Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Employment Diversification and Urban Mobility Disparities: A Multi-scale Analysis of U.S. Core-Based Statistical Areas

Zeyu Wu

zeyu.wu@nyu.edu

New York University

Marta Gonzalez

University of California, Berkeley

Research Article

Keywords: Economic Complexity Index, Urban Mobility, Income Inequality, Simp son's Paradox, Core-Based Statistical Areas, Computational Social Science

Posted Date: February 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8769603/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Employment Diversification and Urban Mobility Disparities: A Multi-scale Analysis of U.S. Core-Based Statistical Areas

Zeyu Wu*

Department of Civil and Urban Engineering
Tandon School of Engineering
New York University
Brooklyn, NY, United States
zeyu.wu@nyu.edu

Marta C. González

Department of Civil and Environmental Engineering
Department of City and Regional Planning
University of California, Berkeley
Berkeley, CA, United States
martag@berkeley.edu

Abstract

The Economic Complexity Index (ECI), originally developed for international trade, is widely associated with *lower* income inequality at the national level. Whether this relationship holds within countries—across cities and commuting zones—remains an open empirical question with direct implications for regional policy. We address this gap by adapting ECI to employment distributions across 121 Core-Based Statistical Areas (CBSAs) in five U.S. states spanning the national income spectrum (California, New York, New Mexico, Louisiana, Mississippi), integrating Replica origin–destination mobility data with socioeconomic indicators from the American Community Survey. The study reveals a *reversal* of the international pattern: CBSAs with the highest economic complexity exhibit the greatest income inequality (Gini = 0.51, Pearson $r = 0.42$, $p < 0.001$), while simultaneously maintaining the most efficient mobility systems (lowest VMT per capita and radius of gyration). Through Principal Component Analysis and bootstrap-validated K-means clustering, we identify four regional typologies—Prosperous Knowledge Hubs, High-Density Mixed Economies,

Rural Resource-Dependent Regions, and Small Industrial Towns—each characterized by distinct trade-off configurations among complexity, mobility, and equity. This "complexity–inequality paradox" is consistent with a Simpson's Paradox in which skill-biased agglomeration economies and occupational polarization dominate at the regional scale, overriding the institutional mechanisms that mediate the ECI–equality relationship internationally. Our findings challenge one-size-fits-all development strategies and demonstrate that employment-based ECI, combined with high-resolution mobility analytics, offers a scalable diagnostic framework for data-driven regional policy.

Keywords: Economic Complexity Index · Urban Mobility · Income Inequality · Simpson's Paradox · Core-Based Statistical Areas · Computational Social Science

1 Introduction

1.1 Background

The 21st century has witnessed the rapid emergence of data-driven approaches to understanding socioeconomic systems, complementing and in many cases transforming the hypothesis-driven methods that have long dominated the social sciences [Blondel et al., 2015, Barbosa et al., 2018]. Large-scale digital traces of human activity—mobile phone trajectories, financial transactions, social media interactions, and administrative employment records—now make it possible to observe the structure and dynamics of regional economies and urban mobility systems at resolutions that were unimaginable a generation ago [González et al., 2008, Pappalardo et al., 2015]. Yet extracting meaningful signals from these massive datasets, and interpreting them within coherent theoretical frameworks, remains a fundamental challenge. This is especially true when the goal is to understand the relationships among multiple dimensions of regional socioeconomic performance—economic sophistication, transportation efficiency, and income equity—that have traditionally been studied in isolation by distinct disciplinary communities.

The geographic distribution of economic opportunity within nations has emerged as one of the defining socioeconomic challenges of the contemporary era. In the United States, the divergence between prosperous coastal knowledge hubs and economically stagnant interior regions has accelerated since the 1980s, generating spatial inequality that rivals the cross-country disparities traditionally studied in development economics [Moretti, 2012, Ganong and Shoag, 2017, Storper, 2018]. Recent evidence from Connor et al. [2025] demonstrates that large cities fuel inequality not only within but also across generations, while Chetty et al. [2014] has shown that the geography of intergenerational mobility varies dramatically across U.S. commuting zones. These findings underscore the urgency of developing analytical frameworks capable of jointly capturing the sophistication of regional economies, the efficiency of their transportation systems, and the equity of their income distributions—three dimensions that interact in ways that single-indicator analyses cannot reveal.

The Economic Complexity Index (ECI), introduced by Hidalgo and Hausmann [2009] and elaborated in the *Atlas of Economic Complexity* [Hausmann et al., 2014], offers a powerful data-driven lens for characterizing economic sophistication. By quantifying the diversity and exclusivity of a country's export basket through the structure of a bipartite country–product

network, ECI captures the embedded knowledge and productive capabilities that underpin long-run growth. A landmark contribution by Hartmann et al. [2017] demonstrated that ECI is not only a predictor of future GDP growth but also a strong correlate of income equality: nations with more complex economies tend to exhibit substantially lower Gini coefficients, even after controlling for income level, education, and institutional quality. This evidence has been replicated across different time periods and samples [Lee and Qu, 2021, Hartmann and Pinheiro, 2022], establishing what appears to be a robust macroeconomic regularity linking productive sophistication to equitable outcomes.

Yet a growing body of evidence suggests that this regularity does not survive disaggregation to sub-national scales. When the unit of analysis shifts from countries to cities and regions within a single nation, the relationship between economic sophistication and income distribution appears to reverse. This paper provides systematic evidence of such a reversal across 121 U.S. Core-Based Statistical Areas (CBSAs) and, by integrating high-resolution digital mobility traces, reveals that the paradox extends to a three-way configuration in which the most complex regions simultaneously exhibit the highest inequality *and* the most efficient commuting systems. We term this configuration "efficient inequality" and argue that it reflects the dominance of skill-biased agglomeration economies at the regional scale—a mechanism that operates within a shared institutional environment and therefore escapes the institutional co-evolution that mediates the complexity–equity relationship across nations.

1.2 Economic Complexity: From National Exports to Regional Employment

The theoretical foundations of economic complexity rest on the insight that economic growth is fundamentally a process of accumulating productive capabilities—tacit knowledge, institutional routines, and organizational competencies that enable the production of increasingly sophisticated goods and services [Hidalgo and Hausmann, 2009, Hausmann et al., 2014]. The ECI operationalizes this insight by analyzing the bipartite network of countries and products, assigning higher complexity scores to countries that export a diverse set of products that few other countries can produce. This network-based approach captures information about productive capabilities that conventional measures such as GDP per capita, human capital indices, or sectoral shares fail to represent [Balland et al., 2022].

The original ECI framework was designed for and validated at the national level using international trade data. However, beginning with the work of Reynolds et al. [2018] on Australian states and Balland et al. [2020] on U.S. metropolitan areas, scholars have increasingly adapted the methodology to sub-national contexts. Balland et al. [2020] demonstrated that complex economic activities—those requiring diverse and exclusive capabilities—disproportionately concentrate in large cities, with city size explaining 40–80% of the variance in the spatial distribution of complex industries across 353 U.S. metropolitan statistical areas. This concentration is driven by agglomeration externalities that facilitate knowledge spillovers, labor market pooling, and input sharing [Duranton and Puga, 2020, Glaeser, 2014].

A critical methodological advance was made by Fritz and Manduca [2021], who computed ECI for U.S. metropolitan areas using NAICS employment data from County Busi-

ness Patterns rather than trade data. This employment-based approach captures the full spectrum of economic activity including non-tradable services—healthcare, education, public administration—which constitute the majority of U.S. employment and are invisible to trade-based measures. Davies and Maré [2021] further refined sub-national complexity analysis by proposing weighted alternatives to the standard binary revealed comparative advantage threshold, arguing that continuous measures better capture gradations of specialization intensity. The present study builds directly on these precedents, extending employment-based ECI to the CBSA level and integrating it with mobility and inequality metrics that have not previously been examined in the economic complexity literature.

1.3 The Complexity–Inequality Nexus: A Scale-Dependent Relationship

The cross-country evidence linking higher economic complexity to lower income inequality is among the most robust findings in the complexity literature. Hartmann et al. [2017] showed that ECI explains a substantial share of cross-national variation in Gini coefficients—a relationship that holds when controlling for GDP per capita, human capital, and institutional quality. The proposed mechanism operates through the composition of employment: complex economies generate a more even distribution of productive, well-compensated jobs across the skill spectrum, whereas simple economies dependent on raw material extraction tend to concentrate rents among a narrow elite [Hartmann et al., 2017, Hausmann et al., 2014]. This view resonates with the broader regional science literature, where economic diversification has long been considered a predictor of resilience and stability [Dissart, 2003, Martin et al., 2016].

However, emerging sub-national evidence paints a strikingly different picture. Sbardella et al. [2017] analyzed U.S. county-level data and found that wage inequality increases monotonically with industrial sophistication, a pattern they attributed to the steep returns to specialized skills in complex sectors. Fritz and Manduca [2021] documented that higher-complexity metropolitan areas experienced larger median income declines between 1998 and 2015, suggesting that complexity gains at the metropolitan level may not translate into broadly shared prosperity. Pinheiro et al. [2022] described this reversal as the "dark side of innovation geography," showing that the spatial concentration of complex activities in European regions produces winner-take-most dynamics that exacerbate territorial inequality.

Hartmann and Pinheiro [2022] explicitly theorized this reversal as a Simpson's Paradox: the negative complexity–inequality correlation observed across countries flips to a positive correlation within countries because the mechanisms at work differ across scales. At the national level, complexity co-evolves with institutions—educational systems, labor regulations, redistributive taxation, and social safety nets—that collectively compress income distributions. At the sub-national level, within a single institutional framework, these institutional mechanisms are held roughly constant, and the dominant forces become those of skill-biased agglomeration: the spatial concentration of high-human-capital workers and knowledge-intensive firms in dense urban areas, which generates productivity premiums that accrue disproportionately to the already-advantaged [Baum-Snow et al., 2018, Moretti, 2012]. Baum-Snow et al. [2018] estimated that approximately 80% of the faster wage inequality

growth in large U.S. cities relative to smaller cities can be attributed to the disproportionate productivity gains accruing to high-skill workers in dense urban environments. Autor et al. [2003] provided early evidence that the skill content of jobs shifted markedly toward cognitive non-routine tasks beginning in the 1980s, and Acemoglu and Autor [2011] formalized how this shift translates into widening wage inequality as returns to education and specialized skills increase.

Additional mechanisms reinforcing the sub-national complexity–inequality link include occupational polarization—the simultaneous growth of high-wage cognitive and low-wage manual service jobs coupled with the hollowing of middle-skill occupations [Autor and Dorn, 2013, Goos et al., 2014]—and housing cost capitalization, whereby productivity gains in complex metros are substantially captured by rising land and housing values, compressing real incomes for renters and lower-wage workers while enriching the property-holding class [Hsieh and Moretti, 2019, Ganong and Shoag, 2017].

1.4 Urban Mobility, Spatial Mismatch, and Socioeconomic Stratification

The spatial mismatch hypothesis, first articulated by Kain [1968], established that geographic disconnection between residential locations and employment centers creates barriers to labor market participation, particularly for minority and low-income workers. Decades of subsequent research have confirmed and extended this insight, demonstrating that transportation access mediates employment outcomes, earnings, and upward mobility across a range of urban contexts [Gobillon et al., 2007, Blumenberg and Manville, 2004, Ihlanfeldt, 2006, Hu, 2017].

The advent of large-scale mobility data derived from mobile phone records and GPS traces has transformed the empirical study of urban movement patterns [Blondel et al., 2015, Barbosa et al., 2018]. González et al. [2008] introduced the radius of gyration (r_g) as a characteristic spatial scale of individual mobility, demonstrating that human movement follows truncated Lévy flight patterns with predictable regularity. Subsequent work by Pappalardo et al. [2015] and Barbosa et al. [2018] has enriched the understanding of mobility signatures across populations, revealing systematic variation in spatial behavior that correlates with socioeconomic status, urban structure, and transportation infrastructure.

The relationship between mobility patterns and socioeconomic status is complex and context-dependent. Xu et al. [2018] compared mobility–income relationships in Singapore and Boston, finding that the direction of the association depends on urban structure: in monocentric cities, higher-income individuals tend to travel farther, whereas in polycentric configurations the relationship may reverse. Moro et al. [2021] showed that mobility patterns in large U.S. cities are strongly associated with experienced income segregation, with residents of different income levels visiting systematically different sets of neighborhoods and establishments. These findings suggest that commuting efficiency is not merely a technical property of transportation systems but a reflection of the deeper socioeconomic sorting processes that organize urban space.

Despite these parallel advances in economic complexity and urban mobility research, no prior study has examined the three-way interaction among economic complexity, mobility

efficiency, and income inequality at the regional level using granular digital mobility data. This gap is consequential because, as our results demonstrate, the three dimensions interact in non-obvious ways: the most complex regions achieve the highest mobility efficiency *and* the deepest inequality, a configuration that challenges assumptions embedded in both the complexity literature and transportation policy.

1.5 Research Gaps, Aims, and Scope

The foregoing review identifies three critical gaps in the existing literature that motivate this study.

The first gap is the absence of a unified empirical framework that jointly examines economic complexity, mobility patterns, and income distribution at the sub-national level. Prior studies have examined pairs of these dimensions—complexity and inequality [Hartmann et al., 2017, Sbardella et al., 2017], mobility and socioeconomic status [Xu et al., 2018, Moro et al., 2021], or mobility and economic structure [Cervero and Duncan, 2006, Ewing and Cervero, 2010]—but none has analyzed the three-way configuration using the kind of large-scale digital traces that characterize the emerging data-driven social sciences. By integrating employment-based ECI with Replica origin–destination mobility data and American Community Survey income distributions across 121 CBSAs, this study provides the first systematic characterization of how economic sophistication, commuting efficiency, and income equity co-vary across functional economic regions.

The second gap concerns the adaptation of ECI methodology to employment data at the CBSA level. While Fritz and Manduca [2021] pioneered employment-based ECI for U.S. metropolitan statistical areas, their analysis focused on temporal trends in income rather than on the cross-sectional relationship with mobility and inequality. Our approach extends this methodological innovation by computing ECI across a strategically selected set of CBSAs spanning the full range of U.S. economic structures—from Silicon Valley knowledge hubs to Mississippi Delta resource economies—and by integrating the resulting complexity scores with granular mobility metrics derived from digital traces of commuting behavior that have not been examined in prior complexity research.

The third gap relates to the policy implications of the sub-national complexity–inequality reversal. If the well-documented international finding that complexity reduces inequality does not hold within countries, then development strategies premised on industrial diversification may produce unintended distributional consequences at the regional level. Our identification of four structurally distinct regional typologies, each characterized by a qualitatively different configuration of complexity, mobility, and inequality, provides an empirical foundation for the differentiated policy frameworks that this paradox demands. The concept of "efficient inequality" that emerges from our analysis—the co-occurrence of mobility efficiency and income dispersion in knowledge-intensive metros—offers a new analytical lens for understanding the distributional dynamics of urban development.

Accordingly, this study pursues three interrelated aims. The first aim is to adapt the Economic Complexity Index from international trade data to sub-national employment distributions, computing employment-based ECI for 121 CBSAs across five U.S. states and validating its capacity to discriminate among qualitatively different regional economic structures. The second aim is to integrate this complexity measure with high-resolution mobility

metrics (VMT per capita, radius of gyration) and income distribution indicators (Gini coefficient, median household income) within a unified dimensionality-reduction and clustering framework, revealing multi-dimensional trade-off configurations that single-indicator analyses cannot detect. The third aim is to characterize and interpret the resulting regional typologies, with particular attention to the paradoxical co-occurrence of high complexity, high inequality, and high mobility efficiency in knowledge-intensive metros, and to draw out the implications for data-driven regional policy.

The scope of the analysis encompasses 121 CBSAs across California, New York, New Mexico, Louisiana, and Mississippi—five states selected to span the U.S. income distribution while capturing the full diversity of American economic geography. The remainder of this paper is organized as follows. Section 2 details the data sources, the adaptation of ECI to employment data, and the statistical methods including PCA, K-means clustering, and bootstrap validation. Section 3 presents results including the principal component structure, four regional typologies, and the complexity–inequality–mobility nexus. Section 4 interprets the findings within the theoretical framework established above and discusses policy implications. Section 5 concludes with directions for future research.

2 Data and Methods

2.1 Analytical Framework

The analysis follows a five-stage pipeline designed to move from raw digital data to interpretable regional typologies (Figure 1). Three conceptually distinct data streams—employment structure, mobility patterns, and income distributions—are first processed independently to generate CBSA-level metrics (Stages 1–3). These metrics are then integrated through dimensionality reduction and unsupervised classification (Stage 4), and the resulting typologies are subjected to validation and interpretation (Stage 5). The framework is modular and scalable: any component can be updated with alternative data sources or extended to additional geographic units without altering the overall pipeline architecture.

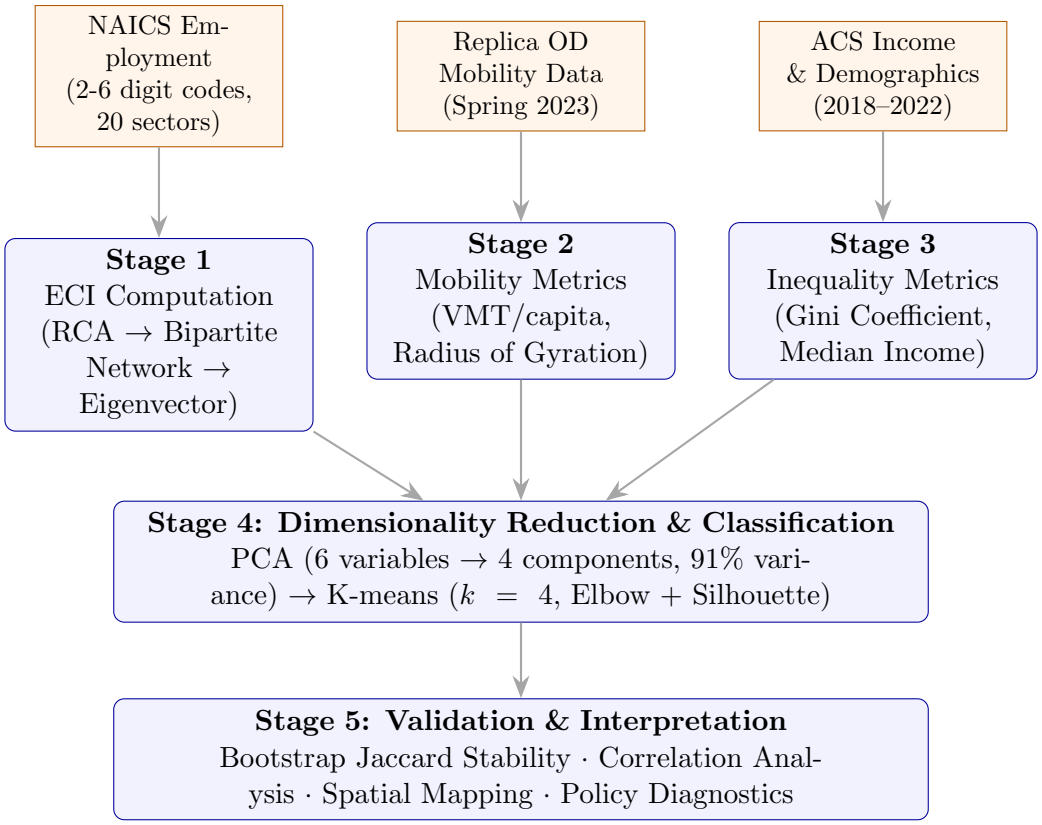


Figure 1: Analytical framework. Three data streams—employment structure, digital mobility traces, and income distributions—are processed through independent metric computation (Stages 1–3), integrated via PCA and clustering (Stage 4), and validated through bootstrap resampling and spatial analysis (Stage 5).

2.2 Study Area and Spatial Unit

This study analyzes 121 CBSAs distributed across five U.S. states selected to span the national income distribution: California (34 CBSAs), New York (26), Mississippi (21), New Mexico (19), and Louisiana (18), with 3 CBSAs crossing state boundaries (Figure 2). We adopt the 2020 CBSA delineations issued by the Office of Management and Budget [Office of Management and Budget, 2020], which define functional economic regions based on county-level commuting thresholds. CBSAs are preferable to counties for this analysis because they avoid the artificial fragmentation of labor markets that arises when administrative boundaries bisect commuting sheds, while maintaining finer resolution than consolidated metropolitan statistical areas [Fowler et al., 2016, 2020]. The five-state selection captures the diversity of American economic geography: technology hubs (Silicon Valley, San Francisco), global financial centers (New York City), petrochemical extraction economies (Lake Charles, Houma-Thibodaux), agricultural regions (California's Central Valley, the Mississippi Delta), military installations (White Sands, Fort Polk), and post-industrial communities (upstate New York, rural Mississippi) [U.S. Census Bureau, 2023].

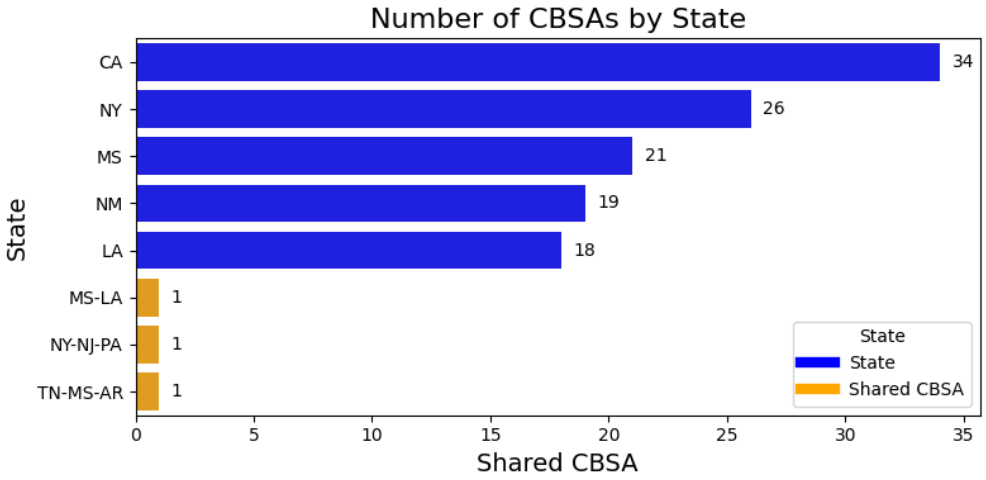


Figure 2: Distribution of 121 CBSAs across five states. Shared CBSAs crossing state boundaries are counted separately.

2.3 Mobility Data

Primary mobility data are drawn from Replica's Spring 2023 Origin–Destination dataset [Replica Inc., 2023], which synthesizes GPS-derived mobile device trajectories, transportation network models, and household travel surveys into a population-level activity-based travel model. This dataset exemplifies the kind of large-scale digital trace data that has transformed the empirical study of human mobility [González et al., 2008, Blondel et al., 2015, Barbosa et al., 2018]. We extract work-related trips for a representative Thursday—capturing typical weekday commuting while avoiding the idiosyncrasies of Mondays and Fridays—for both in-person and remote workers, and compute two CBSA-level metrics: Vehicle Miles Traveled (VMT) per capita, defined as total residential VMT divided by CBSA population, which captures automobile dependence; and the mean radius of gyration ($\bar{r}_g$), which measures the average spatial extent of individual commuting patterns (detailed in Section 2.6).

A note on data quality is essential for contextualizing our results. Replica's traffic volume estimates have been independently validated against permanent count stations by the Texas A&M Transportation Institute, which reported a mean absolute percentage error of 12.1% and $R^2 = 0.97$ for average annual daily traffic [Texas A&M Transportation Institute, 2024]. However, this validation covers aggregate traffic volumes rather than individual-level trip attributes such as trip purpose, mode, or demographic characteristics, all of which are modeled rather than directly observed. Furthermore, cell phone–derived mobility data is known to exhibit systematic demographic biases. Wesolowski et al. [2013] documented that mobile phone ownership—and hence representation in GPS-derived datasets—is disproportionately concentrated among male, educated, and higher-income individuals. Coston et al. [2021] found systematic undersampling in areas with higher concentrations of lower-income, non-white, and elderly populations. These biases may attenuate observed differences in mobility metrics across CBSAs with different demographic compositions, suggesting that our estimates of the mobility–complexity relationship are, if anything, conservative. We flag this limitation explicitly and interpret mobility patterns with appropriate caution throughout

the analysis.

2.4 Socioeconomic and Employment Data

Income data are derived from Replica's population synthesis model, which is calibrated against American Community Survey (ACS) 5-year estimates for 2018–2022 [U.S. Census Bureau, 2023]. This calibration grounds the individual-level income distributions in nationally representative survey data, from which we compute the median household income and Gini coefficient (Section 2.7) for each CBSA. Population density, measured as persons per square mile, is derived from Census boundary files and ACS population estimates.

Industry-level employment data use 2–6 digit NAICS codes from the U.S. Census Bureau, aggregated into 20 major sectors that serve as the "products" in our ECI computation. For interpretive purposes, these 20 sectors are further grouped into two macro-categories: Group 1 (Goods and Resource-Based Industries), comprising agriculture, mining, construction, manufacturing, wholesale trade, retail trade, transportation and warehousing, and utilities, which together account for approximately 40% of total employment across the sample; and Group 2 (Service and Knowledge-Based Industries), comprising information, finance and insurance, real estate, professional and technical services, management, education, healthcare, administrative services, accommodation and food services, other services, and public administration, accounting for approximately 60%. Figure 3 displays the full employment composition across all 121 CBSAs, revealing pronounced heterogeneity in industrial structure that ranges from CBSAs dominated by mining and agriculture to those in which professional services and information technology constitute the plurality of employment.

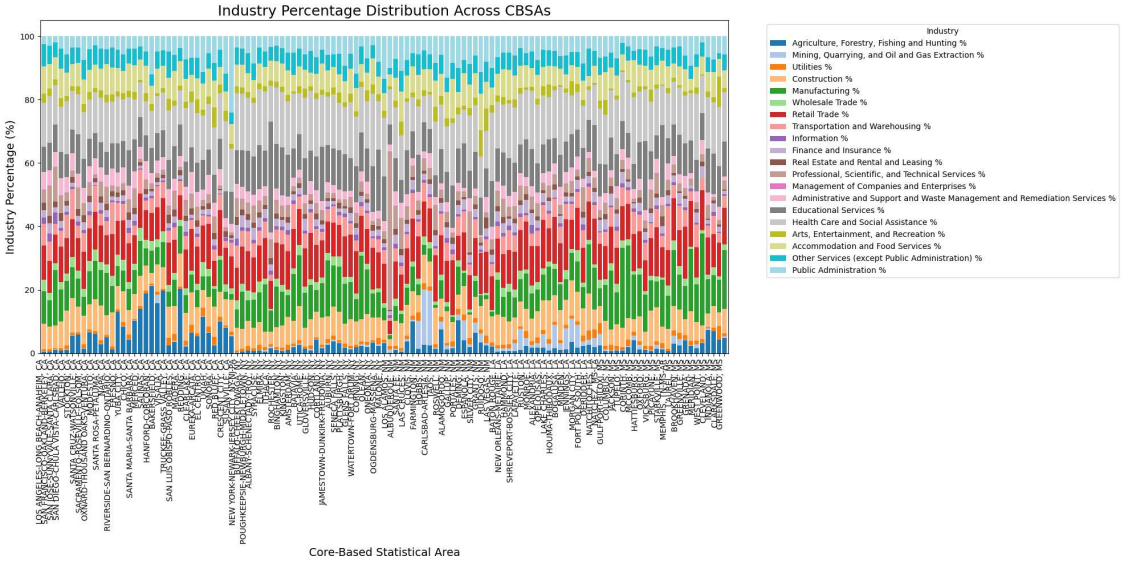


Figure 3: Stacked employment shares across 121 CBSAs, sorted by Group 2 (service and knowledge) share. The pronounced heterogeneity in industrial composition motivates the use of economic complexity analysis to capture structural differences that simple sector shares cannot represent.

2.5 Economic Complexity Index Computation

We adapt the Economic Complexity Index methodology [Hidalgo and Hausmann, 2009] from international trade to regional employment, following the sub-national precedent established by Fritz and Manduca [2021] and the methodological guidance provided by Balland et al. [2022]. The computation proceeds through six steps that construct a bipartite CBSA–industry network and extract eigenvector centrality as a measure of economic sophistication.

The first step defines the employment matrix M_{ci}, where each entry represents total employment in industry i within CBSA c. The second step computes revealed comparative advantage (RCA), which measures the degree to which a CBSA is more specialized in a given industry than the five-state average:

$$\mathrm{RCA}_{ci} = \frac{M_{ci} / \sum_i M_{ci}}{\sum_c M_{ci} / \sum_{c,i} M_{ci}} \tag{1}$$

An RCA exceeding unity indicates that CBSA c has a disproportionate share of its employment in industry i relative to the reference population. In the third step, the continuous RCA matrix is binarized: $\tilde{M}_{ci} = 1$ if $\mathrm{RCA}_{ci} > 1$ and $\tilde{M}_{ci} = 0$ otherwise. This binarization, while standard in the complexity literature, discards information about the intensity of specialization; we return to this limitation in Section 4.5, noting that continuous alternatives proposed by Davies and Maré [2021] represent a promising direction for future work.

The fourth step computes two summary statistics from the binary matrix: diversity $k_{c,0} = \sum_i \tilde{M}_{ci}$, which counts the number of industries in which CBSA c is specialized, and ubiquity $k_{i,0} = \sum_c \tilde{M}_{ci}$, which counts the number of CBSAs specialized in industry i. Diversity captures the breadth of a region's productive capabilities, while ubiquity captures exclusivity—an industry present in only a few CBSAs signals higher complexity than one found everywhere. The fifth step applies the Method of Reflections [Hidalgo and Hausmann, 2009], iterating between CBSA and industry complexity scores until convergence:

$$k_{c,n} = \frac{1}{k_{c,0}} \sum_i \tilde{M}_{ci}\, k_{i,n-1}, \qquad k_{i,n} = \frac{1}{k_{i,0}} \sum_c \tilde{M}_{ci}\, k_{c,n-1} \tag{2}$$

This iteration converges to the eigenvector associated with the second-largest eigenvalue of the bipartite adjacency matrix, capturing the joint structure of diversity and ubiquity across the network. In the sixth and final step, the resulting eigenvector entries are standardized to yield the ECI for each CBSA:

$$\mathrm{ECI}_c = \frac{\vec{K}_c - \langle \vec{K} \rangle}{\sigma(\vec{K})} \tag{3}$$

where $\langle \vec{K} \rangle$ and $\sigma(\vec{K})$ denote the sample mean and standard deviation, respectively. The resulting index is centered at zero, with positive values indicating above-average complexity and negative values indicating below-average complexity.

Figure 4 displays the ECI distribution across all 121 CBSAs. San Jose–Sunnyvale–Santa Clara exhibits the highest complexity score, consistent with its concentration of information technology, professional services, and research and development activities. At the other

extreme, rural CBSAs in Mississippi and Louisiana dominated by agriculture, extraction, and low-skill services exhibit the most negative ECI values. It is worth emphasizing that this employment-based ECI captures non-tradable services—healthcare, education, public administration, accommodation—that constitute the majority of U.S. employment and are invisible to trade-based ECI, providing a more comprehensive picture of regional economic structure than either trade or patent data in a service-dominated economy [Brökel, 2019, Fritz and Manduca, 2021].

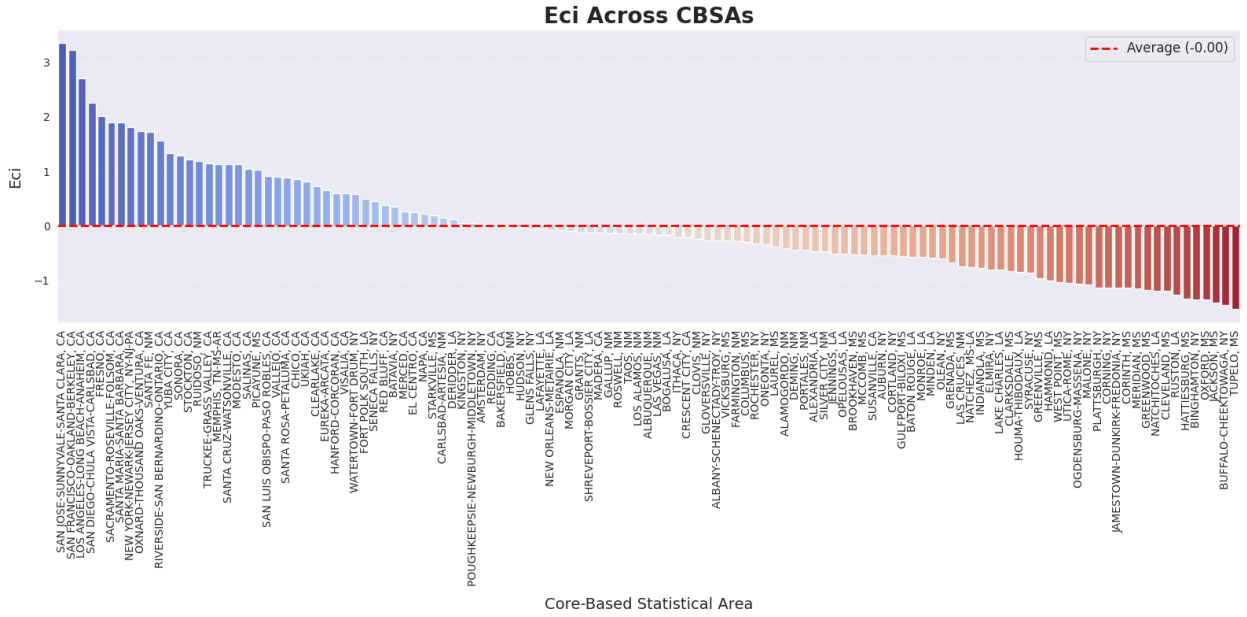


Figure 4: Employment-based Economic Complexity Index across 121 CBSAs, sorted in descending order. The dashed line indicates the sample mean (≈ 0). Knowledge-intensive California and New York metros dominate the upper tail, while resource-dependent CBSAs in Mississippi and Louisiana occupy the lower tail.

2.6 Radius of Gyration

The radius of gyration quantifies the characteristic spatial scale of an individual's mobility pattern [González et al., 2008]:

$$r_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\vec{r}_i - \vec{r}_{\text{cm}})^2} \tag{4}$$

where $\vec{r}_i$ denotes the geographic coordinates of the i-th work trip destination, $\vec{r}_{\text{cm}}$ is the center of mass of all visited locations, and N is the total number of recorded work trips. We restrict the computation to work-related trips to isolate commuting behavior from recreational or personal travel, and aggregate individual r_g values to obtain the CBSA-level mean ($\bar{r}_g$). Higher values of $\bar{r}_g$ indicate greater average spatial separation between homes and workplaces, reflecting more extensive commuting sheds and, potentially, more severe spatial mismatch

[Xu et al., 2018, Pappalardo et al., 2015]. Figure 5 presents the distribution of $\bar{r}_g$ across the 121 CBSAs, revealing variation from compact rural New Mexico CBSAs ($\bar{r}_g < 2$ km) to polycentric metros like New York-Newark ($\bar{r}_g > 20$ km).

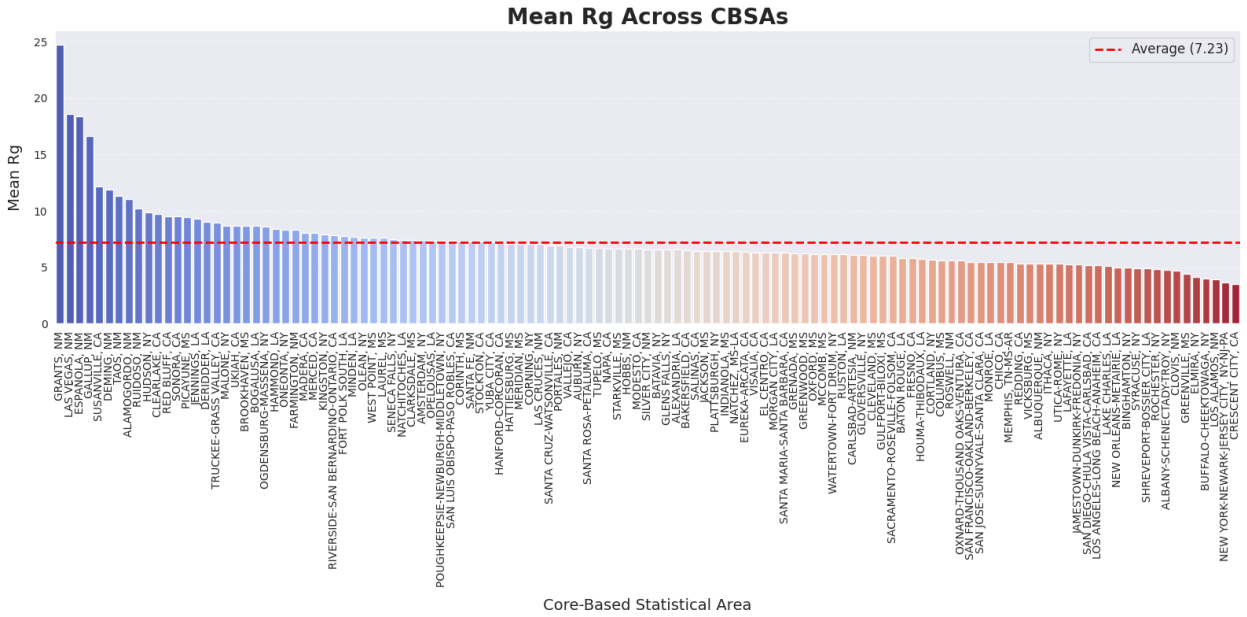


Figure 5: Mean radius of gyration across 121 CBSAs. The dashed line indicates the sample average (7.23 km). CBSAs with compact, monocentric structures exhibit the lowest values, while those with polycentric or spatially dispersed employment exhibit the highest.

2.7 Gini Coefficient

Income inequality within each CBSA is measured using the Gini coefficient [Gini, 1921]:

$$G = 1 - \sum_{i=1}^{n} (X_{i+1} + X_i)(Y_i - Y_{i-1}) \tag{5}$$

where X_i represents the cumulative population share and Y_i the cumulative income share at quantile i. We compute the Gini from the individual-level income distributions provided by Replica's population synthesis model, calibrated against ACS estimates [U.S. Census Bureau, 2023]. Figure 6 maps the spatial distribution of Gini coefficients across the five states, revealing a clear geographic gradient: the highest values (Gini > 0.49) cluster in major urban centers—New York City, San Francisco, Los Angeles—while more moderate values (≈ 0.42–0.46) characterize smaller cities and rural areas. This gradient is consistent with the urban inequality premium documented by Baum-Snow et al. [2018] and Florida [2017].

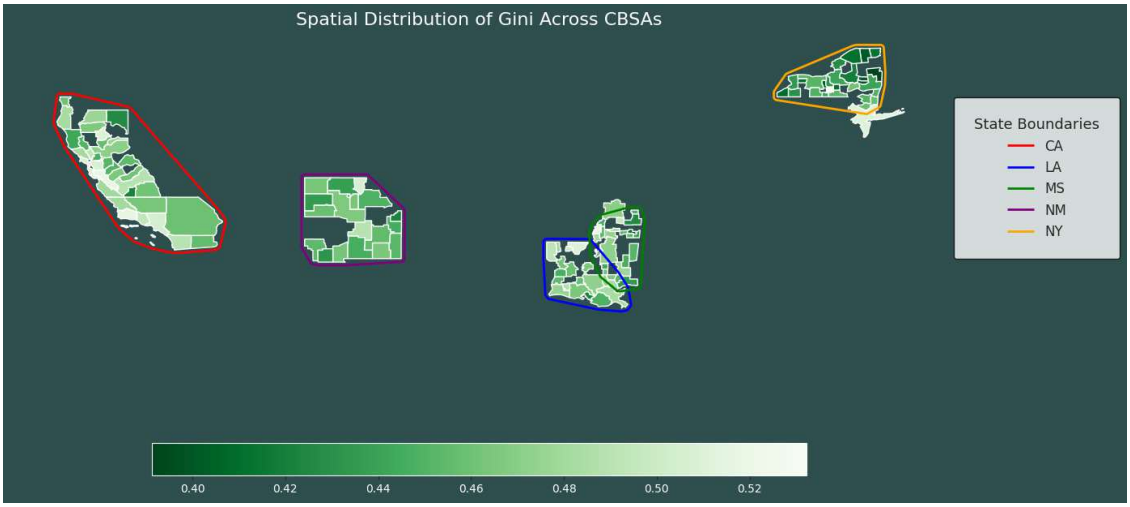


Figure 6: Spatial distribution of Gini coefficients across the five states. Urban centers in California and New York exhibit systematically higher inequality than rural and small-city CBSAs.

2.8 Dimensionality Reduction and Clustering

To identify latent structure in the joint distribution of complexity, mobility, and inequality, we apply a two-step statistical procedure: Principal Component Analysis for dimensionality reduction followed by K-means clustering for typology identification.

PCA is applied to six standardized (zero-mean, unit-variance) variables: population density, median household income, mean r_g, residential VMT per capita, Gini coefficient, and ECI. Standardization ensures that no single variable dominates the principal component structure by virtue of measurement scale [Kaiser, 1960]. Components are retained using the Kaiser criterion (eigenvalue > 1), confirmed by scree plot inspection.

The principal component scores then serve as inputs to K-means clustering. The optimal number of clusters is selected through two complementary criteria: the elbow method, which identifies the inflection point in within-cluster sum of squares (WCSS), and silhouette analysis, which maximizes mean silhouette width as a measure of cluster cohesion relative to separation [Rousseeuw, 1987]. Both methods converge on $k = 4$ as optimal (Figure 7).

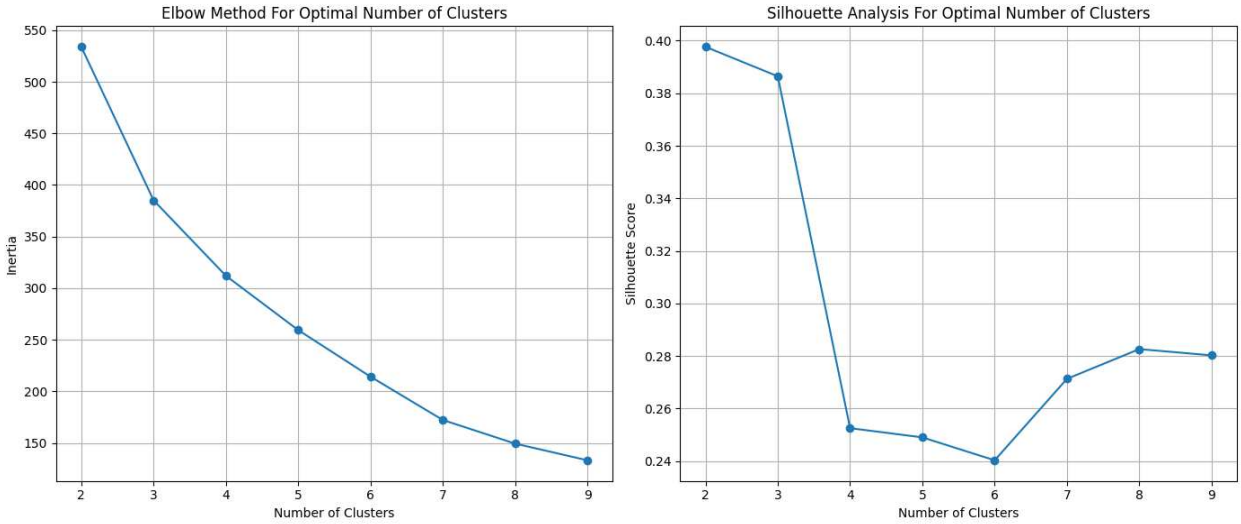


Figure 7: Cluster number selection. Left: Elbow plot with WCSS inflection at $k = 4$. Right: Mean silhouette width peaking at $k = 4$.

Given the moderate sample size ($N = 121$) and unequal cluster sizes—particularly the small Knowledge Hub cluster with $n = 9$—we implement three validation procedures. First, following Hennig [2007], we conduct bootstrap resampling with $B = 1000$ iterations, re-running K-means on each bootstrap sample and computing the Jaccard similarity coefficient between bootstrap and original cluster assignments; clusters with mean Jaccard ≥ 0.75 are considered stable. Second, we compare K-means results with Ward-linkage hierarchical agglomerative clustering to assess sensitivity to the clustering algorithm. Third, we fit a Gaussian Mixture Model (GMM) with $k = 4$ components, which provides probabilistic cluster assignments and allows identification of boundary observations whose classification is uncertain—particularly important for the smallest cluster.

3 Results

3.1 Principal Component Structure

The first four principal components capture 91% of total variance in the six-variable dataset (Table 1), reducing the dimensionality from six to four while preserving the vast majority of the information content. Each component captures a substantively interpretable dimension of regional variation.

Table 1: PCA loadings for the first four principal components. Bold values indicate dominant loadings ($|w| > 0.40$). Variance explained by each component is shown in parentheses.

Variable	**PC1 (41%)**	**PC2 (24%)**	**PC3 (16%)**	**PC4 (10%)**
Population Density	**0.47**	0.29	0.09	**–0.49**
Median Income	0.37	0.06	**0.70**	**0.60**
Mean r_g	**–0.44**	**0.53**	0.12	0.10
Residential VMT/capita	**–0.47**	**0.47**	0.12	0.05
Gini Coefficient	0.31	0.32	**–0.68**	**0.56**
ECI	0.36	**0.55**	0.06	–0.27

The first principal component (PC1), which accounts for 41% of total variance, represents an Urban Complexity–Efficiency axis. It loads positively on population density (0.47), ECI (0.36), and the Gini coefficient (0.31), and negatively on VMT per capita (–0.47) and mean r_g (–0.44). This loading pattern encapsulates the core empirical finding of this study: density, economic sophistication, inequality, and mobility efficiency co-vary along a single dominant dimension. CBSAs scoring high on PC1 are dense, complex, unequal, and mobility-efficient; those scoring low are sparse, simple, more equal, and automobile-dependent. The fact that this configuration accounts for the largest share of variance—more than the next two components combined—underscores its centrality to the structure of regional variation in the United States.

The second principal component (PC2, 24% of variance) captures Complexity with Spatial Dispersion. High positive loadings on ECI (0.55), mean r_g (0.53), and VMT per capita (0.47) identify regions that have achieved economic sophistication despite spatial dispersion. These are typically polycentric metropolitan areas where knowledge-intensive employment is distributed across suburban nodes, generating extensive commuting even among highly productive workers. This component reveals that the complexity–efficiency link captured by PC1, while dominant, is not universal: some complex economies remain spatially inefficient.

The third component (PC3, 16%) captures an Income–Equality Trade-off, with median income (0.70) and Gini coefficient (–0.68) loading in opposition. This dimension identifies smaller prosperous CBSAs—typically college towns or specialized professional centers—where the absence of extreme income tails produces low Gini coefficients despite above-average median incomes. The near-zero ECI loading (0.06) indicates that this dimension operates independently of economic complexity. The fourth component (PC4, 10%) represents Dense Deprivation: the combination of negative density loading (–0.49) with positive income (0.60) and Gini (0.56) loadings identifies regions where moderate density coexists with both relatively high average incomes and pronounced inequality, potentially capturing gentrifying or dual-economy areas.

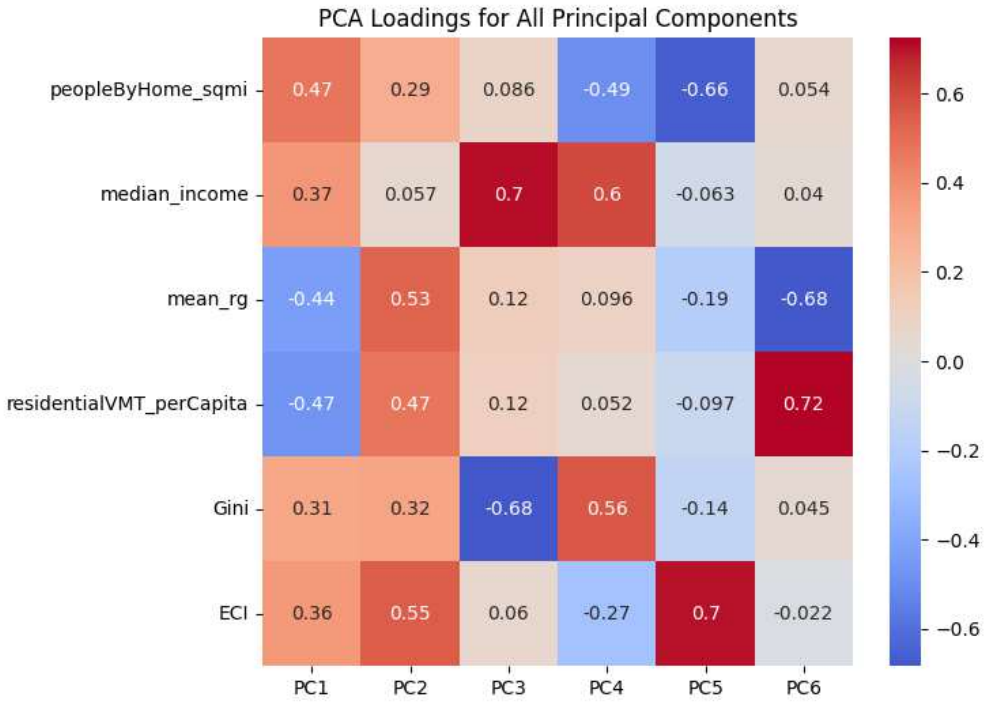


Figure 8: Heatmap of PCA loadings across all six components, visualizing the multi-dimensional structure of regional variation. The first four components capture 91% of total variance.

3.2 Regional Typologies

K-means clustering applied to the principal component scores identifies four structurally distinct regional typologies (Figure 9, Table 2).

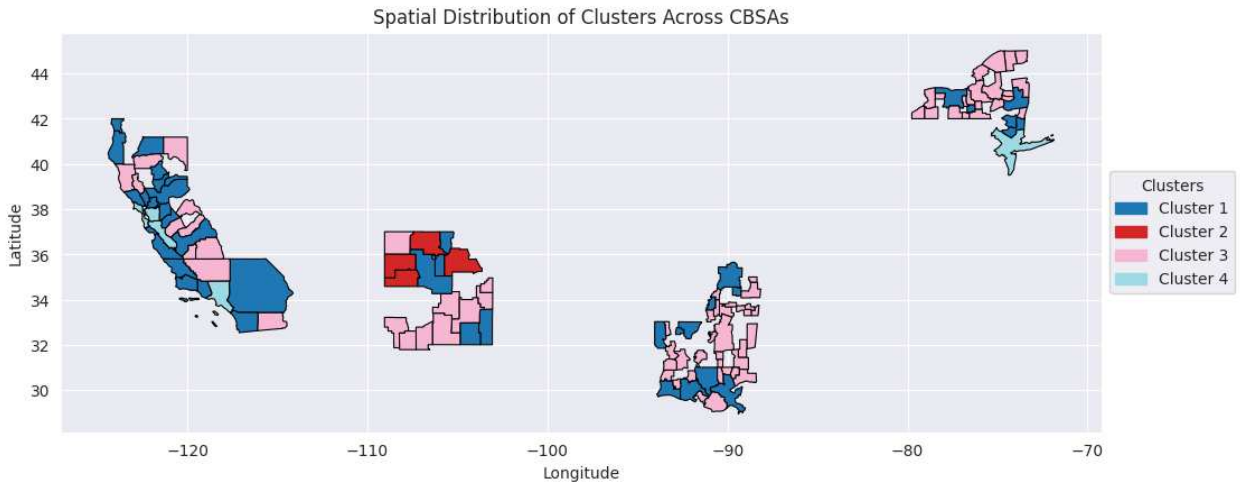


Figure 9: Spatial distribution of the four cluster types across the five states. Cluster 4 (Knowledge Hubs) concentrates in coastal California and the New York metropolitan area; Cluster 2 (Small Industrial Towns) dominates Louisiana and Mississippi; Cluster 3 (Rural Resource) spans all five states; Cluster 1 (High-Density Mixed) occupies mid-tier California and New York metros.

Table 2: Summary statistics by cluster. Medians are reported for continuous variables. Jaccard coefficients computed from $B = 1000$ bootstrap resamples; values ≥ 0.75 indicate stable clusters [Hennig, 2007].

Cluster	n	**Med. Inc.**	**Gini**	**ECI**	**VMT/cap**	$\bar{r}_g$ **(km)**	**Jaccard**
1: High-Density Mixed	31	\$47k	0.48	0.80	14.2	6.8	0.82
2: Small Industrial	42	\$37k	0.45	–0.01	30.5	19.2	0.89
3: Rural Resource	39	\$40k	0.45	–0.50	18.1	7.9	0.84
4: Knowledge Hubs	9	\$64k	0.51	2.40	10.8	5.2	0.78

The 31 CBSAs in Cluster 1 (High-Density Mixed Economy Regions)—including Sacramento, Fresno, Buffalo, and Albany—represent mid-tier urban economies combining moderate economic complexity (ECI = 0.80) with relatively efficient mobility systems (VMT per capita of 14.2, mean r_g of 6.8 km). Their employment structures are moderately diversified, with Group 1 industries accounting for 38% and Group 2 for 62%, and their population densities (median 89 persons per square mile) are sufficient to support some public transit infrastructure, contributing to below-average VMT. However, inequality in these regions is elevated (Gini = 0.48), approaching that of the Knowledge Hub cluster. This pattern is consistent with the "secondary city" phenomenon described by Florida [2017], in which mid-tier metros absorb some of the agglomeration benefits—and inequality penalties—that spill over from primary knowledge hubs through migration, branch-office expansion, and housing market linkages.

Cluster 2 (Small Industrial Towns) is the largest group ($n = 42$) and encompasses low-density, economically undiversified CBSAs such as Lake Charles (Louisiana), Farmington

(New Mexico), and Pine Bluff (Arkansas). With a near-zero ECI (–0.01) and the highest mobility burden in the sample—VMT per capita of 30.5 and mean r_g of 19.2 km, nearly four times the Knowledge Hub average—these regions embody the challenges of spatial mismatch in dispersed, automobile-dependent economies [Kain, 1968, Gobillon et al., 2007]. The even split between Group 1 (50%) and Group 2 (50%) industries masks concentration within each group, with individual CBSAs often dominated by a single employer such as a refinery, military installation, or regional hospital. The moderate Gini coefficient (0.45) reflects compressed wage distributions characteristic of economies with few high-paying professional positions rather than equitable prosperity; median household income ($37,000) is the lowest in the sample, consistent with the limited productive capabilities that Hartmann et al. [2017] associate with low-complexity economies.

The 39 CBSAs in Cluster 3 (Rural Resource-Dependent Regions)—including Gulfport-Biloxi (Mississippi), Las Cruces (New Mexico), and Monroe (Louisiana)—exhibit the lowest average complexity in the sample (ECI = –0.50), reflecting specialization in resource extraction, agriculture, military installations, and low-skill services. Unlike Cluster 2, these CBSAs tend to have relatively compact settlement patterns that produce moderate mobility metrics (VMT per capita of 18.1, mean r_g of 7.9 km), partially offsetting the absence of formal transit infrastructure through short distances between residential areas and employment sites. Median income ($40,000) and inequality (Gini = 0.45) are both moderate, consistent with Sbardella et al. [2017]'s finding that low-complexity regions exhibit compressed income distributions producing moderate Gini values not because of equitable prosperity but because of the absence of a high-earning professional class.

The nine CBSAs in Cluster 4 (Prosperous Knowledge Hubs)—San Francisco-Oakland, San Jose-Sunnyvale-Santa Clara, New York-Newark, Los Angeles-Long Beach, and their metropolitan peers—exhibit the most extreme configuration across every dimension. These regions combine the highest population density (median 420 persons per square mile), the highest median household income ($64,000), and the highest ECI (2.40) with the *highest* income inequality (Gini = 0.51) and the *most efficient* mobility systems (VMT per capita of 10.8, mean r_g of 5.2 km). This three-way configuration—peak complexity, peak inequality, minimum mobility burden—constitutes the empirical signature of the complexity–inequality paradox. The employment structure is heavily tilted toward Group 2 industries (65%), with information, professional services, finance, and management substantially over-represented, generating the bimodal income distribution characteristic of knowledge hubs: a large concentration of high-earning professionals alongside a substantial low-wage service workforce, with relatively few middle-skill, middle-wage positions [Autor and Dorn, 2013, Goos et al., 2014].

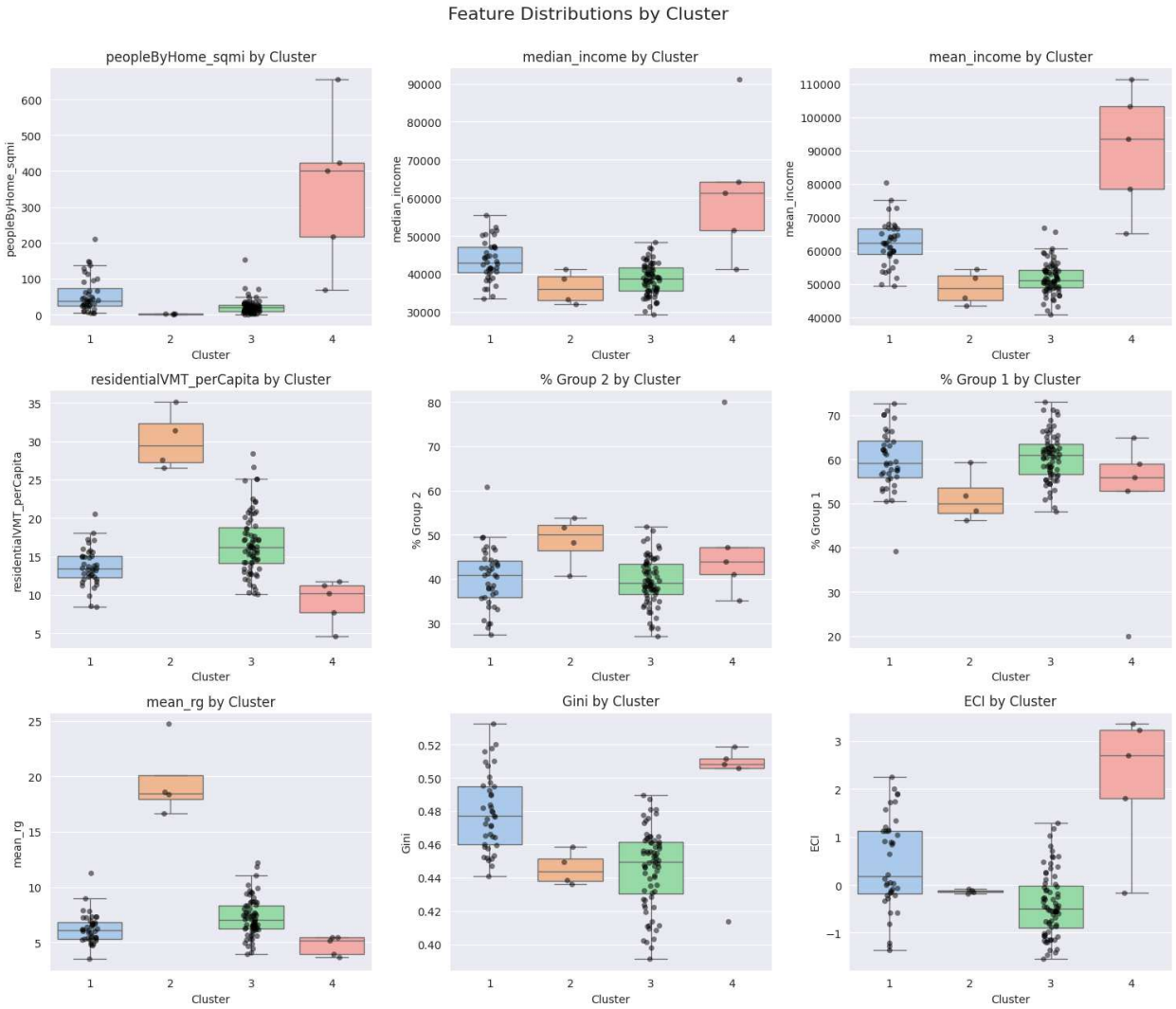


Figure 10: Distribution of key variables across the four clusters. The Knowledge Hub cluster (Cluster 4) occupies the extreme position on every dimension: highest density, income, ECI, and Gini; lowest VMT per capita and mean r_g.

All four clusters exceed the 0.75 Jaccard stability threshold [Hennig, 2007]: Cluster 2 is the most stable (0.89), consistent with its large size and distinctive profile; Clusters 3 and 1 achieve 0.84 and 0.82, respectively; and Cluster 4, despite its small size, achieves 0.78, above the stability threshold but below the "highly stable" benchmark of 0.85. Ward-linkage hierarchical clustering recovers the same four-group structure with 87% assignment agreement, with discrepancies concentrated at the Cluster 1/3 boundary rather than involving the distinctive Cluster 4. A Gaussian Mixture Model with four components confirms that all Cluster 4 observations have posterior membership probabilities exceeding 0.90, indicating a genuinely distinct mode in the data distribution. These validation results support the substantive interpretability of the typology while cautioning that Cluster 4 results characterize a regional *type* rather than providing high-precision within-cluster estimates.

3.3 The Complexity–Inequality Paradox

The central empirical finding is a statistically significant positive correlation between economic complexity and income inequality across the 121 CBSAs (Pearson $r = 0.42$, $p < 0.001$; Spearman $\rho = 0.39$, $p < 0.001$). This directly reverses the well-established negative cross-country relationship documented by Hartmann et al. [2017], who found that nations with higher ECI exhibit systematically lower Gini coefficients. The sign reversal is robust to the choice of correlation measure, survives the exclusion of the nine Knowledge Hub CBSAs ($r = 0.29$, $p < 0.01$ for the remaining 112 CBSAs), and is confirmed by the PCA structure in which ECI and Gini load with the same sign on the dominant first component.

This reversal is consistent with the scale-dependent Simpson's Paradox theorized by Hartmann and Pinheiro [2022]. At the national level, economic complexity co-evolves with institutional factors—educational systems, labor market regulations, redistributive taxation, and social safety nets—that collectively compress income distributions [Hartmann et al., 2017, Lee and Qu, 2021]. At the sub-national level, within the shared institutional framework of a single country, these mechanisms are held roughly constant. The dominant forces instead become skill-biased agglomeration and spatial sorting: the geographic concentration of high-human-capital workers and knowledge-intensive firms in dense urban areas generates productivity premiums that accrue disproportionately to the already-advantaged [Baum-Snow et al., 2018, Moretti, 2012]. This finding echoes Kemeny and Osman [2015], who showed that the wider impacts of high-technology employment in U.S. cities include not only aggregate productivity gains but also widening wage gaps between skilled and unskilled workers, and Lee and Rodríguez-Pose [2016], who documented that innovation-driven knowledge spillovers are spatially bounded and disproportionately benefit workers already positioned to capture them.

The paradox is reinforced by occupational polarization in complex economies: simultaneous growth at the high-wage (technology, finance, professional services) and low-wage (personal services, food preparation, building maintenance) ends of the distribution, with hollowing of middle-skill occupations [Autor and Dorn, 2013, Goos et al., 2014]. This bimodal structure mechanically elevates Gini coefficients even when median incomes are high. Housing cost capitalization provides a further channel, as Hsieh and Moretti [2019] and Ganong and Shoag [2017] have documented that productivity gains in complex metros are substantially captured by rising land values, compressing real incomes for non-owners while enriching the property-holding class.

3.4 Mobility Efficiency and Economic Structure

Economic complexity exhibits a strong inverse association with mobility burden: the Pearson correlation between ECI and VMT per capita is $r = -0.58$ ($p < 0.001$), and between ECI and mean r_g it is $r = -0.51$ ($p < 0.001$). These correlations indicate that more complex economies are associated with substantially lower per capita driving and more compact commuting patterns, a relationship that holds across all four clusters and is not driven solely by the Knowledge Hub CBSAs.

The structural foundations of this efficiency advantage lie in the spatial organization of complex economies. Knowledge-intensive industries cluster in dense, mixed-use urban envi-

ronments where the co-location of diverse employers creates a favorable jobs–housing balance [Cervero and Duncan, 2006]. The population densities characteristic of Cluster 4 far exceed the thresholds identified by Guerra et al. [2012] as necessary for frequent, high-ridership transit service. Ewing and Cervero [2010], in a meta-analysis of built environment studies, found that the mixed-use, walkable development patterns typical of knowledge hubs are associated with 10–30% VMT reductions relative to single-use suburban environments. Furthermore, the industries defining complex economies have the highest remote work potential in the U.S. economy [Dingel and Neiman, 2020], enabling telecommunications-based substitution for physical commuting.

The contrast with Cluster 2 (Small Industrial Towns) is stark. These CBSAs, at 30.5 VMT per capita—nearly three times the Cluster 4 average—are characterized by dispersed settlement, concentrated employment at specific extraction or manufacturing sites, and near-total absence of transit alternatives. The mean r_g of 19.2 km indicates that the average worker commutes across a spatial range roughly four times that of a Knowledge Hub worker, imposing substantial time, money, and emissions costs. This mobility penalty falls disproportionately on lower-income workers, who have the fewest resources to overcome spatial barriers [Blumenberg and Manville, 2004, Hu, 2017], compounding the economic disadvantages of low-complexity employment structures.

3.5 Pairwise Correlation Structure

Table 3 presents the full pairwise correlation matrix, providing a comprehensive view of the associations underpinning the PCA and cluster results.

Table 3: Pairwise Pearson correlation matrix ($N = 121$ CBSAs). All values with $|r| > 0.18$ are significant at $p < 0.05$.

	Density	**Med. Inc.**	$\bar{r}_g$	**VMT/cap**	**Gini**	**ECI**
Density	1.00					
Med. Income	0.48	1.00				
Mean r_g	–0.31	–0.15	1.00			
VMT/capita	–0.44	–0.21	0.72	1.00		
Gini	0.39	0.22	–0.08	–0.16	1.00	
ECI	0.45	0.52	–0.51	–0.58	0.42	1.00

ECI is the most broadly connected variable in the matrix, exhibiting strong correlations with density ($r = 0.45$), income ($r = 0.52$), and Gini ($r = 0.42$) on the positive side, and with VMT per capita ($r = -0.58$) and r_g ($r = -0.51$) on the negative side. This confirms that economic complexity is not a proxy for any single dimension of regional variation but is simultaneously associated with density, prosperity, mobility efficiency, and inequality. The high VMT–r_g correlation ($r = 0.72$) confirms that these metrics capture overlapping aspects of mobility burden. Critically, the Gini–r_g correlation is near zero ($r = -0.08$), indicating that income inequality and commuting extent are largely independent dimensions—a finding that underscores the value of multi-dimensional analysis, since a study examining only

inequality or only mobility would miss the fact that the most unequal regions are those with the *shortest*, not the longest, commutes.

4 Discussion

4.1 The Scale Dependence of the Complexity–Inequality Relationship

The positive correlation between economic complexity and income inequality documented in this study stands in direct contrast to the negative cross-country correlation reported by Hartmann et al. [2017] and confirmed in subsequent international analyses [Lee and Qu, 2021, Hartmann and Pinheiro, 2022]. This reversal adds to a growing body of sub-national evidence—including the U.S. metro-level findings of Fritz and Manduca [2021], the county-level analysis of Sbardella et al. [2017], and the European regional study of Pinheiro et al. [2022]—establishing that the complexity–inequality relationship is fundamentally scale-dependent.

The theoretical explanation centers on the different mechanisms dominating at each scale. Across countries, economic complexity is correlated with institutional development: nations that have accumulated capabilities to produce complex goods typically also develop the educational systems, labor protections, and redistributive programs that compress income distributions [Hartmann et al., 2017, Hartmann and Pinheiro, 2022]. Within a single country, institutions are approximately constant across regions, and the dominant mechanism becomes spatial sorting: the geographic concentration of high-skill workers and knowledge-intensive firms in dense urban agglomerations [Balland et al., 2020, Moretti, 2012]. This concentration generates the superlinear scaling of productivity with city size documented by Bettencourt et al. [2007], but productivity gains accrue disproportionately to the already-skilled, producing the positive complexity–inequality correlation we observe.

Our results provide quantitative texture for this theoretical account. The 2.9-unit ECI gap between Cluster 4 (2.40) and Cluster 3 (–0.50) corresponds to a 0.06-unit Gini gap (0.51 vs. 0.45)—a difference that, while modest in absolute terms, represents a qualitatively different distributional form. In Knowledge Hub CBSAs, the elevated Gini reflects not merely wider income spread but a bimodal distribution: a pronounced right tail of very high earners in technology, finance, and professional services, coupled with a dense concentration of low-wage service workers, and a thinning of the middle-income occupations that traditionally formed the backbone of urban labor markets [Autor and Dorn, 2013, Goos et al., 2014]. This is precisely the distributional signature predicted by models of skill-biased technological change operating within geographically bounded labor markets [Acemoglu and Autor, 2011, Baum-Snow et al., 2018]. The finding that this paradox survives the exclusion of Cluster 4 ($r = 0.29$, $p < 0.01$ for the remaining 112 CBSAs) confirms that it is not an artifact of a small number of outlier metros but reflects a systematic pattern across the complexity gradient.

4.2 Efficient Inequality: The Mobility Paradox in Knowledge Economies

The co-occurrence of the highest income inequality and the most efficient mobility systems in Cluster 4 presents a paradox for transportation policy frameworks that assume a positive link between accessibility improvements and equity outcomes [Levinson, 2008, Litman, 2017]. Rodriguez et al. [2020] found that mobility improvements in U.S. micropolitan areas generate economic benefits that are unevenly distributed across income groups, a finding consistent with our observation that efficient mobility and high inequality coexist in complex economies. If the most unequal regions are also the most mobility-efficient, then transportation investment may be a less effective lever for reducing inequality than conventionally assumed.

We propose the concept of "efficient inequality" to describe this configuration. In dense knowledge economies, the spatial organization of employment—concentrated in accessible, transit-served, mixed-use urban cores [Ewing and Cervero, 2010]—generates aggregate mobility efficiency as measured by VMT per capita and radius of gyration. However, this same spatial concentration produces intense competition for proximate housing, driving prices to levels that sort residents by income: high earners occupy locations with short, efficient commutes, while lower-income workers are displaced to less accessible peripheries or crowded into substandard housing within the core [Hsieh and Moretti, 2019, Ganong and Shoag, 2017]. The aggregate efficiency metric thus masks a distributional pattern in which mobility advantages are unequally shared.

This interpretation finds support in Moro et al. [2021], who used mobile phone data to show that experienced income segregation in large U.S. cities is substantially higher than residential segregation alone would suggest, because high-income and low-income residents visit systematically different locations even when living in the same neighborhoods. Our finding that Gini and mean r_g are nearly uncorrelated ($r = -0.08$) at the CBSA level is consistent with this disaggregated picture: aggregate commuting efficiency does not predict aggregate income equality because efficiency is distributed unequally across the income spectrum. This result carries a cautionary methodological implication for the broader data science community: aggregate mobility metrics derived from digital traces, while powerful for characterizing regional-level patterns, may obscure within-region heterogeneity that is critical for equity analysis. Disaggregated approaches that link individual mobility traces to socioeconomic characteristics—as demonstrated by Xu et al. [2018] and Moro et al. [2021]—are essential complements to the CBSA-level analysis presented here.

The policy implication is that transportation efficiency and income equity are not complementary goals in knowledge economies but dimensions that can vary independently or inversely. Improving transit service or reducing VMT per capita in San Francisco or New York will not, by itself, address the distributional forces—skill premiums, housing costs, occupational polarization—that generate inequality. Effective equity-oriented policy requires interventions targeting the distributional mechanisms directly: inclusionary housing linked to transit stations [Schuetz et al., 2011], workforce development for middle-skill occupations [Holzer, 2017], or progressive local revenue structures capturing agglomeration surplus for redistribution [Piketty, 2014].

4.3 The Mobility Penalty in Low-Complexity Regions

While the complexity–inequality paradox in Knowledge Hubs is the most novel finding, the mobility challenges facing low-complexity regions (Clusters 2 and 3) are equally consequential for policy and for data-driven understanding of regional socioeconomic systems. The 2.8-fold VMT per capita gap between Cluster 2 (30.5) and Cluster 4 (10.8) represents a substantial economic burden on households. At national average fuel prices and vehicle operating costs, this differential translates into thousands of additional dollars per capita per year in transportation expenditure—a regressive burden that falls heaviest on the lowest-income households, who spend a disproportionate share of income on transportation [Blumenberg and Manville, 2004, Litman, 2017].

The spatial structure of employment is the fundamental driver. Where Knowledge Hub economies offer diverse employment distributed across dense environments, small industrial towns concentrate employment at specific sites—refineries, plants, military installations—tens of kilometers from residential areas. The absence of transit alternatives, a consequence of densities too low for fixed-route service [Guerra et al., 2012], forces virtually all workers into private vehicle commuting regardless of their ability to bear the costs. Hu [2017] found that lower-income workers in such environments experience the largest negative effects of poor job accessibility, compounding the economic disadvantages of low-complexity employment structures.

The divergence between Clusters 2 and 3 offers an additional insight. Despite similar ECI values (–0.01 vs. –0.50), Cluster 3's mean r_g of 7.9 km is less than half of Cluster 2's 19.2 km, a difference attributable to settlement compactness rather than economic structure. This suggests that land use configuration partially mitigates mobility challenges even in the absence of economic diversification—an insight with direct implications for planning in resource-dependent regions. Effective mobility strategies for low-complexity regions require fundamentally different approaches than those for knowledge hubs: demand-responsive micro-transit [Shaheen and Cohen, 2016], broadband investment enabling remote work [Gallardo, 2020], strategic economic diversification creating employment closer to residential concentrations [Delgado et al., 2016], and in the most spatially dispersed regions, consolidation of services in accessible regional centers [Henderson and Gloy, 2016].

4.4 Toward Differentiated Regional Policy

The four-cluster typology demonstrates that effective development policy must be structurally differentiated. Knowledge Hub regions face the challenge of managing prosperity-driven inequality without undermining agglomeration economies. Evidence from Schuetz et al. [2011] on inclusionary zoning, Holzer [2017] on workforce development, and Glaeser and Kerr [2010] on broadening entrepreneurial access suggests that targeted interventions can moderate inequality while preserving dynamism. Small Industrial Towns require integrated strategies combining diversification through cluster development [Delgado et al., 2016] with digital infrastructure [Gallardo, 2020] and demand-responsive transport [Shaheen and Cohen, 2016]. Rural Resource-Dependent Regions need longer-term structural transformation through renewable energy [Morrison, 2019], telemedicine [Drake et al., 2019], and strategic service consolidation [Henderson and Gloy, 2016]. High-Density Mixed Economy Regions,

occupying a transitional position, require strategies strengthening knowledge-economy linkages while maintaining employment diversity.

The broader lesson is that the complexity–inequality paradox makes universal prescriptions—"diversify your economy," "invest in transit," "attract technology firms"—potentially counterproductive when applied without attention to distributional consequences. Policymakers in complex regions must anticipate that increasing sophistication may deepen inequality, while those in simple regions must recognize that low inequality often reflects limited opportunity rather than equitable prosperity. This differentiated understanding, grounded in the systematic analysis of digital traces of economic activity and human mobility, exemplifies the kind of data-driven insight that can inform more nuanced and effective regional governance.

4.5 Limitations and Identification Challenges

Several limitations warrant acknowledgment. The cross-sectional design captures associations at a single time point and cannot establish causal direction. The positive ECI–Gini correlation might reflect complexity generating inequality through skill-sorting mechanisms, but could equally reflect reverse causality or confounding by omitted variables such as historical industrial structure or geographic endowments. Future research should exploit temporal variation to implement difference-in-differences designs, or employ shift-share (Bartik) instruments following Goldsmith-Pinkham et al. [2020] to isolate exogenous complexity changes.

The five-state sample, while strategically selected, excludes the broad middle of the U.S. income distribution and may exaggerate between-cluster contrasts. Extension to all U.S. CBSAs ($n \approx 930$) would strengthen statistical power, improve cluster stability, and test whether the typology generalizes. The binary RCA threshold discards variation in specialization intensity; continuous alternatives proposed by Davies and Maré [2021] should be tested. The Knowledge Hub cluster ($n = 9$), while exceeding the bootstrap stability threshold, is small, and its results should be interpreted as characterizing a type rather than providing precise within-cluster estimates. Finally, the documented demographic biases in cell phone–derived mobility data [Wesolowski et al., 2013, Coston et al., 2021, Chen et al., 2019] may attenuate observed mobility differences, rendering our complexity–mobility estimates conservative.

5 Conclusion

This study documents a complexity–inequality paradox across 121 U.S. Core-Based Statistical Areas: the regions with the most sophisticated employment structures simultaneously exhibit the highest income inequality and the most efficient mobility systems. By adapting the Economic Complexity Index from international trade to sub-national employment distributions and integrating it with high-resolution digital mobility traces, we reveal a reversal of the well-established national-level relationship between economic complexity and income equality—a Simpson's Paradox driven by skill-biased agglomeration, occupational polarization, and housing market dynamics operating within a shared institutional context.

The four structurally distinct regional typologies—Prosperous Knowledge Hubs, High-Density Mixed Economies, Rural Resource-Dependent Regions, and Small Industrial Towns—face qualitatively different development challenges demanding differentiated policy responses.

The configuration we term "efficient inequality," in which complex economies minimize mobility externalities while maximizing income dispersion, challenges assumptions that diversification and transportation improvement naturally promote equity. In knowledge-intensive regions, aggregate efficiency metrics derived from digital traces mask distributional patterns in which the benefits of proximity, density, and accessibility are unequally shared.

Three implications emerge for data-driven regional governance. First, aggregate indicators—GDP per capita, unemployment rates, or even ECI alone—fail to capture the multi-dimensional trade-offs defining regional trajectories; integrated frameworks combining economic structure, mobility, and distributional metrics are essential. Second, the scale dependence of the complexity–inequality relationship means that international lessons cannot be mechanically transferred to sub-national contexts. Third, the analytical pipeline demonstrated here—employment-based ECI, digital mobility analytics, PCA, and bootstrap-validated clustering—is replicable with publicly available data and offers a scalable diagnostic tool for evidence-based regional planning.

Future research should extend this framework longitudinally, tracking CBSA trajectories to enable causal identification; geographically, expanding to all U.S. CBSAs and international comparisons; methodologically, incorporating individual-level mobility–income linkages to decompose aggregate patterns; and substantively, integrating environmental dimensions to capture the full sustainability profile of regional development pathways.

Abbreviations

ACS	American Community Survey
CBSA	Core-Based Statistical Area
ECI	Economic Complexity Index
GMM	Gaussian Mixture Model
GPS	Global Positioning System
NAICS	North American Industry Classification System
OD	Origin–Destination
OMB	Office of Management and Budget
PCA	Principal Component Analysis
RCA	Revealed Comparative Advantage
r_g	Radius of Gyration
VMT	Vehicle Miles Traveled
WCSS	Within-Cluster Sum of Squares

Declarations

Ethics approval and consent to participate

Not applicable. This study uses only aggregated, de-identified secondary data sources (U.S. Census Bureau employment statistics, American Community Survey estimates, and commercially licensed mobility data from Replica Inc.) and does not involve human subjects research requiring ethics approval.

Consent for publication

Not applicable.

Availability of data and materials

Employment data from the U.S. Census Bureau's County Business Patterns are publicly available (https://www.census.gov/programs-surveys/cbp.html). American Community Survey estimates are accessible via https://data.census.gov. Replica origin–destination mobility data are available under license from Replica Inc. (https://replicahq.com). The processed analytical dataset and replication code for ECI computation and cluster analysis are available from the corresponding author upon reasonable request.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Funding

The authors received no specific funding for this work.

Authors' contributions

ZW: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft & editing, Visualization. **MCG**: Supervision, Project administration, Methodology, Writing – review & editing. All authors read and approved the final manuscript.

Acknowledgements

The authors gratefully acknowledge the U.S. Census Bureau for maintaining publicly accessible employment and demographic data, and Replica Inc. for access to mobility data. We thank the anonymous reviewers for their constructive feedback on earlier versions of this manuscript.

References

Daron Acemoglu and David Autor. Skills, tasks and technologies: Implications for employment and earnings. *Handbook of labor economics*, 4:1043–1171, 2011.

David H Autor and David Dorn. The growth of low-skill service jobs and the polarization of the us labor market. *American Economic Review*, 103(5):1553–1597, 2013.

David H Autor, Frank Levy, and Richard J Murnane. The skill content of recent technological change: An empirical exploration. *The Quarterly journal of economics*, 118(4):1279–1333, 2003.

Pierre-Alexandre Balland, Cristóbal Jara-Figueroa, Sergio G Petralia, Mathieu PA Steijn, David L Rigby, and César A Hidalgo. Complex economic activities concentrate in large cities. *Nature Human Behaviour*, 4(3):248–254, 2020.

Pierre-Alexandre Balland, Tom Broekel, Dario Diodato, Elisa Giuliani, Ricardo Hausmann, Neave O'Clery, and David Rigby. The new paradigm of economic complexity. *Research Policy*, 51(3):104450, 2022.

Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.

Nathaniel Baum-Snow, Matthew Freedman, and Ronni Pavan. Why has urban inequality increased? *American Economic Journal: Applied Economics*, 10(4):1–42, 2018.

Luís M A Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104(17):7301–7306, 2007.

Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):1–55, 2015.

Evelyn Blumenberg and Michael Manville. Beyond the spatial mismatch: Welfare recipients and transportation policy. *Journal of Planning Literature*, 19(2):182–205, 2004.

Tom Brökel. Using structural diversity to measure the complexity of technologies. *PLoS ONE*, 14(5):e0216856, 2019.

Robert Cervero and Michael Duncan. Which reduces vehicle travel more: jobs-housing balance or retail-housing mixing? *Journal of the American Planning Association*, 72(4): 475–490, 2006.

Cynthia Chen, Jingtao Ma, Yusak Susilo, Yu Liu, and Menglin Wang. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation research part C: emerging technologies*, 100:538–555, 2019.

Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. Where is the land of opportunity? the geography of intergenerational mobility in the united states. *The Quarterly Journal of Economics*, 129(4):1553–1623, 2014.

Dylan S Connor, Peter Kedron, Shiqi Xie, Gaurav Jain, Jinhyung Jang, Yifang Yu, Amy Frazier, and Tom Kemeny. Big cities fuel inequality within and across generations. *PNAS Nexus*, 4(1):pgae587, 2025.

Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Leveraging administrative data for bias audits: Assessing disparate coverage with mobility data for equitable transportation. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 389–400, 2021.

Benjamin Davies and David C Maré. Relatedness, complexity and local growth. *Regional Studies*, 55(3):479–494, 2021.

Mercedes Delgado, Michael E Porter, and Scott Stern. Defining clusters of related industries. *Journal of Economic Geography*, 16(1):1–38, 2016.

Jonathan I Dingel and Brent Neiman. How many jobs can be done at home? *Journal of Public Economics*, 189:104235, 2020.

Jean-Christophe Dissart. Regional economic diversity and regional economic stability: research results and agenda. *International Regional Science Review*, 26(4):423–446, 2003.

Coleman Drake, Tyler Lian, Brooke Cameron, Katy Medynskaya, Heather M Yarish, Korana Perera, and Jessica Paulus. Utilizing telemedicine in small rural hospitals: administrative and clinical considerations. *Current Emergency and Hospital Medicine Reports*, 7:115–122, 2019.

Gilles Duranton and Diego Puga. The economics of urban density. *Journal of Economic Perspectives*, 34(3):3–26, 2020.

Reid Ewing and Robert Cervero. Travel and the built environment: A meta-analysis. *Journal of the American planning association*, 76(3):265–294, 2010.

Richard Florida. *The new urban crisis: How our cities are increasing inequality, deepening segregation, and failing the middle class—and what we can do about it*. Basic Books, 2017.

Christopher S Fowler, Danielle C Rhubart, and Leif Jensen. Reassessing and revising commuting zones for 2010: History, assessment, and updates for us 'labor-sheds' 1990–2010. *Population Research and Policy Review*, 35:263–286, 2016.

Christopher S Fowler, Nathan Frey, David C Folch, Nicholas Nagle, and Seth Spielman. The contribution of metropolitan area migration to us county population change: 2000–2010. *Population Research and Policy Review*, 39:443–463, 2020.

Benedikt S L Fritz and Robert A Manduca. The economic complexity of US metropolitan areas. *Regional Studies*, 55(7):1299–1310, 2021.

Roberto Gallardo. Broadband and local growth. *Journal of Rural and Community Development*, 15(2), 2020.

Peter Ganong and Daniel Shoag. Why has regional income convergence in the us declined? *Journal of Urban Economics*, 102:76–90, 2017.

Corrado Gini. Measurement of inequality of incomes. *The Economic Journal*, 31(121): 124–126, 1921.

Edward Glaeser. *Triumph of the city: How our greatest invention makes us richer, smarter, greener, healthier, and happier*. Penguin Books, 2014.

Edward L Glaeser and William R Kerr. Clusters of entrepreneurship. *Journal of Urban Economics*, 67(1):150–168, 2010.

Laurent Gobillon, Harris Selod, and Yves Zenou. The mechanisms of spatial mismatch. *Urban studies*, 44(12):2401–2427, 2007.

Paul Goldsmith-Pinkham, Isaac Sorkin, and Henry Swift. Bartik instruments: What, when, why, and how. *American Economic Review*, 110(8):2586–2624, 2020.

Marta C González, César A Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

Maarten Goos, Alan Manning, and Anna Salomons. Explaining job polarization: Routine-biased technological change and offshoring. *American Economic Review*, 104(8):2509–2526, 2014.

Erick Guerra, Robert Cervero, and Daniel Tischler. Half-mile circle: Does it best represent transit station catchments? *Transportation Research Record*, 2276(1):101–109, 2012.

Dominik Hartmann and Flávio L Pinheiro. Economic complexity and inequality at the national and regional level. In *Handbook of Complexity Economics*. Routledge, 2022.

Dominik Hartmann, Miguel R Guevara, Cristian Jara-Figueroa, Manuel Aristarán, and César A Hidalgo. Linking economic complexity, institutions, and income inequality. *World Development*, 93:75–93, 2017.

Ricardo Hausmann, César A Hidalgo, Sebastian Bustos, Michele Coscia, Alexander Simoes, and Muhammed A Yildirim. *The Atlas of Economic Complexity: Mapping Paths to Prosperity*. MIT Press, Cambridge, MA, 2014.

Jason Henderson and Brent A Gloy. Will service follow the plow? the impact of broadband on rural and urban businesses. *Regional Science Policy & Practice*, 8(4):225–245, 2016.

Christian Hennig. Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, 52(1):258–271, 2007.

César A Hidalgo and Ricardo Hausmann. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106(26):10570–10575, 2009.

Harry J Holzer. The us labor market in 2030: A scenario based on trends and policies. *The ANNALS of the American Academy of Political and Social Science*, 673(1):123–142, 2017.

Chang-Tai Hsieh and Enrico Moretti. Housing constraints and spatial misallocation. *American Economic Journal: Macroeconomics*, 11(2):1–39, 2019.

Lingqian Hu. Job accessibility and employment outcomes: which income groups benefit the most? *Transportation*, 44:1421–1443, 2017.

Keith Ihlanfeldt. Re-examining spatial mismatch. *Citizens and Scholars*, 2006.

John F Kain. Housing segregation, negro employment, and metropolitan decentralization. *The Quarterly Journal of Economics*, 82(2):175–197, 1968.

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

Thomas Kemeny and Taner Osman. The wider impacts of high-technology employment: Evidence from us cities. *Research Policy*, 44(9):1828–1842, 2015.

Keun Lee and Di Qu. Exogenous shocks and the dynamics of economic complexity. *Research Policy*, 50(7):104282, 2021.

Neil Lee and Andrés Rodríguez-Pose. Innovation, cities, and spatially bounded knowledge spillovers. *Regional Studies*, 50(6):880–894, 2016.

David Levinson. Density and dispersion: the co-development of land use and rail in london. *Journal of Economic Geography*, 8(1):55–77, 2008.

Todd Litman. Evaluating public transportation health benefits. *Journal of Public Transportation*, 10(2):55–85, 2017.

Ron Martin, Peter Sunley, Ben Gardiner, and Peter Tyler. How regions react to recessions: Resilience and the role of economic structure. *Regional Studies*, 50(4):561–585, 2016.

Enrico Moretti. *The new geography of jobs*. Houghton Mifflin Harcourt, 2012.

Esteban Moro, Dan Calacci, Xiaowen Dong, and Alex Pentland. Mobility patterns are associated with experienced income segregation in large US cities. *Nature Communications*, 12(1):4633, 2021.

Robbie Morrison. Energy system modeling: Public transparency, scientific reproducibility, and open development. *Energy Strategy Reviews*, 26:100409, 2019.

Office of Management and Budget. 2020 standards for delineating core based statistical areas. Federal Register Notice, 2020. 85 FR 39244.

Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. *Nature Communications*, 6(1):8166, 2015.

Thomas Piketty. *Capital in the twenty-first century*. Harvard University Press, 2014.

Flávio L Pinheiro, Pierre-Alexandre Balland, Ron Boschma, and Dominik Hartmann. The dark side of the geography of innovation: Relatedness, complexity and regional inequality in Europe. *Regional Studies*, 56(1):111–128, 2022.

Replica Inc. Replica places: Spring 2023 data release. Replica Data Platform, 2023.

Christian Reynolds, Manju Agrawal, Ivy Lee, Chen Zhan, Jiangyu Li, Peter Taylor, Tim Mares, Julia Davison, John Mayes, Melissa McMillan, et al. A sub-national economic complexity analysis of Australia's states and territories. *Regional Studies*, 52(5):715–726, 2018.

Daniel A Rodriguez, Asad J Khattak, and Kelly R Evenson. Mobility and economic opportunity: Evidence from us micropolitan areas. *Urban Studies*, 57(3):598–619, 2020.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Angelica Sbardella, Emanuele Pugliese, and Luciano Pietronero. Economic development and wage inequality: A complex system analysis. *PLoS ONE*, 12(9):e0182774, 2017.

Jenny Schuetz, Rachel Meltzer, and Vicki Been. Inclusionary zoning: can it work in california? *Housing Policy Debate*, 21(1):25–48, 2011.

Susan A Shaheen and Adam P Cohen. Shared mobility: Current practices and guiding principles. *Access Magazine*, 1(49):15–25, 2016.

Michael Storper. Separate worlds? explaining the current wave of regional economic polarization. *Journal of Economic Geography*, 18(2):247–270, 2018.

Texas A&M Transportation Institute. Independent evaluation of replica traffic data. TTI Technical Report, 2024.

U.S. Census Bureau. American community survey 5-year estimates. U.S. Census Bureau Data, 2023.

Amy Wesolowski, Nathan Eagle, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society Interface*, 10(81):20120986, 2013.

Yanyan Xu, Alexander Belyi, Iva Bojic, and Carlo Ratti. Human mobility and socioeconomic status: Analysis of Singapore and Boston. *Computers, Environment and Urban Systems*, 72:51–61, 2018.