DENSITY AND THE SPATIAL ANALYSIS of Principal Components Derived from Mobility-Related Socio-Economic Variables in New England

Devon Lechtenberg
Capitol Region Council of Governments
Hartford, Connecticut

ABSTRACT

Principal component analysis (PCA) is a widely used multivariate data analysis technique for variable reduction and analysis of variable relationships, among other purposes. In this current study set in New England, PCA is used together with local indicators of spatial autocorrelation (LISAs) to examine regional patterns of spatial dependence between population density and synthetic variables (i.e. principal components) derived from mobility-related socio-economic variables. A secondary analysis using geographically weighted principal component analysis (GWPCA) is used to identify which of the original contributing variables are most important for different locations within the region. Using a local bivariate indicator of spatial autocorrelation, overall results show that most resulting principal components are positively spatially correlated with population density, although this correlation is not a direct point-to-point one and instead relates the principal component to the spatial lag of density for a given location. These findings are relevant for understanding the relationship between agglomeration (as represented by density) and the triad of mobility, accessibility, and connectivity. The results of this study could aid and inform future research and efforts at modeling travel behavior.

Key words: Mobility, Population Density, Principal Component Analysis, Geographically Weighted Principal Component Analysis, Bivariate Local Moran's I, New England, NCHS 2013 County Typology

Introduction

Regional travel-to-work patterns are shaped through the conceptual triad of accessibility, mobility, and connectivity, which is anchored by agglomeration. Agglomeration, which can be represented by population density, is understood here to include the geographic concentration of people, jobs, infrastructure, shopping, institutions, and recreation that form urbanized areas. Given the many factors encompassed by agglomeration, the positive correlation between density on the one hand and accessibility and connectivity on the other hand comes as no surprise.

The relationship between agglomeration and mobility is somewhat more difficult to articulate as some mobility measures are evidence for *previous* travel such as vehicles miles traveled (VMT) or indications of potential travel, such as the numerous socio-economic variables used in transportation analyses. Additionally, the number and diversity of socio-economic variables complicate efforts to understand the relationship between these variables and density (agglomeration). As these socio-economic variables shape mobility, they thus also affect, and can be affected by, accessibility and connectivity. Furthermore, socio-economic variables have both attribute values (quantitative or qualitative) and spatial properties (georeferenced locations). Within multivariate data analysis, a standard method of variable reduction in the face of many potentially interrelated variables is principal component analysis (PCA). Synthetic variables (known as principal components) are created out of the original group of variables and have the important property of explaining high levels of variance within the original data set (while using fewer variables) and being mutually orthogonal and unrelated. Traditional PCA is non-spatial but can be used in tandem with measures of global and local spatial autocorrelation to understand their patterns of geographic distribution (see Anselin 2020a). Geographically weighted principal component analysis (GWPCA), a relatively new method, combines analysis of both attribute and spatial heterogeneity (as opposed to autocorrelation), thus highlighting the variability of attributes across space. This study uses a combination of traditional PCA and indicators of spatial (auto-) correlation as well as GWPCA in an exploratory analysis of mobility-related socio-economic variables and their relationship to population density in New England.

New England offers a unique case study for several reasons: Firstly, the region is well-defined and largely self-contained, bordered by New York state and Canada while consisting of the six states of Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont. Secondly, New England states have almost no unincorporated land, which enables the straightforward designation of contiguity regimes between (town) boundaries, a key part of spatial analysis. Thirdly, New England offers an almost textbook example of a coastal dense pattern of urban development contrasting with the low-density hinterland, which obviously lends itself to a density-based analysis. In fact, this contrast may be the most prominent feature of the population distribution in New England (see Figure 1). The density gradient runs diagonally from just north of Boston down to the Connecticut-New York State line, with the ten largest regional urban populations all located to the south of the gradient. Nevertheless, there are isolated population centers in northern New England, notably the Burlington, VT area (home to the three largest Vermont towns) and the scattered patterns found in Maine. Settlement and mobility patterns are also related to topography: New England is characterized by mountains, hills, rivers, and extensive coastline and not only is the population concentrated in comparatively flat, low-lying areas (e.g. coastal plain, Connecticut River valley, Champlain Valley in Vermont), but travel between locations, especially more remote ones, are impeded by difficult terrain even though modern paved roads are found everywhere in the region.

This study is concerned with understanding spatial and attribute patterns of mobility-related socio-economic variables and their relationship to density within New England but must employ a means to reduce the number of variables and simplify the analysis. The principal components derived from these variables can be used in a spatial analysis that more efficiently gets at the relationship between the original socio-economic factors — now combined in the form of synthetic variables and essentially representing multifaceted socio-economic profiles — and density. Thus, the concrete research question for this study is: ***Are the principal components derived from the original socio-economic data (positively) spatially dependent with population density? And if so, where?*** In other words, do spatial patterns of density impact the spatial patterns of these socio-economic variables? It would be expected with accessibility and connectivity, but it is less clear with socio-economic variables and takes on significance if these synthetic variables were to be used in comparative and/or predictive analyses involving measures of accessibility and connectivity. The primary analysis described above has the limitation that it is still global in nature, meaning in this case that the relationship between contributing socio-economic variables within a PCA is determined for the entire study area and is not sensitive to local variations. As a complement to the primary analysis, a secondary analysis consisting of a GWPCA provides a needed local view within this study since there is every reason to believe that these relationships are not static over space.

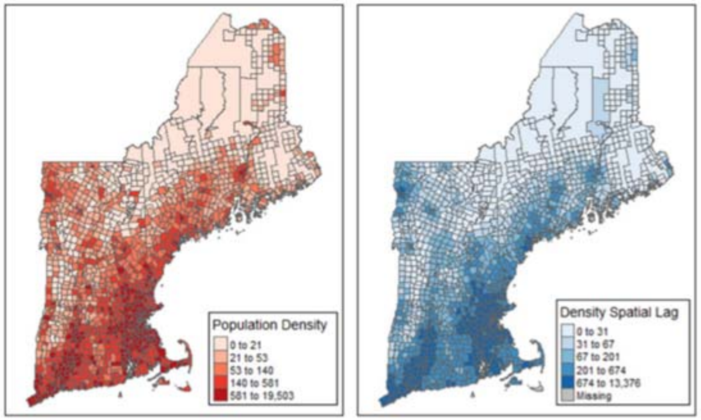


Figure 1: Population Density and Its Spatial Lag in New England.
(Source: Author using 'tmap' package in R).

Results from the traditional PCA, used together with univariate and bivariate local indicators of spatial autocorrelation (LISAs), show that there is positive spatial correlation between the synthetic variables and population density in all cases but one. The intensity of the spatial correlation is low but is nevertheless statistically significant. The subsequent GWPCA demonstrates that the importance of original contributing variables within individual principal components was not spatially constant, with different variables being the most important contributors in different locations, something that is unknowable from traditional "global" PCA methods. The practical significance of these findings are as follows: 1) it strongly suggests the role of agglomeration in shaping mobility via the spatial dependence between density and mobility-related socio-economic characteristics, 2) the derived synthetic variables (principal components) can be used in subsequent analyses involving variables that may be highly correlated with density without the risk of major complications stemming from multicollinearity, 3) there is clear variation between low density (rural) locations and higher density (urban) locations with respect to these synthetic variables, and 4) results of GWPCA offer clues as to which contributing variables are the most important in shaping their principal components in given locations. There are several caveats to this study. Firstly, no list of socio-economic variables is complete, and variable reductions comes at the price of potentially omitting relevant variables. PCA results can vary significantly depending on which variables are included, although some relatively stable patterns emerge. Secondly, these results from an aggregate level analysis are not directly transferable to individual-level analyses. Thirdly, any formal articulation of relationships between these principal components and measures of accessibility and connectivity will have to wait for more rigorous future modeling analyses. Thus, the socio-economic profiles created by each principal component and their relationship to transportation patterns are a potential future avenue of research. This research is of potential interest to transportation geographers, planners, spatial scientists, and those with a regional interest in New England.

Literature Review

For the last thirty years, Paul Krugman's (1991) new economic geography has had considerable influence over debates surrounding the unequal allocation of industry, populations, and attendant phenomena such as infrastructure within and between regions. This massing of interrelated things rooted in a spatial concentration of industry and population can be termed ***agglomeration***. New economic geography holds that regions can develop into core and peripheries as a result of agglomeration creating beneficial economies of scale, transport costs being reduced, and a certain distribution of manufacturing. Krugman (1991) noted that there was general agreement that economic production concentrates or 'localizes' in part because of access to common labor pool and information spillovers, which were first discussed as causes by the English economist Alfred Marshall in the late 1800s. These benefits of agglomeration, generally held to be positive, combined with economies of scale are captured in the term economies of agglomeration (or agglomeration economies). From a transportation perspective, one very important observation put forth by Krugman (1991) was that decreasing transport costs can in fact reinforce the consolidation of core and peripheral areas. Lafourcade and Thisse (2011) echoed this point. Agglomeration manifests itself as urban form and one of the most basic and indeed common means of measuring this is ***population density*** (Clifton, Ewing, Knaap, and Song 2008). Density is usually positively correlated with jobs, infrastructure, trip generation, and higher levels of various socio-economic indicators. Thus, density can justifiably be treated as a stand-in for agglomeration.

Density strongly impacts the relative accessibility of a location, as can be seen in the fact that accessibility tends to reflect the spatial distribution of population density (Sohn 2005). This is illustrative of ***place-based accessibility***, which as a concept can be defined as "the ease of reaching desired locations" (Clifton et al. 2008, 28). Both elements of mass (a desired location, usually made attractive by its size) and frictions (in the form of transportation costs and their inversely proportional relationship) are present in this definition. This sort of conception of accessibility is derived from the gravity model, which arose as an analogy to Isaac Newton's mathematical description of gravity. Alan G. Wilson (1967, 2001) reformulated the classic gravity model into a spatial interaction model (SIM) which is a far more flexible implementation of the gravity concept. The SIM has three potential forms, production-constrained, destination constrained, and doubly (production and destination) constrained. Alternatively, another place-focused conception of accessibility is given by Bruinsma and Rietveld (1996), where it represents "the potential opportunities for interaction" for regional economic purposes. This definition is attractive as it broadens the ways in which place-based accessibility can be defined. In contrast to both place-focused definitions discussed above, there is also ***people-based accessibility***, which refers to "how easily a person...can reach activity sites" (Hanson 2004, 5). People-based accessibility is fundamentally shaped by factors of location and socio-economic attributes, which are in the first instance measures of mobility. Thus, socio-economic attributes, more properly categorized as measures of mobility, can influence people-based accessibility. They could also influence place-based accessibility when included in a spatial interaction model as additional variables.

Mobility can be thought of as the ability of people to physically move around in space. In the context of commuting, mobility attributes represent the either the aggregate or individual capacity to travel to their place of work. Socio-economic attributes tend to be best thought of as factors that affect mobility as opposed to observed (or sometimes estimated) measures of mobility, such as vehicle miles traveled (VMT). A number of factors influence trip production among individuals, including income, vehicles owned, household characteristics, family size, land value, residential density, and accessibility (Ortuzar and Willumsen 2011, 126), although more could certainly be added. These factors can be stratified into groups such as income brackets in order to further differentiate the population. Scholarly research has as well identified other important factors such as gender, race/ ethnicity, level of education, and transit availability. Susan Hanson, for example, has published extensively on the role of gender in mobility (Hanson 2010). Race and ethnicity as factors are also causes of differentiated mobility, including when they are compounded by gender differences (Hu 2020). For example, ethnic and racial minority groups often face long (duration) and complex commutes involving multiple transit connections owing to spatial mismatch, i.e. the residence of these groups far away from their places of work. Thus, they face serious challenges not only related to mobility, but also to accessing the benefits from connectivity and accessibility that would be available were it not for a lack of mobility. The intersection of transit-usage and low-income status takes on special significance where the absence of a public transportation option decisively curtails the person mobility of low-income persons (see Giuliano 2005). The type of employment and level of education can also significantly influence mobility patterns, with high-income, highly educated individuals often commuting much longer distances to work, or alternatively, being in the economic position to afford expensive urban housing closer to their work. The connectivity (or lack thereof) of transportation infrastructure can severely impede personal mobility, even where socio-economic factors are otherwise favorable (Bjarnason 2014). Mobile phone data has been increasingly used in the last decade to produce detailed observational data of personal mobility. Some research has managed to couple this observational data together with socio-economic data to create especially solid analyses of travel behavior (Xu et al. 2017). What is exceedingly clear is that there are a very large number of potential socio-economic variables to choose from when performing an analysis. However, it can be unclear as to which ones are the most important for various analyses.

The concept of ***connectivity*** is the last part of the triad whose first two elements are accessibility and mobility. Connectivity is the measure of 'linkages' between origins and destinations and is best understood in terms of networks, which fundamentally consist of nodes and links, each with their own attributes. Network connectivity is generally described by measures such as degree of node, centrality index, etc... Transportation infrastructure is a particular type of physical network where both nodes (i.e. intersections, interchanges, major transshipment points) and links (i.e. roadways, railways) are tangible and exist to facilitate movement people and goods. ***Transportation infrastructure*** contrasts with commuter flow networks where nodes and links are not both tangible[1] . Transportation networks arise and densify in the presence of agglomeration. The densification of the network leads to greater accessibility as the transportation costs are lowered by better infrastructure. However, although accessibility and infrastructure are generally correlated positively, Marcińczak and Bartosiewicz noted that at some point, increasing the density of the built environment actually decreases accessibility as it forms an impedance (2018). Greater transportation network connectivity does not stymie the benefits of personal mobility as would be the case in conditions of low connectivity, which limit options of people desiring to travel even if they are otherwise able (Vickerman 1996). Strategically chosen new transportation infrastructure can even improve mobility of the population (see Bjarnason 2014). Thus, transportation network connectivity, place-based accessibility, and personal mobility enter a self-reinforcing cycle. Despite the seemingly positive returns of increasing connectivity, regional inequalities resulting from agglomeration can be exacerbated by new and better infrastructure, as workers will move out of the hinterlands to towards urbanized areas but retain easy access to family and friends back home facilitated by said infrastructure. Transportation infrastructure networks can be measured in a number of ways including: total length of facilities in a given area, density of facilities, total length of facility types (i.e. Interstates), interchanges, intersections, among others. Black (2003) details many applicable measures from graph theory, while Labi et al (2019) go further and include additional measures that combine characteristics of network connectivity with accessibility and mobility measures.

Tools are needed that lend greater structure and clarity to the analysis of multiple variables that characterize accessibility, mobility, and connectivity. ***Principal component analysis (PCA)*** is a standard method of multivariate data analysis wherein variable reduction creates uncorrelated (orthogonal) synthetic variables from linear combinations of potentially related *real* variables. These synthetic variables are called *principal components* (PCs) and are equal to the number of original real variables. PCs are in fact eigenvectors given in reverse rank order of most variance explained to least (signified by eigenvalues), with the first few PCs usually explaining the vast majority of variance within the dataset. The steps and their corresponding mathematical notations are lengthy and are only briefly summarized in Appendix A. However, more complete explanations can be found in Joliffe (2002) or Abdi and Williams (2010). Demšar et al. (2013) offer a thorough review of the use of PCA within geography over the last century. The unique qualities of spatial data were discussed as well. Although there have been notable periods of interest in PCA as an analytical tool in its own right, such as during the 'quantitative revolution' of the 1960s, substantial scholarly interest in PCA among geographers has never remained consistent. This is regrettable given the specific analyses which can be performed with PCA as described by Jeffers (1967), cited in Harris et al. (2011, 1717):

1) examination of the correlations between variables of a selected set;
2) elimination of variables that contribute relatively little information;
3) examination of the group of individuals in n-dimensional space;
4) determination of the weighting of variables in the construction of indices;
5) allocation of individuals to previously demarcated groups;
6) recognition of misidentified individuals;
7) orthogonalization of regression calculations; and
8) reduction of the basic dimensions of variability in the measured set.

Furthermore, PCA's multiple potential applications for geographic subject matter are identified by Gould (1967)[2], and also cited in Harris et al. (2011). Standard PCA is a non-spatial global analysis, as it takes neither spatial autocorrelation nor spatial heterogeneity directly into account. Although this can be partly remedied by a spatial analysis performed on the PCs, the analysis would remain fundamentally global as local variations in the data are not addressed.

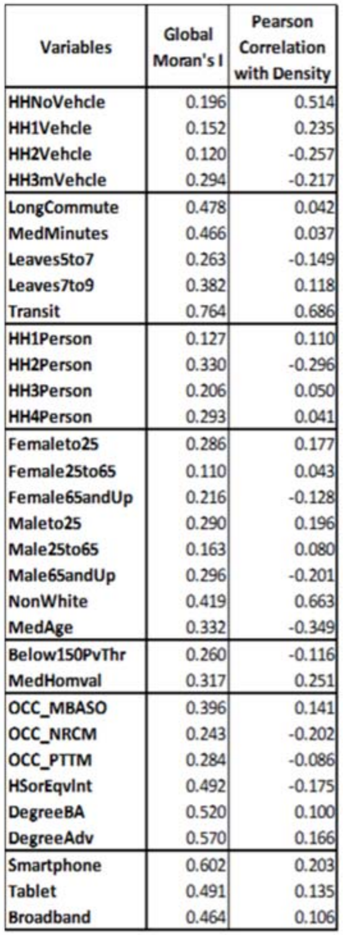
Geographically weighted principal component analysis (GWPCA), a method for incorporating spatial heterogeneity, was first briefly introduced by Fotheringham et al. (2002) and presented in more detail by Harris et al. (2011), who as seen above, strongly endorsed the use of PCA as an analytical technique based on the justifications provided by earlier scholars. Gollinni et al. (2015) detailed an R package, *"GWmodel"*, which could perform GWPCA, among many other novel geographically weighted analyses. In contrast to the geographically weighted approach of dealing with spatial heterogeneity, a PCA accounting for spatial autocorrelation was developed by Jombart et al. (2008). The use of traditional PCA with spatial data has also been featured in online documentation for GeoDa software, where local indicators of spatial autocorrelation (LISAs) are presented as tools for analyzing principal components (Anselin 2019). There have been several recent examples of PCA and/ or GWPCA being used in the analysis of regional economic development, including a combined standard PCA-GIS approach by Petrişor et al. (2012) for different sets of variables in various Romanian regions, a GWPCA used by Li, Cheng, and Wu (2016) for Jiangsu Province in China, and a very recent effort by Cartone and Postiglione (2020) to augment PCA using spatial filtering to account for both spatial autocorrelation and spatial heterogeneity in regional development in the Italian province of Rome. This methodological flexibility when using PCA is well-suited for describing the complex relationships between mobility-related socio-economic variables and population density. The use of PCA and spatial analysis to better understand the relationship between socio-economic characteristics of mobility and the spatial organizing influence of agglomeration would be a methodological contribution to transportation studies, geography, and other spatial sciences.

Methods

The data analyzed in this study consist of 2018 American Community Survey (ACS) 5-year estimates. Shapefiles for minor county divisions (MCDs, equivalent to towns for most of New England), counties, and state boundaries were obtained from the U.S. Census Bureau's website. The 2013 six-way classification typology for U.S. counties by National Center for Health Statistics (NCHS) researchers (see Ingram and Franco 2014) was utilized. Using ESRI ArcGIS, separate shapefiles for MCD in the six New England states were merged into one large shapefile for the whole region, containing 1606 locations. Census data from the ACS were reformatted in Microsoft Excel, where appropriate, to be expressed as population proportions so as to prevent size effects from dominating the data. In the rare cases where there were missing data, the proportion for the variable for the county to which the MCD belongs was substituted, something that mostly occurred in far-northern Maine near the geographically large but sparsely populated towns. A GAL (GenePix Array List) for queen-contiguity adjacency of all towns in New England was obtained using GeoDA, an open-source exploratory spatial econometrics software (see Anselin et al. 2006, Anselin and Rey 2014). The GAL file serves as a portable spatial weights file for determining neighboring spatial unit and being relevant to spatial analysis in multiple methodological approaches. The analysis was carried out using R for the Pearson correlations, traditional PCA, and GWPCA, GeoDA for both the global and local versions of univariate and bivariate Moran's indices for spatial correlation, and R for final mapping of the results.

Thirty-two variables (see Table 1) were selected from an initial list of nearly fifty candidates. Variables were eliminated based on a number of criteria, including: negligible attribute correlation with population density, negligible and statistically insignificant spatial autocorrelation as measured by the Global Moran's I, and anticipated redundancy with other variables. Correlation with density and individual spatial autocorrelation were considered important criteria owing to the central role played by both in this study. If potential variables were not expected to react (either positively or negatively) with density, then there was little point to retaining them in this analysis. Also, a lack of statistically significant positive or negative spatial autocorrelation (clustering) would make these variables difficult to analyze utilizing the methods proposed below. An initial PCA was performed, and variables found to be making little contribution to any of the selected PCs were also used to inform individual decisions. There were some exceptions made even when the aforementioned criteria seemed to warrant removal. Variables removed included: service and sales occupations, educational attainment levels of some college and associate degree, median age, age dependency, old age dependency, and child dependency ratios, female and male age groupings of 0-25, 25-65, and 65 and up, desktop and laptop ownership (not exclusive of other devices such as smartphones and tablets), lack of computer ownership, and finally dial-up Internet. The resulting data set consisted of 1,606 observations (towns in New England) of 32 variables. Of note is the definition of the variable LongCommute as the percentage of workers with a forty-five minute or longer commute. The occupation variables OCC_MBASO, OCC_NRCM, OCC_PTTM represent the respective summary census categories of management, business, arts, and science; natural resources, construction, and maintenance; and finally production, transportation, and material moving.

A traditional principal component analysis (PCA) was performed on the thirty-two variables (seen in Table 1) with the *"FactoMineR"*, *"factoextra"*, and *"corrplot"* R packages[3] . The most important principal components were chosen by consulting the Kaiser criterion (1961), which selects components based on those having an eigenvalue of greater than 1. The selected PCs were analyzed for spatial correlation using both global and local versions of univariate and bivariate Moran's I for identifying spatial clustering (see Appendix B for Equations) in GeoDA. The resulting cluster patterns could then be mapped in R using the *"tmap"* package by Tennekes (2018). The local univariate and bivariate spatial clustering patterns of the selected principal components were then analyzed and interpreted. Use of the NCHS county typology also contributed to the analysis. Owing to the limitations of traditional PCA, even when accompanied by spatial indices of autocorrelation, a geographically weighted principal component analysis in the *"GWmodel"* package in R was performed on the data set as well. The GWPCA output of the localized "winning" variable for each selected principal component was especially helpful here.

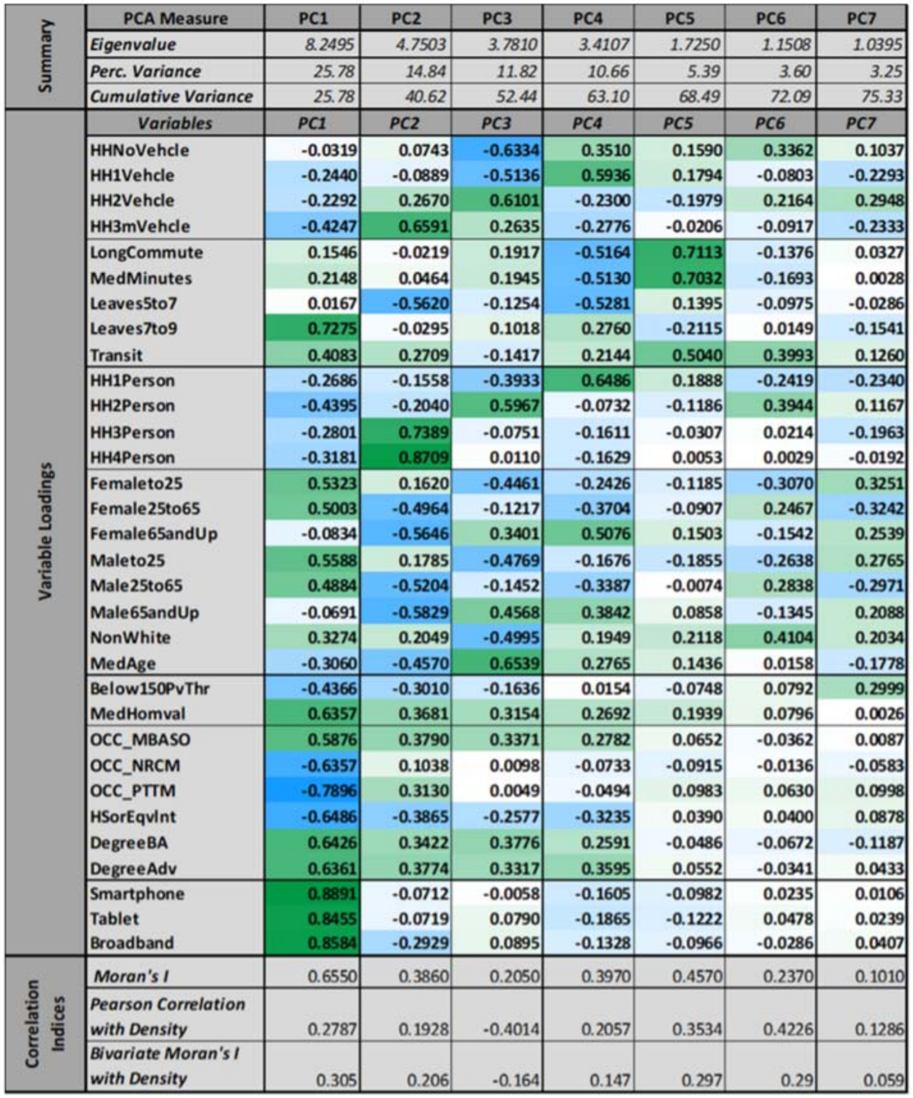


Variables	Global Moran's I	Pearson Correlation with Density
HHNoVehcle	0.196	0.514
HH1Vehcle	0.152	0.235
HH2Vehcle	0.120	-0.257
HH3mVehcle	0.294	-0.217
LongCommute	0.478	0.042
MedMinutes	0.466	0.037
Leaves5to7	0.263	-0.149
Leaves7to9	0.382	0.118
Transit	0.764	0.686
HH1Person	0.127	0.110
HH2Person	0.330	-0.296
HH3Person	0.206	0.050
HH4Person	0.293	0.041
Femaleto25	0.286	0.177
Female25to65	0.110	0.043
Female65andUp	0.216	-0.128
Maleto25	0.290	0.196
Male25to65	0.163	0.080
Male65andUp	0.296	-0.201
NonWhite	0.419	0.663
MedAge	0.332	-0.349
Below150PvThr	0.260	-0.116
MedHomval	0.317	0.251
OCC_MBASO	0.396	0.141
OCC_NRCM	0.243	-0.202
OCC_PTTM	0.284	-0.086
HSorEqvInt	0.492	-0.175
DegreeBA	0.520	0.100
DegreeAdv	0.570	0.166
Smartphone	0.602	0.203
Tablet	0.491	0.135
Broadband	0.464	0.106

Table 1: Description of Basic Socio-Economic Variables Used in Analysis. (Source: Author)

Results

The results section consists of two parts: a traditional principal component analysis (PCA) accompanied by a spatial analysis of the selected principal components and a brief geographically weighted principal component analysis (GWPCA) to add to the findings.



	PCA Measure	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Summary	*Eigenvalue*	8.2495	4.7503	3.7810	3.4107	1.7250	1.1508	1.0395
	Perc. Variance	25.78	14.84	11.82	10.66	5.39	3.60	3.25
	Cumulative Variance	25.78	40.62	52.44	63.10	68.49	72.09	75.33
Variable Loadings	*Variables*	*PC1*	*PC2*	*PC3*	*PC4*	*PC5*	*PC6*	*PC7*
	HHNoVehcle	-0.0319	0.0743	-0.6334	0.3510	0.1590	0.3362	0.1037
	HH1Vehcle	-0.2440	-0.0889	-0.5136	0.5936	0.1794	-0.0803	-0.2293
	HH2Vehcle	-0.2292	0.2670	0.6101	-0.2300	-0.1979	0.2164	0.2948
	HH3mVehcle	-0.4247	0.6591	0.2635	-0.2776	-0.0206	-0.0917	-0.2333
	LongCommute	0.1546	-0.0219	0.1917	-0.5164	0.7113	-0.1376	0.0327
	MedMinutes	0.2148	0.0464	0.1945	-0.5130	0.7032	-0.1693	0.0028
	Leaves5to7	0.0167	-0.5620	-0.1254	-0.5281	0.1395	-0.0975	-0.0286
	Leaves7to9	0.7275	-0.0295	0.1018	0.2760	-0.2115	0.0149	-0.1541
	Transit	0.4083	0.2709	-0.1417	0.2144	0.5040	0.3993	0.1260
	HH1Person	-0.2686	-0.1558	-0.3933	0.6486	0.1888	-0.2419	-0.2340
	HH2Person	-0.4395	-0.2040	0.5967	-0.0732	-0.1186	0.3944	0.1167
	HH3Person	-0.2801	0.7389	-0.0751	-0.1611	-0.0307	0.0214	-0.1963
	HH4Person	-0.3181	0.8709	0.0110	-0.1629	0.0053	0.0029	-0.0192
	Femaleto25	0.5323	0.1620	-0.4461	-0.2426	-0.1185	-0.3070	0.3251
	Female25to65	0.5003	-0.4964	-0.1217	-0.3704	-0.0907	0.2467	-0.3242
	Female65andUp	-0.0834	-0.5646	0.3401	0.5076	0.1503	-0.1542	0.2539
	Maleto25	0.5588	0.1785	-0.4769	-0.1676	-0.1855	-0.2638	0.2765
	Male25to65	0.4884	-0.5204	-0.1452	-0.3387	-0.0074	0.2838	-0.2971
	Male65andUp	-0.0691	-0.5829	0.4568	0.3842	0.0858	-0.1345	0.2088
	NonWhite	0.3274	0.2049	-0.4995	0.1949	0.2118	0.4104	0.2034
	MedAge	-0.3060	-0.4570	0.6539	0.2765	0.1436	0.0158	-0.1778
	Below150PvThr	-0.4366	-0.3010	-0.1636	0.0154	-0.0748	0.0792	0.2999
	MedHomval	0.6357	0.3681	0.3154	0.2692	0.1939	0.0796	0.0026
	OCC_MBASO	0.5876	0.3790	0.3371	0.2782	0.0652	-0.0362	0.0087
	OCC_NRCM	-0.6357	0.1038	0.0098	-0.0733	-0.0915	-0.0136	-0.0583
	OCC_PTTM	-0.7896	0.3130	0.0049	-0.0494	0.0983	0.0630	0.0998
	HSorEqvlnt	-0.6486	-0.3865	-0.2577	-0.3235	0.0390	0.0400	0.0878
	DegreeBA	0.6426	0.3422	0.3776	0.2591	-0.0486	-0.0672	-0.1187
	DegreeAdv	0.6361	0.3774	0.3317	0.3595	0.0552	-0.0341	0.0433
	Smartphone	0.8891	-0.0712	-0.0058	-0.1605	-0.0982	0.0235	0.0106
	Tablet	0.8455	-0.0719	0.0790	-0.1865	-0.1222	0.0478	0.0239
	Broadband	0.8584	-0.2929	0.0895	-0.1328	-0.0966	-0.0286	0.0407
Correlation Indices	*Moran's I*	0.6550	0.3860	0.2050	0.3970	0.4570	0.2370	0.1010
	Pearson Correlation with Density	0.2787	0.1928	-0.4014	0.2057	0.3534	0.4226	0.1286
	Bivariate Moran's I with Density	0.305	0.206	-0.164	0.147	0.297	0.29	0.059

Table 2. Summary of Traditional Principal Component Analysis. Interpretation Notes: Variable loadings may be either positive or negative and it is rather the absolute value, i.e. its distance from 0 that is most important. Thus, negative loadings can just as strongly characterize a principal component as a positive loading. The choropleth scheme is employed to reinforce this point. Correlation summaries are given at the bottom. (Source: Author)

Part 1: Results of Principal Component Analysis

Seven out of thirty-two principal components (PCs) were initially selected for further analysis based on the Kaiser criterion of having an eigenvalue of greater than one. These seven PCs collectively explain 75 percent of the variance in the data set. These PCs are in effect synthetic variables to which the original thirty-two variables are correlated either positively or negatively, over a range of magnitudes. The higher the absolute value of a loading, the more closely identified with the PC a given variable is. A complete summary of the seven PCs, including their variable loadings, is given in Table 2, which additionally contains correlation indices for the PCs. These indices show global spatial autocorrelation (i.e. spatial correlation of a PC with itself), attribute correlation with density, and global spatial correlation between the PCs and density, or rather, the spatial lag of density. Since multiple variables can be strongly correlated with a single PC, it is useful to think of these PCs as complex socio-economic profiles. A description of these profiles is given in Table 3 for future reference and for the sake of brevity, *PC 1*, *PC 2*, *PC 3*, etc... will mostly be used instead of the descriptor in the rest of the results section.

Principal Component	***Description***
PC 1	Tech-Ed-MBASO
PC 2	Larger Professional Households
PC 3	Older, 2-Person, 2-Vehicle Households
PC 4	Older, 1-Person Households
PC 5	Long Commutes
PC 6	Non-White, 2-Person Households, Transit
PC 7	Economically Distressed, 2-Vehicle Households, Non-White

Table 3. Brief Descriptions of Selected Principal Components. Interpretation Notes: Consult Table 2 for Explanation of Principal Component Characteristics (Source: Author).

The initial interpretation of the seven principal components has been informed both by a careful reading of the PCA summary provided in Table 2 and the mapped clusters of spatially autocorrelated PCs in Figure 2. PC1 accounts for over a quarter (25.78 percent) of the variance in the original data set, and showing significant spatial autocorrelation, with statistically significant high-high value clustering near large urban areas and in some rural settings where several higher education institutions are located (i.e. the Five Colleges area in western Massachusetts). PC1's spatial clustering patterns are not surprising given its profile (refer back to Tables 2 and 3). PC2 is moderately spatially autocorrelated and clustered *near* the major cities of Boston and New York City (along the "Gold Coast" of Connecticut). PC3 is weakly autocorrelated and far more geographically dispersed in its clustering, as it is in fact more representative of lower density and even some rural areas than urban ones. PC4 shows an interesting pattern of dispersal across the region that is harder to interpret, as it

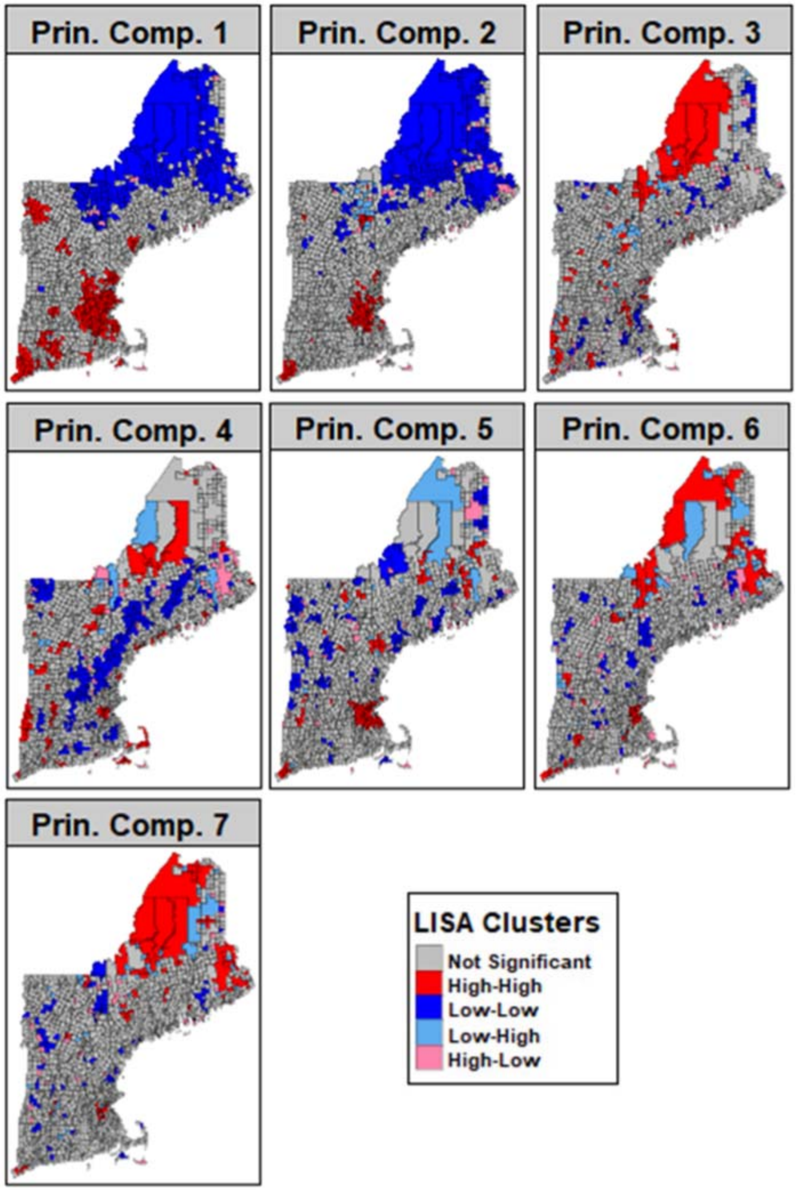


Figure 2. Spatial Autocorrelation of First Seven Traditional Principal Components from PCA. (Source: Author using GeoDA software, Results mapped in 'tmap' package in R).

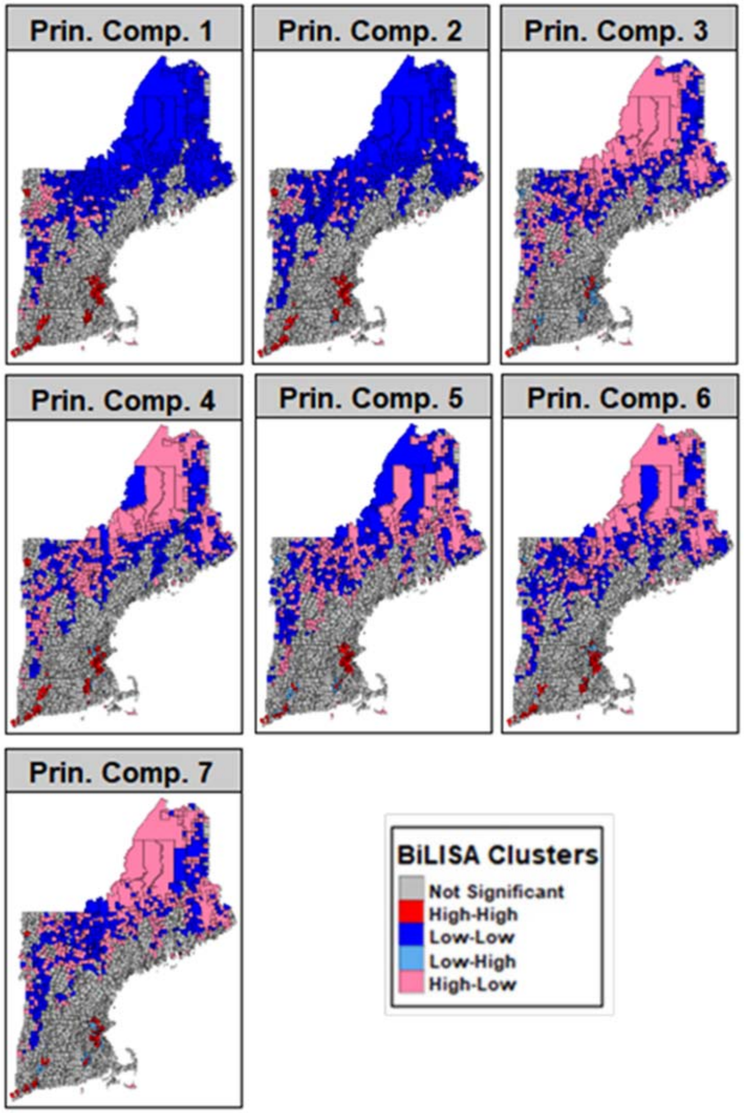


Figure 3. Bivariate Local Moran's I of First Seven Traditional Principal Components from PCA and Population Density. Interpretation Notes: The spatial correlation seen below are between the principal components and the spatial lag of population density. It does not represent a one-to-one spatial correlation analysis between a principal component and density. Rather, it shows the spatial relationship of spatial dependence between the PC in question and general area conditions of density as measured by the lag at each location. (Source: Author using GeoDA software, Results mapped in 'tmap' package in R).

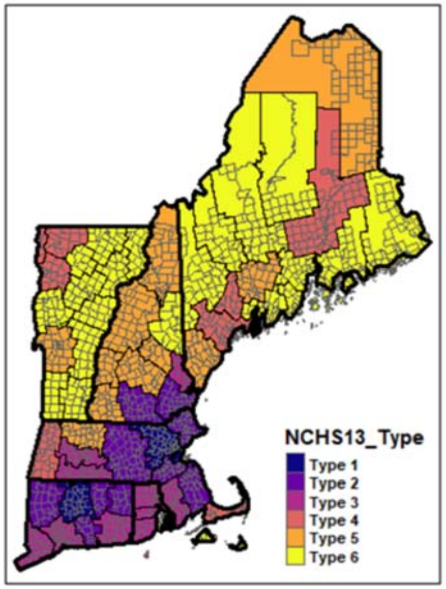


Figure 4. National Center for Health Statistics (NCHS) 2013 Six-way County Typology. (Source: Author using 'tmap' package in R).

appears in both in some urban areas as well as more rural tourist destinations such the Berkshire mountains, Cape Cod, and the White Mountains of New Hampshire. PC5 is squarely defined by longer commuting behavior and thus can be found both on the periphery of urban areas as well as rural ones. PC6 is clustered in urbanized areas but does not register as being strongly spatially autocorrelated, perhaps an artefact concentration of its profile in a relatively small number of towns vis-à-vis the whole of New England. PC7's pattern of clustering is very difficult to discern given its only negligible positive autocorrelation.

Local Spatial Dependence between PCs and Population Density

The principal components' profiles (Table 3) and patterns of spatial autocorrelation (Figure 2) have described internal relationships between original variables and their geographic concentration. Given the known spatial patterns of density, particularly the spatial lag of density (refer back to Figure 1), and the measured spatial patterns of autocorrelation among the PCs, it was mostly anticipated that there would be identifiable clustering of high values of the PCs with high values of population density. In fact, the PCs' statistical relationships demonstrate mildly positive relationships with density in all but one, PC3. *The Pearson's correlation coefficient and the bivariate global Moran's indices given above (see Table 2) establish that a statistical and spatial relationship exists* but does not demonstrate exactly where this correlation is found. The bivariate local Moran's I of the PCs and population density is mapped in Figure 3. A cautious summary and interpretation of the spatial dependency between the PCs and what amounts to the spatial lag of density follows. PC1 and PC2 are clearly positively spatially correlated with the general density characteristic of urbanized areas and have a conversely negative correlation with low density areas of rural northern New England. High-High correlation clusters between PC3 and density are rather scant, with more mixed but statistically significant clustering occurring in the north. The remaining PCs demonstrate a pattern that has more emphasis on High-High clustering in urbanized areas while at the same time having mixed clustering in the north.

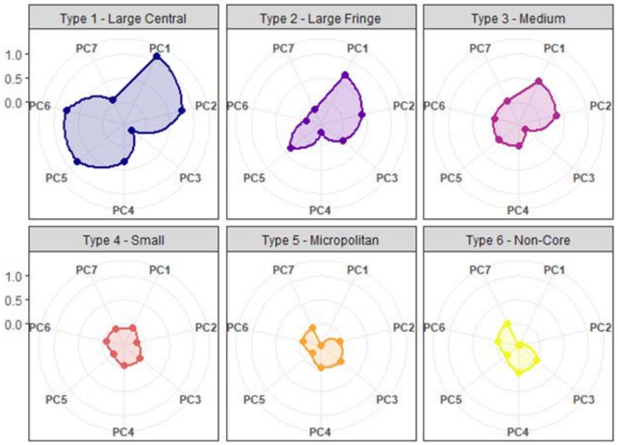


Figure 5. Polar Coordinates Chart of Relative PC Importance for Six NCHS County Types (Refer to Figure 4) in New England. (Source: Author using 'ggplot2' package in R).

Geography has been related *to* principal components as part of the analysis thus far. Conversely, the perspective can be reversed, and multiple PCs related to a geography type. Using the 2013 NCHS County Typology (see Figure 4), the relative importance of PCs to towns falling within a given county type can be assessed (see Figure 5). Large metropolitan counties (Type 1) such as Hartford County in Connecticut and Suffolk County (home of Boston) in Massachusetts, are highly defined by PC1, PC2, PC5, and PC6. Briefly considering these PCs' profiles: there is a prominent role played by educated professionals, their families, technological consumption in urban and suburban areas. Long commutes and non-white, transit using, two-person households (PC6) are also typical, although presumably with slightly different spatial distributions within each county (outer suburbs vs inner suburbs and central city). It can be seen PC1, PC2, and PC5 retain at least some measure of their importance through all but the smallest county types. PC3 and PC4 have alternating levels of relative importance in larger county types but stabilize into moderate importance in smaller ones. As the relative influence of PC1 and PC2 recede in Small (Type 4), Micropolitan (Type 5), and Non-core (Type 6) counties, the relative influence of other PC profiles, such as PC7, begins to be felt. An interesting observation is that small metro counties (Type 4) combine many characteristics of all PCs in moderate fashion but appear to skew to a slightly younger cohort (PC1 vs PC2). These counties are located, for example, around Bangor, Maine, Berkshire Mountains in western Massachusetts, the Burlington, Vermont area, the Cape Cod peninsula, and near Lewiston, Maine. All these counties have notable colleges and universities.

The principal components analyzed above are "global" in nature and reflect a set of constant relationships between contributing variables and their representative principal components. Although these PCs and their relationship to density display spatial

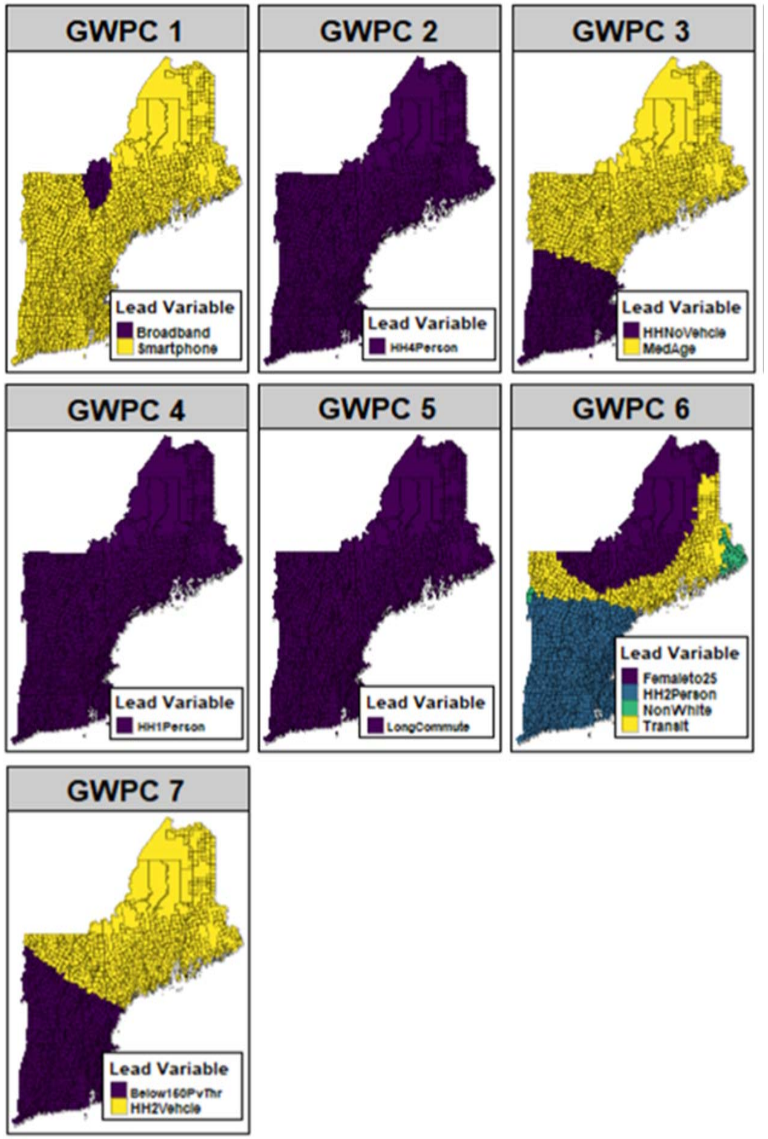


Figure 6. Winning Variable Results for Each Principal Component in Geographically Weighted Principal Component Analysis. (Source: Author Using 'GWmodel' package in R, Mapping Results using 'tmap' package in R).

dependence, this analysis was not fully spatial in the sense that it does not provide for a means of identifying heterogeneity in attribute values across space. One possible approach to this problem is by extending the PCA fully into the spatial domain using a geographically weighted principal component analysis.

Part 2: Geographically Weighted Principal Component Analysis (GWPCA)

A GWPCA offers a more complex analysis of the spatial patterns underlying the PCs and their contributing mobility-related variables than traditional PCA. The GWPCA was performed using the "GWModel" package in R on the same set of variables as the traditional non-spatial PCA above. All data were scaled (z-score normalization) before being put in the model. A non-adaptive kernel function of "exponential" was chosen. A GWPCA consists of both a traditional PCA (that should yield nearly identical results to any other PCA on the same data set) and a spatial one. The outputs of GWPCA include ranges of variance across space as well as the contribution of variables. In effect, each location has its own local PCA, with the "local area" being determined by the kernel function. Despite the sophistication of the analytical tool, the visual output is rather simple. Several types of maps can be produced, including one that displays the *spatial variation* of total cumulative variation of PC1 to a subsequent PC, and the second type displays the "winning" variable for each location, something that helps address the question of which contributing variable for a given PC matters the most for a location. Only "winning" variable maps will used in this analysis.

The results of GWPCA (especially when viewing their mapped patterns) must be treated with caution as they are greatly affected by which variables are included (just as is the case with traditional PCA) but also by the bandwidth selection. Given the potential for variability, it is best to use this method as an exploratory one. With this precaution in mind, the mapped results of the GWPCA on the thirty-two variables in this analysis will be discussed as an additional consideration to the traditional PCA rather than formally analyzed in their own right (see Figure 6). In contrast to global PCA where the "winning" variable for a PC would remain constant, in GWPCA the "winning" variable is allowed to vary over space. The variable with the highest absolute loading for a given location is mapped. As such, it is all the more telling that for GWPC1, GWPC2, GWPC4, and GWPC5 one variable dominates (or very nearly so...) the whole region. GWPC3, GWPC6, and GWPC7 see greater spatial variability in which contributing original variable is the most important. The winning variables from this GWPCA are all important in their corresponding traditional PC, being at or nearly at the top ranking as determined by their absolute value. Substantively they show that these variables were regionally significant in the analysis of the data set used here.

This GWPCA discussion demonstrates that the relative importance of variables can change over space and adds to the discussion of the traditional PCA by confirming the importance of top contributing variables. However, there are two caveats to bear in mind: 1) is the aforementioned potential for instability in geographically weighted results (owing to which variables are included and bandwidth), and 2) is the fact that there may be little quantitative difference between, for example three candidate variables (say, A, B, C) for being the "winner" with hypothetical loadings of 0.7356, -0.7451, 0.7196, respectively, but the variable with the highest absolute loading (e.g. variable B with | -0.7451|) will be named the winner. There is a strong possibility that many important variables would not be considered as a result. Despite these concerns, GWPCA is a valuable addition to the analytical toolkit of multivariate spatial analysis since it highlights the diversity relationships between variables across space.

Discussion

The seven selected principal components all displayed statistically significant spatial dependence with population density. A closer look at their profiles makes this unsurprising. University-educated, technologically connected, professionals and students in their younger and working years comprised PC1. They lived in more expensive homes and are able to begin their commute to work at a reasonable hour as distances traveled and time spent commuting are not that great. This profile is significantly concentrated around urban areas and college towns in New England. A similar profile was seen in PC2, only it is associated with larger families and can be understood as suburban professional families with children. These two profiles were among the most associated with the higher general density of urbanized areas. Thus, even where the original contributing variables were selected to be related to density, but not too closely as would have been the case with a measure of place-based accessibility or commuter network connectivity, the top two PCs are clearly marked by an affiliation with density. The mildly negative correlation of PC3 consisting of older, two-person, and -vehicle households and density stands out among the other PCs, but not decisively so. Long commutes (PC5) are found in both urban and rural areas but do not constitute an especially weighty PC overall. The GWPCA highlighted that a number of contributing variables can be considered the most influential vis-à-vis their PC depending upon location. Smartphones and broadband Internet could both claims to be the most important contributing variable for different areas within the PC1 profile. Other GWPCA results on winning variables can be interpreted and informative towards the traditional PCA analysis in a similar manner. A number of conclusions can be drawn from the forgoing analysis: 1) Even for socio-economic variables selected for the purpose of not being too closely associated with density, density still matters a great deal. 2) There appears to be a more dynamic relationship between urbanized (cities, suburbs, and urbanized corridors) areas and socio-economic indicators dealing with education, occupation, wealth than is the case with areas with far lower density. 3) Variables may be more important in some locations than others.

The usefulness of a PCA analysis lay in identifying important variables and creating new synthetic ones that can be used in regression modeling or the creation of indices. The seven PCs created from this analysis could certainly be used for this purpose. Additionally, the traditional PCA could also act as a filter of potential variables to include in a model and be aided in this effort by the results of a GWPCA that would explicitly name “winning” variables for given locations. Both of these applications are directly related to fitting regression models and future research could pursue them, especially in the context of travel behavior. A third benefit of PCA is in exploring the relationships between variables within the data set, performed on both a geographic and attribute level in this analysis. This is the more immediate contribution of this research as large numbers of socio-economic variables are used in modeling transportation behavior and sometimes the relationship between them is far from clear. A PCA is only a snapshot of these relationships for a given set of specific inputs, yet nevertheless, some trends in the data relevant to the relationship between mobility, accessibility, and connectivity, all anchored by agglomeration (represented here by population density) do emerge. Although the close relationship of density to place-based accessibility (derived from a SIM) is well known, the most important socio-economic variables within this data (selected for their association with mobility) are also highly related to density, since they are most frequently found in dense, urbanized areas. All of these areas also happen to have reasonably high measures of accessibility and connectivity. It may be unclear as to how much population attributes such as wealth (home value), level of educational attainment, occupation would directly translate into higher observed levels of mobility (such as VMT), but they would certainly augment existing advantages of highway network connectivity and accessibility, thus affecting a measure such as VMT. It is clear that the spatial element of studying principal components cannot be minimized or ignored. There is pronounced spatial variation in where different PCs are important. Likewise, there is pronounced spatial variation in where constituent variables with a single PC are important. In the realm of transportation studies, this would reaffirm that the spatial distribution of socio-economic factors contributing towards mobility must be studied just as spatial variance in accessibility and connectivity.

Conclusion

The relationship between the socio-economic variables that influence mobility and population density was shown to be statistically significant. Principal component analysis allowed for a faster, unified, and more comprehensive analysis of these variables as they were too numerous to study individually and their combined influence in the form of individual PCs pointed to larger impacts that could not be described by one variable alone. Both the improved understanding of these socio-economic variables as well as the direct outputs of PCA such as the PCs and variables identified by their importance can be put towards efforts to model transportation behavior and transportation systems in future research. One specific avenue may be how socio-economic variables can help model connectivity within a regional commuter-flow network when place-based accessibility does not adequately explain this alone. Another would be to investigate the cumulative variance of geographically weighted principal components across a region and how it relates to the spatial clustering found in global principal components. This research should be of interest to scholars engaged with the study of travel behavior, especially at larger regional scales. Some practitioners in the field of transportation planning and modeling may be interested as well as this points towards methods of reducing the number of potential variables to be including in customized travel demand models. Finally, it may also be of interest to geographers with an interest in the New England region as it addresses important socio-economic patterns across the region that have application both within and beyond travel-to-work matters.

Appendix A: Traditional and Geographically Weighted Principal Component Analysis

Principal component analysis involves complex matrix operations and will only be very briefly summarized here using notation found in Harris et al. (2011). All input data considered in PCA can be summarized in *m* (number of variables) *by n* (number of observations) matrix **X**. The subsequently derived variance-covariance matrix **Σ** has dimensions *m by m* and is calculated from $\frac{\mathbf{XX^T}}{\boldsymbol{n-1}}$, after original variables in **X** have been mean centered, and ideally standardized so as to eliminate size effects and distortions caused by different units of measurement, where T is the matrix transpose function. The variance-covariance matrix **Σ** is described by the equation $\mathbf{LVL}^{\mathrm{T}}$=**R** where **R** = **Σ** and is a positive definite matrix, **L** is a matrix containing eigenvectors (in this case, the loadings of each variable on the corresponding principal component), and **V** is a diagonal matrix of eigenvalues. The matrix product of **XL** creates the component scores for each observation *n*.

In contrast to the single global nature of traditional PCA, geographically weighted regression (GWPCA) enables *n* local PCAs across the study area. Each observation in the data set matrix **X** is a geographic location *i* (a point or centroid within an areal unit) with coordinates (*u, v*). A weights matrix **W** is calculated for each location i and its neighbors using a kernel function, in the case of this analysis: $w_{ij}=exp(-d_{ij}/r)$ where the weighting w_{ij} is equal to the exponentiated quotient of negative distance $-d_{ij}$ (between locations i and j) and the bandwidth r. Thus, in calculating the variance covariance matrix **Σ** in GWPCA, the weight matrix **W**, as defined above, is used in the equation:

$$\boldsymbol{\Sigma}(u, v) = \mathbf{X}^{\mathrm{T}}\mathbf{W}(u, v)\mathbf{X},$$

Having obtained the geographically weighted variance covariance matrix, eigenvalues and eigenvectors can be calculated for each location *i* as:

$$\mathbf{LVL}^{\mathrm{T}}|(u_i, v_i) = \boldsymbol{\Sigma}(u_i, v_i),$$

Given that each location *i* has a small PCA, it is impractical to communicate GWPCA in the same way as traditional PCA. Outputs such as local cumulative variance and identification of "winning" original variables are provided instead.

Appendix B: Indicators of Spatial Autocorrelation

The global version of Moran's index of spatial autocorrelation, referred to here as Global Moran's I, is expressed by the following equation (using notation borrowed from Anselin 2019):

$$I = \frac{\sum_i \sum_j w_{ij} z_i . z_j / S_0}{\sum_i z_i^2 / n},$$

Where the index *I* is equal the sum of the cross-product of observations of variable *x* at location *i* $(x_i - \bar{x}) = z_i$ with observations of variable *x* at location j $(x_j - \bar{x}) = z_j$ over the variance at location *i*, $\sum_i z_i^2 /n$, where *n* is the number of observations. The binary (0,1) weights matrix w_{ij} filters out cross products of locations that are not classified in advance as neighbors, either through contiguity measures or some other scheme, by multiplying the cross product of location pairs either by a 1 if they are designated neighbors, or 0 if they are not. The sum of all the locational weights $\sum i \sum j\ w_{ij} = S_0$ which weights the deviation scores of z_j. Owing to the filtering of neighbors vs non-neighbors, this measures is essentially one of the correlation of variable x to the spatial lag itself.

The Global Bivariate Moran's I is similarly calculated as the univariate Moran's I above, however it differs in that it is instead a correlation between variable *x* and the spatial lag (average of neighbors) of variable *y*. The formula is given as:

$$I_B = \frac{\sum_i (\sum_j w_{ij} y_j . x_i)}{\sum_i x_i^2},$$

Where the index I_B is the sum of the filtered (by weight matrix w_{ij}) cross product of observation *i* of variable *x* with the spatial lag of variable y over the variance of x_i.

Anselin (1995) devised local variants of these global measures of spatial autocorrelation so that the clustering of statistically significant like values could be viewed on a map. The equation for the univariate Local Moran's I is given as:

$$I_i = cz_i \sum_j w_{ij}\, z_j,$$

Where the index *I* at location *i* is equal to the product of $(x_i - \bar{x}) = z_i$ and the weighted sum of $(x_j - \bar{x}) = z_j$, i.e. the spatial lag of *x*. The variance, $\sum_i z_i^2$, is abbreviate to *c*. Similarly, the Bivariate Local Moran's I is given as:

$$I_i^B = cx_i \sum_j w_{ij}\, y_j,$$

Where the index I^B at location *i* is equal to the product of $(x_i - \bar{x}) = x_i$ and the weighted sum of $(y_j - \bar{y}) = y_j$, i.e. the spatial lag of *y*, and the variance $\sum_i z_i^2$ abbreviate to *c*. Statistically significant clusters of index values for both I_i and I_i^B can be mapped according to the scheme of high-high, low-low, low-high, and high-low to reflect the relationship between a observation at location *i* compared to its neighbors, high value of *x* at *i* surrounding by low values etc.

Devon Lechtenberg is a Senior Transportation Planner at the Capitol Region Council of Governments in Hartford, Connecticut. Email: dlechtenberg@crcog.org. He holds a PhD in Geography from the University of Illinois at Urbana-Champaign (2014) and a professional Master of Urban Planning & Policy from the University of Illinois at Chicago (2017). He has professional and research interests in the application of spatial statistics and econometrics to regional transportation and land use issues.

Acknowledgements

I would like to offer my gratitude to the two anonymous reviewers and the editor of the Northeastern Geographer for their thoughts and comments that were extremely helpful in shaping this final product. I am also grateful to colleagues working in academia that gave me feedback on early drafts of this paper.

Disclaimer

The research presented here is my own and does not represent the opinion, position, or intentions of my employer, the Capitol Region Council of Governments (CRCOG).

Endnotes

[1] The nodes are physical locations (i.e. residential origins and work destinations) but the links are conceptual (i.e. mapped straight-line flows of commuters of varying volume reaching work by some indeterminate physical route).

[2] These applications were first stated by Gould (1967) but were cited in Harris et al. (2011) as: "(a) measures of terrain roughness; (b) the varying spatial nature in the connectivity of towns; (c) orientations of physical features and transport networks; (d)characteristics of mean information fields (Hägerstrand 1967); (e) classification; (f) homogeneity of architectural features; (g) measures of residential desirability; and (h) the interpretation of mental maps." (1718).

[3] The "FactoMineR" package was developed by Le Sebastien and Husson (2008). It contains the algorithms for PCA. Kassambara and Mundt (2020) specifically developed "factoextra" to work alongside "FactoMineR" and visually communicate the complex PCA output. The R package "corrplot" (Wei and Simko 2017) is also used for visualization. Kassambara (2017) encouraged and demonstrated the coordinated use of these three packages in a practical introductory text to multivariate analysis.

References

Abdi, H., and L. Williams. 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* 2 (4): 433-459. *https://doi.org/10.1002/wics.101*.

Anselin, L. 1995. Local indicators of spatial association - LISA. *Geographical Analysis* 27: 93-115. doi:*10.1111/j.1538-4632.1995.tb00338.x.*

Anselin, L. 2019. Global Spatial Autocorrelation (1). GeoDa: An Introduction to Spatial Data Analysis. *https://geodacenter.github.io/workbook/5b_global_adv/lab5b.html.* (last accessed 28 February 2021).

________. 2020a. Dimension Reduction Methods (1). GeoDa: An Introduction to spatial data analysis. *https://geodacenter.github.io/workbook/7aa_dimensionreduction/lab7aa.html.* (last accessed 28 February 2021).

________. 2020b. Local Spatial Autocorrelation (1). Geo Da: An Introduction to spatial data analysis. *https://geodacenter.github.io/workbook/6a_local_auto/lab6a.html.* (last accessed 28 February 2021).

Anselin, L., I. Syabri and Y. Kho. 2006. GeoDa: An Introduction to spatial data analysis. *Geographical Analysis* 38 (1): 5-22. https://doi.org/10.1111/j.0016-7363.2005.00671.x.

Anselin, L., and S. Rey. 2014. Modern spatial econometrics in practice: *A Guide to Geoda, Geodaspace and Pysal*. Chicago: GeoDa Press LLC.

Bjarnason, T. 2014. The effects of road infrastructure improvement on work travel in northern Iceland. *Journal of Transport Geography* 41 (December): 229.

Black, W. R. 2003. *Transportation: a geographical analysis*. New York: Guilford.

Bruinsma, F. R., and P. Rietveld. 1996. *A stated preference approach to measure the relative importance of location factors*. Amsterdam: Tinbergen Institute.

Cartone, A., and P. Postiglione. 2020. Principal component analysis for geographical data: the role of spatial effects in the definition of composite indicators. *Spatial Economic Analysis* 16 (2): 126-147..

Clifton, K ., R. Ewing, G. Knaap, and Y. Song. 2008. Quantitative analysis of urban form: a multidisciplinary review, *Journal of Urbanism: International Research on Placemaking and Urban Sustainability* 1 (1): 17-45, DOI: 10.1080/17549170801903496.

Demšar, U., P. Harris, C. Brunsdon, A. S. Fotheringham, and S. McLoone. 2013. Principal component analysis on spatial data: An Overview. *Annals of the Association of American Geographers*. 103 (1): 106-128.

Environmental Systems Research Institute (ESRI). 2019. *ArcGIS Release 10.7.1*. Redlands, CA

Fotheringham, A. S., C. Brunsdon, and M. Charlton. 2003. *Geographically weighted regression: The analysis of spatially varying relationships*. West Sussex: Wiley.

Giuliano, G. 2005. Low income, public transit, and mobility. *Transportation Research Record*. (1927): 63-70.

Gollini, I., B. Lu, M. Charlton, C. Brunsdon, and P. Harris. 2015. GWmodel: An R package for exploring spatial heterogeneity using geographically weighted models. *Journal of Statistical Software* 63 (17): 1-50. https://doi.org/10.18637/jss.v063.i17.

Gould, P. R. 1967. On the geographical interpretation of eigenvalues. *Transactions of the Institute of British Geographers* 42: 53–86. https://doi.org/10.2307/621372.

Hägerstrand, T. 1967. *Innovation diffusion as a spatial process*. Chicago: University of Chicago Press.

Hanson, S. 2004. Chapter 1: The context of urban travel: Concepts and recent trends. *In The Geography of Urban Transportation*, eds. S. Hanson and G. Giuliano, 3-29. New York: Guilford Press.

________. 2010. Gender and mobility: new approaches for informing sustainability. *Gender, Place & Culture* 17 (1): 5-23.

Harris, P., C. Brunsdon, and M. Charlton. 2011. Geographically weighted principal components analysis. *International Journal of Geographical Information Science* 25 (10): 1717-1736.

Hu, L. 2021. Gender differences in commuting travel in the U.S.: interactive effects of race/ethnicity and household structure. *Transportation* 48: 909-929.

Ingram, D. D. and S. J. Franco. 2014. 2013 NCHS Urban–Rural Classification Scheme for Counties. National Center for Health Statistics. *Vital Health Statistics* 2 (166).

Jeffers, J. N. R. 1967. Two case studies in the application of principal component analysis. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 16 (3): 225–236. https://doi.org/10.2307/2985919.

Jolliffe, I. T. 2002. *Principal Component Analysis*. New York: Springer.

Jombart, T., S. Devillard, A. Dufour, and D. Pontier. 2008. Revealing cryptic patterns in genetic variability by a new multivariate method. *Heredity* 101: 92–103.

Kaiser, H. F. 1960. The application of electronic computers to factor analysis. *Educational and Psychological Measurement* 20 (1): 141–151.

Kassambara, A. 2017. *Practical guide to principal component methods in R*. Marseille: STHDA.

Kassambara, A. and F. Mundt. 2020. factoextra: Extract and visualize the results of multivariate data analyses (R package Version 1.0.6.). https://CRAN.R-project.org/package=factoextra.

Kolaczyk, E. D., and G. Csárdi. 2014. *Statistical analysis of network data with R*. New York: Springer.

Krugman, P. 1991. Increasing returns and economic geography. *Journal of Political Economy* 99 (3), 483-499. https://doi.org/10.1086/261763.

Labi, S., A. Faiz, T. U. Saeed, B. N. T. Alabi, and W. Woldemariam. 2019. Connectivity, accessibility, and mobility relationships in the context of low-volume road networks. *Transportation Research Record* 2673 (12): 717–27. doi:10.1177/0361198119854091.

Lafourcade, M. and J. Thisse, 2011. Chapter 4: New economic geography: The role of transport costs. In *Handbook of Transport Economics*. eds. A. de Palma, R. Lindsey, E. Quinet, and R. Vickerman 67- 96. New York: Edward Elgar Publishing.

Le Sebastien, J. J. and F. Husson. 2008. FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software* 25 (1): 1-18. https://doi.org/10.18637/jss.v025.i01.

Li, Z., J. Cheng, and Q. Wu. 2015. *Analyzing regional economic development patterns in a fast developing province of China through geographically weighted principal component analysis*. Letters in Spatial and Resource Sciences.

Marcińczak, S., and B. Bartosiewicz. 2018. Commuting Patterns and Urban Form: Evidence from Poland. *Journal of Transport Geograph*y 70 (June): 31- 39.

Moran, P. A. P. 1950. Notes on continuous stochastic phenomena. *Biometrika* 37 (1): 17-23.

Ortúzar, J., and L. G. Willumsen. 2011. *Modelling transport*, 4th edition. West Sussex: Wiley.

Petrişor, A., I. Ianoş, D. Iurea, and M. Văidianu. 2012. Applications of principal component analysis integrated with GIS. *Procedia Environmental Sciences* 14: 247-256.

Sohn, J. 2005. Are commuting patterns a good indicator of urban spatial structure? J*ournal of Transport Geography 13 (4): 306–17. https://doi.org/10.1016/j.jtrangeo.2004.07.005*.

Tennekes, M. 2018. tmap: Thematic Maps. R. *Journal of Statistical Software* 84 (6): 1-39.

Vickerman, R. 1996. European transport: Problems and policies. *Journal of Transport Geography* 4 (2): 137-138.

Wei, T. and V. Simko. 2017. corrplot: Visualization of a correlation matrix (R Package Version 0.84). Available from *https://github.com/taiyun/corrplot*.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Wilson, A. G. 1967. A Statistical theory of spatial distribution models. *Transportation Research* 1 (3): 253-269. https://doi.org/10.1016/0041-1647(67)90035-4.

________. 2000. *Complex spatial systems: The modelling foundations of urban and regional analysis*. Harlow, Great Britain: Prentice Hall.

Xu, M., J. Xin, S. Su, M. Weng, and Z. Cai. 2017. Social inequalities of park accessibility in Shenzhen, China: The role of park quality, transport modes, and hierarchical socioeconomic characteristics. *Journal of Transport Geography* 62: 38-50.