trials reveals that the largest group it participates in contains 12 tracts and the smallest contains 3 tracts. Figures 13b and 13c depict...
samples of the largest and smallest regions in which the tract participates. The tract participates in small regions despite its high degree, implying high variability in the tract membership of regions to which it belongs.

Figure 14a illustrates the same information for a tract with a low degree, possessing just 9 partner tracts. The largest region containing the tract consists of 6 tracts (Figure 14b), and it is an outlier among the results. Figure 14c displays a region containing the same source tract but only 3 total tracts, and is a more typical result for this tract. 37 out of the 100 regions involving the tract are composed of just 3 tracts. Figure 14b is an anomaly, whereas Figure 14c is a more typical representation. One anomalous region can increase a tract’s degree substantially, but for a
tract with a low degree such anomalies must be few and other partnerships consistent.

The examples in Figures 13 and 14 demonstrate how degree is a supportive but imperfect measure of tract stability, and help contextualize region size. Tracts with a high degree have a wider range of possible region formations, while tracts with a lower degree are more limited in their connections and thus consistent in region membership. The initial results from assessing degree suggests small regions are common among regionalization results but their membership is variable, which bodes poorly for region stability.
**Frequency**

The 614 nodes in the general scenario network form 11,636 edges, meaning that there are 11,636 pairs of tracts that share membership in a region at least once in 100 trials, excluding reciprocal pairs (i.e. tract A to B and tract B to tract A count as the same edge). Figure 15 illustrates the distribution of tracts across the possible frequency range, 0 to 100. Most edges concentrate in the low frequency range, with a long, diminishing rightward tail towards higher frequency connections. High frequency connections are necessary for stable regions, so the small number of tract pairs in the small rightward tail are of most interest for this analysis.

Table 5 translates the values in the histogram into frequency thresholds. Of the 11,636 edges in the network, 73 meet the high frequency threshold of occurring in 90 or more trials, accounting for 0.63% of all edges. 576 edges, or 4.95%, occur in 60 or more trials, indicating moderate frequency pairings. Most pairs are unstable and arise through random combinations in shared regions. 82.38% of all edges occur in 25 or fewer trials.

![Figure 15: Distribution of Edge Frequency Among Tract Pairs in the Generalization Scenario](image)

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Edges</th>
<th>Percent of all Edges</th>
<th>Nodes</th>
<th>Percent of all Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>5</td>
<td>&lt; 0.01%</td>
<td>10</td>
<td>1.63%</td>
</tr>
<tr>
<td>≥ 90</td>
<td>73</td>
<td>0.63%</td>
<td>125</td>
<td>20.36%</td>
</tr>
<tr>
<td>≥ 75</td>
<td>279</td>
<td>2.40%</td>
<td>332</td>
<td>54.07%</td>
</tr>
<tr>
<td>≥ 60</td>
<td>576</td>
<td>4.95%</td>
<td>508</td>
<td>82.74%</td>
</tr>
<tr>
<td>≤ 25</td>
<td>9586</td>
<td>82.38%</td>
<td>614</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

**Table 5:** Nodes and Edges Meeting Frequency Thresholds for General Scenario
Interpreting the small number of high and moderate frequency tract pairs in terms of the number of census tracts in the region that participate in those pairs conveys a more promising picture. Despite the high degree of randomness in the composition of regions indicated by the large proportion of low-frequency pairings, 125 tracts, or 20.36% of all tracts in the region, participate in a pair that occurs in 90 or more trials. 508 tracts, 82.74% of all tracts in the region, participate in 576 moderate-frequency pairs (60 or greater occurrences across 100 trials). The high volume of random, low frequency pairs contrasts with a small volume of highly or moderately stable pairs that involve most tracts in the region. More than four fifths of all tracts share a region with one other tract 60 or more times in 100 trials, implying that there are some tracts that are likely to share a region regardless of the algorithm’s random starting point. Amidst mostly random pairings, consistent shared region membership exists.

Distilling the original network graph into only nodes with high frequency connections highlights potential stable regions. Figure 16 depicts a network graph representing only the 73 high frequency edges occurring in 90 or more trials. The graph contains sparse groups of two, three, and four linked nodes. Isolated pairs represent two tracts that share the same region in 90 or more trials. In seven instances, three tracts connect through one central node, and one cluster of four nodes is fully interconnected, each tract sharing a high frequency pair with every other tract in the group (identified in red, Figure 16). No group of connected tracts in the high-frequency network contains more than four nodes, demonstrating that only small groups exist as high-frequency partnerships.
Figure 16: Network Graph for Tract Pairs that Occur in 90 or More Regionalization Trials

Visualizing the same high frequency tracts on a map of the Denver area reveals that all high-frequency pairs represent adjacent tracts (Figure 17). Some of the pairs are spatially isolated whereas others border another high-frequency pair, creating the appearance of high stability regions composed of four or more tracts, but lacking a high frequency edge connecting them in the network graph. The high frequency pairs are relatively dispersed, occurring in high and low population density tracts (roughly approximated by the size of the tract).
The sole group of four interconnected tracts in the high frequency network represents the largest highly stable regionalization output and bears further examination as a high stability region. The group is circled in red in Figure 17 (inset) and will be referred to as cluster A. Cluster A has a gamma index of 1.0, indicating that it is interconnected to the maximum extent possible. A more extensive analysis of this region follows in the neighborhood assessment section below.

An isolated pair of high frequency tracts, although smaller than the average data-driven region, could adequately capture a realistic neighborhood size. Many of Denver’s statistical neighborhoods are composed of just one or two census tracts, and registered neighborhood
organizations are sometimes even smaller. However, small regions encounter the constraints of data uncertainty, raising the question of whether any of the high frequency pairs contain low enough sampling error to suffice as independent regions.

Relaxing the frequency constraint to consider moderate frequency pairs expands the potential for finding stable regions. Figure 18 shows a network graph for moderate frequency tract pairs, those occurring in at least 60 of 100 trials. Groups larger than two tracts are more common and larger groups exhibit more varied network structures. Moderate frequency groups range in size from 2 to 28 tracts, and include nodes that are highly interconnected and others dependent on connections through intermediary nodes.

**Figure 18:** Network Graph for Tract Pairs that Occur in 60 or More Regionalization Trials
Visualizing the moderate frequency regions on the Denver MSA map (Figure 19) illustrates the earlier finding that most tracts in the Denver MSA are involved in a moderate frequency pair. Delineations of pairs or groups is more difficult in the moderate frequency map because so many tracts are involved, necessitating comparing the structured groups in the network graph to the map.

**Figure 19:** Census Tracts in Denver MSA that Participate in Moderate Frequency Pairs

Understanding region stability requires using both the map and network graph to tease apart the relationship between structure, geography, and stability. The aspatial configuration of nodes in the network graph can be manipulated to adopt limitless formations as long as the links remain in place. Conversely, the map anchors nodes with spatial coordinates but does not illustrate the nature of relationships between tracts. Like cluster A in the high frequency graph, the following two tract clusters identified in the moderate frequency network graph represent highly interconnected groups. They will also serve as case studies for the neighborhood assessment that follows.
In cluster B (Figure 20), a group of ten tracts displays high interconnectivity in the network graph representation. The cluster contains 33 edges among 10 nodes, resulting in a moderately high gamma index of 0.73. The tracts form a partially compact cluster on the map although the compactness of the region is reduced by idiosyncracies in the input tract shapes. Three member tracts share only one boundary with the region in the map view, but the network view indicates that all but one tract have at least four moderate frequency connections to other tracts in the cluster.

Cluster C (Figure 21), like cluster A, is perfectly bonded with a gamma index of 1.0. All
tracts have moderate frequency connections with other tracts. The four vertically aligned tracts in the map view are also involved in high frequency pairs (> 90 shared membership) as two separate pairs. Cluster C gives evidence for the relationship between high frequency pairs and moderate frequency groups, illustrating a continuum along which isolated stable pairs grow to become larger groups as the frequency constraint is relaxed.

The prior examples use 90 and 60 as cutoff weights to examine pair frequency. Cluster A remains a group of just four tracts in the moderate frequency scenario, while other tracts like those in cluster C accumulate connections as the edge weight cutoff is relaxed. The threshold for frequency strength can be adjusted to target groups of varying sizes, serving as another user-input to adjust results and utilize the algorithm as a zone-design tool.

**Entropy**

Tract entropy values range from 0.44 to 0.91, with 50% of values concentrated between 0.76 and 0.84 (Figure 22). Outlying values skew towards the low end of the range but overall entropy is high, mirroring visible patterns of high degree values and low frequency, and indicating high variability among pair partner formation.

![Entropy Distribution of Tracts in General Scenario](image)

**Figure 22:** Boxplot of Entropy for Census Tracts in General Scenario
Mapping entropy for the entire region provides modest visual evidence of spatial patterns (Figure 23).

![Entropy Map of Denver MSA, General Scenario](image)

**Figure 23**: Entropy Map of Denver MSA, General Scenario

Entropy is especially high in the central downtown area (Figure 23, inset) and in a cluster of tracts southeast of central Denver. A few modest clusters of low entropy exist north and west of central Denver. Using Moran’s I to quantify spatial autocorrelation across the study area yields a Moran’s I statistic of 0.433 compared to an expected value of -0.002 (p < 0.05), indicating the distribution of entropy values is not random across the study area. A positive z-score of 19.00
implies spatial clustering of values in the study area.

Local Moran’s I clusters mapped for the study area illustrate a few large concentrations of high entropy in the central, south, and southeast portions of the metro area and concentrations of low entropy values in the north and in smaller clusters east of central Denver (Figure 24). Comparing the clusters to the high frequency tract pairs outlined in Figure 17 suggests no clear relationship between entropy and high frequency tract pairs. The high frequency tracts are located in areas of statistically significant high and low concentrations of entropy, and equally as often in areas with no clustering of values. The local clustering visible in Figure 24 could be a skewed by the high range of entropy values across all tracts, 0.44 to 0.91, meaning average entropy is also relatively high and moderate entropy values appear low.

**Figure 24:** Local Moran’s I Cluster for General Scenario Entropy Values
Unfortunately, as with degree and frequency, examples of region stability are sparse. All three measures of degree, frequency, and entropy paint a consistent picture of relatively low region stability across trials. Specific groups of tracts show promise as stable regions when all three measures are combined, but the patterns are faint, requiring detailed detective work to extract relatively few significant regions.

**Correlation**

To understand how all three measures relate across tracts, correlations between entropy, degree, and frequency are shown in Table 6 and illustrated graphically in Figure 25.

<table>
<thead>
<tr>
<th>Independent</th>
<th>Dependent</th>
<th>R²</th>
<th>Coefficient</th>
<th>Standard Error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree</td>
<td>Entropy</td>
<td>0.257</td>
<td>0.002</td>
<td>0.000</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Maximum Frequency</td>
<td>Entropy</td>
<td>0.144</td>
<td>-0.002</td>
<td>0.00</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Mean Frequency</td>
<td>Entropy</td>
<td>0.001</td>
<td>-0.000</td>
<td>0.001</td>
<td>&gt;0.05</td>
</tr>
<tr>
<td>Degree</td>
<td>Maximum Frequency</td>
<td>0.009</td>
<td>-0.094</td>
<td>0.039</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>Degree</td>
<td>Mean Frequency</td>
<td>0.204</td>
<td>-0.142</td>
<td>0.011</td>
<td>&lt;0.05</td>
</tr>
</tbody>
</table>

**Table 6: OLS Correlation Statistics for Frequency, Degree, and Entropy**

Evaluating the relationship between entropy and degree for all tracts yields an R-squared value of 0.257, demonstrating modest correspondence between the measures. As degree increases entropy values rapidly increase and then plateau just below 0.9, nearing the relative entropy maximum of 1.0 (Figure 25a). Many tracts defy the broad pattern. The highest entropy tracts have below average degree values between 20 and 30. The only clear outcome demonstrated in the regression graph is that tracts with a high degree (50 or greater) consistently also have high entropy (greater than 0.7).
Entropy and maximum frequency display a very weak trend in the reverse direction with an R-squared of 0.144. As the maximum edge weight for a tract increases, entropy tends to decline (Figure 25b). The clearest pattern evident is that at low frequency values entropy is consistently greater than 0.7. Any relationship disappears completely when entropy is compared to the mean frequency for a tract. The R-squared value is nearly zero and not significant at p < 0.05, which is reinforced by the lack of any visible trend in Figure 25c.

Based on the expectation that a low degree and high frequency reflect likely pair stability, an inverse relationship between the two in a correlation analysis might be expected. In fact, there is no relationship when the maximum frequency is applied. With an R-squared value of 0.009 and no evident pattern in the scatterplot (Figure 25d), the two measures appear to function independently. In contrast to the relationships between entropy and frequency, using the mean frequency instead of the maximum frequency produces better correspondence (R-squared = 0.204). The highest mean frequency values correspond with low degree but the relationship becomes more tenuous as degree increases (Figure 25e).

**Figure 25(a-e):** Regression Plots for Degree, Frequency, and Entropy Measures for General Scenario

**Figure 25a:** Degree vs. Entropy
Attempts to transform the variables in search of any better fit yielded no notable improvements to the weak correlations exhibited here. The lack of correspondence between measures does not nullify the possibility that stable regions would exhibit correspondence. The data yielded few stable regions to examine, and so the information captured by the metrics largely reflects the high degree of randomness in region formation and tract relationships. While
select tracts model stability well, most fluctuate on broad scales that lead to variable relationships between their measures of degree, frequency, and entropy.

**Neighborhood Assessment**

The cumulative results of degree, frequency, and entropy indicate weak stability across trials and few cohesive regions for comparison to neighborhoods exist. Capitalizing on the three stable regions that were identified (clusters A, B, & C), the following assessment explores their coherence as meaningful units of study.

**Figure 26a:** Aerial View of Cluster A  
**Figure 26b:** Entropy Map of Cluster A

Cluster A represents the mostly tightly bound group evident throughout the regionalization trials. The physical landscape surrounding cluster A helps explain why it is a strongly bound cluster (Figure 26a). Two tracts directly north and west of the cluster are excluded from regionalization because they do not contain a residential population. The large tract to the north houses the Rocky Mountain National Wildlife Refuge. The tract to the west is an industrial park with no residences. The four tracts in the cluster are predominantly residential...
within a landscape of single-family homes. They form the bulk of an isolated residential community sandwiched between I-70 to the south and the wildlife refuge to the north. Entropy is moderate for all four tracts, and distinctly lower than entropy values for neighboring tracts. All four nodes have degree values ranging from 10-13, much lower than the average of 38 across all tracts.

In its physical appearance and on all three quantitative measures, cluster A exemplifies evidence that would support the argument that Spielman & Folch’s regionalization algorithm can help identify neighborhoods. But does it correspond to any actual common neighborhood labels? Returning to the map of Denver County statistical neighborhoods as assigned by the City of Denver (Figure 5), cluster A is part of a larger neighborhood called Montbello. Although cluster A does not capture all tracts within the neighborhood of Montbello, it consistently captures four of the six tracts that currently comprise the neighborhood. A recent Denver Post article proclaimed Montbello the “nation’s hottest suburban housing market” due to recent home price appreciation (Svaldi, 2017). A 1991 Denver city planning document describes Montbello as possessing a “suburban character” (Planning and Community Development Office, 1991, p. 2) and at the time of the report was Denver’s largest neighborhood both in population and in acreage. The area has a relatively short history among Denver neighborhoods, having been annexed into Denver in the mid-1960s to be designed as a master planned community. Montbello is in most respects a strong example of regionalization output that produces meaningful boundaries in terms of demographic data, common perceptions, and the physical landscape.
Cluster B is highly interconnected, but as a moderate frequency region its strength as a meaningful unit of analysis is unclear. It is a large region, containing 10 tracts that are mostly residential but the housing varies in density and development patterns. The region also contains business parks and a golf course, and borders the southern edge of Cherry Creek State Park. A portion of the region’s tracts exhibit lower than average entropy, but others have high entropy, making any clear conclusion difficult. Most of the region coincides with the city of Centennial but the city’s boundaries are erratic, explaining the idiosyncratic tract boundary lines.
The city’s boundaries make it difficult or impossible to identify specific neighborhoods. The City of Centennial provides a map of “HOAs, Civic Association and Neighborhoods” that are much smaller regions than cluster B (City of Centennial, n.d.b). Most are no more than one or two census tracts in size, whereas cluster B covers nearly the entire eastern half of Centennial. Cluster B roughly follows a few consistent borders of the larger city landscape. For instance, its southern edge ends where suburban development transitions into open space, and the region carves out the small town of Foxfield that exists as a hole within Centennial city limits. Overall, the landscape lacks clear patterns and the short history of the region provides little to reference: Foxfield only became an official town in 1994 (Kenney, n.d.), and Centennial was officially established as a city in 2001 (City of Centennial, n.d.a). Cluster B raises questions about how to judge a statistical unit’s value when the underlying population and landscape exhibit no clear patterns.

Cluster C is located in inner Aurora, east of Denver and slightly north of cluster B.

Cluster C forms a perfectly interconnected cluster at the moderate frequency level, combines
two high frequency pairs. Comparing cluster C to the Aurora neighborhood map shows that the five tracts in the region encompass five designated neighborhoods: Willow Park to the west, and Center Pointe, Rocky Ridge, Horseshoe Park, and Kingsborough in the four tracts running North to South (City of Aurora, n.d.a). The neighborhoods closely align with census tracts. Like much of Aurora and the Denver region, cluster C contains mostly suburban, single-family homes but also apartment complexes and a long park threading through the region. There are no obvious distinctions between cluster C and neighboring tracts, although cluster C avoids most local commercial development. Neighborhoods have clear boundary lines according to the city, but information reinforcing neighborhood identities is sparse. Little exists to distinguish one from another. Cluster C may represent a case of imposed neighborhood units that exhibit similar demographic patterns relative to their surroundings, and thus tend to form high frequency pairs and a moderate frequency cluster.

**Scenario Comparison**

Comparing regionalization outcomes for 100 trials of the regionalization algorithm applied to the general, poverty, transportation, and housing scenarios illustrates the impact of variable selection upon regionalization outcomes. Overall, high variability in regionalization results are consistent between scenarios. Small differences arise in patterns of pair formation, implying spatial patterns specific to certain variables can affect regionalization outcomes.

**Degree**

The range of tract degree values is much greater among the poverty, transportation, and housing scenarios than for the general scenario, suggesting that the general scenario limits pair
formation (Table 7). Degree statistics for the poverty scenario skew particularly high. The poverty scenario mean degree of 50 far eclipses the average values for the general, housing and transportation scenarios, indicating regionalization outcomes for the poverty scenario are unstable or error is high and thus large regions are necessary to reduce error sufficiently. The transportation scenario features a high concentration of tracts with a degree between 0 and 10, though its mean is similar to the general scenario.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>General</th>
<th>Housing</th>
<th>Poverty</th>
<th>Transportation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>38</td>
<td>35</td>
<td>50</td>
<td>34</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>17</td>
<td>23</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td>Range</td>
<td>7 - 88</td>
<td>1 - 115</td>
<td>11 - 130</td>
<td>1 - 104</td>
</tr>
</tbody>
</table>

Table 7: Degree Summary Statistics for Variable Scenarios

Frequency

Frequency distributions are similarly concentrated at low frequencies among all scenarios, with small differences in magnitude (Table 8). The poverty scenario again demonstrates substantial variation in regionalization outcomes as shown by its large proportion of pairs with low edge weights and a high total number of edges in the network. The housing and transportation scenarios form fewer edges cumulatively, but the housing scenario generates the highest number of high frequency pairs while the transportation scenario generates the fewest. Despite these differences, the implications are the same. All scenarios have a high fraction of pairs that occur infrequently relative to a small percent of pairs that occur frequently across scenarios.
Table 8: Count and Percent of Tract Pairs Meeting Frequency Thresholds for All Scenarios

Figure 30 contains network graphs for the housing, poverty and transportation scenarios filtered for high frequency edges. The graphs resemble the general scenario high frequency graph with a few exceptions. The housing scenario (Figure 29a) contains more large, connected groups including a fully interconnected cluster of five tracts and another group of seven. Housing patterns may provide better fodder for stable region identification, perhaps because of the strength in spatial trends related to home ownership and home value. Despite the high variability in pair formation seen in the degree and frequency tables (Tables 7, 8), the poverty scenario high frequency network graph (Figure 29b) also resembles the others, containing numerous pairs, a few trios, and one group of five tracts.

Figure 29a: Network Graph for Tract Pairs that Occur in 90 or More Regionalization Trials, Housing Scenario  
Figure 29b: Network Graph for Tract Pairs that Occur in 90 or More Regionalization Trials, Poverty Scenario
The transportation high frequency network graph (Figure 29c) contains fewer nodes and edges. Still, at least one group of five fully interconnected nodes emerges. Examining the high frequency clusters within each scenario might yield a few more promising regions like Montbello but they will remain just a fraction the Denver MSA landscape.

**Figure 29c**: Network Graph for Tract Pairs that Occur in 90 or More Regionalization Trials, Transportation Scenario

*Entropy*

Average entropy values are high across all scenarios, visualized as boxplots in Figure 31. The range of entropy is notably small and high in the poverty scenario. In contrast, entropy in the transportation scenario ranges widely from 0.14 to 1.00 and contains the largest proportion of low entropy values. To understand how the distribution of entropy in the transportation scenario differs in geographic space, Figure 32 maps transportation entropy values for tracts. A clear pattern of low entropy in the outer MSA and high entropy in the inner MSA emerges, with variation in the intermediary zones representing fringe suburban areas. As with housing’s high
frequency pairs, transportation estimates and uncertainty could contain explicit patterns that are captured in regionalization and manifest in pair formation tendencies across the study region.

**Figure 30**: Entropy Distribution of Tracts in All Scenarios

**Figure 31**: Entropy Map of Denver MSA, Transportation Scenario
DISCUSSION

Spielman & Folch’s regionalization algorithm achieves what it sets out to do: produce regions with higher quality data than the input tracts. The algorithm does not achieve the secondary objective imposed upon it and tested in this analysis: create regions with substantive meaning beyond reducing uncertainty. The primary contributor to the lack of findings is the high variability in region output reflected in tract-level measures of degree, frequency, and entropy. Setting different starting seeds for the algorithm leads to unstable regionalization outcomes, possibly resulting from a lack of strong underlying spatial patterns in the input data.

Cluster A, the neighborhood of Montbello, exemplifies an ideal outcome of the regionalization process. Cluster A demonstrates strong stability as viewed through degree, frequency, and entropy. The four tracts within Cluster A correspond well to boundaries and patterns in the physical landscape, and match historical evidence of the evolution of the neighborhood. Though the region of Montbello outlined by the City of Denver is slightly larger than cluster A (six tracts versus four), the resemblance is strong enough to match the region and neighborhood. Other examples of stable regions formed under the general scenario moderate frequency threshold, like clusters B and C, do not produce equally compelling narratives as meaningful units of study. Their borders are erratic, and few indicators exist to distinguish the regions from neighboring places. Perspectives from people familiar with those areas would provide much better context for understanding how the regions do or do not reflect the experience of being there, but from an outsider’s perspective there is little to suggest region boundary significance.

The absence of consistent identifiable regions does not appear to result from poor metrics for evaluating stability. Despite the lack of statistical correspondence between degree, frequency
and entropy when all tracts are considered, all three measures provide consistent evidence for high variation among regionalization results. In rare cases of region stability like clusters A, B and C, the metrics better align. Other measures of clustering could complement the network graph statistics applied here, however the weakness of evidence accumulated suggests additional analysis would not unearth drastically different results.

**Towards a Flexible Spatial Statistical Unit**

Despite the lack of evidence for stable data-driven regions, the results bring attention to the importance of developing appropriate spatial statistical units for neighborhood-based research. One issue is the discrepancy between the size of regions necessary for sufficient data quality and the size of real neighborhoods. Like debates over optimal city size, arguments can be made regarding the ideal population or area of a neighborhood, but it is unlikely any clear benchmarks will emerge. Comparing data-driven regions to practical evidence of neighborhoods in the Denver MSA illustrates a mismatch between neighborhood size and regionalization results. Neighborhoods are typically small, no more than one or two census tracts and sometimes smaller than one tract. Regions are also modest, four tracts on average, but consistently larger than one or two tracts. Since these regionalization results are primarily driven by sampling error, sampling error in ACS data clearly prevents measuring communities based on the size of actual neighborhoods as defined internally by residents and externally by other groups.

When ACS data is aggregated to control for sampling error, neighborhoods must be combined. The result is a statistical unit on a scale between tracts and counties that has no clear equivalent in theoretical discussions of urban spatial organization. In Kevin Lynch’s description of a city composed of paths, edges, districts, nodes and landmarks, regions produced by the
regionalization algorithm most resemble districts, implying some level of cohesion and character without necessarily the connotations of a well-defined neighborhood. Lynch and others’ frameworks viewing the city through nested spatial hierarchies is particularly relevant here.

Instead of trying to assign theoretical weight to data-driven statistical units as attempted in this analysis, researchers should assign less theoretical significance to areal units of analysis. For studies based on ACS data, we should articulate spatial units as districts, or in Openshaw’s framework, zones, rather than trying to equate them to neighborhoods. Tools like Spielman & Folch’s algorithm afford options to test zone suitability against other study priorities and constraints. Given the arbitrariness of imposing census tracts as neighborhoods when their boundaries exhibit no concrete relationships to the underlying population, imposing other units with more useful attributes, including sufficient data quality, can only improve circumstances. Freeing analytical units from the theoretical implications of neighborhoods, while having the potential to significantly alter interpretations of neighborhood-based research, could encourage more flexible approaches to zone design, and more comparative testing of results using different zone configurations. An interesting subsequent analysis would be to examine how different regionalization solutions from the 100 general scenario trials affect measured trends in demographic groups across the Denver MSA. Such an experiment could highlight MAUP effects upon analytical outcomes.

What Makes a Neighborhood?

The lack of stable regions in the output from Spielman & Folch’s algorithm does not negate data-driven approaches to region identification. However, this study illustrates the challenges in reconciling such approaches with theoretical foundations of neighborhood
formation, and potentially other units of sociological interest. Part of the challenge in this study arose in reconciling the few moderately stable regions with sparse evidence to describe the underlying neighborhoods. Some neighborhoods better represent theoretical constructs describing neighborhoods. In the case of cluster B located in Centennial, little evidence exists to understand how internal or external groups interpret local neighborhoods. Centennial’s city boundaries are highly irregular, the urban landscape lacks strong patterns, and historical precedent offers little to evaluate neighborhood formation and change over time. In contrast, Montbello’s longer history of development, compact structure, and isolation from other neighborhoods makes it much easier to identify in practice and in data. The presence of tracts excluded from regionalization bordering two sides of Montbello imposed additional constraints that likely influenced region formation.

Seeing the potential utility of constraining region formation with additional information, one path to corral high variation in regionalization results might be to incorporate other non-ACS variables that relate to other measures of neighborhoods. As noted earlier, the variables incorporated in the four scenarios do not capture all aspects of neighborhood formation. For instance, the physical landscape plays a large role in census tract boundaries. Incorporating measures of accessibility or imposing constraints based on travel routes might limit the formation of regions in the same way that such features serve as barriers separating actual neighborhoods. Indicators of social cohesion, business activity, and historical patterns of neighborhood development could further improve a data-driven regionalization approach to identifying neighborhoods, acknowledging the difficulty of quantifying some of these concepts. Conversely, imposing additional constraints might further diffuse any recognizable patterns, leading to even less evidence of region stability. A simpler approach would be to examine the
behavior of one variable in isolation in the regionalization algorithm, to understand the effects of any single measure upon the outcome.

Comparing regionalization stability results across the four variable sets – general, housing, poverty, and transportation – illustrates that variable selection matters at the margins. Changing scenarios did not change the paucity of evidence in favor of region stability. Changing scenarios did produce metrics with some notable distinctions that spark questions regarding how patterns in the underlying estimates and MOE lead to differences in degree, frequency, and entropy. The housing and transportation scenarios hinted at more conclusive variable patterns that could relate to the strength of housing and transportation trends across metropolitan regions. Though very modest, the demonstrated impacts of variable selection support the idea that analytical units should have some basis in information relevant to the research subject. Instead of consuming statistical information in generic areal unit packages like tracts, applying some form of data-driven geographies specific to the subject of interest can affect study outcomes in the same way that variable selection plays an important role. Attribute selection is inherently linked to zone design.

Conclusion

Neighborhoods have real conceptual and practical meaning for people within and outside of them, but they will always be difficult to define. It is easy to articulate neighborhoods as either the products of data or theory, not both. Spielman & Folch’s regionalization algorithm is just one example of a data-driven regionalization process among other possible methods, and it targets the problem of ACS uncertainty. Designing a regionalization algorithm with the explicit purpose of approximating neighborhoods would be a separate and daunting task that could result in very
different outcomes from those observed in this analysis. Any alternative approach, however, would still need to consider data quality implications. Despite the challenges of trying to reconcile data-driven and theory-driven approaches to defining neighborhoods, the intersection of both highlights tradeoffs, mismatches, and gaps in knowledge that either alone might overlook.
BIBLIOGRAPHY


Green, H. W., & Truesdell, L. E. (1934). Census Tracts in American Cities: A Brief History of
the Census Tract Movement, with an Outline of Procedure and Suggested Modifications
Analysis. In S. Wenzhong, B. Wu, & A. Stein (Eds.), Uncertainty Modelling and Quality
Control for Spatial Data (pp. 3–9). CRC Press. Retrieved from
Grira, J., Y. Bédard, and S. Roche. 2010. Spatial data uncertainty in the VGI world: Going from
Acmqueue, 8(5), 1–22.
International. Washington, D.C. Retrieved from
http://www.upi.com/Top_News/2001/11/26/Court-No-missionaries-in-
census/19251006790111/
[London]: Springer.
https://www.littletongov.org/building-development/historic-preservation/neighborhood-
lookup-map
MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., & Hetzler,
We Need to Know. Cartography and Geographic Information Science, 32(3), 139–160.
https://doi.org/10.1559/1523040054738936


U.S. Const. art. I, § 2.


Utah et al. v. Evans, Secretary of Commerce, et al., No. No. 01—714 (Supreme Court of the United States June 20, 2002). Retrieved from https://www.law.cornell.edu/supct/html/01-714.ZS.html


## Scenario Variable Tables

### Table A1. General Scenario Variables

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Numerator</th>
<th>Denominator</th>
<th>Derivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average rooms per housing unit</td>
<td>ACS12_5yr_B25019001</td>
<td>ACS12_5yr_B25001001</td>
<td>ratio</td>
</tr>
<tr>
<td>Average household income</td>
<td>ACS12_5yr_B19025001</td>
<td>ACS12_5yr_B25002002</td>
<td>ratio</td>
</tr>
<tr>
<td>Average persons per household</td>
<td>ACS12_5yr_B25008001</td>
<td>ACS12_5yr_B25002002</td>
<td>ratio</td>
</tr>
<tr>
<td>Percent of housing units occupied</td>
<td>ACS12_5yr_B25002002</td>
<td>ACS12_5yr_B25002001</td>
<td>proportion</td>
</tr>
<tr>
<td>Percent of households with married-couple family</td>
<td>ACS12_5yr_B11001003</td>
<td>ACS12_5yr_B11001001</td>
<td>proportion</td>
</tr>
<tr>
<td>Percent of population age 25+ with bachelor's degree or higher</td>
<td>ACS12_5yr_B16010041</td>
<td>ACS12_5yr_B16010001</td>
<td>proportion</td>
</tr>
<tr>
<td>Percent of population that lived in the same house 1 year ago</td>
<td>ACS12_5yr_B07003004:</td>
<td>ACS12_5yr_B07003001</td>
<td>proportion</td>
</tr>
<tr>
<td>Percent of population non-Hispanic white</td>
<td>ACS12_5yr_B03002003</td>
<td>ACS12_5yr_B03002001</td>
<td>proportion</td>
</tr>
<tr>
<td>Percent of population non-Hispanic black</td>
<td>ACS12_5yr_B03002004</td>
<td>ACS12_5yr_B03002001</td>
<td>proportion</td>
</tr>
<tr>
<td>Percent of population Hispanic</td>
<td>ACS12_5yr_B03002012</td>
<td>ACS12_5yr_B03002001</td>
<td>proportion</td>
</tr>
<tr>
<td>Percent of population under 18</td>
<td>ACS12_5yr_B09001001</td>
<td>ACS12_5yr_B01003001</td>
<td>proportion</td>
</tr>
<tr>
<td>Percent of population over 65</td>
<td>ACS12_5yr_B09020001</td>
<td>ACS12_5yr_B01003001</td>
<td>proportion</td>
</tr>
<tr>
<td>Total population</td>
<td>ACS12_5yr_B01003001</td>
<td></td>
<td>count</td>
</tr>
</tbody>
</table>

### Table A2. Poverty Scenario Variables

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Numerator</th>
<th>Denominator</th>
<th>Derivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Housing cost to income ratio - owners (annual)</td>
<td>ACS12_5yr_B25089001_ann</td>
<td>ACS12_5yr_B25120002</td>
<td>ratio</td>
</tr>
<tr>
<td>Housing cost to income ratio - renters (annual)</td>
<td>ACS12_5yr_B25065001_ann</td>
<td>ACS12_5yr_B25120005</td>
<td>ratio</td>
</tr>
<tr>
<td>Percent of children above poverty</td>
<td>ACS12_5yr_B17006016</td>
<td>ACS12_5yr_B17006001</td>
<td>proportion</td>
</tr>
<tr>
<td>Percent of population above poverty</td>
<td>ACS12_5yr_B17001031</td>
<td>ACS12_5yr_B17001001</td>
<td>proportion</td>
</tr>
<tr>
<td>Percent of population employed</td>
<td>ACS12_5yr_B23025004</td>
<td>ACS12_5yr_B23025003</td>
<td>proportion</td>
</tr>
<tr>
<td>Total population</td>
<td>ACS12_5yr_B01003001</td>
<td></td>
<td>count</td>
</tr>
</tbody>
</table>

### Table A3. Transportation Scenario Variables

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Numerator</th>
<th>Denominator</th>
<th>Derivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicles per person</td>
<td>ACS12_5yr_B25046001</td>
<td>ACS12_5yr_B01003001</td>
<td>ratio</td>
</tr>
<tr>
<td>Average travel time to work</td>
<td>ACS12_5yr_B08013001</td>
<td>ACS12_5yr_B08134001</td>
<td>ratio</td>
</tr>
<tr>
<td>Percent of population who drove to work alone</td>
<td>ACS12_5yr_B08101000</td>
<td>ACS12_5yr_B08101001</td>
<td>proportion</td>
</tr>
<tr>
<td>Percent of population who take public transit to work</td>
<td>ACS12_5yr_B08101025</td>
<td>ACS12_5yr_B08101001</td>
<td>proportion</td>
</tr>
<tr>
<td>Total population</td>
<td>ACS12_5yr_B01003001</td>
<td></td>
<td>count</td>
</tr>
</tbody>
</table>
### Table A4. Housing Scenario Variables

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Numerator</th>
<th>Denominator</th>
<th>Derivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average rooms per housing unit</td>
<td>ACS12_5yr_B25019001</td>
<td>ACS12_5yr_B25001001</td>
<td>ratio</td>
</tr>
<tr>
<td>Average home value of owner-occupied units</td>
<td>ACS12_5yr_B25082001</td>
<td>ACS12_5yr_B25003002</td>
<td>ratio</td>
</tr>
<tr>
<td>Average rent</td>
<td>ACS12_5yr_B25065001</td>
<td>ACS12_5yr_B25003003</td>
<td>ratio</td>
</tr>
<tr>
<td>Percent of housing units occupied</td>
<td>ACS12_5yr_B25002002</td>
<td>ACS12_5yr_B25002001</td>
<td>proportion</td>
</tr>
<tr>
<td>Percent of housing units owner-occupied</td>
<td>ACS12_5yr_B25003002</td>
<td>ACS12_5yr_B25003001</td>
<td>proportion</td>
</tr>
<tr>
<td>Percent of housing units renter-occupied</td>
<td>ACS12_5yr_B25003003</td>
<td>ACS12_5yr_B25003001</td>
<td>proportion</td>
</tr>
<tr>
<td>Percent of housing units that are single detached units</td>
<td>ACS12_5yr_B25024002</td>
<td>ACS12_5yr_B25024001</td>
<td>proportion</td>
</tr>
<tr>
<td>Total population</td>
<td>ACS12_5yr_B01003001</td>
<td>ACS12_5yr_B01003001</td>
<td>count</td>
</tr>
</tbody>
</table>