Title
A methodology to develop a geospatial transportation typology

Authors
Popovich, Natalie
Spurlock, C Anna
Needell, Zachary
et al.

Publication Date
2021-05-01

DOI
10.1016/j.jtrangeo.2021.103061

A methodology to develop a geospatial transportation typology

Authors:

Natalie Popovich[1], C. Anna Spurlock[1], Zachary Needell[1], LingJin[1], TomWenzel[1], ColinSheppard[1], MonaAsudegi[2]

[1]Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, United States

[2]Federal Highway Administration, 1200 New Jersey Ave., SE, Washington, DC 20590, United States

Energy Analysis and Environmental Impacts Division
Lawrence Berkeley National Laboratory

April 2021

This is the Accepted version of a paper published in the Journal of Transport Geography.

Citation and license:

Popovich, N., Spurlock, C. A., Needell, Z., Jin, L., Wenzel, T., Sheppard, C., & Asudegi, M. (2021). A methodology to develop a geospatial transportation typology. *Journal of Transport Geography*, 93, 103061.

DOI: https://doi.org/10.1016/j.jtrangeo.2021.103061

This research was funded by the Federal Highway Administration (FHWA) Office of Transportation Policy Studies, under Interagency Agreement 693JJ318N300068 with the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 to Lawrence Berkeley Na-tional Laboratory. Any statement or results in this paper reflects the authors' perspectives and not necessarily the point of view of FHWA.

ABSTRACT:
We introduce a methodology to develop a geo-typology (geotype) that categorizes each location in the United States in terms of their main drivers of transportation demand and supply. We develop the first comprehensive set of geotypes for both urban and rural areas across the entire United States. This typology is designed to facilitate national level modeling of multi-modal transportation system's response to alternative investment strategies differentiated across different types of locations. We develop a two-stage clustering procedure to systematically and quantitatively characterize the ways in which locations across the nation are similar or different with respect to their potential response to investment strategies of interest. First, we cluster all 73,057 census tracts, using factor analysis and the CLARA clustering algorithm into "microtypes" based on their street network and economic characteristics. Then we cluster regions (core-basic statistical areas and counties) into "geotypes" using PAM clustering according to their commute configurations, polycentricity and density. The resulting set captures both local and regional variation. These microtypes and geotypes are comparable across all locations, enabling a national level perspective, while maintaining sufficient heterogeneity to support a variety of transportation analyses capturing critical geographic variation.

Keywords: Geo-Spatial, Neighborhood, Typology, Geotype, Transportation Investment, Machine Learning, Clustering Algorithm

1. Introduction

While a growing body of work examines innovative transportation design in dense urban environments, the interaction among transportation performance objectives, new service solutions, and dynamic supply and demand across different types of environments, and not only urban areas, is not yet well understood. We develop a geospatial and geo-economic typology using publicly available data to enable the modeling of shared transportation outcomes for prototypical regions that share common transportation attributes. This set of prototypical system configurations, with associated generalizable transportation supply and demand, are designed to facilitate quantitative macro-level analyses of transportation outcomes. Prior efforts have made progress in enabling the micro-level modeling of prototypical transportation systems to examine urban mobility outcomes (Fielbaum et al., 2017; Oke et al., 2019). However, these efforts have been limited to identifying prototypical urban systems and have largely ignored variation in rural forms. They have historically also relied on proprietary datasets. Here, we extend that concept to include both rural and urban locations.

One of the challenges facing transportation planning in the U.S. is the lack of a high-level transportation modeling tool that can capture key regional heterogeneity without relying on data-intensive, regional travel demand models. Existing travel demand models are typically computationally burdensome, data-intensive, unavailable in rural locations, and highly context-specific. Furthermore, they are trending in a direction of greater detail, such as agent-based-modeling (Auld et al., 2016; Sheppard et al., 2016), rather than generalizability and speed. A classification scheme of transportation systems is a powerful tool to enable effective modeling of transport scenarios at a national level, while still capturing relevant variation and allowing for different tradeoffs in different types of locations. We provide a preliminary step toward this type of modeling effort by defining such a classification scheme in the form of a nationally-comprehensive set of representative location types. This approach has application in fields outside of transportation, but is particularly powerful for modeling applications requiring large geographic coverage, but with critical inputs needed that are only available for a subset of locations, and thereby require a rigorous framework that enables appropriate aggregation of different locations or extrapolation of findings from one location type to another.

In our application, we identify fundamental factors that define the differences in the underlying local transportation systems and travel patterns across different types of locations from a national perspective. The use case of this specific typology is geared toward a national level transportation model with transportation cost and traffic system dynamics models specified to capture heterogeneity across different types of regions and neighborhoods. The model could be used to, for example, examine tradeoffs between the allocation of right-of-way to different modes in different types of locations. The inputs to our typology categorize localities based on their local constraints and potential to use different types of modes. This typology could then enable a model built, for example, to identify a more efficient way of serving demand through an alternative modal mix. Defining our typology as a function of fundamental geo-economic and geospatial factors, and not only current travel choices, allows for a broader understanding of the likely impact of future investments on travel demand and supply in different locations, as the results are not limited by current mode choices. For this purpose, we conduct a two-stage clustering approach in which we account for both local and regional variation in geo-economic attributes in our typology.

Building on a long history of spatial analysis and typological efforts, this paper presents the first tractable yet comprehensive typology that is informed by primary drivers of travel demand and transportation supply constraints, without relying on current transportation outcomes, with an emphasis on economic and geospatial characteristics of locations, and that cover all areas of the U.S., from rural to urban. To our knowledge, it is the first clustering effort to address variation at two levels of resolution, the census tract and greater regional area within which the tract is situated. The resulting typology provides a foundation for modeling the dynamics of transportation supply and demand for large portions of the country at a time. Building such a model is the focus of continuing work leveraging the typology methodology described here. In addition to the typology, this work generated a unique and extensive integrated database of heretofore disparate publicly available datasets that can facilitate similar analyses for different types of applications. For example, city or regional planners, air quality researchers or oversight bodies, transportation safety policy makers, among others, could use the database to tailor new typologies with different sets of inputs targeted to their applications and use cases.

This paper proceeds as follows. Section 2 provides an overview of other typology efforts relevant to transportation. Section 3 explains the conceptual framework for this work, drawing from a wide range of previous research. Section 4 describes the data selected as inputs for the clustering procedures, the factor analysis, and the cluster analysis. Section 5 presents the results and Section 6 concludes. Section 7 provides a discussion of the contributions and possible extensions of this work.

2. Existing typologies and their applications

The U.S. convention of using population densities to define location types as "urban" or "rural" is widespread (Ratcliffe et al., 2016). However, the different thresholds used for these definitions often result in the conflicting assignment of locations (Isserman, 2005), causing controversy, for example, over the allocation of transportation funding (Federal Transit Administration, 2020). While informative for comparing one urban region to another, or locations within urban regions, this urban-suburban-rural distinction falls short of providing a comprehensive method to capture nuance across a variety of dimensions contributing to differences across location types. Sociologists and demographers have emphasized that the urban-suburban-rural definitions commonly used are inadequate to capture variation in access to opportunities (van Eupen et al., 2012) or the primary function of the transportation system (Lowe et al., 2018). Decades of research has worked to address this limitation and provide nuance and complexity to definitions of urban and rural systems, providing a more accurate description of the variation within these region types, often through the use of typological methods. These efforts incorporate multifaceted aspects of economic and geographic interactions that collectively determine access to opportunities. Of particular relevance here are typology efforts that have focused on different spatial scales of street patterns and network structures to describe 'typical' urban forms at the regional (Angel and Blei, 2016; Fielbaum et al., 2017; Huang et al., 2007; Sarzynski et al., 2014; Van Der Laan, 1998) or neighborhood (Bagley et al., 2002; Song and Knaap, 2007) level. Some research efforts have extended this thinking to leverage their typologies to predict neighborhood-level responses to transportation policies such as transit investment (Lutin et al., 2008; Nilsson and Delmelle, 2018) in different location types. None of these prior efforts, however, have considered both neighborhood and regional aspects of development patterns.

Other typologies have been developed specifically to predict travel outcomes, such as mode choice, vehicle-miles-traveled, or emissions, as a function of the built environment and street network attributes (Bagley and Mokhtarian, 2002; Handy, 1996; McCormack et al., 2012; McIntosh et al., 2014). These efforts inform the choice of variables we use in our clustering procedure. However, their methods differ in one important way. To the extent these typologies provide meaningful, categorical distinctions, they do so to better predict current outcomes, and therefore, endogenize current travel choices. In contrast, we develop our typology according to primary determinants of trip generation and the constraints on transportation supply, which in turn determine travel demand and costs for different modes. Our intention is that this framework provides maximum flexibility, enabling modeling of the *potential* for locations to respond to different transportation investment strategies, regardless of their previous degree of reliance on alternative transportation options.

The transportation-related typologies most relevant to our application are those that have set out to define prototypical urban configurations to aid in regional transportation planning, such as Fielbaum et al. (2017), Oke et al. (2019), and Thomson (1978). These efforts have demonstrated the value of defining representative systems upon which detailed models can be run. The main limitations of typologies such as Oke et al., (2019) is that they are defined only for urban systems, rely on current mode shares, and are designed specifically to provide inputs to data-intensive, agent-based models. Typologies such as Fielbaum et al. (2017), provide an extensive treatment of the spatial pattern of employment centers, but omit other important attributes of the urban system. While recognizing the large contribution of these papers to the field, we attempt here to expand those approaches to include non-urban regions as well, while also omitting current travel outcomes from the definition.

Finally, we draw from research that employs two-stage clustering methods to accommodate both spatial and temporal attributes to define not only geographical locations, but also their trajectories over time (Delmelle, 2017, 2016; Wei and Knox, 2014). While these approaches provide rich information about specific locations over time, they are unable to provide sufficient scope to generate a classification approach to compare large numbers of locations with each other. To capture key characteristics for all types of locations, we adapt this two-stage spatial-temporal clustering procedure to a bi-level sequential spatial clustering procedure that represents every location in the United States, except territories.

3. Defining the transportation system

Our approach to generate a typology hinges on the conceptualization of location-specific factors that underpin determinants of travel demand and costs. Formulating the definition of transportation demand and supply constraints requires a recognition that some of these are relevant at a highly localized level, and some are shared by a broader region. The motivation for our two-stage, two-level, approach is that some factors affect relative cost of using alternative modes, for example, at a highly local scale (e.g., street network configurations affecting traffic dynamics and thereby vehicle travel times) while other characteristics affect these relative costs at a broader regional level (e.g., transit is only cost-effective in regions with a certain population density, and is particularly well-suited to locations with highly directional travel flows (Guerra and Cervero, 2011)). We refer to characteristics with higher resolution local variation as *neighborhood determinants* and attributes with relevant variation at the regional-level as *regional determinants*, some of which build from different configurations of neighborhood determinants. Importantly, this research focuses on attributes of the built environment and geography that are not expected to change in the short run. We thus do not include hyper-local attributes, such as the current land use mix or zoning codes.

Error! Reference source not found. Figure 1 summarizes the categories of inputs to our typology. Broadly, demand determinants capture the rate of trips and factors likely to influence that rate, while supply determinants capture the system constraints that reflect the extent to which a location faces restrictions in accommodating the demand, including street network structures and topography. The demand determinants reflect the exogenous drivers of travel, while the supply constraints—or cost determinants—reflect the location-specific exogenous factors mediating the resulting cost of travel, including user travel times, system costs, and external costs (or "externalities") (Applied Research Associates, 2018; Batelle, 2019). Together, cost and demand determinants define the transportation system within which individuals make travel choices, including where, when, and how they travel. The typology developed here is intended to enable subsequent modeling efforts to examine how investment strategies interact in different types of locations to influence outcomes in the transportation system. In this section, we provide a theoretical overview of our choice of inputs, followed by a detailed description of the data available to construct these attributes in Section 4.

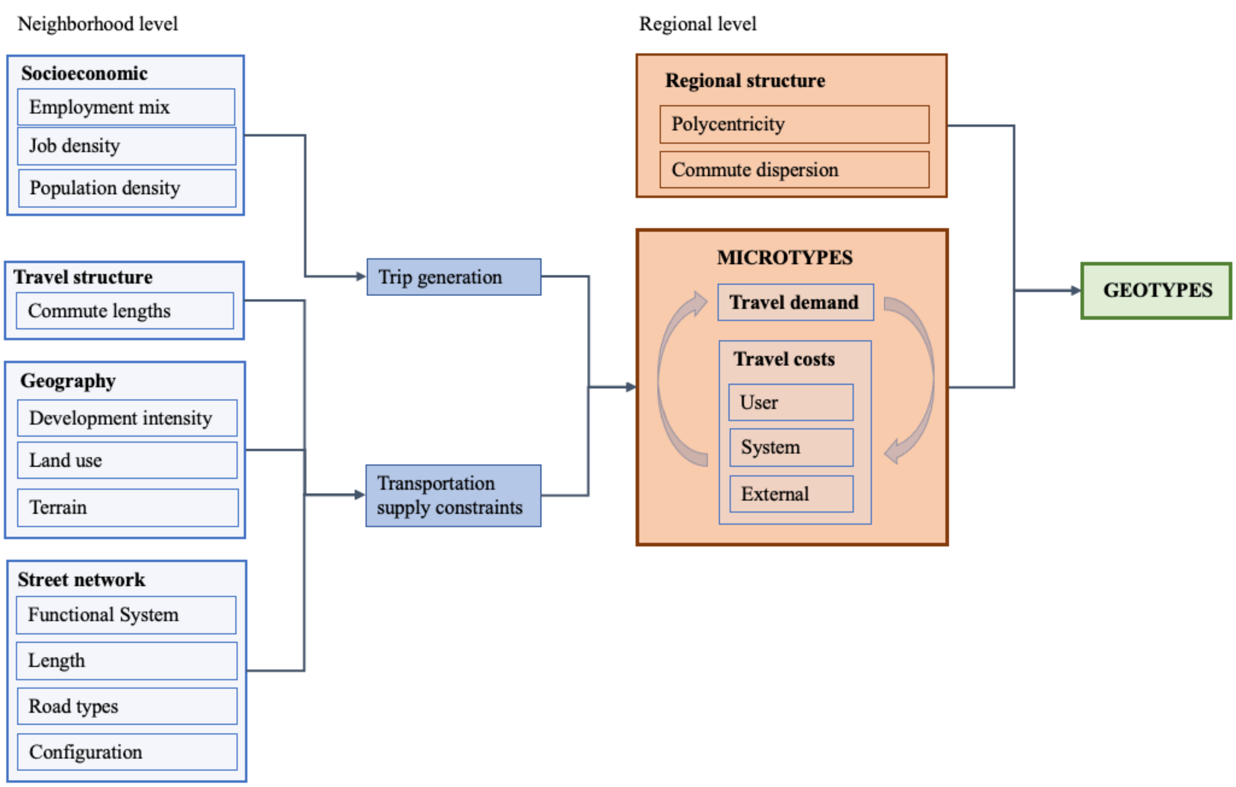

Figure 1 Transportation system and typology structure

3.1 Neighborhood demand for travel

Travel demand is jointly determined by trip generation rates and the costs of travel associated with those trips. We begin with the premise that neighborhood trip generation rates are the product of population density and employment density. In addition, all else equal, locations with similar macro-economic attributes, such as the distribution of employment types, are expected to generate similar travel behavior. This relationship has been made especially clear in light of the COVID-19 pandemic, in which different employment sectors have vastly different opportunities for telecommuting (Alon et al., 2020). Internet access can also eliminate the need to travel for certain activities, such as shopping (Weltevreden, 2007), though whether online activities actually decrease total travel demand is far from clear (Jaller and Pahwa, 2020).

3.2 Neighborhood costs of travel

Travel costs comprise user, system, and external costs, and they are jointly determined by standard trip attributes, such as distance, as well as the geography, built environment, network attributes along the route, and other users in the system. In other words, geospatial attributes constrain the potential transportation supply and dictate the costs of travel that users and non-users of the system face. Hence, we consider both elements of the system that constrain the potential allocation or expansion of road space (street network and geography characteristics) as well as the estimated distribution of travel on the existing infrastructure (travel structure).

User costs are costs borne by the traveler. At the most basic level, user costs consist of time and money. All travel costs increase with trip length. However, there are important elements that determine variation in travel costs for a trip of a given distance. Street network attributes influence travel times for different modes, transportation system resiliency, and access to different destinations. For a given straight-line distance between and origin and a destination, the number of accessible destinations and time required to reach those destinations varies as a function of the density of activities (job density) and the configuration of the street network along the route, including road type, and total right-of-way (ROW) available (street network). For example, the higher the demand for travel per unit of road space, the higher are the expected costs of travel, including travel times, system costs, and external costs. For a given mode and trip distance, user costs increase with road grade (Boriboonsomsin and Barth, 2009) and pavement deterioration (Islam and Buttlar, 2012; Thigpen et al., 2015).

System costs include capital and maintenance costs of modal infrastructure and are typically paid for fully or in part by public entities. System costs, such as ROW allocation and expansion, vary directly according to the built environment surrounding the road, such as development intensity, road grade, and terrain, as well as the functional class of the roadway (U.S. Federal Highway Administration Office of Highway Policy Information, 2016). For example, it may be cost-prohibitive to reallocate ROW on highways with full access control to active modes, whereas existing roads with no access control are less costly to reallocate. Other determinants of system costs include freight demand, which reduces the effective amount of ROW that can be allocated to other modes and increases congestion in the system. Maintenance and resurfacing costs, in particular, increase with the level of freight traffic (Bai et al., 2009), owing to the significant weight of trucks. The provision of transit services is also typically more expensive in dense locations, but requires a minimum level population and employment density to be viable at all (Guerra and Cervero, 2011).

External costs are those imposed by travelers on other users and non-users of the transportation system, and may include things such as air pollution, crashes, noise, greenhouse gas emissions, or congestion. For a given mode and trip distance, the external costs of travel increase with road grade (Boriboonsomsin and Barth, 2009), intersection density (Batterman et al., 2010), pavement deterioration (Wang et al., 2012), population density via exposure response rates, and baseline levels of pollution (Goodkind et al., 2019). External costs for motorized vehicles also vary with travel speeds (National Academies of Science, Engineering, and Medicine, 1995), which are a function of the number of other users in the system (population density and freight travel), and functional system of the roadway, which is highly correlated with speed limits (U.S. Federal Highway Administration Office of Highway Policy Information, 2016).

3.3 Regional structure of travel

Understanding local generators of trips and costs only provides half of the story for understanding the impacts of transportation investments. To capture the expected impacts of new transit investments or dedicated bicycle lanes, it is also necessary to know how such travel and costs are distributed throughout the region. The overall density of the system determines whether a region is able to support high-capacity transit, for example. Another important determinant of the applicability of different modes is how dispersed travelers are within the region. Polycentricity captures the degree of dispersion of commute trip destinations in the region and is widely acknowledged as playing a key role in determining optimal transportation infrastructure investments (Angel and Blei, 2016; Schwanen et al., 2001). For example, large, fixed-guideway transit modes are most cost-effective in locations where many people travel along the same corridor (Pushkarev and Zupan, 1977). The availability of high-resolution data on commute trips lends itself to defining polycentricity most often in terms of employment opportunity location or access (Anas et al., 1998; Craig and Ng, 2001; Giuliano and Small, 1999; McDonald, 1987; Sarzynski et al., 2014). In addition to the way that origins and destinations are oriented, it is important to understand how travel is dispersed between origins and destinations. For instance, there may be locations in which very few people live or work, but which serve as a major corridor between origins and destinations, receiving a disproportionately high share of total person-miles-travelled, even if they are not especially important trip sources or sinks.

4. Data and Methods

We capture neighborhood determinants at the census tract level. To accommodate both rural and urban areas, we capture regional structure at the Core-Based Statistical Area (CBSA) level, where applicable, and county otherwise. Broadly, we cluster census tracts in the first stage based on neighborhood socioeconomic, street network, land use, and geography determinants, the output of which we call microtypes. In the second stage, we cluster regions according to their polycentricity, travel dispersion, and density, the output of which we call geotypes. The final output of this sequential clustering procedure is a set of microtypes and geotypes, with each tract assigned to a microtype according to its first stage cluster and each county/CBSA assigned to geotype in the second stage clustering. We select inputs to the clusters based on existing literature and our guiding conceptual framework, as described in the previous section. The remainder of this section describes the data sources used to generate the inputs to the clustering at these two stages, falling into the categories summarized in **Error! Reference source not found.**and captured at either the microtype or geotype level.

4.1 First stage cluster inputs

The basis for the first stage clusters are 73,056 census tracts as defined by the 2010 census boundaries (U.S. Census, 2019b). We omit 266 tracts whose areas are 100 percent water from the first stage clustering for a final sample of 72,794 tracts. Some inputs capture aspects of multiple cost categories.

4.1.1. Travel demand determinants

Table 1 summarizes the trip generation determinants that enter the first stage clusters. We calculate population density as the number of residents per square mile of land area and job density using the number of jobs listed in the U.S. Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) database (U.S. Census Bureau, 2017). We capture the traditional economic sector with the percent of jobs in mining and manufacturing, each as a proportion of total private sector employment. In practice, the percent of land area made up of agricultural land and the agricultural employment mix are highly correlated, so we chose not to include the latter and instead allow agricultural land use to proxy for agricultural employment representation. We include job-housing balance to approximate job competition, which mediates trip rates. We include the percent of households with broadband access (Federal Communications Commission (FCC), 2019) as a proxy for potential trip substitution. Finally, we include a variable called "trip-sink magnitude," which provides an estimate of the relative draw of that census tract, in terms of number of work trip destinations divided by work trip origins, drawing on the notion of polycentricity as a function of in-flows and out-flows put forth by Sarkar et al. (2020). A location with a trip source magnitude greater than one has more work trips that terminate in the tract than work trips that begin in the tract.

Table 1 Trip generation inputs to the first stage clusters

Variable	Definition	Mean	Std. Dev	Source
Population density	People per sq. mile	2,097	4,683	U.S. Census Bureau (2019a); Dewitz (2019)
Job density	Jobs per sq. mile	964	6,821	U.S. Census (2017); Dewitz (2019)
Pct. jobs manufacturing	Pct. of jobs in NAICS 31-33	0.006	0.037	U.S. Census Bureau (2017)
Pct. jobs mining	Pct. of jobs in NAICS 21	0.075	0.135	
Pct. ag land	Pct. of area in classes 81 and 82 (crops, pasture, ag)	0.12	0.208	Dewitz (2019)
Jobs-housing balance	Jobs per capita	0.9	26.3	U.S. Census Bureau (2019a); U.S. Census (2017)
Trip-sink magnitude	Number of work trip destinations divided by number of work trips origins (homes)	1.357	6.811	U.S. Census Bureau (2017)
Broadband access	Pct. of households with broadband access (categorical ranging from 1 to 5)	4.312	0.896	Federal Communications Commission (FCC) (2019)

4.1.2 Travel costs determinants

Table *2* summarizes the neighborhood determinant inputs to the first stage clusters capturing user, system, and external costs. The largest determinants of user and system costs are derived from the road network characteristics. The Bureau of Transportation Statistics National Transportation Atlas Database (NTAD) provides detailed attributes of each road link in the country. From these, we include: the distribution of road types (e.g., arterials, highways, or local roads), road grade, International Roughness Index (IRI), percent of roads with controlled access, and estimates for average annual daily traffic (AADT) of combination trucks per lane-mile and the percent of total AADT that is freight traffic per lane-mile of ROW in each tract (Bureau of Transportation Statistics, 2018). When road grade data are not available, we use average slope, derived from the R packages *elevatr* and *raster*. To enable comparability across regions, ROW quantities are normalized by population and land area in the clustering inputs (lane mile density and lane-miles-per-capita).

ROW represents the existing area available for surface transportation modes to meet travel demand. The surrounding terrain provides an indication of how expensive it may be to expand this ROW to meet demand. In addition to the NTAD variables, we capture important terrain attributes that constrain future transportation supply. The U.S. Geological Survey National Land Cover Database (NLCD) provides the amount of impervious surface area, which we use as a proxy for development intensity, and water area of each tract (Dewitz, 2019). We capture external costs with characteristics that contribute to exposure rates, including population density and the number of criteria pollutants for which a census tract is in non-attainment status (Environmental Protection Agency (EPA), 2018).

We capture exogenous, location-specific variation in determinants of user costs with inputs that contribute to variation in travel time and experiential trade-offs between modes of a given trip. The street network inputs relevant to user costs capture the extent to which location-specific exogenous factors mediate travel time for a trip of the same straight-line distance. These inputs include average circuity, dead-end proportion, intersection density, self-loop proportion, street density and average street length, from a public database of street network attributes (Boeing, 2017). We include a binned distribution of commute trip lengths for each tract estimated from the LEHD database (U.S. Census Bureau, 2017) to identify locations that might be suitable for micromobility modes.[1]

[1] We recognize that commute trips are only a small subset of trips, but trips of other purposes are not available at such a fine scale with national coverage.

Table 2 Transportation supply constraint inputs to the first stage clusters

Variable	Definition	Mean	Std. Dev	Source
Non-attainment count	No. of pollutants for which tract is in non-attainment status	0.854	1.438	Environmental Protection Agency (2015)
Pct. water	Pct. of area that is water	0.036	0.099	Dewitz (2019)
Development intensity	Pct. of area that is impervious	0.281	0.239	
Avg. circuity	total edge length /sum of great circle distances between the nodes indecent to each edge	1.073	0.06	Boeing (2017)
Dead-end proportion	proportion of nodes that are dead-ends	0.203	0.114	
Intersection density	intersection count / area	34.7	33.3	
Self-loop proportion	proportion of edges with single incident node	0.008	0.016	
Street density	Street length (m) / area (km^2)	7,287	5,067	
Avg street length	mean edge length (m) in unidirected representation of network	240	210	
Avg. IRI	Mean IRI of all lane miles	202.9	80.1	Bureau of Transportation Statistics (2018)
Pct. full access control	Pct. lane miles with full access control	0.101	0.183	
Pct. partial access control	Pct. lane miles with partial access control	0.041	0.123	
Pct. highways	Pct. lane miles of functional system 1 or 2	0.076	0.171	
Pct. ADDT truck	Percent of total average annual daily traffic that is combination trucks	0.035	0.051	
Truck AADT per lane mile	Avg. combination truck AADT / total lane miles	653	1,264	
Lane mile density	Lane miles / area (km^2)	6.6	7.6	
Lane meters per capita	Lane meters / population	36.4	249.7	
Pct. arterials/collectors	Pct. lane miles of functional system 3,4, or 5	0.231	0.253	
Pct. local roads	Pct. lane miles functional system 6 or 7	0.016	0.07	
Avg. road grade	average road grade of shortest commute trip (or average slope if missing)	1.511	1.112	(Graphhopper, n.d.); USGS (2019); U.S. Census Bureau
Pct. trips < 1.3 miles	Distribution of home to work distances	0.085	0.089	U.S. Census Bureau (2017)
Pct. trips 1.3 –3 miles		0.111	0.084	
Pct. trips 3–8 miles		0.244	0.109	
Pct. trips > 8 miles		0.559	0.155	

4.2 Second stage cluster inputs

A region is defined as a CBSA, when applicable, and county otherwise. Of the 3,142 counties, 1,825 are part of one of the 933 CBSAs and the remaining 1,317 counties are included at the county level, for a total of 2,250 observations in the second stage sample. While the inputs to the first stage focus on drivers of the magnitude of local travel demand and costs, inputs to the second stage clusters capture the spatial structure of that travel demand, including patterns of flow and the dispersion of key employment centers in the region. Table 3 summarizes the regional determinant inputs to the second stage clustering.

First, we generate a metric of regional *commute dispersion* using the origins and destinations from the LEHD data (U.S. Census Bureau, 2017). Based off of the normalized Herfindahl-Hirschman Index (HHI), this measure captures the extent to which commutes linking the origins and destinations flow through other tracts within the region. Regional commute dispersion is defined as $H_r = \sum_{i=1}^{N_r} s_{ir}^2$, where r indexes the region, s_{ir} is the proportion of person-miles traveled (PMT) per lane mile (LM) that tract i demands across all N_r tracts in region r. We normalize the HHI, to account for the differing number of tracts in each region, to produce a value between zero and one. The HHI is equal to 1 if all demand for person miles traveled (PMT) per lane mile is concentrated in a single tract ($N_r = 1$) and approaches zero the more uniformly distributed trips are through census tracts.

Next, we derive two inputs to the geotypes from the results of the first stage microtypes. The first input is the region's overall density, defined as the proportion of total census tracts that are in the densest cluster from the first-stage output, representing the breadth of the densest areas of the region. While the proportion of tracts in the densest microtype is highly correlated with the number of jobs, the dispersion of those jobs is important for transportation investments. Hence, we calculate the degree of *employment polycentricity* of each region. We estimate polycentricity by dissolving all spatially contiguous tracts in the first-stage cluster with the highest employment density into a single polygon in QGIS. Then we count the number of non-contiguous employment centers in each region. The higher the number of non-adjacent polygons of microtype 1, the more polycentric is the region.

Table 3 Regional structure inputs to the second stage clusters

Variable	Definition	Mean	Std. Dev	Source
Commute dispersion	normalized HHI of distribution of commute PMT per lane-mile in each tract	0.20	0.30	US Census Bureau (2017)
Polycentricity	# of non-contiguous employment centers	0.65	3.79	Derived from first-stage clusters
Density	Share of tracts in the region assigned to the first-stage cluster with the highest population density	0.018	0.95	

4.3 Factor Analysis

We conduct factor analysis on the 32 raw inputs of the first stage to remove redundancy in the data and improve cluster efficiency (Steinbach et al., 2004). To accommodate the different ranges over which each variable is observed, we first normalize the raw inputs using z-scores such that all values are nationally relative. We choose to normalize values at the national, rather than regional, level because unlike certain socio-demographic variables, such as household income, which are most interpretable in the context of their immediate surroundings, the variables in our analysis are selected to facilitate a direct comparison of locations at the national level. We then group the initial set of normalized raw variables into a smaller number of factors to reduce the number of inputs to the clustering procedure with Exploratory Factor Analysis (EFA) using the R package *psych*. EFA assumes there exists some underlying relationship between the raw variables that can be expressed in a condensed structure. Refer to Appendix A for details. The factor loadings are obliquely rotated based on the assumption that latent factors are correlated (Fabrigar et al., 1999) and extracted using an Ordinary Least Squares (OLS) procedure. We drop observations with systematically correlated missing inputs (.8% of the data). We impute missing values for the remaining observations with randomly missing inputs (.75% of the data) using median values. Parallel analysis suggests that 11–13 factors is optimal to explain the observed variation in the data (Figure A.1 in Appendix A). Weighing tradeoffs between variation explained and interpretability with respect to travel demand and costs, we select 12 factors.

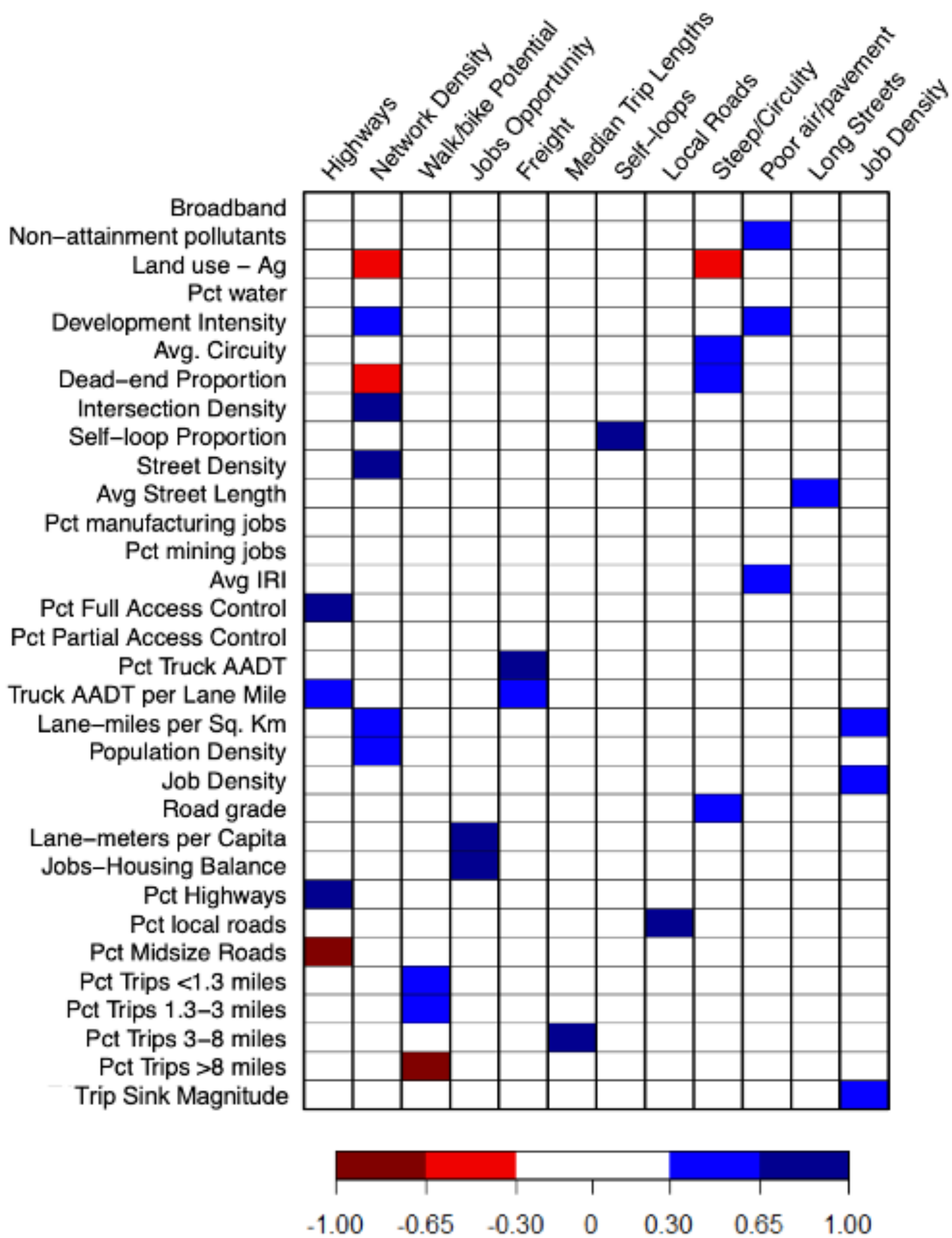

Figure 2 First stage input factors (x-axis) and their loadings (y-axis). *Only loadings with values greater than .3 are depicted.*

Figure 2 depicts the resulting factors (x-axis) and their loadings (y-axis) with an absolute value greater than 0.3 for each of the 12 factors uncovered from the EFA process. Loadings with a large magnitude, negative or positive, indicate a larger contribution of that input to that factor. These factors are grouped according to their relevance to our intended outcomes of travel demand and costs in **Table 4**.

Table 4 Description of factors uncovered

Factor	Components	Travel Demand	Travel Costs		
			User	System	External
Median Trip Lengths	trips 3-8 miles	X	X		
Job Opportunity	high ratio of jobs and ROW per capita	X			
Freight	high freight demand	X	X		
Walk/Bike Potential	many trips under 3 miles	X	X		
Network Density	grid-like, high jobs/population density, high ROW coverage	X	X	X	X
Job density	dense jobs and road coverage	X		X	X
Poor Air/Pavement	high pollution, extensive paved areas, poor pavement condition		X	X	X
Highway	high proportion of highways, heavy freight traffic		X	X	X
Steep/Circuity	dead-ends, steep roads, low agriculture		X		X
Long Streets	long streets		X		
Self-Loops	self-looping streets		X		
Local Roads	mostly local roads		X		

Most of the driving variation comes from the type and configuration of roads in the network. Except for job and population densities, the attributes selected for travel demand do not contribute very significantly to any of the factors. Four inputs are surprisingly in their own factor, indicating that there may be no underlying structure relating these inputs to the others. As they capture important attributes of the street network, we still choose to include them in the cluster analysis. The full loadings for each factor are available in Table C.1 of Appendix C.

4.5 Cluster Analysis

One of the primary objectives of this research is to identify and understand the variation in and co-location of transportation-related geophysical and geo-economic attributes of the built environment across the country. To do so, we select inputs to the clustering procedure that are known to impact trip generation rates and transportation supply constraints, as described in Section 3. In an effort to balance the merits of unsupervised clustering approaches with causal inference methods, we intentionally select inputs demonstrated to influence the demand for and costs of travel, while using a method that does not require defining a structural relationship between the inputs and outputs. Using an unsupervised method allows us to uncover heretofore unknown patterns between the inputs, generating new insights with respect to how certain attributes known to influence travel might be spatially interdependent.

We thus use CLARA, a centroid-based, unsupervised clustering procedure (Kaufman and Rousseeuw, 1990), with the R package *Cluster.* CLARA is an efficient extension of the conventional partitioning around medoids (PAM) clustering technique. CLARA also accommodates outliers better than other centroid-based clustering methods, such as k-means, and works well with continuous variables (Swarndepp and Pandya, 2016). We generate clusters using the reduced dataset of 72,469 tracts by 12 factor scores, the output of the factor analysis. Weighing the results from the silhouette metric (Figure B.1 in Appendix B) with interpretability of the output, we select six first-stage clusters, or "microtypes."

As the goal of the second stage clustering is to group locations according to similarity for use in national-level models, rather than to uncover latent relationships between the inputs, we allow the final set of inputs to enter directly into the second stage clustering procedure without any dimensionality reduction. For consistency with the first-stage clustering methods, we use the PAM clustering procedure to cluster the 2,250 regions. The significantly smaller dataset in the second stage does not require the use of the more-efficient extension of PAM, CLARA. Weighing the results from the silhouette metric and inverse DBI (Figure B.1 in Appendix B) with interpretability, we select six second stage clusters.

5. Results

The first-stage cluster analysis produces six microtypes, labeled 1 through 6, that define all census tracts in the U.S according to their geo-economic, geospatial, and street network attributes that contribute to travel demand and costs. The second stage cluster analysis produces six geotypes, labeled A through F, that define all CBSA and rural counties in the U.S. based on their density, polycentricity, and dispersion of person-miles. While it is not feasible to describe in detail the specific attributes of each of the 36 elements of the resulting typology (six microtypes by six geotypes) in this paper, we highlight some of the features that distinguish clusters from each other. Further details are available from the corresponding author upon request.

5.1 First-stage clusters (microtypes)

Figure 3 summarizes the microtypes. The center of each radar chart in each case represents the minimum median value of each factor score observed across all microtypes and the outer ring represents the maximum. The values are comparable across charts for each factor.

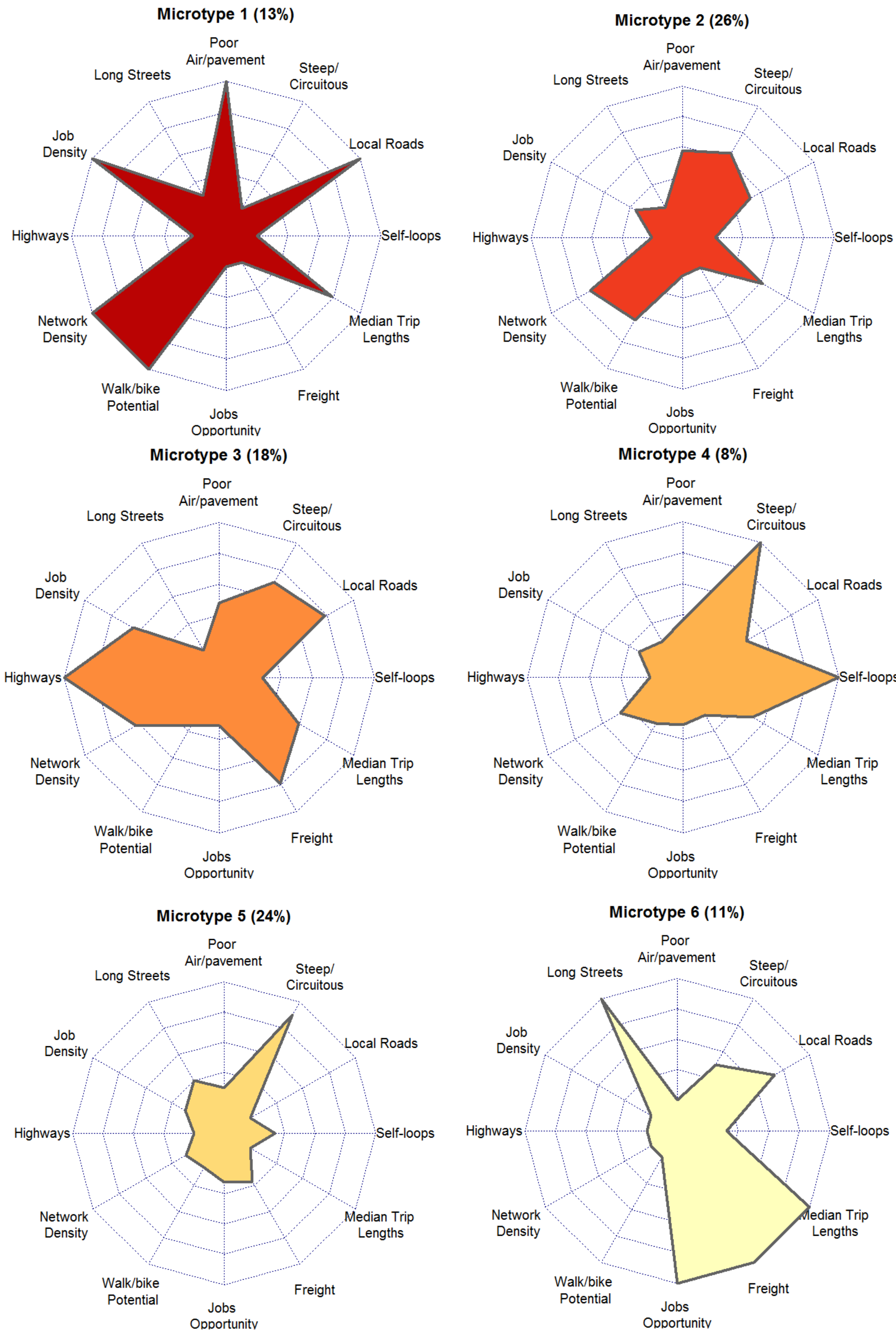

***Figure 3 Factor score medians by first stage microtype**. The center of each radar chart in each case represents the minimum median value of each factor score observed across all microtypes and the outer ring represents the maximum. The values are comparable across charts.*

Microtype 1 is characterized by the highest concentrations of jobs, people, and intersections, and, on average, worst air and pavement quality. These areas represent locations with a low ratio of jobs per capita, potentially in-part due to the high population density, and the highest external travel costs, also resulting from its high population densities, as well as low ambient air quality, and over-utilized pavement. These locations are most often in urban centers. Their grid-like street networks and short commute lengths make them potentially most amenable to multi-modal transportation. Owing to its high concentration of jobs, Microtype 1 is used to denote employment centers for the polycentricity metric in the second stage. Microtype 2 represents the second most populous, second densest street networks, and second-shortest share of commute trips under 3 miles, with few highways or land dedicated to agriculture. Tracts in this microtype are often adjacent to Microtype 1 tracts and do not rank among the highest or lowest observed medians of any factors. It has the largest proportion of census tracts in the country (26%). Microtype 3 has the largest proportion of highways as well as significant freight travel. These locations typically represent highway corridors in urbanized areas and may be amenable to the re-allocation of road space to long-distance transit modes. **Figure 4** depicts the microtype results surrounding the Nashville, TN area, and, in particular, the arrangement of microtype 3 (dark orange) Viewed from above, these tracts often serve as highway and corridors between urbanized locations. Microtype 4 is characterized by its circuitous and self-looping street network. These locations are also the steepest, on average, with very little agriculture, few highways, long commute trips, and sparse populations. They experience some of the highest costs for capital improvements as well as the lowest access to amenities. Microtype 5 is quite similar to Microtype 4, except with fewer dead-ends and somewhat more freight traffic. It represents about one-quarter of all census tracts in the U.S. Microtype 6 has the longest average street lengths, highest freight demand, highest proportion of median commute trip lengths (3-8 miles), and the highest number of jobs per capita, given the low population density in these locations. These tracts tend to cover large area with low population densities. By conventional definitions, these locations would be considered the most rural. Approximately 11% of tracts are in this cluster.

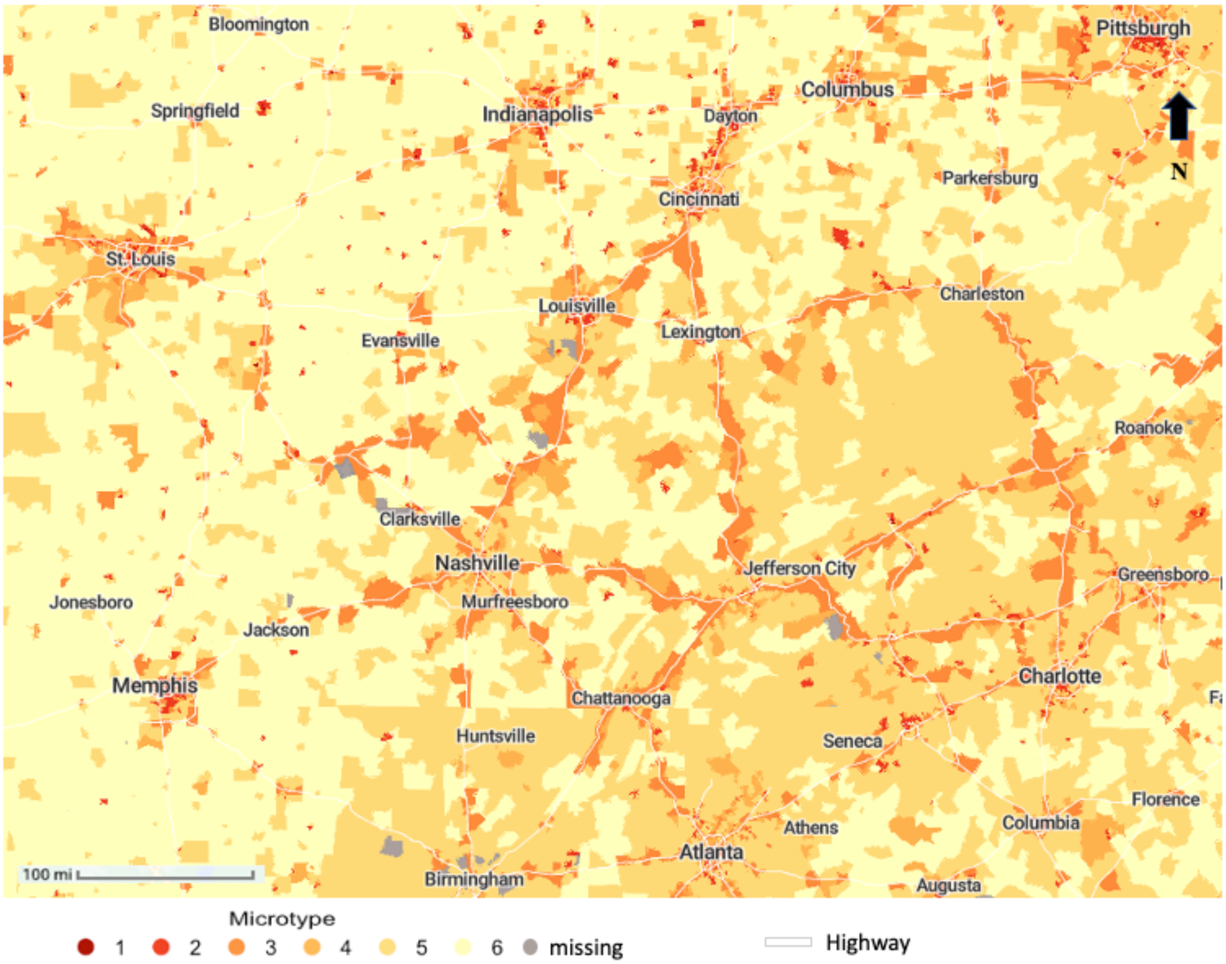

Figure 4 Example microtypes depicting highway corridors (microtype 3) near Nashville, TN.

Compared to existing urban and rural distinctions often used for national transportation planning, the microtypes provide significant improvement in explaining the variation within urban areas and rural areas. For reference, Figure 5 displays the urbanized area that includes Seattle, WA, defined as a region with a population greater than 5,000 individuals (U.S. Federal Highway Administration Office of Highway Policy Information, 2016) as well as the census tracts in the greater Seattle, WA area and their microtype assignment. While population densities drive a large portion of the variation in the microtypes, outside of core downtown areas, the distinction across types is driven more by differentiation in street network characteristics and road types. These different representations of network characteristics imply different tradeoffs in transportation investment strategies.

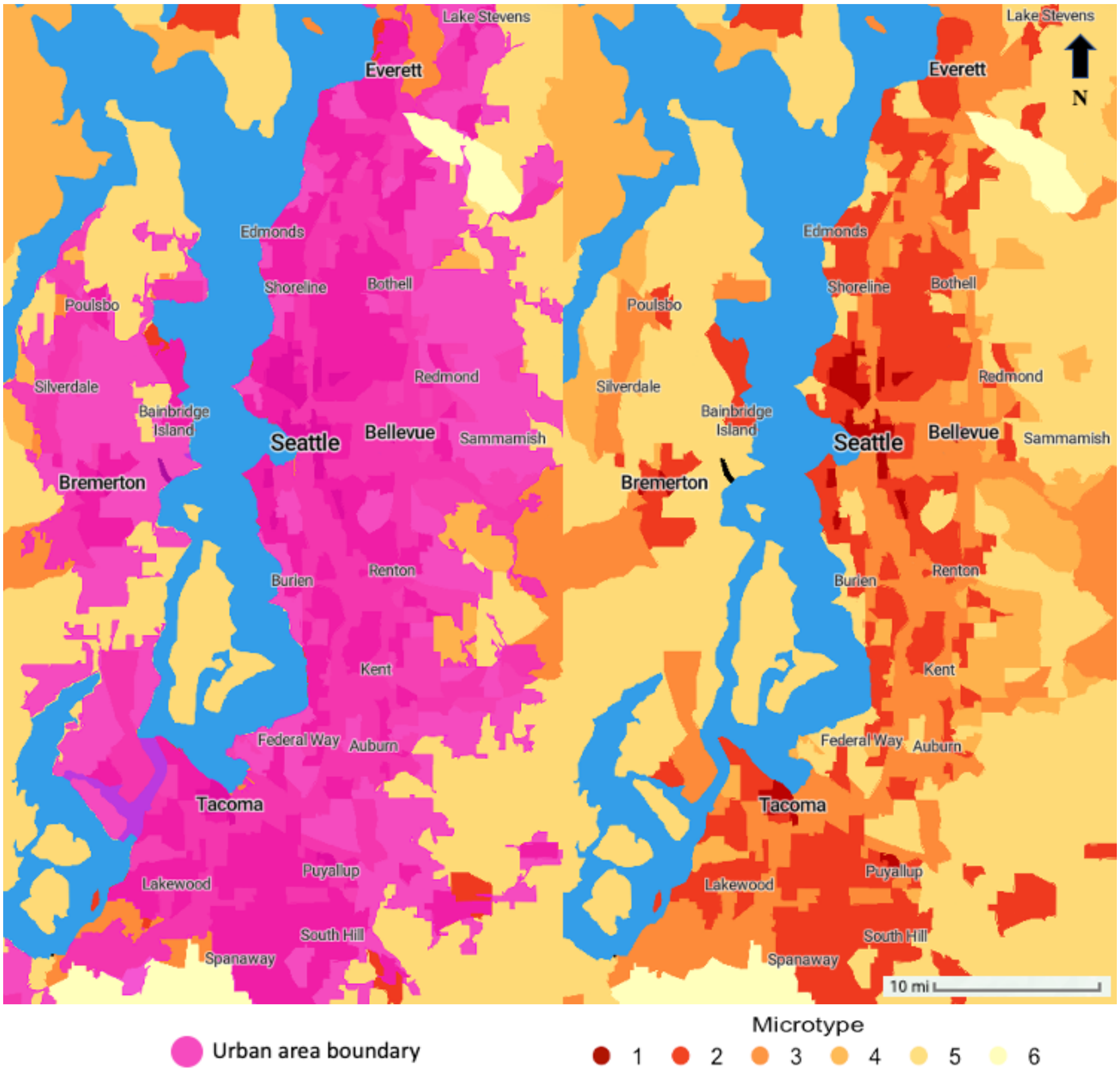

Figure 5 Urban area boundary (left) and microtype assignment of census tracts (right) in Seattle, WA.

5.2 Second-stage clusters (geotypes)

The second stage produces six geotypes, labeled A through F, with the raw inputs depicted in Figure 6. The value for polycentricity is normalized for plotting purposes, such that values closer to one represent a higher number of employment subcenters. Commute dispersion is the normalized HHI index, where a value of zero represents the most dispersed PMT and a value of one represents the most concentrated PMT. Geotype A is comprised of the largest, densest CBSAs, with high polycentricity. On average, regions in this geotype consist of about 20% employment centers (Microtype 1). Commute travel is highly dispersed, with person-miles-traveled spread fairly equally throughout the census tracts in the region. A total of 19 regions are in this geotype with no non-CBSA counties represented. Geotype B represents the next most polycentric locations, in terms of job density, and includes many of the smaller CBSAs. Geotypes C, D, E, and F are defined almost entirely by their relative dispersions of commute travel, as they represent regions with few to no dense employment centers. Geotype F has highly concentrated travel, owing mostly to the fact that these rural counties have only a handful of census tracts. In line with earlier studies of metropolitan commute behavior (Angel and Blei, 2016), we find that a large majority of CBSAs (89%) are characterized by polycentric configurations with dispersed travel, Geotypes A, B, and C.

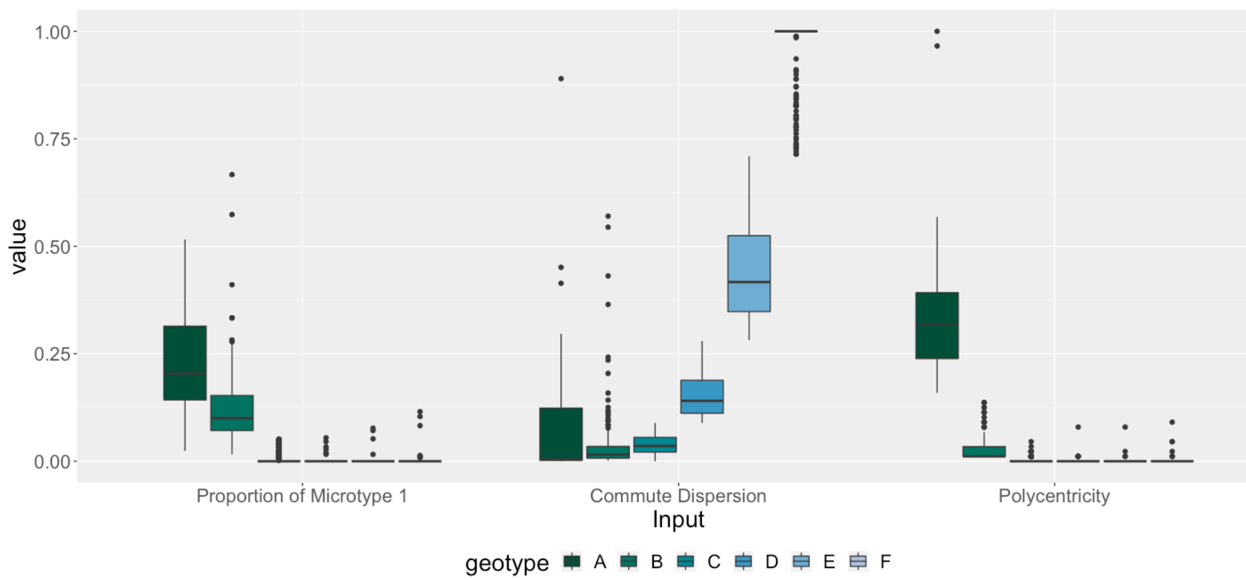

Figure 6 Median input values by Geotype. *The center line represents the median value observed and the bounds of the boxes depict the 25th and 75th percentile. Polycentricity is defined by the number of employment centers in the region (values are normalized for plotting).*

Considering the results of the microtypes and geotypes together, we observe patterns of development across the U.S. At the regional level, we can identify which locations are amenable to different types of existing and emerging modes, for example. Taken together, the microtype and geotype assignment of a location with a high proportion of dense neighborhoods (Microtype 1) and relatively concentrated commute travel (Geotypes E, F) could be able to support transit modes such as light rail.

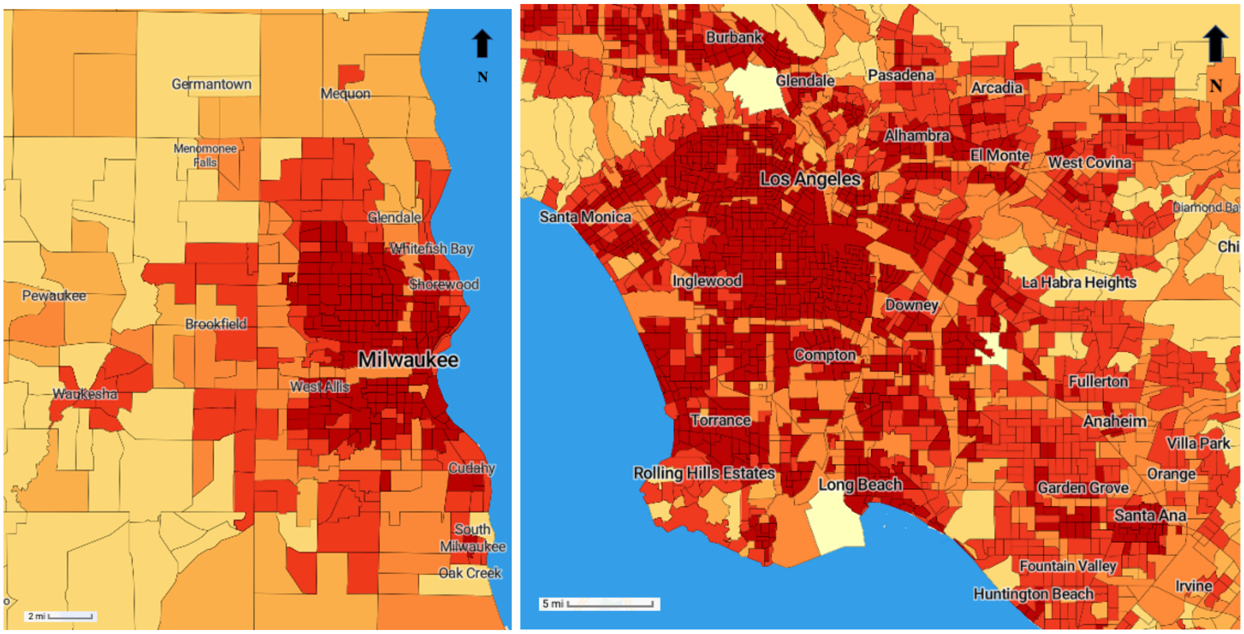

Figure 7 Monocentric (left) versus polycentric (right) employment distribution in Milwaukee, WI (left) and Los Angeles, CA (right)

Milwaukee, WI, and Los Angeles, CA each have the same proportion of tracts that are dense job centers (Microtype 1 represents 42% of each CBSA region). However, the location of these employment centers is highly concentrated in Milwaukee and highly dispersed in Los Angeles (Figure 7). Naturally, larger regions will be correlated with higher numbers of employment centers simply because they have more census tracts. We consider this correlation acceptable because the spatial units of analysis (CBSA or county) roughly capture the full commute sheds for a given location.

6. Discussion

To our knowledge, this paper produces the first tractable and comprehensive transportation typology of all locations across the United States, including locations that are traditionally categorized as both urban and rural. In contrast to existing typologies at the neighborhood (e.g. Song and Knaap, 2007) or urban level (e.g. Oke et al., 2019; Sarzynski et al., 2014), our typology describes locations in terms of both local and regional attributes. It also provides a method to categorize every single location in the U.S., unlike other typologies which focus only on urban (Delmelle, 2017) or rural (van Eupen et al., 2012) locations at a time. One of the primary benefits of constructing this typology is to enable the possibility of modeling plausible transportation scenarios for multiple regions that share common characteristics together. Defining our typology in terms of fundamental geo-economic and geospatial factors could allow for a model, built from this typology, to increase our understanding of the likely impact of future investment strategies on travel demand and supply in different location types. Prior efforts have made considerable progress in enabling the micro-level modeling of prototypical transportation systems to examine urban mobility outcomes (Fielbaum et al., 2017; Oke et al., 2019). However, these efforts have been limited to identifying prototypical urban locations and have largely ignored variation in rural forms. We extend that same concept to include both rural and urban locations and develop a typology for the country as a whole. In a similar vein as (Oke et al., 2019), which demonstrated the value of using a prototypical city to extrapolate the results of transportation modeling scenarios to other similar cities, our two-level typology could be enable a model to be calibrated to observed travel outcomes (e.g. vehicle ownership, mode shares, and demographic data) and thereby used to predict outcomes of different transportation investment strategies in different location types. This is the objective of future work building from this typology.

There are a few key limitations to our analysis and results. One of the limitations of using an unsupervised clustering approach is that we risk losing some of the homogeneity within each cluster, relative to a case in which we cluster locations specifically in terms of their explanatory power on travel outcomes. As this approach is largely already explored in the existing literature, one of the goals of our work is to identify the extent to which such an unsupervised approach can generate meaningful and usable typologies. One of the limitations of interpreting our results is their reliance on work trip data. Focusing on work trips to define polycentricity and trip lengths ignores trip-chaining and all non-work trips, which systematically biases against the travel needs of women, who complete the lion's share of non-work trips (Boarnet and Hsu, 2015). Researchers are collecting more and more non-work trip data, such as the National Household Travel Survey, but there is no existing dataset comparable to the LEHD data for commute trips. Future efforts could expand on this research to incorporate non-work travel when defining location types. Incorporating information on fuel options and regional energy mixes would also provide insight into the energy implications of different transportation investment strategies.

7. Conclusion

In this paper we set out to create a geospatial typology of the United States that captures both local drivers of travel demand and supply constraints as well as the regional structure of travel. We used a two-stage unsupervised clustering analysis to define this typology in terms of the structure of their underlying factors. We found that U.S. census tracts can be categorized into six microtypes according to their geo-economic, geospatial, and road network attributes, each with different implications for transportation investment and mode choice. We found that U.S. CBSAs and rural counties can be categorized into six geotypes that describe the extent of their density, polycentricity, and travel dispersion throughout the region. The resulting set of 36 location types provide preliminary prototypical configurations that could be used as inputs to or constraints in a model to examine different mobility scenarios.

This typology provides a critical first step toward a national level transportation modeling tool upon which the impacts of specific investment strategies on transportation system performance metrics such as travel times or accessibility could be examined in different prototypical regions. The similarities captured here suggest that key components of travel demand and costs are co-located and can be distilled into relatively few location types. This typology can enable co-modeling or extrapolation of calibrated models from one location type to another to leverage relatively sparse, but potentially necessary, modeling data for broader regional or national model coverage. Understanding the distribution of commute trip lengths, in conjunction with the current strain on existing road space, can enable focused alternative investment strategies to increase, for example, micromobility infrastructure, in locations with amenable travel demand characteristics. The ability to isolate specific characteristics of every census tract in the nation and to identify other locations that share similar characteristics enables the testing of investment strategies to be targeted to ensure representation of all types of locations, while capturing critical factors that differ across these types of locations. In addition to the typology, this work generated a unique and extensive integrated database of heretofore disparate publicly available datasets that can facilitate further analyses. For example, city or regional planners, air quality researchers or oversight bodies, transportation safety policy makers, among others, could use the database to tailor new typologies with different sets of inputs targeted to their applications and use cases.

REFERENCES

Alon, T., Doepke, M., Olmstead-Rumsey, J., Tertilt, M., 2020. the Impact of Covid-19 on Gender Equality (No. 26947), NBER Working Paper Series. Cambridge.

Anas, A., Arnott, R., Small, K.A., 1998. Urban Spatial Structure. J. Econ. Lit. 36, 1426–1464. https://doi.org/10.1007/978-1-349-04537-2_6

Angel, S., Blei, A.M., 2016. The spatial structure of American cities: The great majority of workplaces are no longer in CBDs, employment sub-centers, or live-work communities. Cities 51, 21–35. https://doi.org/10.1016/j.cities.2015.11.031

Applied Research Associates, 2018. Task Order #31-10-14043: Update to Typical Improvement Costs per Lane Mile Matrix for HERS.

Auld, J., Hope, M., Ley, H., Sokolov, V., Xu, B., Zhang, K., 2016. POLARIS: Agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations. Transp. Res. Part C Emerg. Technol. 64, 101–116. https://doi.org/10.1016/j.trc.2015.07.017

Bagley, M.N., Mokhtarian, P.L., 2002. The impact of residential neighborhood type on travel behavior: A structural equations modeling approach. Ann. Reg. Sci. 36, 279–297. https://doi.org/10.1007/s001680200083

Bagley, M.N., Mokhtarian, P.L., Kitamura, R., 2002. A methodology for the disaggregate, multidimensional measurement of residential neighbourhood type. Urban Stud. https://doi.org/10.1080/00420980220119525

Bai, Y. (University of K., Schrock, S.D.. (University of K., Mulinazzi, T.E. (University of K., Hou, W., Liu, C., Firman, U., 2009. Estimating Highway Pavement Damage Costs Attributed to Truck Traffic. Analysis.

Batelle, 2019. External Costs of Highway Users. Columbus.

Batterman, S.A., Zhang, K., Kononowech, R., 2010. Prediction and analysis of near-road concentrations using a reduced-form emission/dispersion model. Environ. Heal. A Glob. Access Sci. Source 9.

https://doi.org/10.1186/1476-069X-9-29
Boarnet, M.G., Hsu, H.P., 2015. The gender gap in non-work travel: The relative roles of income earning potential and land use. J. Urban Econ. 86, 111–127. https://doi.org/10.1016/j.jue.2015.01.005
Boriboonsomsin, K., Barth, M., 2009. Impacts of road grade on fuel consumption and carbon dioxide emissions evidenced by use of advanced navigation systems. Transp. Res. Rec. 21–30. https://doi.org/10.3141/2139-03
Craig, S.G., Ng, P.T., 2001. Using quantile smoothing splines to identify employment subcenters in a multicentric urban area. J. Urban Econ. 49, 100–120. https://doi.org/10.1006/juec.2000.2186
Delmelle, E.C., 2017. Differentiating pathways of neighborhood change in 50 U.S. metropolitan areas. Environ. Plan. A 49, 2402–2424. https://doi.org/10.1177/0308518X17722564
Delmelle, E.C., 2016. Mapping the DNA of urban neighborhoods: Clustering longitudinal sequences of neighborhood socioeconomic change. Ann. Am. Assoc. Geogr. 106, 36–56. https://doi.org/10.1080/00045608.2015.1096188
Fabrigar, L.R., MacCallum, R.C., Wegener, D.T., Strahan, E.J., 1999. Evaluating the use of exploratory factor analysis in psychological research. Psychol. Methods. https://doi.org/10.1037/1082-989X.4.3.272
Federal Transit Administration, 2020. Formula [WWW Document]. URL https://www.transit.dot.gov/taxonomy/term/2496 (accessed 12.10.20).
Fielbaum, A., Jara-Diaz, S., Gschwender, A., 2017. A Parametric Description of Cities for the Normative Analysis of Transport Systems. Networks Spat. Econ. 17, 343–365. https://doi.org/10.1007/s11067-016-9329-7
Giuliano, G., Small, K.A., 1999. The determinants of growth of employment subcenters. J. Transp. Geogr. 7, 189–201. https://doi.org/10.11436/mssj.15.250
Goodkind, A.L., Tessum, C.W., Coggins, J.S., Hill, J.D., Marshall, J.D., 2019. Fine-scale damage estimates of particulate matter air pollution reveal opportunities for location-specific mitigation of emissions. Proc. Natl. Acad. Sci. U. S. A. 116, 8775–8780. https://doi.org/10.1073/pnas.1816102116
Graphhopper, n.d. GraphHopper Directions API [WWW Document]. URL https://www.graphhopper.com/products/ (accessed 3.1.20).
Guerra, E., Cervero, R., 2011. Cost of a ride. J. Am. Plan. Assoc. 77, 267–290. https://doi.org/10.1080/01944363.2011.589767
Handy, S., 1996. Urban Form and Pedestrian Choices: Study of Austin Neighborhoods. Transp. Res. Rec. J. Transp. Res. Board 1552, 135–144. https://doi.org/10.3141/1552-19
Huang, J., Lu, X.X., Sellers, J.M., 2007. A global comparative analysis of urban form: Applying spatial metrics and remote sensing. Landsc. Urban Plan. 82, 184–197. https://doi.org/10.1016/j.landurbplan.2007.02.010
Islam, S., Buttlar, W., 2012. Effect of pavement roughness on user costs. Transp. Res. Rec. 47–55. https://doi.org/10.3141/2285-06
Isserman, A.M., 2005. In the national interest: Defining rural and urban correctly in research and public policy. Int. Reg. Sci. Rev. 28, 465–499. https://doi.org/10.1177/0160017605279000
Jaller, M., Pahwa, A., 2020. Evaluating the environmental impacts of online shopping: A behavioral and transportation approach. Transp. Res. Part D Transp. Environ. https://doi.org/10.1016/j.trd.2020.102223
Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics), Eepe.Ethz.Ch.

Lowe, C., Stanley, John, Stanley, Janet, 2018. A broader perspective on social outcomes in transport. Res. Transp. Econ. 69, 482–488. https://doi.org/10.1016/j.retrec.2018.03.006

Lutin, J.M., Krykewycz, G.R., Hacker, J.F., Marchwinski, T.W., 2008. Transit Score Screening Model for Evaluating Community Suitability for Transit Investments. Transp. Res. Rec. J. Transp. Res. Board 115–124. https://doi.org/10.3141/2063-14

McCormack, G.R., Friedenreich, C., Sandalack, B.A., Giles-Corti, B., Doyle-Baker, P.K., Shiell, A., 2012. The relationship between cluster-analysis derived walkability and local recreational and transportation walking among Canadian adults. Heal. Place 18, 1079–1087. https://doi.org/10.1016/j.healthplace.2012.04.014

McDonald, J.F., 1987. The identification of urban employment subcenters. J. Urban Econ. https://doi.org/10.1016/0094-1190(87)90017-9

McIntosh, J., Trubka, R., Kenworthy, J., Newman, P., 2014. The role of urban form and transit in city car dependence: Analysis of 26 global cities from 1960 to 2000. Transp. Res. Part D Transp. Environ. 33, 95–110. https://doi.org/10.1016/j.trd.2014.08.013

National Academies of Science Engineering and Medicine, 1995. Expanding Metropolitan Highways: Implications fo Air Quality and Energy Use - Special Report 245. The National Academies Press, Washington DC. https://doi.org/10.17226/9676

Nilsson, I., Delmelle, E., 2018. Transit investments and neighborhood change: On the likelihood of change. J. Transp. Geogr. 66, 167–179. https://doi.org/10.1016/j.jtrangeo.2017.12.001

Oke, J., Aboutaleb, Y., Akkinepally, A., Azevedo, C.L., Han, Y., Zegras, P.C., Ferreira, J., Ben-Akiva, M.E., 2019. A novel global urban typology framework for sustainable mobility futures. Environ. Res. Lett.

Pushkarev, B., Zupan, J., 1977. Public Transportation and Land Use Policy. Indiana University Press.

Ratcliffe, M., Burd, C., Holder, K., Fields, A., 2016. Defining Rural at the U.S. Censusu Bureau - ACSGEO-1. U.S. Censusu Bur. 1–8.

Sarkar, S., Wu, H., Levinson, D.M., 2020. Measuring polycentricity via network flows, spatial interaction and percolation. Urban Stud. 57, 2402–2422. https://doi.org/10.1177/0042098019832517

Sarzynski, A., Galster, G., Stack, L., 2014. Typologies of sprawl: Investigating United States metropolitan land use patterns. Urban Geogr. 35, 48–70. https://doi.org/10.1080/02723638.2013.826468

Schwanen, T., Dieleman, F.M., Dijst, M., 2001. Travel behaviour in Dutch monocentric and policentric urban systems. J. Transp. Geogr. 9, 173–186. https://doi.org/10.1016/S0966-6923(01)00009-6

Sheppard, C.J.R., Harris, A., Gopal, A.R., 2016. Cost-Effective Siting of Electric Vehicle Charging Infrastructure with Agent-Based Modeling. IEEE Trans. Transp. Electrif. 2, 174–189. https://doi.org/10.1109/TTE.2016.2540663

Song, Y., Knaap, G.J., 2007. Quantitative classification of neighbourhoods: The neighbourhoods of new single-family homes in the Portland metropolitan area. J. Urban Des. 12, 1–24. https://doi.org/10.1080/13574800601072640

Steinbach, M., Ertöz, L., Kumar, V., 2004. The Challenges of Clustering High Dimensional Data, in: New Directions in Statistical Physics. https://doi.org/10.1007/978-3-662-08968-2_16

Swarndepp, S.J., Pandya, S., 2016. An Overview of Partitioning Algorithms in Clustering Techniques. Int. J. Adv. Res. Comput. Eng. Technol. 5, 1943–1946. https://doi.org/10.1017/CBO9781107415324.004

Thigpen, C.G., Li, H., Handy, S.L., Harvey, J., 2015. Modeling the impact of pavement roughness on bicycle ride quality. Transp. Res. Rec. https://doi.org/10.3141/2520-09

Thomson, J.M., 1978. Great cities and their traffic, Peregrine books.
U.S. Federal Highway Administration Office of Highway Policy Information, 2016. Highway Performance Monitoring System Field Manual.
Van Der Laan, L., 1998. Changing urban systems: An empirical analysis at two spatial levels. Reg. Stud. 32, 235–247. https://doi.org/10.1080/00343409850119733
van Eupen, M., Metzger, M.J., Pérez-Soba, M., Verburg, P.H., van Doorn, A., Bunce, R.G.H., 2012. A rural typology for strategic European policies. Land use policy 29, 473–482. https://doi.org/10.1016/j.landusepol.2011.07.007
Wang, T., Lee, I.S., Kendall, A., Harvey, J., Lee, E.B., Kim, C., 2012. Life cycle energy consumption and GHG emission from pavement rehabilitation with different rolling resistance. J. Clean. Prod. 33. https://doi.org/10.1016/j.jclepro.2012.05.001
Wei, F., Knox, P.L., 2014. Neighborhood Change in Metropolitan America, 1990 to 2010. Urban Aff. Rev. 50, 459–489. https://doi.org/10.1177/1078087413501640
Weltevreden, J.W.J., 2007. Substitution or complementarity? How the Internet changes city centre shopping. J. Retail. Consum. Serv. https://doi.org/10.1016/j.jretconser.2006.09.001

Datasets

[dataset] Boeing, G. (2017). U.S. Street Network Analytic Measures. *Harvard Dataverse*. https://doi.org/https://doi.org/10.7910/DVN/F5UNSK

[dataset] Bureau of Transportation Statistics (BTS). (2018). Direct Download of National Transportation Atlas Database (NTAD) Geospatial Files. https://www.bts.gov/geography/geospatial-portal/NTAD-direct-download

[dataset] Dewitz, J.(2019) National Land Cover Database (NLCD) 2016 Products: U.S. Geological Survey data release, https://doi.org/10.5066/P96HHBIE.

[dataset] Environmental Protection Agency (EPA). (2018) Green Book GIS Download. https://www.epa.gov/green-book/green-book-gis-download

[dataset] Federal Communications Commission (FCC). (2019). Fixed Broadband Deployment Data: June, 2019 Status V1. https://broadbandmap.fcc.gov/#/data-download

[dataset] U.S. Census Bureau. (2017). Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES). Version 7. https://lehd.ces.census.gov/data/

[dataset] U.S. Census Bureau. (2019a). *2014-2018 American Community Survey 5-year Estimates Detailed Tables. Total Population*. *Table*

B01003. https://data.census.gov/cedsci/table?q=population%20census%20tract&g=0400000US01.140000&hidePreview=true&tid=ACSDT5Y2018.B01003&vintage=2018

[dataset] U.S. Census Bureau. (2019b) *2019 TIGER/LINE Shapefiles: Census Tracts.* https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2019&layergroup=Census+Tracts

[dataset] U.S. Geological Survey (USGS). (2019). 1 meter Digital Elevation Models (DEMs). USGS National Map 3DEP Downloadable Data Collection. https://catalog.data.gov/dataset/usgs-national-elevation-dataset-ned-1-meter-downloadable-data-collection-from-the-national-map-

[dataset] U.S. Geological Survey (2011). U.S. Conterminous Wall-to-Wall Anthropogenic Land Use Trends (NWALT) https://water.usgs.gov/GIS/metadata/usgswrd/XML/ds948_NWALT.xml

Appendix A. Exploratory Factor Analysis and Parallel Analysis

EFA creates a representation of the M tracts in terms of their latent response variables that represent the underlying structure of the data. The vector of M latent response variables y is defined as $y = \nu + \lambda\eta + \epsilon$, where ν is a vector of variable means, λ is an $M \times T$ matrix of factor loadings, T is the number of factors, η is a $T \times 1$ vector of variable scores on each factor (i.e. factor scores), and ϵ is an $M \times 1$ vector of error terms. The resulting reduced dataset is comprised of $72{,}794 \times T$ inputs, one column of factor loadings for each factor uncovered for each tract. Parallel analysis suggests around 11-13 factors.

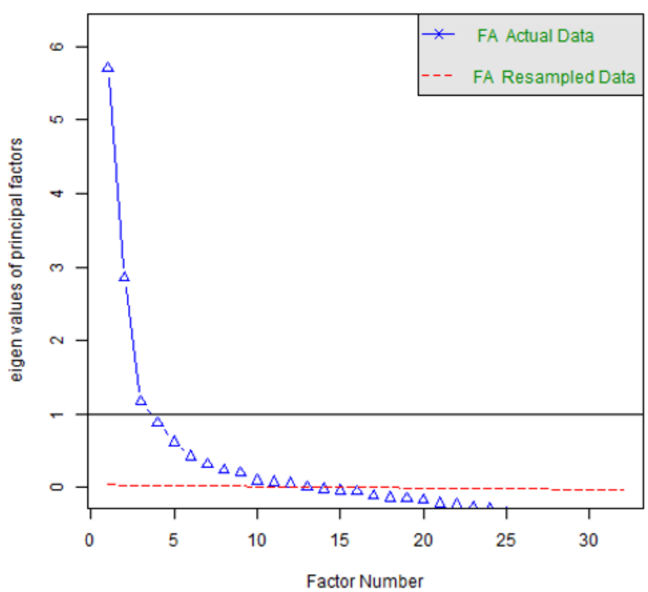

Figure A.1 Scree plot results of parallel analysis

Appendix B. Robustness Checks

We check the robustness of our clusters according to definition and interpretability. Cluster definition refers to whether there is meaningful distinction between clusters and sufficient within-cluster homogeneity. The Silhouette and Inverse DBI indices in Figure B suggest that the clusters are best defined with six to eight groups and the region types with between eight and 10 groups.

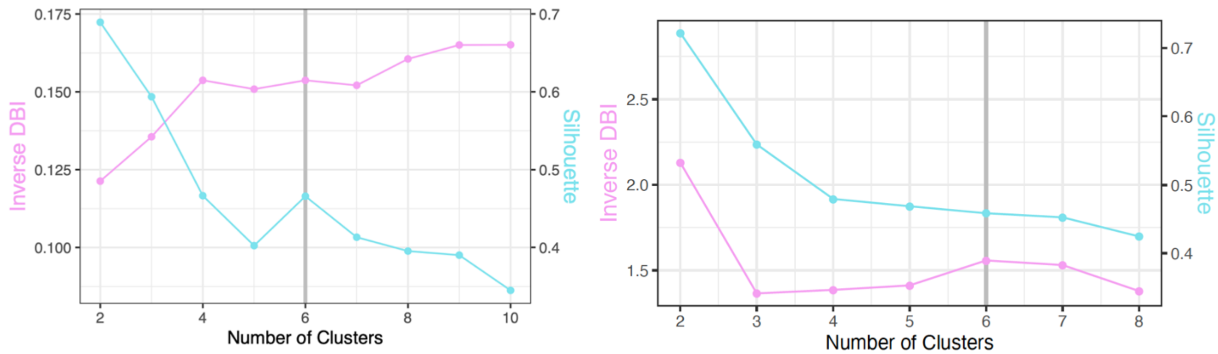

Figure B.1 Metrics to select optimal number of microtypes (left) and geotypes (right)

Appendix C. Factors and their loadings

	Highway	Network Density	Walk/ Bike	Job Opps	Freight	Median Trips	Self-Loops	Local Roads	Steep/ Circuity	Poor Air/ Pavement	Long Streets	Job density
broadband	0.087	0.076	0.092	-0.067	-0.150	-0.053	0.082	0.029	0.248	0.016	-0.171	-0.081
pollutant_count	0.081	-0.098	-0.003	0.003	-0.072	0.022	-0.006	-0.019	-0.011	0.489	-0.038	-0.017
pct_ag_land	-0.023	-0.424	0.056	-0.017	0.137	-0.033	0.010	-0.006	-0.339	-0.085	0.292	-0.018
pct_water	0.001	0.032	0.025	0.023	-0.068	-0.006	0.023	0.014	0.068	-0.167	-0.097	0.021
dev_intensity	0.019	0.513	-0.015	0.000	-0.063	0.059	-0.038	0.018	-0.149	0.359	-0.131	0.106
circuity_avg	0.016	-0.094	-0.096	0.024	-0.001	0.014	0.291	0.011	0.480	-0.095	-0.010	-0.017
dead_end_proportion	0.009	-0.443	-0.064	-0.011	0.026	-0.004	-0.049	0.025	0.445	-0.101	0.026	-0.058
intersection_density_km	-0.054	0.875	0.065	-0.005	-0.009	0.002	-0.027	0.001	-0.020	0.052	0.013	0.026
self_loop_proportion	-0.005	0.002	0.005	-0.001	-0.003	-0.001	1.003	-0.002	-0.009	0.010	-0.002	0.005
street_density_km	-0.024	0.875	0.026	0.002	-0.035	0.016	-0.031	0.001	-0.034	0.017	-0.133	-0.058
street_length_avg	-0.007	-0.197	-0.158	0.076	0.130	0.071	-0.026	-0.033	0.007	-0.073	0.617	-0.028
pct_jobs_manuf	-0.037	0.028	-0.062	0.028	0.072	-0.028	-0.038	-0.011	0.105	-0.016	0.172	-0.008
pct_jobs_mining	-0.006	-0.286	-0.068	-0.008	0.065	0.032	-0.026	-0.017	-0.213	0.008	-0.018	0.036
avg_iri	-0.211	-0.058	-0.006	0.004	-0.114	-0.022	-0.091	-0.099	0.047	0.418	-0.018	0.005
pct_controlf	0.927	-0.037	0.007	0.003	0.042	-0.006	0.003	-0.112	0.002	0.047	-0.008	-0.003
pct_controlp	-0.033	0.117	0.022	0.004	0.045	0.006	-0.008	0.015	0.013	-0.257	0.071	0.026
pct_aadt_combi	-0.020	0.004	0.005	0.007	0.938	-0.014	-0.003	0.009	-0.004	-0.022	0.036	0.005
aadt_combi_per_lm	0.467	0.029	-0.044	-0.024	0.469	0.015	0.008	-0.044	0.025	0.077	-0.207	0.018
lane_miles_sqkm	0.206	0.586	-0.013	-0.028	-0.048	0.036	-0.027	-0.048	-0.059	-0.028	0.026	0.348
pop_density	-0.009	0.398	0.161	-0.039	-0.011	-0.020	0.011	0.012	0.024	0.294	0.256	0.172
job_density	-0.089	0.040	0.008	0.006	0.029	0.003	-0.003	0.022	0.069	0.035	0.022	0.569
road_grade	0.003	-0.049	0.069	-0.009	-0.012	-0.048	-0.030	-0.022	0.414	0.187	0.097	0.180
lanemeters_per_capita	0.028	0.078	-0.024	0.795	0.025	-0.008	0.001	-0.013	0.017	-0.009	0.157	-0.091
jobs_housing_bal	-0.010	-0.050	0.044	0.802	-0.022	-0.011	0.000	0.007	-0.010	0.014	-0.127	0.094
pct_hiway	0.993	0.009	-0.002	0.007	-0.007	0.000	-0.001	-0.112	0.003	-0.032	0.008	-0.005
pct_local_roads	0.001	-0.010	0.000	-0.004	0.004	-0.001	-0.001	0.901	0.000	-0.006	-0.005	0.007
pct_mid_roads	-0.956	0.001	0.005	-0.005	0.004	0.001	0.001	-0.253	0.000	0.020	-0.012	-0.001
pct_trips_bin1	-0.053	0.216	0.415	0.015	-0.020	-0.231	-0.011	0.069	0.018	0.160	-0.087	-0.164
pct_trips_bin2	0.012	-0.082	0.597	-0.029	-0.034	-0.290	-0.016	-0.041	-0.061	-0.110	-0.013	0.108
pct_trips_bin3	-0.007	0.006	0.072	-0.009	-0.011	0.988	-0.004	-0.001	-0.002	-0.002	0.009	0.004
pct_trips_bin4	0.000	-0.016	-0.953	-0.010	0.004	-0.173	0.008	-0.007	-0.001	-0.019	0.021	0.018
trip_sink_mag	-0.028	-0.148	-0.086	0.253	0.003	0.053	-0.009	0.019	-0.087	-0.010	-0.185	0.392