WorkReach: Modeling Urban Work Location Choices Through Economic Complexity, Informality, and Mobility Data

Ollin D. Langle-Chimal[1*], Steffen Knoblauch[2,3] and Marta C. González[1,4*]

[1*]Department of Civil and Environmental Engineering, University of California, Berkeley, Berkeley, 94720, CA, USA.
[2]Interdisciplinary Centre of Scientific Computing (IWR) Heidelberg University, Heidelberg Germany.
[3]Heidelberg Institute for Geoinformation Technology Heidelberg University, Heidelberg Germany.
[4]Department of City and Regional Planning, University of California, Berkeley, Berkeley, 94720, CA, USA.

*Corresponding author(s). E-mail(s): ollin@berkeley.edu; martag@berkeley.edu;

Abstract

Human mobility in urban areas largely depends on the spatial distribution of economic opportunities, yet our understanding of how workers balance the trade-off between proximity and job quality is still limited. Existing mechanistic models of trip distributions capture physically interpretable effects such as distance decay or intervening opportunities, but they do not account for how socioeconomic factors influence commuting decisions. Here we introduce the WorkReach model, which, grounded in discrete choice theory, reveals how aggregate commuting patterns emerge when workers behave as if maximizing a location-choice utility. By incorporating for the first time residential informality and the economic complexity of jobs into its utility function, the model quantifies how workers trade distance for job quality depending on their socioeconomic context. The results reconstruct flows with accuracy comparable to widely used benchmarks with added information in the choice mechanisms. Applied to four cities in the U.S., Mexico, and Brazil, the framework highlights a consistent preference across the four cities studied for economically sophisticated areas, while also revealing marked regional differences in how workers balance this trade-off. Moreover, by defining

accessibility as the perceived utility of opportunities rather than only physical proximity, WorkReach offers new insights into assessing urban inequality.

Keywords: Human Mobility, Economic Complexity, Urban Science, Informal Labor

1 Introduction

The decision of where to live and work is a fundamental aspect of urban life and has great consequences at the individual and collective urban dynamics. This choice is often a complex trade-off, balancing the desire for proximity with the pursuit of better economic opportunities, all within a landscape defined by varying labor market conditions [1]. Understanding these trade-offs is crucial for addressing spatial inequalities in employment accessibility, particularly in metropolitan regions characterized by significant socioeconomic heterogeneity. Factors such as the prevalence of informal labor and the spatial distribution of the economic desirability of work opportunities can create large disparities in the access to quality jobs, which greatly affects economic stability and social mobility [2–4]. Regardless of the importance of these socioeconomic dimensions in shaping urban mobility patterns, current mobility flow modeling approaches do not take them into account.

Traditional spatial interaction models, particularly gravity and radiation [5] models are known for their intuitive formulation and parsimony [6–9], which has led to their popularity in urban and transportation planning. Gravity models, as the name suggests, are inspired by Newtonian gravity, modeling mobility flows between origins and destinations as a function of their population (or "mass") and a distance-decay component. Radiation models, in contrast, are parameter-free and rely on the number of intervening opportunities located closer to the origin [10]. While the accuracy and general interpretability of these models have long been studied and described [7, 11], they provide limited information on the behavioral mechanisms that drive these flows, particularly when socioeconomic factors are taken into account. Recent advances in the field, including deep learning approaches such as deep gravity [12] characterized by being data intensive and other based on symbolic regression such as the gravity-like models found by the Bayesian machine scientist (BMS) [13], have pushed the limits of predictive accuracy. However, this increased accuracy doesn't necessarily offer new interpretability of the underlying decision-making processes.

In order to address these limitations, we present WorkReach, a model designed to provide behavioral insights into urban work location choices. WorkReach accounts for the trade-offs commuters make between distance and work quality by conditioning work location choices on residential locations and their associated socioeconomic characteristics. We apply this framework to analyze commuting patterns across four urban areas; the San Francisco Bay Area, Los Angeles, Mexico City, and Rio de Janeiro. These cities represent different regions, physical characteristics, and structures of the urban labor market.

This study prioritizes model interpretability and the extraction of actionable insights. To this end, we ground our approach in discrete choice methods, which rely

on the theory of Random Utility Models (RUM), by integrating key socioeconomic variables directly into a utility function that commuters are assumed to maximize, so that the resulting parameters approximate the aggregate trade-offs reflected in observed commuting patterns [14, 15]. Specifically, we incorporate two critical, yet not commonly studied, dimensions in flow modeling: economic complexity and labor informality. Economic complexity, a measure of the diversity and sophistication of a region's economy [16, 17], serves as a useful proxy for the economic quality of work opportunities. Although earlier work has used economic complexity to study cities as a whole [18], its application at the sub-city level is just starting to be explored [19–23], revealing fine-grained differences in urban economies. The trade-off between commuting distance and wages is well established in the labor economics literature [24, 25], but wages capture only current compensation. $\text{ECI}^{\text{employment}}$, by contrast, is derived from the productive structure of each zone and therefore reflects location-level advantages that current pay does not internalize. Agglomeration theory states that zones where diverse and complementary industries co-locate generate productivity externalities from sectoral complementarities, lower job-search costs by concentrating a wider range of employment opportunities, and facilitate human capital accumulation through exposure to a broader set of economic activities [26, 27]. In addition to these conceptual advantages, we compare WorkReach against a variant that substitutes $\text{ECI}^{\text{employment}}$ with a wage-based covariate and show that both variables carry complementary predictive information (Section S4.5). Labor informality is defined as the employment outside formal regulatory frameworks and lacking social protections [28]. It is a widespread characteristic of urban economies, particularly in Latin America [2]. However, despite its documented presence, informality remains less systematically studied in high-income countries such as the United States [3, 4]. By incorporating economic complexity at the work location and informality at the home location into a utility function, our model quantifies how these factors, along with commuting distance, shape work location choices. This approach let us estimate a set of parameters that are interpretable and useful to quantify the trade-offs workers are willing to make. For example, we can measure how much additional distance a worker will travel for a work in a more economically complex area, and how this willingness is modulated by the informality context of their home location. These behavioral insights are crucial to understanding how the aggregation of individual decisions shapes macro-level mobility patterns and the socioeconomic stratification of a city [29–31].

Another key advantage of the WorkReach model is that it allows us to derive economically meaningful work accessibility metrics based on a perceived satisfaction of the decision maker. Accessibility metrics usually focus on the number of opportunities reachable from a specific location within a certain travel time or physical distance [6, 32, 33]. Discrete choice models, on the other hand, directly derive accessibility through consumer-surplus, which reflects the expected maximum utility an agent gains from the whole set of available work choices [34–37]. These measures provide a more holistic understanding of disparities in access to opportunities by not only accounting for proximity but also the perceived quality of the choices [38, 39].

Our results reveal both fundamental differences and striking similarities between the cities and regions studied. For example, we found that while workers in the U.S.

cities weigh economic opportunity across the whole distance spectrum, commuters in the Latin American cities operate within a 'proximity-first' regime. This divergence can be explained with the role of informality in the utility function. In Mexico City and Rio de Janeiro, informal workers perceive a utility gain by undertaking longer commutes to access high-complexity jobs, whereas in U.S. cities, this behavior is observed among workers from low-informality areas. Our framework allows for a quantification of these behavioral nuances, showing, for instance, that a worker in Mexico City is willing to trade nearly four times the distance for a one-unit gain in economic complexity compared to a worker in Los Angeles, or that a reduction of economic complexity of 1% in a work location in the Bay Area is associated with a 0.804% reduction in the probability of a commuter to select that specific location as their work choice. Despite these regional differences, our utility-based accessibility metric shows that home locations from high-informality workers experience a lower consumer-surplus accessibility in the four cities, translating into an overall lower satisfaction with the work market.

2 Results

2.1 Data Overview

2.1.1 Study Areas

Our study focuses on four major urban areas that are comparable in population size: Mexico City and the municipality of Rio de Janeiro in Latin America, and Los Angeles and the San Francisco Bay Area in the United States. These areas of interest represent diverse urban landscapes and economic structures with different characteristics in their labor markets, giving us a rich comparative context. Mexico City and Rio de Janeiro are two megacities in middle income economies that have very high levels of informal employment and population density[40]. Los Angeles and the Bay Area, in contrast, are major economic hubs in a high income country, with substantial socioeconomic disparities and distinct patterns of suburbanization and polycentricity [41, 42]. Informality is less systematically tracked or studied in the U.S. compared to Latin America, yet California is one of the States where researchers have estimated its prevalence [3]. For Mexico City, Los Angeles, and Rio de Janeiro, the study area corresponds to the city proper, while the San Francisco Bay Area encompasses multiple cities within a multi-county region. This choice prioritizes comparability in population scale and number of spatial units across the four cities. The composition of these urban areas (Table 1) allows us to examine the differences and similarities that make up the economic decision of commuters to choose their work location based on distance from their home location, economic opportunity, and the prevalence of informality of the job.

2.1.2 Economic Complexity Framework

The four cities exhibit not only spatial heterogeneity in their social and economic composition but also in the spatial distribution of economic units (business or establishments that employ workers). We study this with the economic complexity

framework, which is built by a bipartite network of locations and products (in this case, economic sectors), where a link exists if a location has a revealed comparative advantage (RCA) in that sector. A location is considered to have an RCA when the ratio of workers in that sector within the area exceeds the ratio of workers in the same sector across the entire economy.

From this network, we derive two key metrics: *diversity*, which represents the number of sectors where a location has a comparative advantage, and *ubiquity*, which is the number of locations where a sector shows a comparative advantage. Using the method of reflections (described in the Methods section 4.2), an iterative process over these two vectors, we calculate the Economic Complexity Index ($\text{ECI}^{\text{employment}}$) for locations and the Product Complexity Index ($\text{PCI}^{\text{employment}}$) for sectors [16]. Because our ECI and PCI are derived from employment data at the sub-city level rather than from trade data at the country level, we denote them $\text{ECI}^{\text{employment}}$ and $\text{PCI}^{\text{employment}}$ throughout to distinguish them from their original formulations [16].

Applying this framework to the 17 two-digit North American Industry Classification System (NAICS) sectors reveals distinct specialization patterns. For example, the Bay Area shows a higher than average $\text{PCI}^{\text{employment}}$ in the information sector, reflecting its tech-driven economy. The product space network, is a projection of the bipartite network, with nodes representing sectors and link weights representing the marginal probabilities of two sectors having a RCA at the same location. As shown in Fig. 2b), highly complex sectors tend to cluster closely in the network, while less complex sectors do the same. The correlation matrix of these co-location patterns across cities, shown in Fig. 2c), reveals regional similarities: the Bay Area and Los Angeles exhibit a high similarity ($r = 0.85$), and Mexico City and Rio de Janeiro show an equally strong correlation ($r = 0.86$). Cross-regional correlations are lower ($r = 0.68$–0.71), indicating that U.S. and Latin American cities share the broad sectoral co-location structure but differ in the finer details of which sectors cluster together.

2.1.3 Variable construction

Informality Rate: Given the varying prevalence and nature of informal employment, we need to utilize context-specific methodologies for each specific region, as there is no universal definition that can be obtained or derived directly from available data. For Mexico City, we estimate informality rates using microdata from the 2020 Population and Housing Census from The National Institute of Statistics and Geography (INEGI), aggregated at the urban Basic Geostatistical Area (AGEB) level. We used affiliation to social security services as a proxy for formality, following the legal requirement in Mexico for formal employers to register workers in the Instituto Mexicano del Seguro Social (IMSS). For this, we calculate the proportion of the economically active population without access to medical insurance services within each AGEB as the informality rate, adjusting for insured dependents (minors covered through a family member) who are not themselves part of the workforce. This proxy is validated using Mexico's National Survey of Occupations and Employment (ENOE, 2022), which tracks formal and informal employment and indicates that 99.95% of formally employed individuals have access to medical insurance, compared to only 11.18% of informal workers.

For Rio de Janeiro, we follow the methodology applied by Moreno-Monroy and Ramos (2021) for the city of São Paulo [2], using census data from 2010 which is the latest available at the Área de Estatística e Planeamento (AEP) scale, which is the highest resolution publicly available. Workers are classified as informal if they are employees without formal registration (*empregado sem carteira assinada*), or self-employed individuals who do not contribute to social security.

For Los Angeles and the Bay Area, measuring informality is more challenging due to its low prevalence and low government tracking. We adapt a method used by the U.S. Census Bureau staff in a study by Graham and Ong's (2007) [3], which compares workforce numbers from the American Community Survey (ACS 5-year estimates) with employment records from the Longitudinal Employer-Household Dynamics (LEHD) dataset. Workers present in ACS data but absent from LEHD records are considered informal as the latter only track individuals in payroll with unemployment insurance contributions, effectively tracking formal employment. This approach, was originally developed using the Census Long Form, which is no longer available but has been replaced by the ACS. This is done at the Census Block Group level.

Economic Complexity Index ($\mathrm{ECI}^{\mathrm{employment}}$**)**: $\mathrm{ECI}^{\mathrm{employment}}$ is computed at the H3 hexagonal zone level by assigning each individual job record to its containing hexagon based on the establishment's geographic coordinates. This differs from the informality rate, which is measured at the census-tract level (Census Block Groups in the U.S., AGEBs in Mexico City, AEPs in Rio de Janeiro) and then area-weighted to H3 zones; for Rio de Janeiro, where AEP polygons are already coarser than H3 cells, we retain the original boundaries for both variables. Computing $\mathrm{ECI}^{\mathrm{employment}}$ directly at the H3 level, rather than aggregating from administrative units, provides a consistent spatial resolution across cities and enables the product-space analyses reported in the SI, including skill relatedness (Section S5), diversity and ubiquity decompositions (Section S4.4), and counterfactual $\mathrm{ECI}^{\mathrm{employment}}$ uplift scenarios (Section S8). For this, we use the number of employed individuals at each location in any of the 17 industry sectors from the 2-digit NAICS, excluding Agriculture and Other Services to retain sectors with clear interpretability in the product-space framework. These sectors are treated as the "products" in Economic Complexity terms [16, 17], and their associated Product Complexity Index ($\mathrm{PCI}^{\mathrm{employment}}$) reflects the complexity of each work sector. We confirm that the zone–sector matrices underlying the $\mathrm{ECI}^{\mathrm{employment}}$ computation are nested in all four cities (SI Section S4.2), a property that supports the reliability of the complexity rankings [43]. For Mexico City, employment data come from INEGI's Directory of Economic Units (DENUE), which provides geocoded establishment records that are spatially joined to H3 hexagons. DENUE reports the number of employees per establishment in categorical size ranges rather than exact counts, so we impute a point estimate by drawing a uniform random integer within each range, capping the open-ended upper category at 500 employees. For Rio de Janeiro, employment records come from Brazil's Annual Social Information Report (RAIS), which includes codes from the National Classification of Economic Activities (CNAE) and are manually harmonized with NAICS. As RAIS reports employment in a similar categorical format as DENUE, we apply the same imputation procedure.

For Los Angeles and the Bay Area, we use synthetic employment data from Replica, which integrates demographic, economic, and transportation data including NAICS sector. Replica provides individual-level records with geographic coordinates, which we spatially join to H3 hexagons via an area-weighted overlay from the Census Block Group level.

Data: In order to understand the connection between home and work location choices we use commuting flows derived from different data sources. For Mexico City and Rio de Janeiro, we used anonymized Location-Based Services (LBS) data from a single commercial provider, covering March-May 2023 and March-May 2024 (six months total), the period for which data are available. We process raw ping data to estimate frequent visitations to specific locations and used a heuristic method based on circadian rhythms to assign a home and workplace location of each individual user, as described in the Methods section 4.5.1. Because we infer only home and work locations rather than modeling the full set of trips, and because inference requires a minimum of 14 and 8 distinct days of observations respectively (Section 4.5.1), the resulting commuting patterns reflect habitual behavior and are robust to short-term seasonal variation. We retain only users whose inferred home and work locations are consistent across both years. For Los Angeles and the San Francisco Bay Area, Replica's synthetic mobility dataset provides average weekday commuting patterns. Multiple years and seasons are available, and we used the generated data for Spring 2023. Across all four cities, once home and workplace locations are identified, they are spatially merged with their respective geostatistical units to assign origin-level informality metrics and destination-level $\text{ECI}^{\text{employment}}$.

To build comparable flow networks across cities, we remap the original geographic units to Uber's H3 hexagonal grid system, which offers a consistent spatial resolution. When administrative or census polygons are smaller than the target H3 cells, we aggregate them using area-weighted averages for $\text{ECI}^{\text{employment}}$ and informality rates, based on their overlap with each hexagon. In contrast, when original zones are larger and encompass multiple H3 cells, we preserve their scale by taking the union of the fully contained hexagons and treating them as a single unit. For Rio de Janeiro, where the original AEP polygons are already coarse, we retain the original boundaries without remapping.

2.1.4 Descriptive Statistics and Spatial Distributions

Table 1 provides key descriptive statistics for the four cities. The number of origin-destination links and unique geostatistical nodes varies, reflecting differences in urban scale and data granularity. Rio de Janeiro has the fewest links and nodes, while Mexico City has the most. In all four cities the number of links corresponds to a complete graph without self-loops: every pair of locations is connected by at least one observed commute. Locations without incoming or outgoing trips are discarded. The median and mean area of these nodes show the spatial resolution of every city; Mexico City has the finest resolution shown by the smaller average area, we used a smaller H3 scale compared to the US urban areas, H3 levels 8 and 7 respectively, due to the small overall diameter of the area covered. Flow counts, commute distances, and populations show wide ranges within and between cities, highlighting the heterogeneous nature

of commute patterns. Los Angeles and the Bay Area exhibit the longest maximum commute distances, consistent with their larger areas and polycentric nature [42].

Table 1 Key descriptive statistics for the four mobility datasets. Total population refers to the sum within the study area boundaries.

	Bay Area	**Los Angeles**	**Mexico City**	**Rio de Janeiro**
Links (OD pairs)	446 892	363 006	626 472	39 402
Nodes (Locations)	669	603	792	199
Area (median/mean [km^2])	5.40 / 27.39	5.77 / 16.59	0.85 / 0.98	2.93 / 6.03
Total Area [km^2]	18 321.06	10 004.74	775.48	1 199.24
Flow range (min–max)	1–40 119	1–21 036	1–3 243	6–9 780
Total observed flows	1 936 002	2 388 812	473 078	357 005
Median OD distance [km]	11.92	10.39	4.12	4.08
Diameter [km]	283.22	145.42	52.62	73.65
Population range (min–max)	39–100 177	5–297 978	9–38 381	3 361–65 846
Total population	7 373 986	10 084 557	8 995 891	6 289 038

The spatial distribution of $\text{ECI}^{\text{employment}}$ and informality rates, shown in Fig. 1 varies across cities. Both variables have statistically significant spatial autocorrelation in all cities, as indicated by Moran's I, which indicates that similar characteristics tend to cluster together, especially in the Latin American region (Table S1 and Figs. S1-S8). However, the spatial correlation between both variables differs greatly by region. Mexico City and Rio de Janeiro show a moderate negative correlation between informality rates and $\text{ECI}^{\text{employment}}$ ($r \sim -0.4$), suggesting spatial segregation where informal workers live far away from high-complexity economic areas. In contrast, the correlation between these two variables in the U.S. cities is close to zero ($r \sim 0$), which suggests that the spatial distribution of informal workers is less systematically excluded from high economically complex areas. This points to different forms of socio-spatial organization: whereas Latin American cities display sharper patterns of economic segregation, U.S. cities exhibit a more mixed spatial arrangement of formal and informal employment.

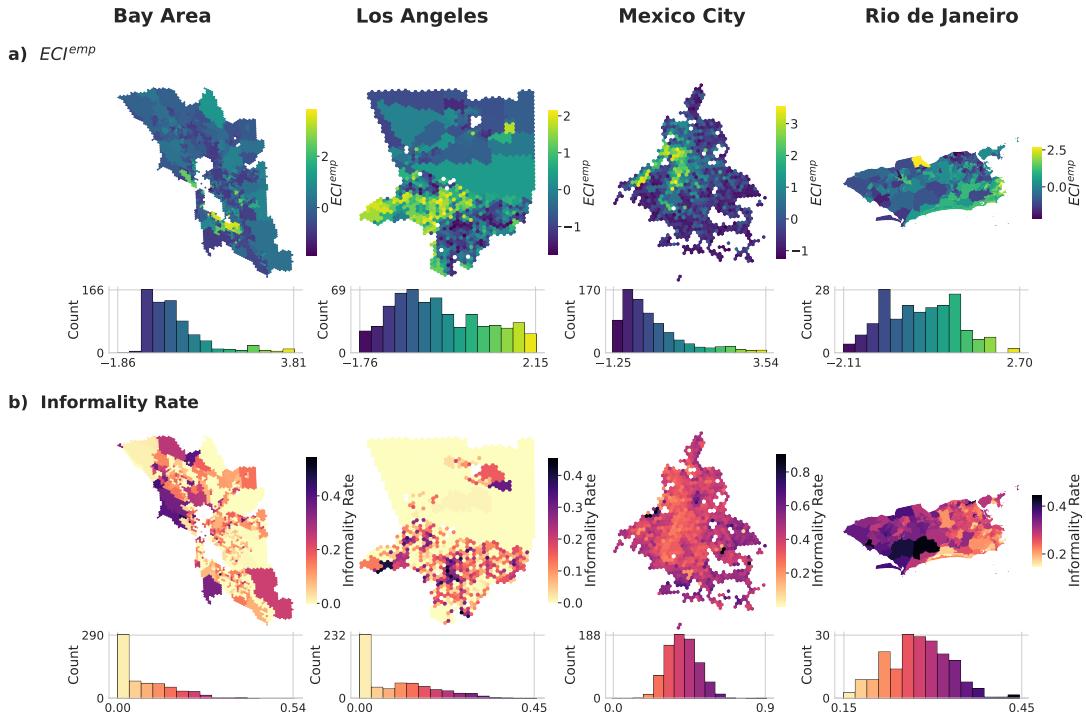

Fig. 1 Choropleth maps depicting the Economic Complexity Index ($ECI^{employment}$) and informality rate for each geographical unit within the study cities. The inset histograms illustrate the overall distributions. **a)** Economic Complexity Index **b)** Informality rate.

2.1.5 Mobility Patterns and Socioeconomic Characteristics

Mobility patterns also vary systematically with socioeconomic characteristics. U.S. cities generally feature longer commutes compared to their Latin American counterparts. However, disaggregating commutes by origin informality and work location complexity reveals an interesting pattern. In all four cities, workers from high-informality origins traveling to low-complexity destinations have the shortest median commutes, but those who face the longest median commutes varies between regions. In Mexico City and Rio de Janeiro, the longest commutes are typically made by workers from high-informality areas traveling to high-$ECI^{employment}$ destinations, indicating that vulnerable populations often endure longer commutes to reach better opportunities. In contrast, in the Bay Area and Los Angeles, the longest commutes are taken by workers from low-informality areas traveling to high-$ECI^{employment}$ destinations, reflecting the behavior of affluent, formally employed suburban commuters. These patterns highlight the importance of understanding how workers, faced with different conditions when choosing where to live and work, respond to varying needs and constraints in their work location decisions across different economies.

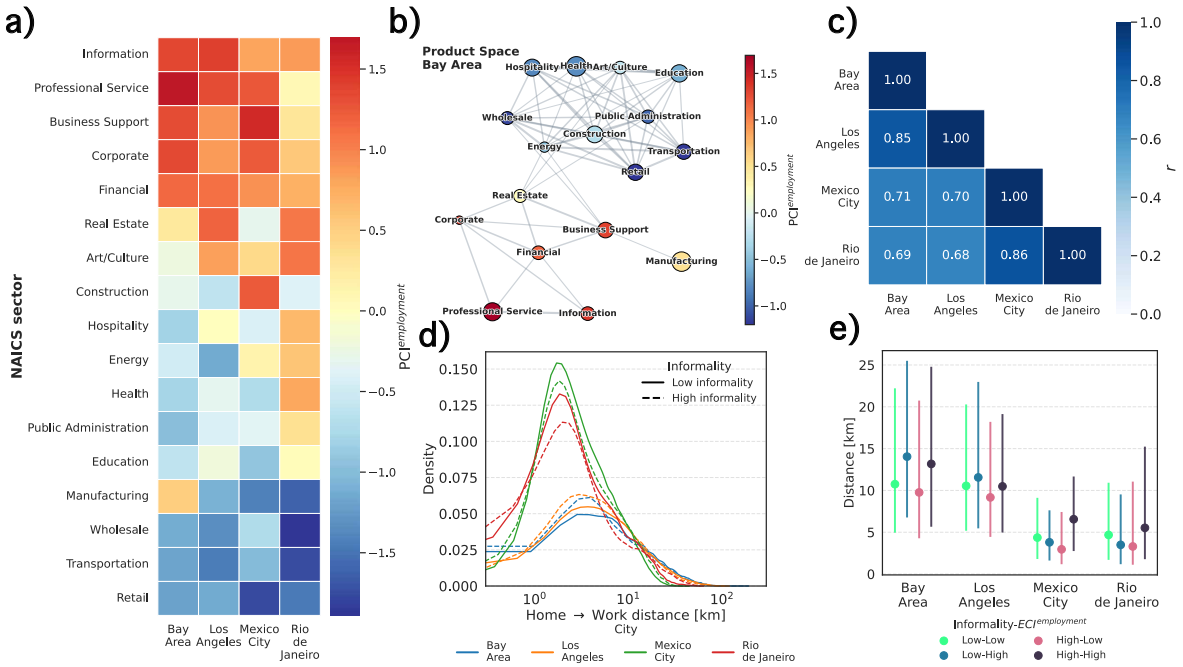

Fig. 2 Multi-scale structure of productive specialization and commuting in four cities of study. **a)** Product-complexity heat-map across 17 NAICS sectors and cities, with colors representing Product Complexity Index ($\text{PCI}^{\text{employment}}$). **b)** Product-space network for the Bay Area, where node size reflects workforce and color indicates $\text{PCI}^{\text{employment}}$; edges represent sector co-location probabilities. **c)** Correlation matrix of sector co-location patterns showing high within-region similarity (U.S. cities $r = 0.85$; Latin American cities $r = 0.86$) and lower cross-regional correlations ($r = 0.68$–0.71). **d)** Commuting distance distributions by origin informality level (solid = low, dashed = high), showing heavier right tails in U.S. cities. **e)** Median commuting distances across informality-$\text{ECI}^{\text{employment}}$ groups reveal contrasting patterns: longest commutes occur in high-informality/high-$\text{ECI}^{\text{employment}}$ areas in Latin America versus low-informality/high-$\text{ECI}^{\text{employment}}$ areas in the U.S., reflecting different socio-spatial segregation patterns.

2.2 Model Definition

To model how people choose where to work relative to where they live, we developed WorkReach, a discrete choice framework based on random utility models [15]. Utility represents the net benefit an individual derives from choosing a given alternative. It is usually described by $U_{pc} = V_{pc} + \varepsilon_{pc}$, where U_{pc} represents the utility of an individual p choosing alternative c, V_{pc} is the systematic utility (the measurable portion of it), and ε_{pc} is the unobserved part of the utility, i.e. the things that we cannot measure with our model. For our model, we assume commuters select a work location j from a set of available alternatives J to maximize their utility U_{ij}, where i is the home location.

The core of our model is the definition of our utility function V_{ij} associated with a worker from home location i choosing work location j:

$$V_{ij} = \beta_{\text{distance}} d_{ij} + w_{ij}(\tau, k) \cdot \left(\beta_{\text{ECI}} \text{ECI}_j^{\text{employment}} + \beta_{\text{informality}} Inf_i \right) , \tag{1}$$

where d_{ij} is the commuting distance, $\text{ECI}_j^{\text{employment}}$ is the Economic Complexity Index of destination j, and Inf_i is the informality rate at origin i. The parameters β_{distance}, β_{ECI}, and $\beta_{\text{informality}}$ are the marginal utilities (coefficients) to be estimated for distance, $\text{ECI}^{\text{employment}}$, and informality, respectively. A key part of this utility function is the transition weight function $w_{ij}(\tau, k)$, which represents perceived remoteness:

$$w_{ij}(\tau, k) = \frac{1}{1 + \exp\left(-k(d_{ij} - \tau)\right)} . \tag{2}$$

This transition weight based on a logistic function introduces a behavioral shift based on distance, as illustrated in Fig. 3b). The shape of this function as calibrated for each city is shown in Fig. S14. This transition is motivated by the previous result that showed that high-informality workers traveling to low-complexity areas generally have shorter commutes while longer commutes are associated to high-complexity areas, which can be interpreted as the work location choice has two regimes, one prioritizing local opportunities and another one prioritizing economic opportunities. In this transition function, τ is a distance threshold that indicates the point around which the work selection behavior transitions from a local to a economic opportunity, and k is a steepness parameter controlling the sharpness of this transition, both τ and k are also parameters to be estimated by our model. For short distances ($d_{ij} \ll \tau$), $w_{ij} \approx 0$, and utility is primarily driven by minimizing distance (convenience). For long distances ($d_{ij} \gg \tau$), $w_{ij} \approx 1$, and the attributes of $\text{ECI}^{\text{employment}}$ at the destination and informality at the origin exert their full influence on utility. The interaction term $w_{ij} \cdot \beta_{\text{informality}} Inf_i$ allows the socio-demographic characteristic Inf_i (which would otherwise be constant for all choices j from a given origin i) to vary across alternatives, allowing it to enter in the utility specification [15].

The probability P_{ij} that a worker from origin i chooses destination j is given by a multinomial logit (MNL) model:

$$P_{ij} = \frac{\exp(V_{ij})}{\sum_{j' \in J_i} \exp(V_{ij'})} , \tag{3}$$

where J_i is the set of all possible work locations considered from origin i. The model parameters $(\beta_{\text{distance}}, \beta_{\text{ECI}}, \beta_{\text{informality}}, \tau, k)$ are estimated by maximizing the log-likelihood of predicting the observed commuting flows T_{ij}, as shown in Fig. 3c). Prior to estimation, we discard origin-destination pairs with fewer than five observed flows and origins with fewer than ten total outgoing flows.

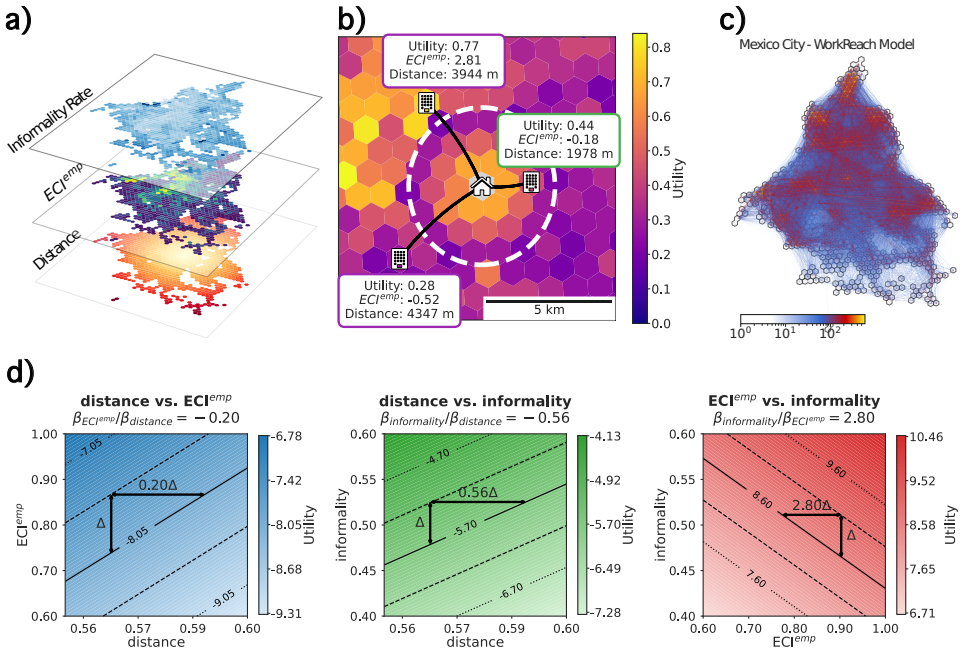

Fig. 3 Conceptual overview of the WorkReach model. **a)** Stacked data inputs: commuting distance, economic complexity ($\text{ECI}^{\text{employment}}$), and informality rate. Distance is illustrated for a given home location. **b)** Utility calculation showing distance threshold τ (white dashed circle) dividing choice set into near-by and far-away regimes. Beyond the threshold, transition weight $w_{ij}(\tau, k)$ modulates distance influence while utility incorporates $\text{ECI}^{\text{employment}}$ and informality effects. **c)** Predicted origin–destination commuting network for Mexico City, with edge intensity representing expected commuter flows. **d)** Iso-utility contours in three attribute planes (distance vs. $\text{ECI}^{\text{employment}}$; distance vs. informality; $\text{ECI}^{\text{employment}}$ vs. informality) showing trade-offs that maintain constant utility in the far-away regime ($w_{ij} = 1$).

2.3 Model Fit Results

2.3.1 Estimated Coefficients

To enhance both the stability of the model and the interpretability of the coefficients, we min-max scaled all three input variables (commute distance, $\text{ECI}^{\text{employment}}$, and informality rate) to the interval $[0, 1]$ (Section S1.2 and Figs. S9-S12). All covariates enter the utility function linearly, so each coefficient directly represents the utility change from moving a covariate across its full observed range.

The estimated coefficients of our calibrated model (Table 2) reveal different behavioral patterns across cities. We compute 95% confidence intervals using a row-wise multinomial bootstrap in which origin-level destination shares are resampled conditional on observed home populations and the model is re-estimated for each replicate. The distance coefficient (β_{distance}) is negative and relatively large in all four cities. This result indicates, as expected, that the utility decreases with distance. This coefficient

has its largest magnitude value in the Mexico City ($\beta_{\text{distance}} = -21.881$) and smallest in Los Angeles ($\beta_{\text{distance}} = -9.601$). The economic complexity coefficient (β_{ECI}) is positive in all four cities, indicating that higher-ECI$^{\text{employment}}$ work locations are systematically more attractive to commuters regardless of the region. Interestingly, the informality coefficient ($\beta_{\text{informality}}$) shows different regional patterns. In Mexico City and Rio de Janeiro, it has a positive value (10.250 and 4.267, respectively), consistent with workers in areas of high-informality traveling longer distances to find complex work opportunities. In contrast, the coefficient is negative in the Bay Area and Los Angeles, which indicates that high-informality areas in these cities are less prone to travel long distances for high-ECI$^{\text{employment}}$ work alternatives. This could be due more spatial integration of informal workers or the willingness of low informality workers to live far away from high economic complexity areas and closer to strictly residential zones. This regional contrast is consistent with high-ECI$^{\text{employment}}$ destinations offering a broader and denser set of employment opportunities, reducing matching costs for workers with fewer local options. In Latin American cities, where high-informality residential zones are spatially segregated from high-ECI$^{\text{employment}}$ employment hubs, workers from those zones are willing to overcome larger geographic barriers to access them (Section S4.1).

The estimated distance threshold τ marks the approximate distance at which the behavioral regime shifts from the local to the economic opportunity regime. In the U.S. cities, τ is strikingly small, suggesting that ECI$^{\text{employment}}$ and informality considerations become relevant almost immediately as distance increases. This likely reflects a combination of socioeconomic conditions and urban shape. In the U.S., better highway infrastructure and widespread access to private vehicles make longer commutes more feasible, while at the same time, there's a preference for single-family housing, often located far from dense work centers. In contrast, τ is substantially larger for Mexico City (16.201 km) and Rio de Janeiro (24.955 km), indicating a longer "convenience-first" regime before ECI$^{\text{employment}}$ and informality significantly weigh in. The steepness parameter k is higher for U.S. cities, especially the Bay Area (128.764), implying a very sharp transition, while it is more gradual for the Latin American cities.

2.3.2 Marginal Substitution Rates and Elasticities

The estimated coefficients allow us to derive interpretable metrics such as substitution rates and elasticities, which quantify how commuting flows vary with different work attributes.

Marginal Substitution Rates (MSR), are calculated as the ratio of the marginal utilities between variables (e.g., $(\partial V_{ij}/\partial X_1)/(\partial V_{ij}/\partial X_2)$), and they reveal the marginal rate at which individuals are willing to trade one attribute for another while maintaining the same level of utility. The actual derivation of the MSR for our model can be found in the methods section 4.2.2. For instance, the $\text{MSR}_{\text{ECI/distance}}$ indicates how much the negative effect of an additional unit of distance can be offset by an increase in ECI$^{\text{employment}}$. A negative value for this ratio, such as -0.193 for Mexico City shown in Table 2, depicts a positive trade-off: given that β_{distance} is negative and β_{ECI} is positive, this MSR means that for every unit increase Δ in ECI$^{\text{employment}}$, individuals are willing to tolerate an increase in distance equivalent to $0.193 \times \Delta$ to

Table 2 Optimized utility-model coefficients, substitution rates, and elasticities for the four cities.

	Bay Area	**Los Angeles**	**Mexico City**	**Rio de Janeiro**
Coefficients:				
β_{distance}	-21.292	-9.601	-21.881	-16.447
95% C.I.	[-21.298, -21.279]	[-9.607, -9.592]	[-21.899, -21.864]	[-16.454, -16.434]
β_{ECI}	2.462	0.482	4.220	6.603
95% C.I.	[2.460, 2.463]	[0.480, 0.484]	[4.211, 4.232]	[6.584, 6.620]
$\beta_{\text{informality}}$	-14.735	-8.702	10.250	4.267
95% C.I.	[-14.767, -14.702]	[-8.721, -8.692]	[10.235, 10.270]	[4.254, 4.284]
τ [km]	0.006	0.006	16.201	24.955
95% C.I.	[0.006, 0.006]	[0.006, 0.006]	[16.179, 16.217]	[24.911, 24.997]
k	128.764	32.413	10.824	9.654
95% C.I.	[128.761, 128.766]	[32.412, 32.416]	[10.799, 10.847]	[9.620, 9.686]
Marginal Rates of Substitution (MRS):				
$\text{MRS}_{\text{ECI/distance}}$	-0.116	-0.050	-0.193	-0.401
95% C.I.	[-0.116, -0.116]	[-0.050, -0.050]	[-0.194, -0.192]	[-0.402, -0.400]
$\text{MRS}_{\text{informality/distance}}$	0.692	0.906	-0.468	-0.259
95% C.I.	[0.690, 0.694]	[0.905, 0.909]	[-0.469, -0.468]	[-0.260, -0.259]
$\text{MRS}_{\text{informality/ECI}}$	-5.985	-18.056	2.429	0.646
95% C.I.	[-6.001, -5.972]	[-18.106, -17.998]	[2.421, 2.434]	[0.644, 0.649]
Elasticities (E):				
E_{distance}	-5.177	-3.238	-3.682	-2.616
95% C.I.	[-5.186, -5.167]	[-3.243, -3.233]	[-3.691, -3.674]	[-2.641, -2.595]
E_{ECI}	0.804	0.215	0.477	0.979
95% C.I.	[0.802, 0.805]	[0.214, 0.215]	[0.476, 0.478]	[0.969, 0.988]
$E_{\text{informality}}$	-2.305	-1.917	2.222	0.666
95% C.I.	[-2.312, -2.296]	[-1.924, -1.911]	[2.219, 2.226]	[0.659, 0.673]

keep the same level of utility. This effectively demonstrates that workers are willing to travel farther for jobs in locations with higher economic complexity. Our results show Rio de Janeiro exhibiting the highest willingness to trade distance for $\text{ECI}^{\text{employment}}$, while Los Angeles displays the lowest among the studied cities.

The interpretation of the $\text{MSR}_{\text{informality/distance}}$ is shaped by the different signs of $\beta_{\text{informality}}$ across regions. For U.S. cities, the positive MSR (0.692 for the Bay Area; 0.906 for Los Angeles), combined with a negative β_{distance} and a negative $\beta_{\text{informality}}$, imply that residing in an area with higher informality amplifies the negative effect of distance when considering distant jobs. In contrast, for Latin American cities, the negative ratios (-0.468 for Mexico City; -0.259 for Rio de Janeiro), coupled with a positive $\beta_{\text{informality}}$, suggest that higher informality in the origin somewhat mitigates the utility cost associated with traveling longer distances to pursue jobs in high-$\text{ECI}^{\text{employment}}$ locations.

While substitution rates describe trade-offs in utility terms, it is also useful to understand how a change in an attribute directly affects the probability of a specific work location being chosen. For this, we use elasticities (E). An elasticity measures the

percentage change in the probability of choosing a particular alternative in response to a 1% change in one of its attributes, holding all else constant (Table 2). In simpler terms, it tells us how sensitive our choice probabilities are to small changes in factors like distance, $\text{ECI}^{\text{employment}}$, or origin informality. A key property of elasticities in multinomial logit models is that a change in an attribute of one alternative affects not only its own choice probability, but also the probabilities of all other alternatives in the choice set, as the sum of probabilities must always equal one. The direct elasticity of the probability of choosing an alternative j from origin i (P_{ij}) with respect to an attribute X_{ij} (an attribute X specific to the alternative j for an individual from i) is given by $E_{X_{ij}}^{P_{ij}} = (\partial V_{ij}/\partial X_{ij}) X_{ij}(1 - P_{ij})$.

Across all cities, commuting distance exhibits the largest (negative) elasticity. For example, in Mexico City, $E_{\text{distance}} = -3.682$. This means that for a worker from a given origin i, if one potential work location j is 1% farther away than an otherwise identical work location j', the probability of choosing j is approximately 3.682% lower than choosing j'. The $\text{ECI}^{\text{employment}}$ term generally shows a smaller, positive elasticity, indicating that choice probabilities increase with higher $\text{ECI}^{\text{employment}}$, but at a proportionally lower rate than the change in the $\text{ECI}^{\text{employment}}$ term itself. The elasticity related to origin informality demonstrates considerable variation with a negative sign for U.S. cities and positive for Latin American cities. This highlights that the socio-demographic composition of a worker's home location, through its interaction with perceived remoteness, has a substantial and directionally distinct impact on average choice probabilities. It is important to recall that for $\text{ECI}^{\text{employment}}$ and informality, the "attribute" X in the elasticity formula effectively corresponds to the interacted terms $w_{ij}\text{ECI}_j^{\text{employment}}$ and $w_{ij}Inf_i$, respectively, as these are the forms in which they directly influence utility in our model, the complete formulation is described in the Methods section 4.2.2.

2.3.3 Flow Estimation and Model Performance

A key application of our utility-based discrete choice model, which estimates the probability that an individual of a given origin will select any particular location for work, is its ability to predict aggregate commute flows between locations. This predictive power is valuable not only for validating the model's behavioral assumptions but also for estimating trips in contexts where empirical data may be incomplete or unavailable. To evaluate this predictive capacity, we compare the commute flows generated by our WorkReach model with the observed flows. Furthermore, we benchmark its performance against three established spatial interaction models: the gravity model with a power-law distance decay function, the extended version of the radiation model [5], and a recently proposed Bayesian Machine Scientist (BMS) gravity-like model in its plausible version [13].

To quantify this performance, we used two standard metrics. The first is the Pearson correlation coefficient (r), which assesses the linear association between the predicted and observed flows. The second metric, which is widely used in mobility studies, is the Common Part of Commuters (CPC) [7, 44]. The CPC measures the degree of overlap between the observed and predicted flow distributions across all origin-destination pairs. Then it can be interpreted as the fraction of commuters whose

trips are correctly allocated by the model to the observed destinations, a mathematical description is described in the Methods section 4.4.1. Similarly to r, the CPC is a number between 0 and 1, 1 being a perfect match, while a value 0 means that there is no overlap between the predicted and observed flows.

Our WorkReach model achieves CPC values that are comparable with those of the benchmark models, and in some cases as for the Bay Area, our model yields the highest CPC (0.389). In Los Angeles and Mexico City, the CPC values are closely aligned with the top-performing gravity models. For Rio de Janeiro, the CPC (0.614) is slightly lower than the other alternatives but still yields a high value. The Pearson correlations have a similar pattern, with our WorkReach model consistently achieving high correlation coefficients. These findings demonstrate that our behaviorally-grounded model, which uniquely incorporates $\text{ECI}^{\text{employment}}$ and informality through an interactive utility structure, can predict commuting flows with an accuracy comparable to established spatial interaction models (Figs. S15-S18). The primary advantage of our approach, therefore, lies not solely in its predictive accuracy but in the enhanced interpretability and deeper understanding it offers regarding the socioeconomic drivers shaping commuting choices. A stratified 5-fold cross-validation on origin zones confirms that these results generalize out of sample: the train–test performance gap is negligible for all models ($\Delta\text{CPC} \leq 0.001$; SI Section S3.5).

Beyond overall fit, we audited WorkReach's predictive fairness by computing the CPC across $\text{ECI}^{\text{employment}}$ and Informality Rate bins (Figs. S19-S20). While CPC remains relatively high across most strata, we observe a drop in the highest informality bins, especially in the Bay Area and Mexico City. This drop may reflect the fact that WorkReach estimates flows at the population level without group-specific optimization, unlike fairness-aware models such as FairMobi-Net[45], which incorporate a fairness-loss term to explicitly address such disparities. However, this audit reveals where future Fairness Enhancement of WorkReach could be most impactful.

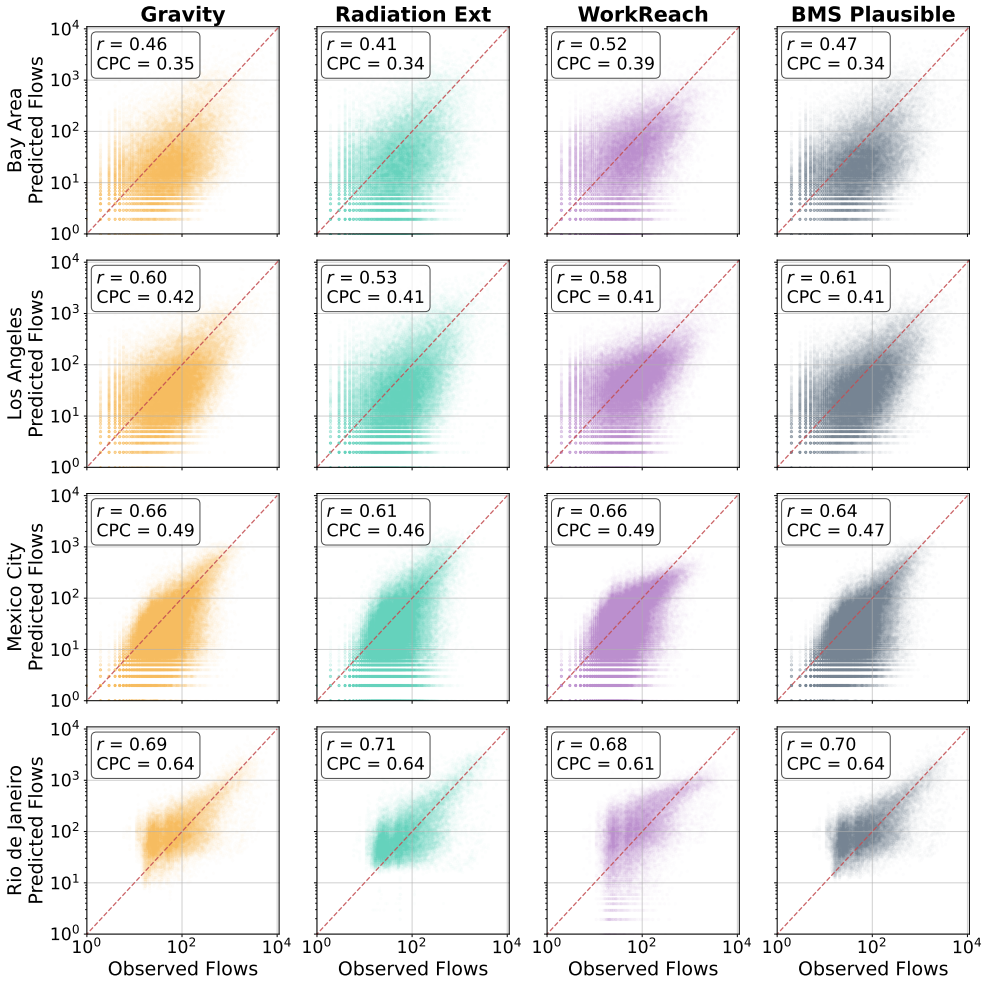

Fig. 4 Scatterplot comparison of model performance. Observed flows (x-axis) versus predicted flows (y-axis) for four models across the four study cities, on a log-log scale. The models are: Gravity (power-law distance decay), Radiation Extended, the WorkReach model, and the BMS Plausible model. Values for Pearson Correlation Coefficient (r) and Common Part of Commuters (CPC) are displayed for each city and model. The WorkReach model demonstrates comparable predictive accuracy to established benchmark models across all cities.

The economic complexity framework also allows us to measure how similar the employment structures of two zones are through a skill-relatedness score, which quantifies the overlap in sector composition between a worker's home zone and a potential destination (Section S5). An alternative WorkReach specification that replaces $\text{ECI}^{\text{employment}}$ with this score reveals that commuters tend to travel to zones offering structurally different employment from what is locally available, consistent with a spatial mismatch interpretation. We retain $\text{ECI}^{\text{employment}}$ in the main specification

because, unlike skill relatedness, it is an intrinsic property of a destination zone and can be paired with the feasibility score [17, 46] to simulate how targeted economic development would affect commuting accessibility (Section S8).

2.3.4 Accessibility

Our calibrated discrete choice model allows us to compute accessibility measures that go beyond a physical proximity measure by incorporating the trade-offs and preferences revealed by commuter behavior. In this study, we use both types of measures: a more conventional distance-weighted accessibility and a consumer-surplus accessibility, also known as logsum accessibility [34, 35]. The latter, defined in terms of random utility theory, provides an assessment of the overall attractiveness of the whole work market as perceived by the origin locations by considering the utility derived from all available employment choices.

Specifically, for an origin i, distance-weighted accessibility (A_i^{DW}) is calculated as the sum of predicted flows T_{ij} to each destination j, weighted by the inverse of distance d_{ij}, and normalized by the total outflow from i:

$$A_i^{DW} = \frac{\sum_j T_{ij} \cdot (1/d_{ij})}{\sum_j T_{ij}} \tag{4}$$

Consumer-surplus accessibility (A_i^{CS}), representing the expected maximum utility from the choice set at origin i, is given by the logsum formulation:

$$A_i^{CS} = \ln \sum_{j \in J_i} \exp(V_{ij}) \tag{5}$$

where V_{ij} is the systematic utility of choosing work location j from origin i, and J_i is the set of all available work locations from i. A detailed derivation of these measures is provided in the Methods section.

Figure 5 compares two accessibility metrics across locations by informality rate (above or below the city median). In Fig. 5**a**, U.S. cities show higher median distance-weighted accessibility for high-informality neighborhoods. In contrast, in Mexico City and Rio de Janeiro, low-informality areas enjoy greater distance-weighted accessibility. This pattern reflects the peripheralization of informal settlements in Latin American metros, which forces residents to live farther from employment centers [2, 33].

Fig. 5**b** shows consumer-surplus accessibility, which measures the net benefit of available work choices by incorporating distance, destination $\text{ECI}^{\text{employment}}$, and origin informality. Here, median consumer-surplus accessibility is consistently lower for high-informality origins in all four cities. In the U.S. cities, this reverses the "distance advantage" observed in Fig. 5**a**: once work quality and informality effects enter utility, high-informality neighborhoods lose their apparent edge. In Mexico City and Rio de Janeiro, the positive informality coefficient ($\beta_{\text{informality}} > 0$) partially offsets raw distance penalties, but high-informality origins still record lower overall surplus. In other words, any benefit from informality's interaction with perceived remoteness does

not fully compensate for longer distances and typically lower $\text{ECI}^{\text{employment}}$ levels of accessible jobs.

These results highlight the value of utility-based accessibility in urban analysis, but no single metric fits every purpose: relying solely on distance-weighted measures can mask quality differences, while using only consumer-surplus might understate disparities faced by informal workers who must accept distant, low-quality jobs for lack of alternatives. A combined view delivers a more behaviorally grounded assessment, revealing how uneven spatial distributions of homes and jobs create socioeconomic inequities in urban labor markets.

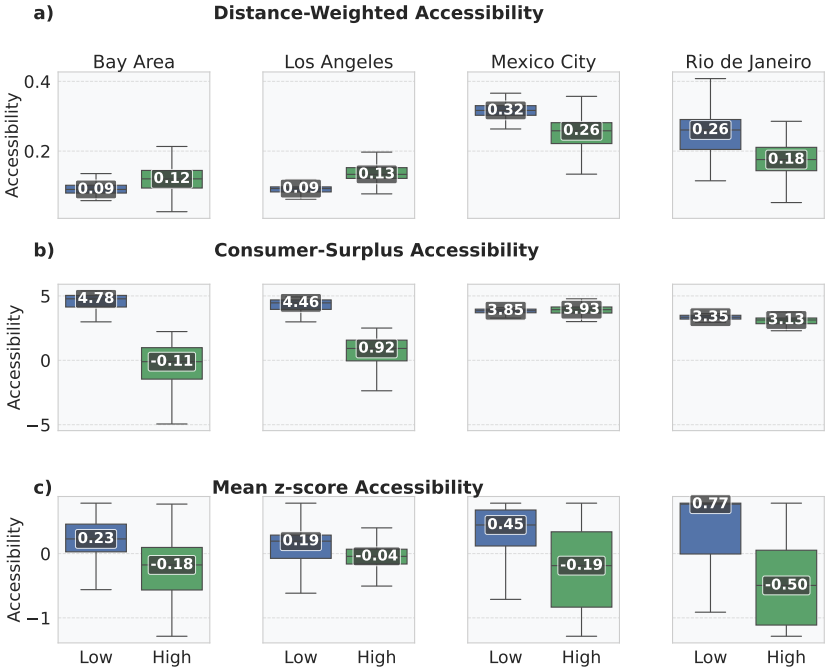

Fig. 5 Accessibility measures by origin informality group. Origins are classified as "low" or "high" informality based on whether their informality rate falls below or above the city-specific median. Numbers indicate median values. **a)** Distance-weighted accessibility boxplots showing higher median values for high-informality areas in U.S. cities but lower values (distance penalty) in Latin American cities. **b)** Consumer-surplus accessibility (expected maximum utility) lower for high-informality areas in three of four cities; Mexico City is the exception, where the positive $\beta_{\text{informality}}$ partially offsets distance penalties. **c)** Combined accessibility index computed as the mean of the z-scored distance-weighted and consumer-surplus measures. High-informality origins have lower combined accessibility in all four cities.

Because only relative utilities matter (absolute levels can even be negative), we standardize both metrics using within-city z-scores and compute a combined index as

their mean (Fig. 5**c**). This combined measure captures the overall accessibility level by averaging the contributions of physical proximity and utility-based attractiveness, so that areas scoring low on both dimensions are identified as the most underserved. Figure 6 maps these z-scores and their combined index. Across the four cities, peripheral locations, especially the western zones of Rio de Janeiro, emerge as the most underserved.

Correlation analysis between informality rates and accessibility measures shows regional differences (Fig. S33). For distance-weighted accessibility, correlations are positive in U.S. cities, meaning that higher informality areas have better distance-based accessibility, while this is negative in Latin American cities, reflecting the peripheralization of higher informal rates away from complex work hubs. Consumer-surplus accessibility, on the other hand, shows negative correlation with informality in all cities but Mexico City, which suggests that the perceived advantage of pursuing high complexity jobs increasingly outweighs the distance penalty for workers in more informal areas. The combined accessibility measure (mean z-score) shows decreasing accessibility with increasing informality in all cities, confirming that high-informality origins face systematic accessibility disadvantages when both dimensions are considered jointly.

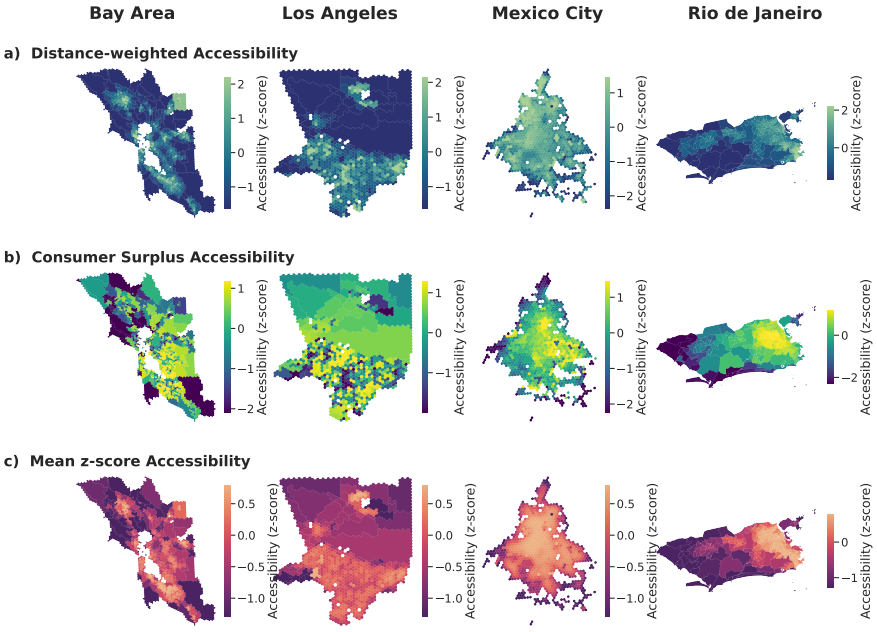

Fig. 6 Spatial distribution of accessibility measures (z-scores). **a)** Distance-weighted accessibility z-scores standardized by city for cross-regional comparison. **b)** Consumer-surplus accessibility z-scores, with discrepancies from a) highlighting impacts of work quality ($ECI^{employment}$) and origin informality beyond distance. **c)** Combined accessibility index (mean of z-scored distance-weighted and consumer-surplus measures), representing the overall accessibility level.

3 Discussion

Understanding commuter behavior is essential for developing urban policies that reduce socioeconomic inequality. Incorporating economic complexity ($ECI^{employment}$) and informality into an interpretable model based on discrete choice theory shows how these factors, alongside distance, shape home–work choices in ways that differ across cities. This broadens the concept of accessibility from a purely spatial measure to one that also reflects perceived economic benefit.

Our results indicate that there is a consistent preference across the four cities studied for economically sophisticated areas, while commuter demographics, represented by the informality rate in our model, vary in behavior across urban contexts. Workers from high-informality neighborhoods in Mexico City and Rio de Janeiro travel farther to reach economically complex job centers, overcoming geographic and socioeconomic barriers. In contrast, in the Bay Area and Los Angeles, workers from low-informality

areas are more likely to travel greater distances, implying that informality in these contexts is more related to social or institutional constraints than to spatial accessibility challenges. WorkReach captures these context-specific dynamics while matching the predictive performance of established parsimonious models, yet gains interpretability by deriving utility coefficients, marginal substitution rates, and elasticities. Moreover, a critical insight from our accessibility analysis is that physical proximity to employment and meaningful access to quality opportunities are fundamentally different. In cases where distance-weighted accessibility presents advantages for high-informality populations, consumer-surplus accessibility reveals the opposite pattern. This exposes that simply being near employment centers does not guarantee access to jobs that maximize worker utility. This distinction highlights that socioeconomic background and job quality matter just as much as physical proximity in determining access to meaningful economic opportunities, revealing disparities that conventional distance-based metrics overlook.

$\text{ECI}^{\text{employment}}$ and residential informality are empirically distinct: their zone-level correlations are near-zero in the U.S. cities and moderate in the Latin American cities ($r \approx -0.40$), confirming that each captures a different spatial dimension (Section S4.1). The positive $\beta_{\text{informality}}$ in Latin American cities is consistent with the spatial structure of these metropolitan areas, where high-informality residential zones tend to be peripheral and offer fewer local employment options. In addition, the employment data used to construct $\text{ECI}^{\text{employment}}$ in these cities records formally registered establishments, so working in a high-$\text{ECI}^{\text{employment}}$ zone tends to carry social security affiliation and statutory benefits even for non-specialized roles, providing an additional incentive for workers from high-informality areas to overcome geographic barriers.

Together, the interpretable parameters and the accessibility metric allow WorkReach to generate policy-relevant insights that other models cannot provide, and illustrate the importance of customizing interventions to local contexts. In Latin American cities, improving connectivity between peripheral, high-informality neighborhoods to central high-$\text{ECI}^{\text{employment}}$ areas may help expand economic opportunities. However, these measures need to be paired with strategies to incentivize local employment, especially formal jobs, and improve local economic complexity to avoid merely shifting burdens onto workers by making them commute long distances. In contrast, in US cities, more effective interventions might be those focused on targeted skill development and reducing discriminatory barriers, as there the physical distance barrier is lower for workers in the informal sector. Our results can directly inform infrastructure investment, as areas with high economic complexity but limited accessibility for disadvantaged groups represent good candidates for the improvement of public transportation. In contrast, regions characterized by high physical accessibility and low economic complexity offer good opportunities for focused economic development programs.

Our analysis has several limitations. The estimated coefficients reflect systematic associations in cross-sectional commuting flows, not experimentally identified preferences; quasi-experimental or longitudinal designs would be needed to establish causal links between changes in $\text{ECI}^{\text{employment}}$ and shifts in commuting behavior. Relatedly, our framework treats residence as fixed and models only workplace choice, a standard

assumption supported by the higher frequency of job transitions relative to residential moves in both the United States [47] and Latin America [48], but one that does not capture the joint residential-workplace decision. Finally, $\text{ECI}^{\text{employment}}$ remains an aggregate index: decomposing which agglomeration channels most strongly attract commuters is an important direction for future work. Data resolution limitations and inconsistencies in measuring informality across regions are difficult to overcome. Standardizing informality metrics and economic complexity indices represents a key area for improvement. Future research should also expand this analysis to additional urban settings, particularly in regions with different economic structures. Longitudinal studies would be beneficial for evaluating the causal effects of particular interventions over time. Furthermore, incorporating additional socioeconomic variables could deepen our understanding of the mechanisms driving urban labor market dynamics.

4 Methods

4.1 Conceptual and Theoretical Background

The study of how individuals choose where to live and work is grounded in economic theories of spatial interaction and individual decision-making. We build upon the random utility maximization (RUM) framework [49], which assumes that individuals choose the alternative (in this case, a work location) that offers the highest utility from a set of available options. Utility is typically a function of attributes of the alternatives and characteristics of the decision-maker. Our work extends this by explicitly incorporating measures of local economic structure ($\text{ECI}^{\text{employment}}$) and labor market conditions (informality) into the utility function, alongside traditional factors like travel distance.

4.2 Economic Complexity Index ($\text{ECI}^{\text{employment}}$)

Economic Complexity provides a quantitative framework to analyze the productive structure of geographic regions by assessing the diversity and sophistication of their economic activities [16]. Central to this approach are two interrelated metrics: the Economic Complexity Index ($\text{ECI}^{\text{employment}}$) for zones (our sub-city zones), and the Product Complexity Index ($\text{PCI}^{\text{employment}}$) for economic activities (here, NAICS industry sectors, treated as "products" in line with the ECI literature [17]).

The calculation begins by building a binary matrix M_{zs} indicating whether a zone z is specialized in sector s. Specialization is determined using the Revealed Comparative Advantage (RCA):

$$RCA_{zs} = \frac{W_{zs} / \sum_{s'} W_{zs'}}{\sum_{z'} W_{z's} / \sum_{z',s'} W_{z's'}},$$

where W_{zs} is the employment in sector s within zone z. If $RCA_{zs} \geq 1$, then $M_{zs} = 1$, indicating specialization; otherwise, $M_{zs} = 0$. This would mean that the proportion of employed persons in sector s in zone z to all of the employed individuals in that given zone is larger than the same proportion for the whole city.

$$\tilde{K}_z^0 = \sum_s M_{zs}$$

$$\tilde{K}_s^0 = \sum_z M_{zs}$$

$$\mathrm{ECI}_z^{\mathrm{employment},\,(N)} = K_z^N = \frac{1}{\tilde{K}_z^{N-1}} \sum_s M_{zs} K_s^{N-1}$$

$$\mathrm{PCI}_s^{\mathrm{employment},\,(N)} = K_s^N = \frac{1}{\tilde{K}_s^{N-1}} \sum_z M_{zs} K_z^{N-1}$$

This iterative process (the method of reflections) starts with $K_z^0 = \sum_s M_{zs}$ (diversity of zone z) and $K_s^0 = \sum_z M_{zs}$ (ubiquity of sector s). The iterations ($N > 0$) refine these initial measures by averaging the complexity of the connected nodes in the bipartite network. The $\mathrm{ECI}^{\mathrm{employment}}$ is typically taken as the second eigenvector of a matrix derived from M_{zs}, which this iterative process converges to. Formally, this eigenvector corresponds to the largest non-trivial eigenvalue of the projection of the bipartite zone-sector network onto the zone space [16]. Its entries partition zones into two groups according to their connectivity patterns in this projected network, analogous to a spectral bisection of the graph: zones with similar entries share denser connections to zones with similar sector structure. Because the eigenvector is defined only up to a sign, its direction is fixed following the convention of [16] by choosing the orientation in which the eigenvector entries correlate positively with diversity. The resulting $\mathrm{ECI}^{\mathrm{employment}}$ values are standardized for use in the WorkReach model. We use employment data from DENUE (Mexico City), RAIS (Rio de Janeiro), and Replica (U.S. cities), harmonized to NAICS sectors, aggregated to H3 hexagonal zones for consistent $\mathrm{ECI}^{\mathrm{employment}}$ calculation across cities. All calculations of the Economic Complexity framework were done via the econci Python package [50].

4.2.1 Discrete Choice Model Specification

The work location choice process is modeled using a multinomial logit (MNL) framework. The utility V_{ij} that a worker residing in origin zone i derives from choosing a work in destination zone j is given by Eq. 1:

$$V_{ij} = \beta_{\mathrm{distance}} d_{ij} + w_{ij}(\tau, k) \cdot \left(\beta_{\mathrm{ECI}} \mathrm{ECI}_j^{\mathrm{employment}} + \beta_{\mathrm{informality}} Inf_i \right) ,$$

where d_{ij} is the min-max standardized Haversine distance between i and j. $\mathrm{ECI}_j^{\mathrm{employment}}$ is the min-max standardized Economic Complexity Index of destination j. Inf_i is the min-max standardized informality rate at origin i. All three variables are min-max scaled to $[0, 1]$ directly, placing them on a comparable scale and improving the efficiency and convergence of the estimation procedure.

The transition weight $w_{ij}(\tau, k)$ is defined by Eq. 2:

$$w_{ij}(\tau, k) = \frac{1}{1 + \exp\left(-k(d_{ij} - \tau)\right)} .$$

Here, d_{ij} is the standardized distance, used specifically for the threshold parameter τ. k is a steepness parameter. Crucially, attributes of the decision-maker (or their origin, like Inf_i) that do not vary over alternatives can only enter the WorkReach model if they are specified in ways that create differences in utility over alternatives [15]. The interaction term $w_{ij}(\tau, k) \cdot \beta_{\text{informality}} Inf_i$ achieves this: because w_{ij} is a function of d_{ij} (an alternative-specific attribute, i.e., distance to work j), the impact of origin informality Inf_i on utility now varies across the different work alternatives j available from origin i. This makes the socio-demographic variable Inf_i identifiable and interpretable within the choice model framework.

The probability P_{ij} of choosing j from i is given by the standard MNL formula (Eq. 3):

$$P_{ij} = \frac{\exp(V_{ij})}{\sum_{j' \in J_i} \exp(V_{ij'})} .$$

The model parameters $\theta = (\beta_{\text{distance}}, \beta_{\text{ECI}}, \beta_{\text{informality}}, \tau, k)$ are estimated by maximizing the log-likelihood function $\mathcal{L}(\theta) = \sum_i \sum_j T_{ij}^{obs} \log P_{ij}$, where T_{ij}^{obs} is the observed number of commuters from i to j. Numerical stability is ensured by subtracting the maximum utility within each choice set before exponentiation.

In order to obtain the model's commuter flows, we multiply the probability of observing a flow going from i to j to the total population from the origin and round the result to the nearest integer:

$$T_{ij} = \text{round}\left(m_{ij} \cdot P_{ij}\right) .$$

4.2.2 Iso-Utility Curves, Substitution Rates, and Elasticities

WorkReach stands out in terms of interpretability and as a discrete choice based model allows for the quantification of trade-offs between work attributes. This framework enables the derivation of key behavioral measures that reveal how workers value and substitute between different work characteristics.

Iso-utility curves represent the combination of attribute levels that maintain a same level of systematic utility V^*. In general, given two attributes X and Y with coefficients β_X and β_Y, the iso-utility condition is $\beta_X X + \beta_Y Y = C$, where C is a constant. For example, in the "far-away" regime where the transition weight $w_{ij}(\tau, k) \approx 1$ (i.e., for distant choices where socioeconomic effects are fully active), and holding informality constant, the iso-utility condition for distance and $\text{ECI}^{\text{employment}}$ approximates $\beta_{\text{distance}} d_{ij} + \beta_{\text{ECI}} \text{ECI}_j^{\text{employment}} = C$. This equation describes how an increase in distance can be compensated for by a relative increase in $\text{ECI}^{\text{employment}}$ in order to maintain the same utility level.

We can measure these trade-off more generally using Marginal Rates of Substitution (MRS) that are the rate at which individuals are willing to exchange one attribute

for another while maintaining constant utility. The MRS between any two attributes is defined as the ratio of their marginal utilities: $MRS_{X,Y} = (\partial V/\partial Y)/(\partial V/\partial X)$.

For attributes modulated by the transition weight ($\text{ECI}^{\text{employment}}$ and informality), the marginal utility takes the form:

$$\frac{\partial V_{ij}}{\partial X_j} = w_{ij}\beta_X \,.$$

For distance, the marginal utility is more complex because distance appears both directly in the utility function and indirectly through the transition weight w_{ij}. The transition weight $w_{ij}(\tau, k) = \frac{1}{1+\exp\left(-k(d_{ij}-\tau)\right)}$, has the derivative with respect to distance:

$$\frac{\partial w_{ij}}{\partial d_{ij}} = k \cdot w_{ij}(1 - w_{ij}) \,.$$

Applying the chain rule to the full utility function, the marginal utility of distance becomes:

$$\frac{\partial V_{ij}}{\partial d_{ij}} = \beta_{\text{distance}} + \frac{\partial w_{ij}}{\partial d_{ij}}(\beta_{\text{ECI}}\text{ECI}_j^{\text{employment}} + \beta_{\text{informality}} Inf_i) \,. \tag{6}$$

This derivative has a direct component (β_{distance}) and an indirect component that captures how distance modulates the importance of socioeconomic factors. Then, the complete MRS between distance and any modulated attribute X takes the form:

$$MRS_{d,X} = -\frac{w_{ij}\beta_X}{\beta_{\text{distance}} + k \cdot w_{ij}(1 - w_{ij})(\beta_{\text{ECI}}\text{ECI}_j^{\text{employment}} + \beta_{\text{informality}} Inf_i)} \,. \tag{7}$$

This expression reveals that trade-offs are not constant across space but vary with distance due to the transition weight effects. For choices far from the distance threshold τ (where $w_{ij} \approx 1$ and $\partial w_{ij}/\partial d_{ij} \approx 0$), the MRS simplifies to the traditional ratio $\beta_X/\beta_{\text{distance}}$.

Elasticities provide a complementary measure of sensitivity, representing the percentage change in choice probability for a 1% change in an attribute, ceteris paribus. For any attribute X, the elasticity of the probability of choosing alternative j is given by:

$$E_X^{P_{ij}} = \frac{\partial V_{ij}}{\partial X} X(1 - P_{ij}) \,, \tag{8}$$

where $\partial V_{ij}/\partial X$ is the marginal utility of the attribute and P_{ij} represents the probability of the pair ij from the multinomial logit model [15].

In our model, elasticities are computed differently for each type of attribute:

- For transition-weight modulated attributes ($\text{ECI}^{\text{employment}}$ and informality), the marginal utility is $w_{ij}\beta_X$, then:

$$E_X^{P_{ij}} = w_{ij}\beta_X X(1 - P_{ij}) \,. \tag{9}$$

- For distance, the elasticity uses the full marginal utility including both direct and indirect effects:

$$E_{d_{ij}}^{P_{ij}} = \Big(\beta_{\text{distance}} + kw_{ij}(1-w_{ij})(\beta_{\text{ECI}}\text{ECI}_j^{\text{employment}} + \beta_{\text{informality}} Inf_i)\Big)\, d_{ij}(1-P_{ij})\,. \tag{10}$$

4.3 Benchmark models

For flow generation accuracy we compare our WorkReach model against three established models in their singly constrained versions to meet the outflow, i.e. $\sum_k T_{ik} = T_i^{tot}$. All of the models were directly implemented in Python using the SciPy library for parameter optimization [51].

- **Gravity Model** (four parameters): The estimated flow T_{ij} from origin i (with population m_i and total observed outflow T_i^{obs}) to destination j (with jobs m_j) is modeled using a singly-constrained (production-constrained) formulation with a power-law distance decay. Formally:

$$T_{ij} = T_i^{obs} \frac{C m_i^{\alpha} m_j^{\beta} / d_{ij}^{\gamma}}{\sum_{l \neq i} (C m_i^{\alpha} m_l^{\beta} / d_{il}^{\gamma})}\,, \tag{11}$$

  where d_{ij} is the distance. The parameters C (scaling constant), α (origin mass exponent), β (destination mass exponent), and γ (distance decay exponent) are estimated by minimizing the negative log-likelihood of observed flows, the same method is used for the optimization of the parameters in all 4 models [52].
- **Extended Radiation Model** (one free parameter): The original radiation model [10] is parameter-free and predicts flows from the spatial distribution of population alone. The extended version introduces a single free parameter α that calibrates the decay of opportunity influence [5]. The predicted flows are given by:

$$T_{ij} \;=\; \gamma\, m_i \frac{P(1 \mid m_i, m_j, a_{ij})}{\sum_k P(1 \mid m_i, m_k, a_{ik})} \tag{12}$$

  where

$$P(1 \mid m_i, m_j, a_{ij}) \;=\; \frac{\left[(a_{ij}+m_j)^{\alpha} - a_{ij}^{\alpha}\right](m_i^{\alpha}+1)}{(a_{ij}^{\alpha}+1)\left[(a_{ij}+m_j)^{\alpha}+1\right]}\,, \tag{13}$$

  and

$$a_{ij} \equiv m_i + s_{ij}\,.$$

  γ being a normalization constant. This extended version lets the decay of opportunity influence be calibrated, improving fit to empirical flows by capturing regional heterogeneity in commuters' sensitivity to intervening opportunities [5, 10].
- **BMS Plausible Gravity** (six parameters): A gravity-like model variant derived from Bayesian symbolic regression [13]. The predicted flow T_{ij} from origin i (with

mass m_i and total observed outflow T_i^{obs}) to destination j (with mass m_j) is:

$$T_{ij} = T_i^{obs} \frac{A\left(1 + B\frac{m_i m_j + C m_j + D}{d_{ij}^{\alpha}}\right)^{\gamma}}{\sum_{k \neq i} A\left(1 + B\frac{m_i m_k + C m_k + D}{d_{ik}^{\alpha}}\right)^{\gamma}}, \tag{14}$$

where A, B, C, D, α, and γ are parameters estimated.

4.4 Parameter Estimation and Optimization

The parameters for all presented models (Utility, Gravity, and BMS Plausible Gravity) are estimated by maximizing the likelihood of observing the empirical commuting flows. This is equivalent to minimizing the negative log-likelihood of the data given the model and its parameters.

Let T_{ij}^{obs} be the observed flow of commuters from origin i to destination j, and let $P_{ij}(\theta)$ be the probability of a commuter from i choosing j, or the predicted proportion of flow from i to j, as defined by a specific model with parameter vector θ. For models predicting flow proportions, the likelihood of observing the set of flows from a given origin i, $T_{i*} = \{T_{i1}^{obs}, T_{i2}^{obs}, ..., T_{iJ}^{obs}\}$, given a total outflow $O_i = \sum_j T_{ij}^{obs}$, follows a multinomial distribution. The log-likelihood for a single origin i is:

$$\mathcal{L}_i(\theta | T_{i*}) = \sum_j T_{ij}^{obs} \log P_{ij}(\theta)\,. \tag{15}$$

The total log-likelihood for the entire dataset is the sum over all origins:

$$\mathcal{L}(\theta | \mathbf{T}^{obs}) = \sum_i \mathcal{L}_i(\theta | T_{i*}) = \sum_i \sum_j T_{ij}^{obs} \log P_{ij}(\theta)\,, \tag{16}$$

where $\mathbf{T}^{obs}$ represents the matrix of all observed flows.

The objective is to find the parameter vector $\hat{\theta}$ that maximizes $\mathcal{L}(\theta | \mathbf{T}^{obs})$, or equivalently, minimizes $-\mathcal{L}(\theta | \mathbf{T}^{obs})$.

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta | \mathbf{T}^{obs}) = \arg\min_{\theta} [-\mathcal{L}(\theta | \mathbf{T}^{obs})]\,. \tag{17}$$

4.4.1 Evaluation Metrics

Model performance is assessed using:

- **Common Part of Commuters (CPC)**: $CPC = \frac{2\sum_i \sum_j \min(T_{ij}^{obs}, T_{ij}^{pred})}{\sum_i \sum_j (T_{ij}^{obs} + T_{ij}^{pred})}$. Measures the proportion of overlapping flows between observed and predicted flows [7].
- **Pearson Correlation (r)**: $r = \frac{\sum_{ij}(T_{ij}^{obs} - \bar{T}^{obs})(T_{ij}^{pred} - \bar{T}^{pred})}{\sqrt{\sum_{ij}(T_{ij}^{obs} - \bar{T}^{obs})^2 \sum_{ij}(T_{ij}^{pred} - \bar{T}^{pred})^2}}$. Measures the strength of the linear relationship between observed and predicted flows.

4.5 Accessibility Measures

We compute two main accessibility measures:

1. **Distance-Weighted Accessibility**: This accessibility measure quantifies the difficulty of reaching job locations from a given origin i, considering the number of predicted flows and the impedance given by distance. This is, the sum of predicted flows T_{ij} to each destination j, weighted by the inverse of the distance d_{ij} to that destination, and then normalized by the total predicted outflow from i. Formally:

$$A_i^{DW} = \frac{\sum_j T_{ij}(1/d_{ij})}{\sum_j T_{ij}}, \tag{18}$$

   where T_{ij} are the predicted commuting flows from origin i to destination j, and d_{ij} is the distance between them (a minimum distance of 0.1 km is enforced to prevent division by zero if d_{ij} is very small). This metric, is widely used in accessibility studies [6, 33, 34]. It reflects that closer opportunities, or opportunities reached by larger flows, contribute more to accessibility.
2. **Consumer-Surplus Accessibility (Logsum Accessibility)**: Derived from RUM, it is the expected maximum utility from i:

$$A_i^{CS} = \frac{1}{\lambda} \ln \sum_{j \in J_i} \exp(V_{ij}),$$

   where V_{ij} is the systematic utility (Eq. 1) and λ is the scale parameter of the Gumbel distribution (typically normalized to 1 in MNL estimation, so $A_i = \ln \sum_j \exp(V_{ij})$). This metric incorporates the attractiveness of all alternatives by using their utility [14, 34]. It represents the overall benefit or "surplus" a commuter gains from the available set of work opportunities.
3. **Combined Accessibility Measure**: In order to combine both types of accessibility in one single index, we derive a composite between the first two using Principal Component Analysis (PCA). Both accessibility measures are first standardized to z-scores within each city, then PCA is applied separately for each city to extract the first principal component ($PC1_i$), which captures the largest shared variance between distance-weighted and consumer-surplus accessibility. This combined measure identifies areas with high benefit from both physical proximity and utility-based attractiveness (positive values) versus locations facing disadvantages in both dimensions (negative values).

These measures are calculated for each origin zone i and analyzed spatially (Fig. 6) and by socioeconomic groups (Fig. 5).

4.5.1 Location inference from LBS data

Raw LBS pings, each with latitude, longitude and timestamp, are processed in six steps.

Home and work inference:

- **Stay-event detection**. For each device, consecutive chains of pings within a radius of 20 meters and with a minimum event total duration of 5 minutes are flagged as stay events.
- **Spatial aggregation**. Stay events are spatially clustered using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [53] with an $\epsilon = 20$m and a minimum of 2 points, producing distinct stop location clusters.
- **Local-time conversion**. The start time of the stay-events are timestamped in UTC; we spatially joined each cluster centroid to a timezone lookup table and converted them to local datetimes using the matched timezone.
- **Home/Work labeling**. For each user–year: a) A cluster is labeled as *home* if it is the most visited one during "home hours" (weekends or weekdays between 19:00–08:00) on at least 14 distinct days. b) A cluster is labeled *work* if it is the most visited on weekdays between 09:00–18:00 on at least 8 distinct days and lies at least 50 meters away from the home cluster.

Flow aggregation and scaling:

- **Origin-Destination matrix generation**. The derived Home and work locations are spatially joined to the city administrative polygons (AGEBs for Mexico City, AEPs for Rio de Janeiro) to generate aggregated origin-destination matrices.
- **Population scaling**. To improve the representativeness and scale of the actual flows, we compute a scaling factor which is the ratio of observed home locations to census population within each administrative unit. This accounts for differences in penetration rates of the mobility data and adjusts the flows to better represent the actual population.

Declarations

- **Funding** The authors acknowledge funding from the HAYSTAC project from I-ARPA.

- **Conflict of interest/Competing interests** The authors have no relevant financial or non-financial interests to disclose.
- **Ethics approval and consent to participate** This study used anonymized and aggregated data; therefore, ethics approval and consent to participate were not required. All mobility data were handled in accordance with privacy regulations and ethical guidelines for human subject research.
- **Consent for publication** Not applicable as the manuscript does not contain any individual person's data in any form.
- **Data availability** The aggregated mobility flow data and derived $\text{ECI}^{\text{employment}}$/informality indices at the H3 hexagonal level that support the findings of this study are available from the corresponding author upon reasonable request. Raw mobility data are proprietary and cannot be shared. Census data are publicly available from national statistical agencies (INEGI for Mexico, IBGE for Brazil, U.S. Census Bureau for USA). Replica data are proprietary.

- **Code availability** Code used to perform the analyses and generate the figures in this study will be made available in a public repository at https://github.com/humnetlab/workreach upon publication.
- **Author contribution** O.D.L-C. and M.G. conceived the study and designed the methodology. O.D.L-C. performed the analysis, and wrote the manuscript. M.G. contributed to the methodological design and interpretation of results. S.K. contributed to gathering, analyzing, and validating mobility data in Rio de Janeiro. M.G. and S.K. edited the manuscript. All authors read and approved the final manuscript.

References

1. Gobillon, L., Selod, H. & Zenou, Y. The Mechanisms of Spatial Mismatch. *Urban Studies* **44,** 2401–2427. ISSN: 0042-0980, 1360-063X. https://journals.sagepub.com/doi/10.1080/00420980701540937 (Nov. 2007).
2. Moreno-Monroy, A. I. & Ramos, F. R. The impact of public transport expansions on informality: The case of the São Paulo Metropolitan Region. *Research in Transportation Economics* **88,** 100928. ISSN: 07398859. https://linkinghub.elsevier.com/retrieve/pii/S0739885920301268 (Sept. 2021).
3. Graham, M. R. & Ong, P. *Social, economic, spatial, and commuting patterns of informal jobholders* tech. rep. (Center for Economic Studies, US Census Bureau, 2007).
4. Theodore, N. & Rondeau, J. L'informalité et la sélectivité stratégique de l'État : la montée de l'emploi précaire dans l'industrie de la construction aux États-Unis. *Lien social et Politiques,* 114–136. ISSN: 1703-9665. http://id.erudit.org/iderudit/1037068ar (July 2016).
5. Yang, Y., Herrera, C., Eagle, N. & González, M. C. Limits of Predictability in Commuting Flows in the Absence of Data for Calibration. *Scientific Reports* **4.** Publisher: Nature Publishing Group, 5662. ISSN: 2045-2322. https://www.nature.com/articles/srep05662 (July 2014).
6. Piovani, D., Arcaute, E., Uchoa, G., Wilson, A. & Batty, M. Measuring accessibility using gravity and radiation models. *Royal Society Open Science* **5,** 171668. ISSN: 2054-5703. https://royalsocietypublishing.org/doi/10.1098/rsos.171668 (Sept. 2018).
7. Lenormand, M., Bassolas, A. & Ramasco, J. J. Systematic comparison of trip distribution laws and models. *Journal of Transport Geography* **51.** arXiv:1506.04889 [physics], 158–169. ISSN: 0966-6923. http://arxiv.org/abs/1506.04889 (Feb. 2016).
8. Erlander, S. & Stewart, F. *The Gravity Model in Transportation Analysis: Theory and Extensions* ().
9. Barthelemy, M. *The role of parsimonious models in addressing mobility challenges* arXiv:2411.04484 [physics]. Nov. 2024. http://arxiv.org/abs/2411.04484.
10. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484.** Publisher: Nature Publishing

Group, 96–100. ISSN: 1476-4687. https://www.nature.com/articles/nature10856 (Apr. 2012).
11. Pappalardo, L., Rinzivillo, S. & Simini, F. Human Mobility Modelling: Exploration and Preferential Return Meet the Gravity Model. *Procedia Computer Science. The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016) / The 6th International Conference on Sustainable Energy Information Technology (SEIT-2016) / Affiliated Workshops* **83,** 934–939. ISSN: 1877-0509. https://www.sciencedirect.com/science/article/pii/S1877050916302216 (Jan. 2016).
12. Simini, F., Barlacchi, G., Luca, M. & Pappalardo, L. A Deep Gravity model for mobility flows generation. *Nature Communications* **12,** 6576. ISSN: 2041-1723. https://www.nature.com/articles/s41467-021-26752-4 (Nov. 2021).
13. Cabanas-Tirapu, O., Danús, L., Moro, E., Sales-Pardo, M. & Guimerà, R. Human mobility is well described by closed-form gravity-like models learned automatically from data. *Nature Communications* **16.** Publisher: Nature Publishing Group, 1336. ISSN: 2041-1723. https://www.nature.com/articles/s41467-025-56495-5 (Feb. 2025).
14. Ben-Akiva, M. & Bierlaire, M. in *Handbook of Transportation Science* (ed Hall, R. W.) 5–33 (Springer US, Boston, MA, 1999). ISBN: 978-1-4615-5203-1. https://doi.org/10.1007/978-1-4615-5203-1_2.
15. Train, K. E. *Discrete Choice Methods with Simulation* ISBN: 978-0-521-74738-7 (Cambridge University Press, Cambridge ; New York, 2009).
16. Hidalgo, C. A. & Hausmann, R. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences* **106,** 10570–10575. ISSN: 0027-8424, 1091-6490. https://pnas.org/doi/full/10.1073/pnas.0900943106 (June 2009).
17. Hidalgo, C. A. Economic complexity theory and applications. *Nature Reviews Physics* **3,** 92–113. ISSN: 2522-5820. https://www.nature.com/articles/s42254-020-00275-1 (Jan. 2021).
18. Balland, P.-A. *et al.* Complex economic activities concentrate in large cities. *Nature Human Behaviour* **4.** Publisher: Nature Publishing Group, 248–254. ISSN: 2397-3374. https://www.nature.com/articles/s41562-019-0803-3 (Mar. 2020).
19. Magalhães, L., Kuffer, M., Schwarz, N. & Haddad, M. Bringing economic complexity to the intra-urban scale: The role of services in the urban economy of Belo Horizonte, Brazil. *Applied Geography* **150,** 102837. ISSN: 01436228. https://linkinghub.elsevier.com/retrieve/pii/S0143622822002089 (Jan. 2023).
20. Juhász, S. *et al.* Amenity complexity and urban locations of socio-economic mixing. *EPJ Data Science* **12** (2023).
21. Kim, J., Yu, D., Choi, H., Seo, D. & Jun, B. *Redefining Urban Centrality: Integrating Economic Complexity Indices into Central Place Theory* arXiv:2407.19762. 2024. https://arxiv.org/abs/2407.19762.
22. Rossi Mori, L., Loreto, V. & Di Clemente, R. Time-space dynamics of income segregation in the city of Milan. *PNAS Nexus* **4,** pgaf283 (2025).
23. Chen, Z., Zhang, X. & van der Wouden, F. Neighborhood clusters and citywide technological diversification. *Research Policy* **55,** 105397 (2026).

24. Manning, A. *Monopsony in Motion: Imperfect Competition in Labor Markets* ISBN: 978-0-691-11312-6 (Princeton University Press, Princeton, NJ, 2003).
25. Le Barbanchon, T., Rathelot, R. & Roulet, A. Gender Differences in Job Search: Trading off Commute against Wage. *The Quarterly Journal of Economics* **136,** 381–426 (2021).
26. Moretti, E. in *Handbook of Labor Economics* (eds Ashenfelter, O. & Card, D.) 1237–1313 (Elsevier, 2011).
27. Glaeser, E. L. & Gottlieb, J. D. The Wealth of Cities: Agglomeration Economies and Spatial Equilibrium in the United States. *Journal of Economic Literature* **47,** 983–1028 (2009).
28. Abowd, J. M. *et al.* The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators. *NBER Chapters.* Publisher: National Bureau of Economic Research, Inc, 149–230. https://ideas.repec.org//h/nbr/nberch/0485.html (2009).
29. González, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding individual human mobility patterns. *Nature* **453,** 779–782. ISSN: 1476-4687 (June 2008).
30. Song, C., Koren, T., Wang, P. & Barabási, A.-L. Modeling the scaling properties of human mobility. *Nature Physics* **6.** arXiv:1010.0436 [cond-mat], 818–823. ISSN: 1745-2473, 1745-2481. http://arxiv.org/abs/1010.0436 (Oct. 2010).
31. Schläpfer, M. *et al.* The universal visitation law of human mobility. *Nature* **593.** Publisher: Nature Publishing Group, 522–527. ISSN: 1476-4687. https://www.nature.com/articles/s41586-021-03480-9 (May 2021).
32. Geurs, K. T. & van Wee, B. Accessibility evaluation of land-use and transport strategies: review and research directions. *Journal of Transport Geography* **12,** 127–140. ISSN: 0966-6923. https://www.sciencedirect.com/science/article/pii/S0966692303000607 (June 2004).
33. Pereira, R. H. M., Banister, D., Schwanen, T. & Wessel, N. Distributional effects of transport policies on inequalities in access to opportunities in Rio de Janeiro. *Journal of Transport and Land Use* **12.** ISSN: 1938-7849. https://www.jtlu.org/index.php/jtlu/article/view/1523 (Oct. 2019).
34. Guzman, L. A., Cantillo-Garcia, V. A., Oviedo, D. & Arellana, J. How much is accessibility worth? Utility-based accessibility to evaluate transport policies. *Journal of Transport Geography* **112,** 103683. ISSN: 09666923. https://linkinghub.elsevier.com/retrieve/pii/S0966692323001552 (Oct. 2023).
35. Ben-Akiva, M. & Lerman, S. R. *Discrete Choice Analysis: Theory and Application to Travel Demand* ISBN: 978-0-262-53640-0 (Cambridge, Massachusetts, 2018).
36. Chorus, C. G. Logsums for utility-maximizers and regret-minimizers, and their relation with desirability and satisfaction. *Transportation Research Part A: Policy and Practice* **46,** 1003–1012. ISSN: 0965-8564. https://www.sciencedirect.com/science/article/pii/S0965856412000729 (Aug. 2012).
37. De Jong, G., Daly, A., Pieters, M. & van der Hoorn, T. The logsum as an evaluation measure: Review of the literature and new results. *Transportation Research Part A: Policy and Practice. Selected Papers on Applications of Discrete Choice Models Presented at the European Regional Science Conference, Amsterdam,*

August 2005 **41,** 874–889. ISSN: 0965-8564. https://www.sciencedirect.com/science/article/pii/S0965856407000316 (Nov. 2007).
38. Hernandez, D. Uneven mobilities, uneven opportunities: Social distribution of public transport accessibility to jobs and education in Montevideo. *Journal of Transport Geography* **67,** 119–125. ISSN: 0966-6923. https://www.sciencedirect.com/science/article/pii/S0966692316303556 (Feb. 2018).
39. Bocarejo S., J. P. & Oviedo H., D. R. Transport accessibility and social inequities: a tool for identification of mobility needs and evaluation of transport investments. *Journal of Transport Geography* **24.** Publisher: Elsevier, 142–154. https://ideas.repec.org//a/eee/jotrge/v24y2012icp142-154.html (2012).
40. Group, W. B. *World Bank Open Data* https://data.worldbank.org.
41. Giuliano, G., Hou, Y., Kang, S. & Shin, E. J. Polycentricity and the evolution of metropolitan spatial structure. *Growth and Change* **53.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/grow.12599, 593–627. ISSN: 1468-2257. https://onlinelibrary.wiley.com/doi/abs/10.1111/grow.12599 (2022).
42. Xu, Y. *et al.* Urban dynamics through the lens of human mobility. *Nature computational science* **3,** 611–620 (2023).
43. Bustos, S., Gomez, C., Hausmann, R. & Hidalgo, C. A. The Dynamics of Nestedness Predicts the Evolution of Industrial Ecosystems. *PLOS ONE* **7,** e49393 (2012).
44. Sørensen, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol Skrifter/Kongelige Danske Videnskabernes Selskab.* **5,** 1 (1948).
45. Liu, Z., Huang, L., Fan, C. & Mostafavi, A. FairMobi-Net: a fairness-aware deep learning model for urban mobility flow generation. *arXiv preprint arXiv:2307.11214* (2023).
46. Hidalgo, C. A., Klinger, B., Barabási, A.-L. & Hausmann, R. The Product Space Conditions the Development of Nations. *Science* **317.** Publisher: American Association for the Advancement of Science, 482–487. https://www.science.org/doi/10.1126/science.1144581 (July 2007).
47. U.S. Bureau of Labor Statistics. *Job Openings and Labor Turnover Survey (JOLTS) Highlights* Federal Reserve Bank of St. Louis (FRED). Monthly quits rate approximately 2.2%. 2024. https://fred.stlouisfed.org/series/JTSQUR.
48. Alaimo, V., Bosch, M., Kaplan, D. S., Pagés, C. & Ripani, L. *Jobs for Growth* (Inter-American Development Bank, Washington, DC, 2017).
49. McFadden, D. in *Frontiers in Econometrics* (ed Zarembka, P.) 105–142 (Academic press, New York, 1974).
50. Soares, P. *econci* https://github.com/phcsoares/econci. 2020.
51. Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17,** 261–272 (2020).
52. Masucci, A. P., Serras, J., Johansson, A. & Batty, M. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows.

Physical Review E—Statistical, Nonlinear, and Soft Matter Physics **88,** 022812 (2013).

53. Ester, M., Kriegel, H.-P., Sander, J., Xu, X., *et al. A density-based algorithm for discovering clusters in large spatial databases with noise* in *kdd* **96** (1996), 226–231.

WorkReach: Modeling Urban Work Location Choices Through Economic Complexity, Informality, and Mobility Data

Supplementary Information

S1 Spatial Properties of Model Inputs

In order to understand the spatial structure of the variables, we analyze their spatial autocorrelation. These analyses confirm that the variables are not randomly distributed but follow underlying spatial patterns, which justifies their use in the model.

S1.1 Spatial Autocorrelation

We calculated Moran's I to test for spatial autocorrelation in the Economic Complexity Index ($ECI^{employment}$) and the Informality Rate for each city. Table S1 shows that both variables in all cities are statistically significant ($p < 0.001$) and that they all show a positive spatial autocorrelation. This result indicates that the spatial distribution of the measured variables is clustered, with high value areas tending to be near other high value areas and vice versa. Figures S1 through S8 represent the spatial patterns of $ECI^{employment}$ and informality rate along the their Moran's I analysis.

Table S1 Moran's I and P-value for Spatial Autocorrelation by City and Variable.

City	Variable	Moran's I	P-value
Bay Area	Informality Rate	0.171	<0.001
Los Angeles	Informality Rate	0.253	<0.001
Mexico City	Informality Rate	0.407	<0.001
Rio de Janeiro	Informality Rate	0.467	<0.001
Bay Area	$ECI^{employment}$	0.364	<0.001
Los Angeles	$ECI^{employment}$	0.519	<0.001
Mexico City	$ECI^{employment}$	0.594	<0.001
Rio de Janeiro	$ECI^{employment}$	0.425	<0.001

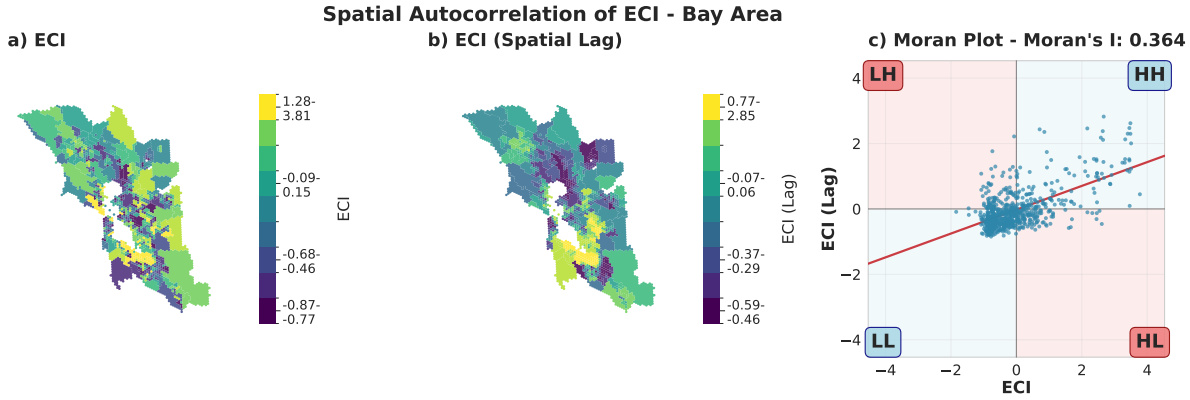

Fig. S1 Spatial autocorrelation analysis of Economic Complexity Index ($ECI^{employment}$) for the Bay Area. The figure shows **a)** the choropleth map of $ECI^{employment}$, **b)** the map of the spatial lag of $ECI^{employment}$, and **c)** the Moran's I scatter plot (Moran's I = 0.364).

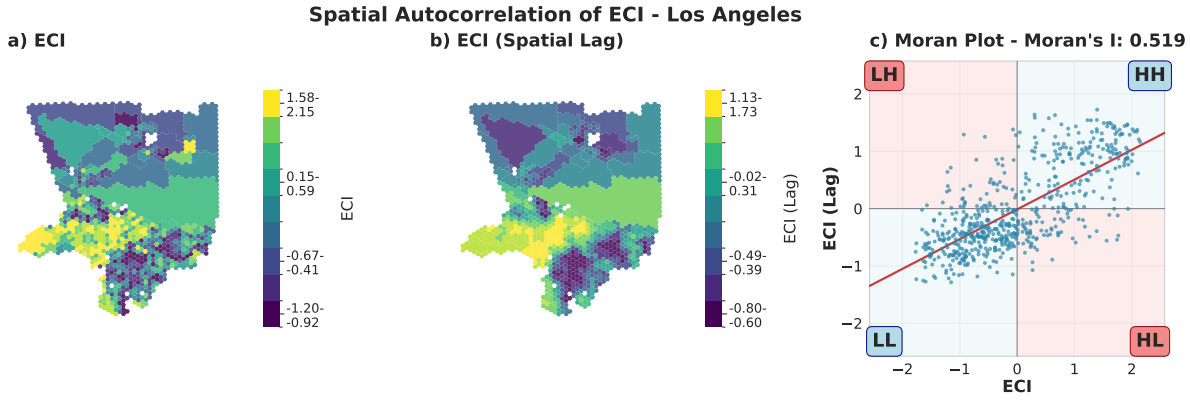

Fig. S2 Spatial autocorrelation analysis of Economic Complexity Index ($ECI^{employment}$) for Los Angeles. The figure shows **a)** the choropleth map of $ECI^{employment}$, **b)** the map of the spatial lag of $ECI^{employment}$, and **c)** the Moran's I scatter plot (Moran's I = 0.519).

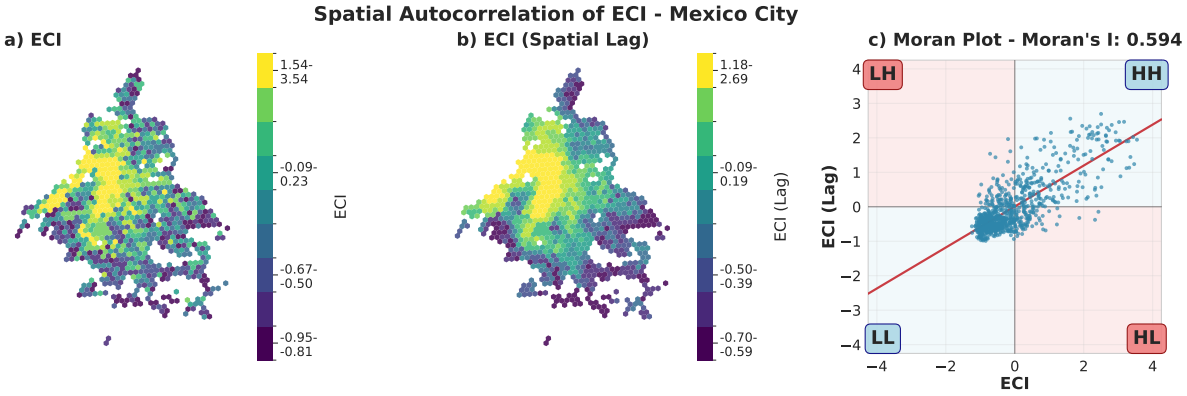

Fig. S3 Spatial autocorrelation analysis of Economic Complexity Index ($ECI^{employment}$) for Mexico City. The figure shows **a)** the choropleth map of $ECI^{employment}$, **b)** the map of the spatial lag of $ECI^{employment}$, and **c)** the Moran's I scatter plot (Moran's I = 0.594).

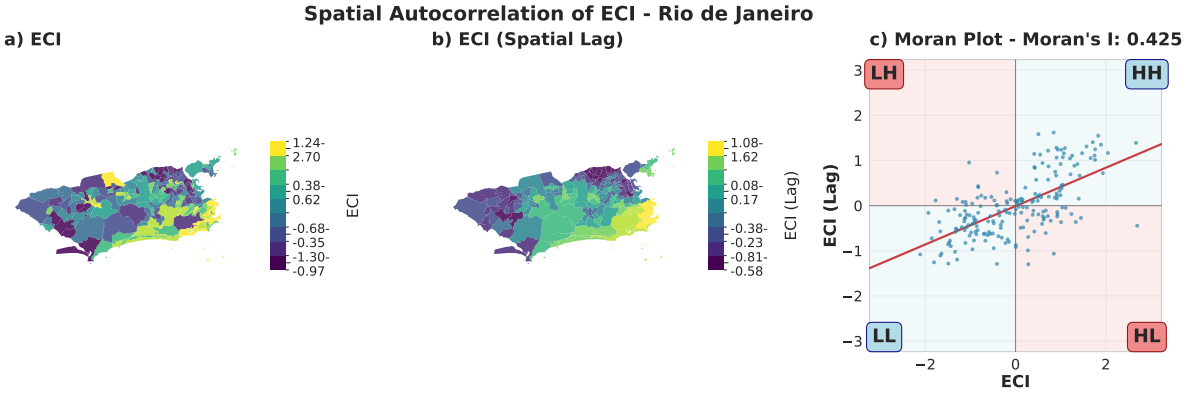

Fig. S4 Spatial autocorrelation analysis of Economic Complexity Index ($ECI^{employment}$) for Rio de Janeiro. The figure shows **a)** the choropleth map of $ECI^{employment}$, **b)** the map of the spatial lag of $ECI^{employment}$, and **c)** the Moran's I scatter plot (Moran's I = 0.425).

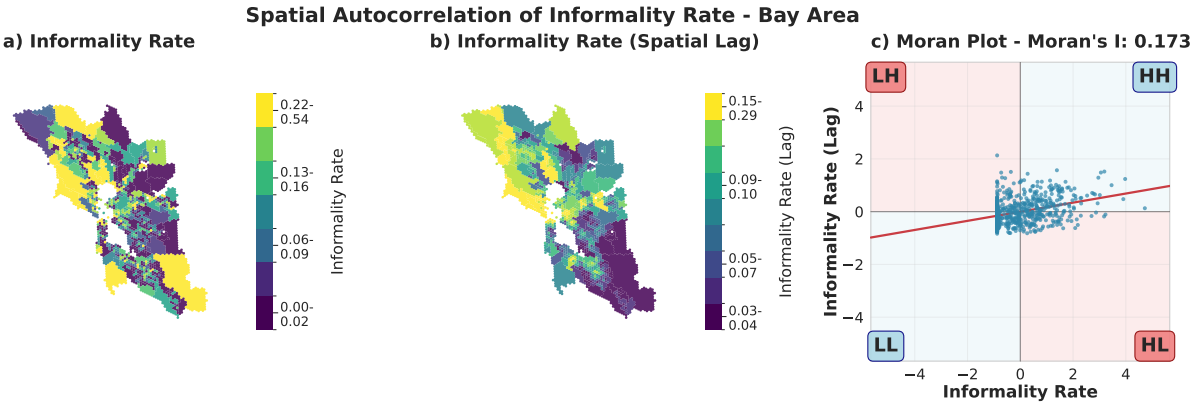

Fig. S5 Spatial autocorrelation analysis of Informality Rate for the Bay Area. The figure shows **a)** the choropleth map of Informality Rate, **b)** the map of the spatial lag of Informality Rate, and **c)** the Moran's I scatter plot (Moran's I = 0.171).

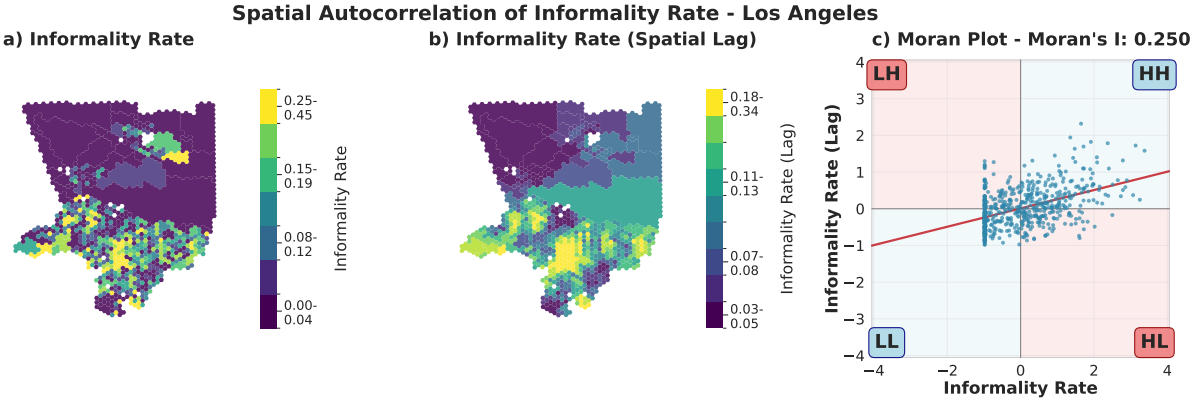

Fig. S6 Spatial autocorrelation analysis of Informality Rate for Los Angeles. The figure shows **a)** the choropleth map of Informality Rate, **b)** the map of the spatial lag of Informality Rate, and **c)** the Moran's I scatter plot (Moran's I = 0.253).

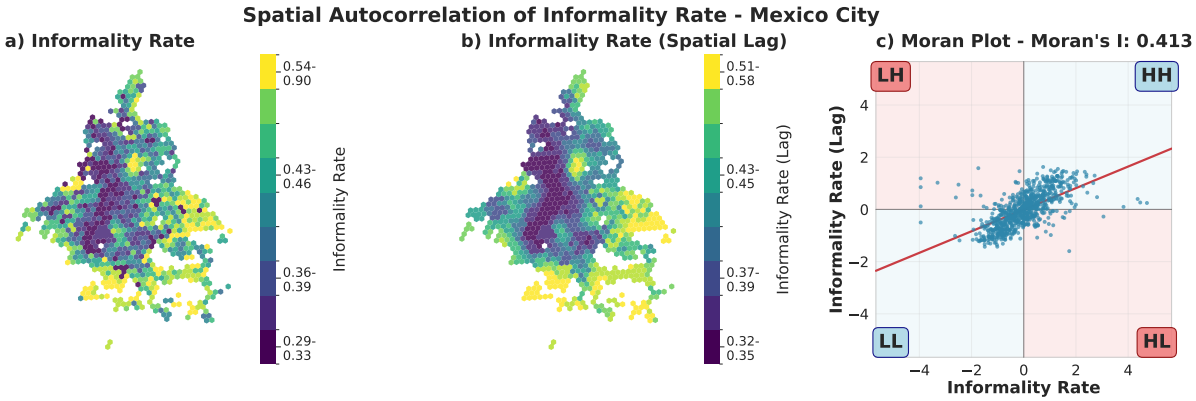

Fig. S7 Spatial autocorrelation analysis of Informality Rate for Mexico City. The figure shows **a)** the choropleth map of Informality Rate, **b)** the map of the spatial lag of Informality Rate, and **c)** the Moran's I scatter plot (Moran's I = 0.407).

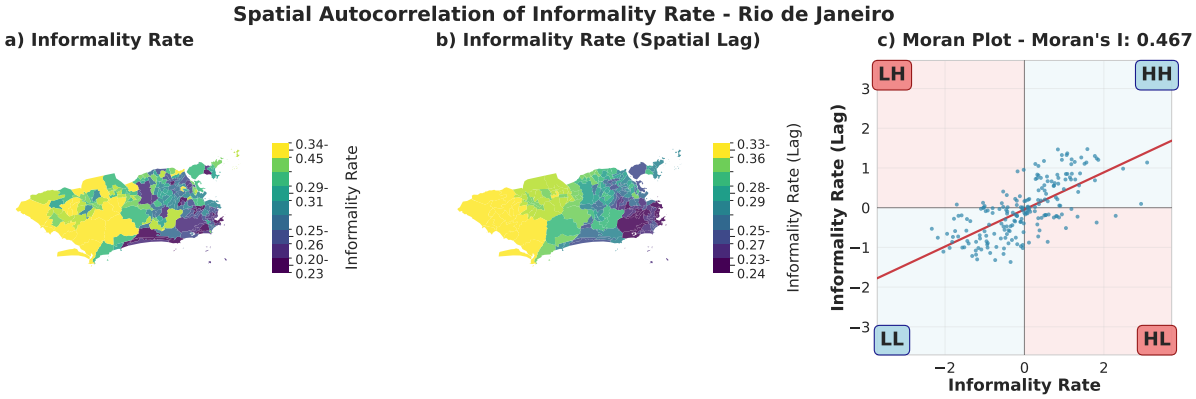

Fig. S8 Spatial autocorrelation analysis of Informality Rate for Rio de Janeiro. The figure shows **a)** the choropleth map of Informality Rate, **b)** the map of the spatial lag of Informality Rate, and **c)** the Moran's I scatter plot (Moran's I = 0.467).

S1.2 Variable Distributions and Preprocessing

The WorkReach model has 3 input variables (Distance, $\text{ECI}^{\text{employment}}$, Informality Rate) which are pre-processed for model stability and interpretability of results. All three variables are min-max scaled to the $[0, 1]$ interval. Figures S9 to S12 show the final distributions of these variables that were used as input to the model.

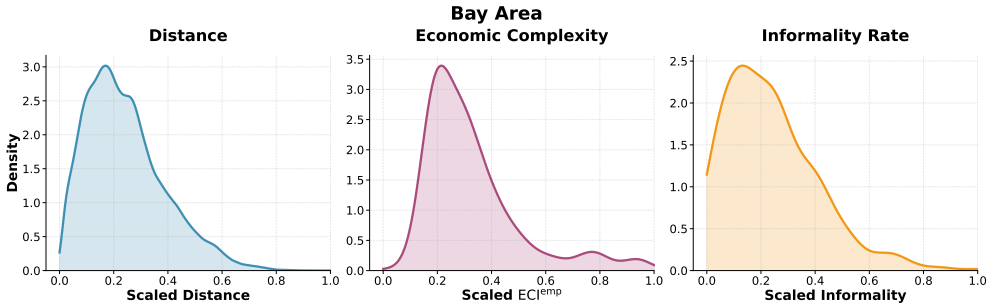

Fig. S9 Distributions of the transformed variables for the Bay Area.

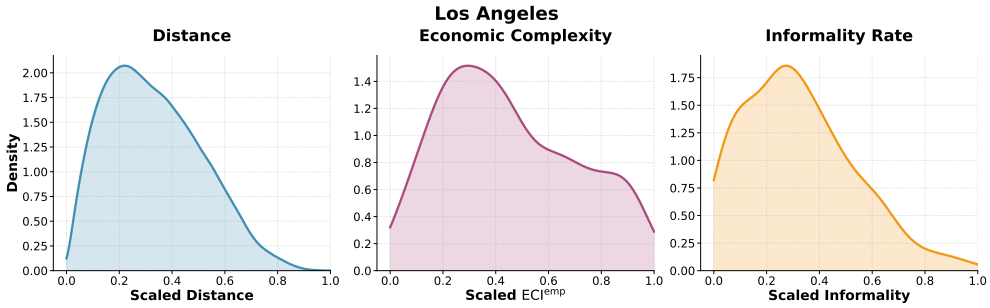

Fig. S10 Distributions of transformed variables for Los Angeles.

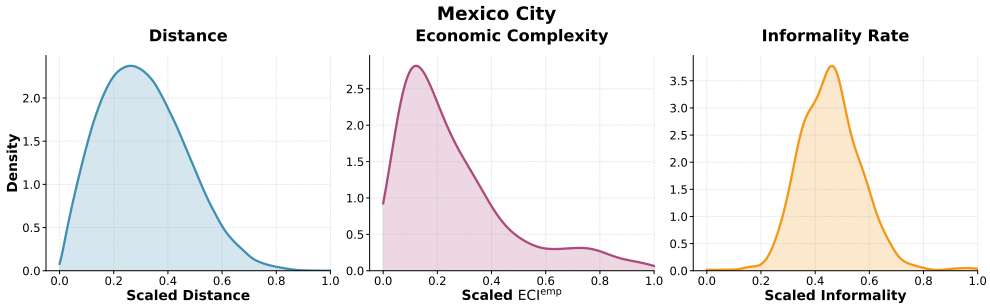

Fig. S11 Distributions of transformed variables for Mexico City.

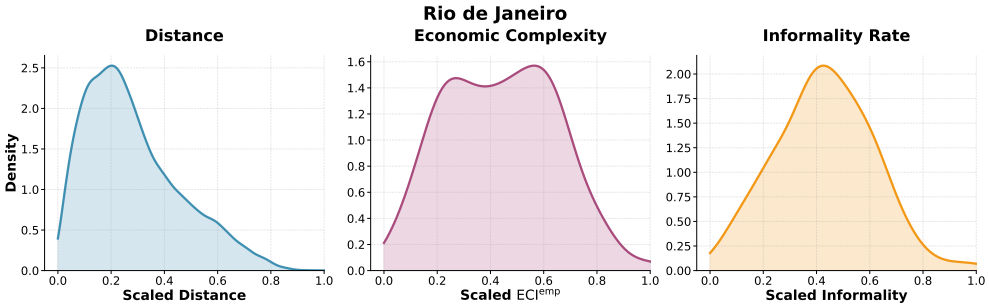

Fig. S12 Distributions of the transformed variables for Rio de Janeiro.

S2 Validation of Commuting Data in Mexico City

To validate our LBS-derived commuting data for Mexico City, we compare aggregate commuting distances against the 2017 Origin-Destination Survey (EOD 2017), conducted by INEGI across the Zona Metropolitana del Valle de México (ZMVM) [1]. The EOD collected trip records for the entire metropolitan area, with geographic detail at the district level. Districts are spatial units defined by the survey to contain roughly equal populations; CDMX proper comprises 85 districts, distinct from the 16 administrative alcaldías. No equivalent OD survey exists for Rio de Janeiro; because both cities rely on the same LBS vendor and processing pipeline, this validation also provides indirect support for the quality of the Rio de Janeiro commuting data.

Distance computation from the EOD survey

The EOD 2017 does not report trip-level distances directly. We compute centroid-to-centroid Haversine distances between origin and destination districts using the geographic center of each administrative unit. The 85 CDMX districts vary considerably in size; the mean minimum bounding circle (MBC) radius is 3.14 km and the median is 2.61 km. For intra-district trips (where origin and destination share the same district), the centroid-to-centroid distance would be zero. We instead assign these trips a distance equal to the mean MBC radius of their district, providing a rough but non-trivial estimate of the average within-district trip length.

Distance comparison

Table S2 summarizes the commuting distances and travel times from both data sources. We subset the EOD's district-level OD matrix to the 85 districts within CDMX proper to match the spatial extent of our LBS data. From this subset, we obtain a mean centroid-to-centroid distance of 7.74 km (median 6.23 km) and a mean travel time of 50.28 minutes (median 45.00 min). Our LBS data at H3 resolution 8 (the level used in our analysis, with cell edge length ≈0.575 km) yields a mean Haversine distance of 6.02 km and a median of 4.30 km for CDMX.

Sensitivity to spatial resolution

To understand how spatial aggregation affects distance estimates, we re-compute the LBS distances at multiple resolutions: the 85 EOD districts, the native AGEB (census tract) level, and four H3 hexagonal grid levels (resolutions 6 through 9). Table S3 reports the cell radius, number of cells covering CDMX, and the resulting mean and median distances for each level.

At H3 resolution 6, the 35 hexagonal cells covering CDMX are most comparable in number and size to the 16 survey districts. At this level, the LBS data yields a mean distance of 7.56 km and a median of 6.88 km, compared to the survey's 7.74 km (mean) and 6.23 km (median) for CDMX proper. The mean difference is only 0.18 km. As resolution increases (smaller cells), distances converge downward toward the AGEB-level values, where finer spatial detail allows the identification of short-range commutes that coarser grids merge into single zones. The stability of estimates

Table S2 Commuting distance and travel time comparison between the EOD 2017 survey and LBS data for Mexico City. Survey distances are computed as centroid-to-centroid Haversine between districts; LBS distances are centroid-to-centroid Haversine between H3 cells. Standard errors are shown in parentheses.

Dataset	Mean	Median	Std. error
EOD 2017 survey			
Distance, ZMVM (km)	10.18	7.39	0.030
Distance, CDMX (km)	7.74	6.23	0.028
Travel time, ZMVM (min)	55.84	50.00	0.131
Travel time, CDMX (min)	50.28	45.00	0.169
LBS data (CDMX)			
District level	7.00	5.09	0.168
AGEB level	6.01	4.20	0.020
H3 resolution 6	7.56	6.88	0.400
H3 resolution 7	6.29	4.53	0.102
H3 resolution 8	6.02	4.30	0.030
H3 resolution 9	5.99	4.21	0.021

Table S3 LBS commuting distance at different spatial resolutions for CDMX. Cell radius refers to the circumscribed radius of each hexagonal cell (or the mean MBC radius for AGEBs). The number of cells, mean distance, and median distance are shown for each level.

Resolution	Cell radius (km)	Cells	Mean (km)	Median (km)
District	3.136	85	7.00	5.09
AGEB	0.478	2,431	5.98	4.20
H3 res. 6	4.023	35	7.56	6.88
H3 res. 7	1.521	250	6.29	4.53
H3 res. 8	0.575	1,766	6.02	4.30
H3 res. 9	0.217	12,325	5.99	4.21

between H3 resolution 8 (our analysis level) and the finest resolution 9 indicates that our working resolution adequately captures the distance distribution without excessive aggregation.

Origin-destination matrix comparison

Beyond aggregate distance metrics, we compare individual zone-to-zone flows in the origin-destination matrices. Both the EOD survey and our LBS data are aggregated to Mexico City's 85 districts. Figure S13 shows the OD heatmaps from the EOD survey and the LBS data alongside a scatter plot of district-level OD flows. The spatial patterns are visually consistent: both matrices exhibit a strong diagonal (intra-district flows) and highlight the attraction of central districts. The Pearson correlation between the two matrices is $r = 0.76$ and the cosine similarity is 0.887, indicating strong concordance in the overall structure of commuting flows despite differences in data source, time period, and collection methodology. The LBS data also captures OD pairs for which the survey reports zero flows, providing broader spatial coverage

than the survey alone. Together with the distance estimates at comparable spatial resolutions (mean difference < 0.2 km at H3 resolution 6), these results confirm that the LBS data reliably represents Mexico City's commuting structure.

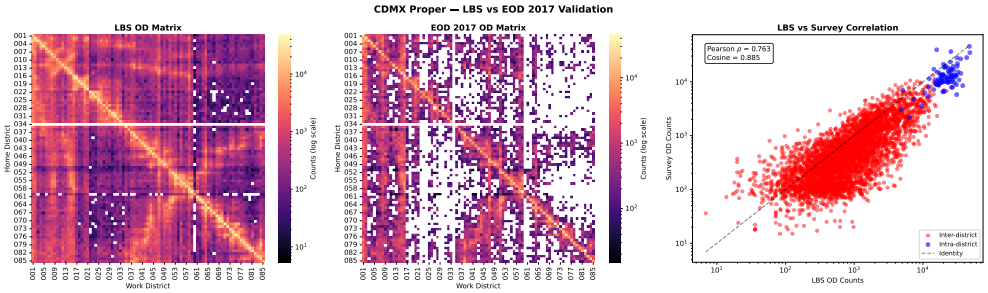

Fig. S13 Validation of LBS commuting data against the EOD 2017 survey for Mexico City proper, aggregated to 85 districts. **a)** OD heatmap from LBS data. **b)** OD heatmap from the EOD survey. **c)** Scatter plot of district-level OD flows (Pearson $r = 0.76$; cosine similarity $= 0.887$).

S3 Model Results

S3.1 Shape of Transition Weight

A core component of our WorkReach model is the transition weight function $w_{ij}(\tau, k)$, which introduces a behavioral shift in job choice based on distance. Figure S14 shows the shape of this transition weight for each city.

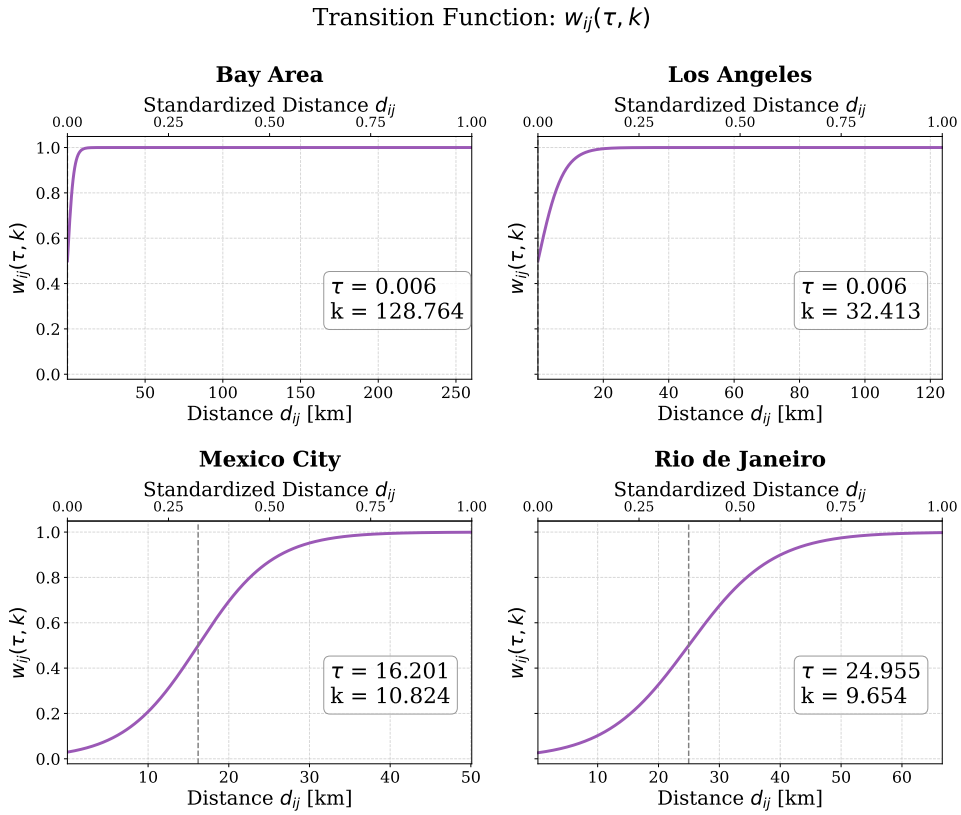

Fig. S14 Shape of the transition weight with the optimized parameters, $w_{ij}(\tau, k)$, for all four cities. The w represents a logistic function which models the behavioral shift in local vs opportunity based job choice. For $d_{ij} > \tau$, the influence of $\text{ECI}^{\text{employment}}$ and informality in the utility function is maximal. The parameters (τ, k) show differences by region. While U.S. cities have lower thresholds (τ) and steeper transitions (k), the Latin American cities have a higher distance threshold and a less steep transition.

S3.2 Flow Map Comparison

We compare the predictive performance of our WorkReach model, by comparing its generated flows against the observed ones and the obtained by other three benchmark

models. Figures S15 to S18 show the flow visualizations for each model and city along with the observed flows.

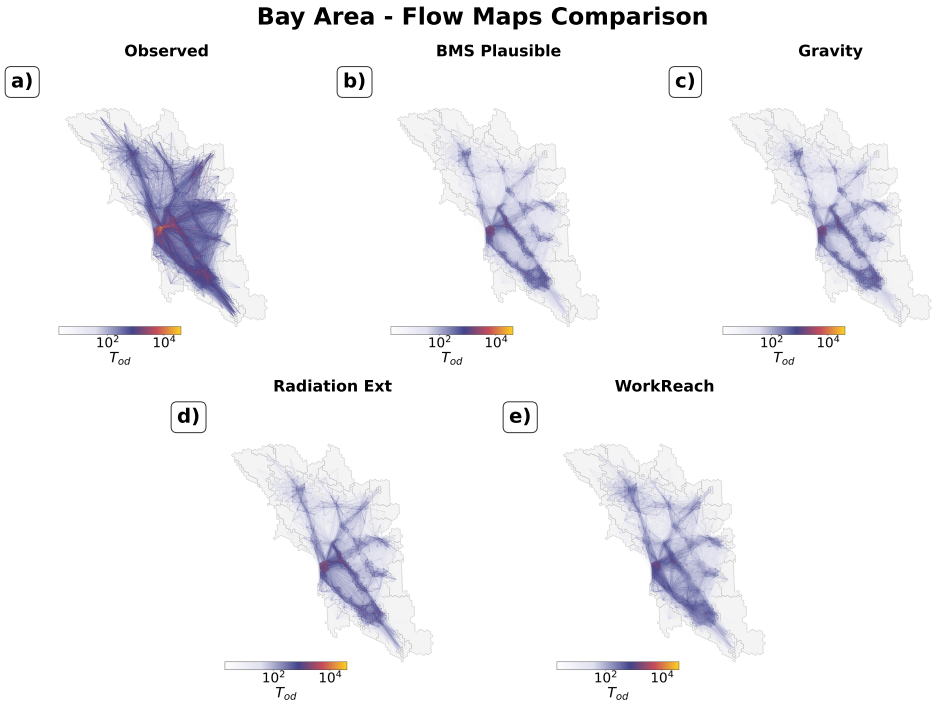

Fig. S15 Commuter flows for the Bay Area. **a)** Observed flows **b)** BMS Plausible, **c)** Gravity, **d)** Extended Radiation, and **e)** WorkReach.

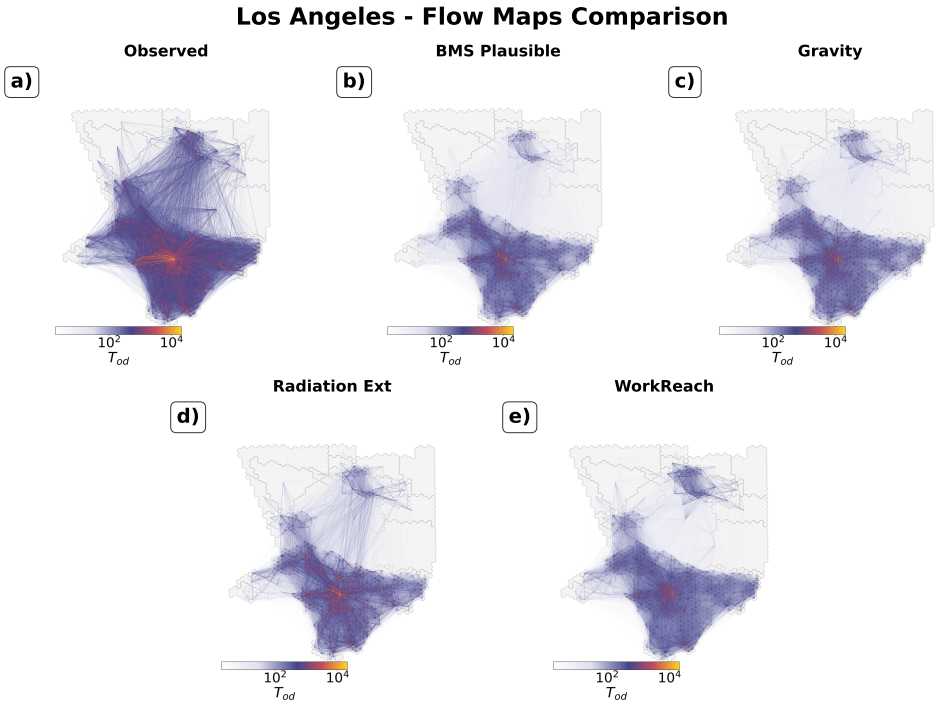

Fig. S16 Commuter flows for Los Angeles. **a)** Observed flows **b)** BMS Plausible, **c)** Gravity, **d)** Extended Radiation, and **e)** WorkReach.

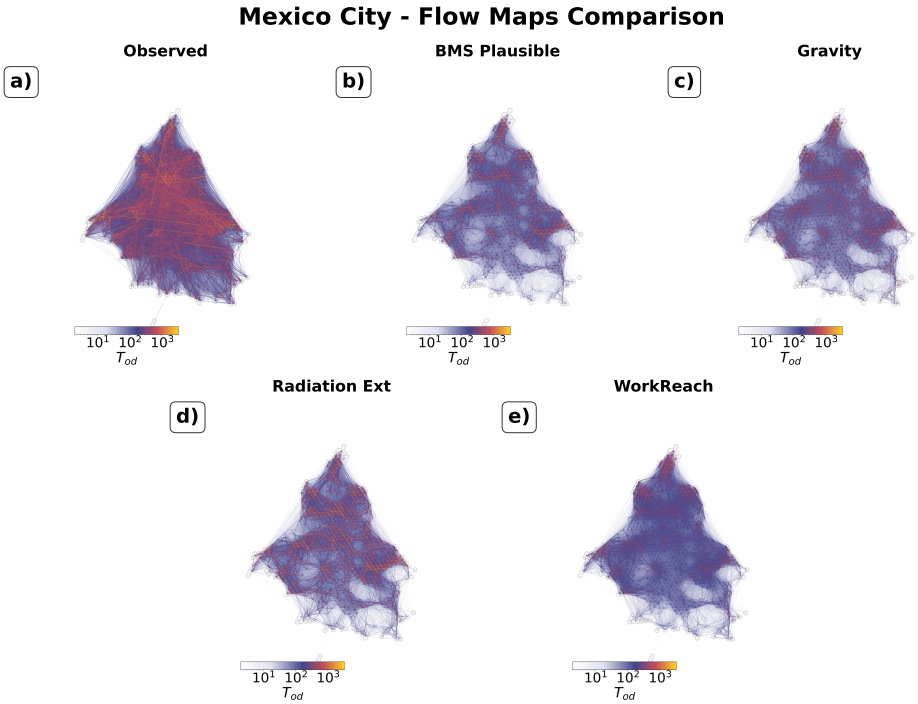

Fig. S17 Commuter flows for Mexico City. **a)** Observed flows **b)** BMS Plausible, **c)** Gravity, **d)** Extended Radiation, and **e)** WorkReach.

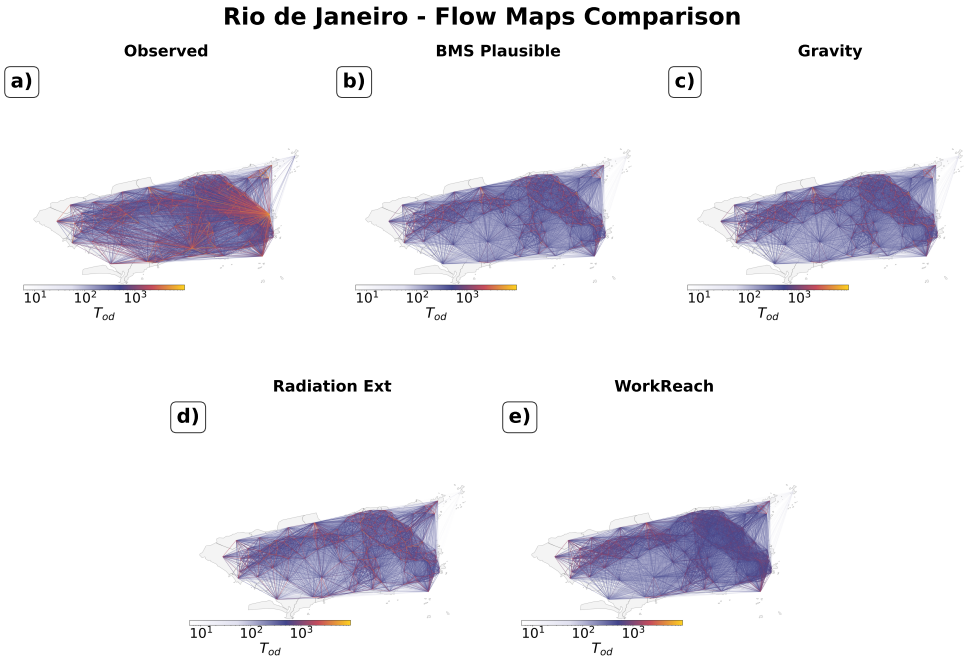

Fig. S18 Commuter flows for Rio de Janeiro. **a)** Observed flows **b)** BMS Plausible, **c)** Gravity, **d)** Extended Radiation, and **e)** WorkReach.

S3.3 Permutation Based Null Model

To test the added value of $\text{ECI}^{\text{employment}}$ and Informality Rate as contributors to WorkReach's predictive power, we performed a permutation analysis. We compare the Common Part of Commuters (CPC) of the full model against versions where the spatial distribution of $\text{ECI}^{\text{employment}}$, Informality Rate, or both are randomly shuffled across zones, preserving the original marginal distributions. For each variant, the model is re-estimated on the shuffled inputs (1000 replications). Figures S19 and S20 show this comparison across variable bins, and Table S4 summarizes the overall results. In all four cities the full model significantly outperforms every randomized variant ($p < 0.001$), confirming that both variables contribute to predictive power independently and jointly.

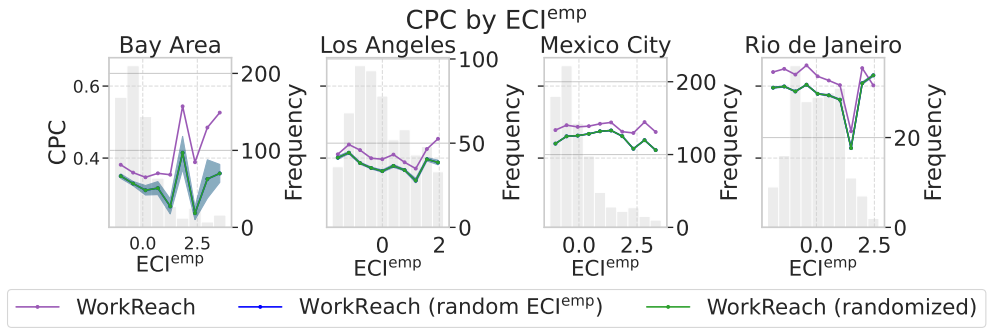

Fig. S19 Model performance (CPC) across bins of Economic Complexity Index ($\text{ECI}^{\text{employment}}$). The purple line shows the full WorkReach model; the blue line shows the model with randomized $\text{ECI}^{\text{employment}}$ only; the green line shows the model with both $\text{ECI}^{\text{employment}}$ and Informality Rate randomized. Shaded bands indicate ±1 standard deviation across 1000 replications.

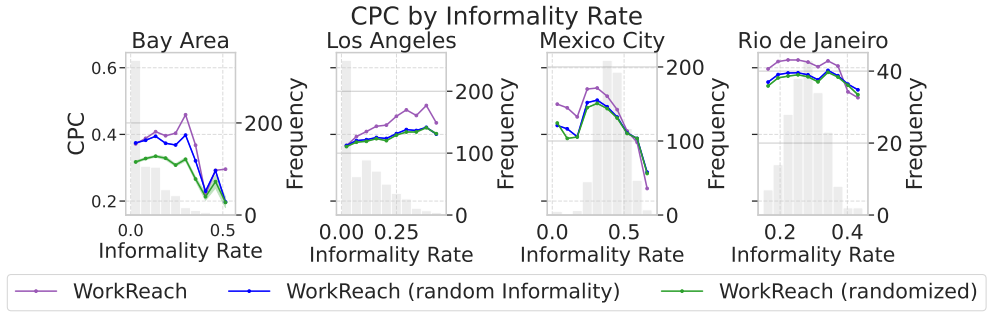

Fig. S20 Model performance (CPC) across bins of Informality Rate. The purple line shows the full WorkReach model; the blue line shows the model with randomized Informality Rate only; the green line shows the model with both $\text{ECI}^{\text{employment}}$ and Informality Rate randomized. Shaded bands indicate ±1 standard deviation across 1000 replications.

Table S4 Permutation analysis of model performance (CPC). Each variant shuffles the spatial distribution of the target variable across zones and re-estimates the model (1000 replications). Standard errors and one-sided permutation p-values test whether the full model significantly outperforms each null.

City	**WorkReach**	**Random** $\text{ECI}^{\text{employment}}$	**Random Inf.**	**Random Both**	p-value
Bay Area	0.389	$0.330 \pm <0.001$	$0.363 \pm <0.001$	$0.330 \pm <0.001$	< 0.001
Los Angeles	0.413	$0.379 \pm <0.001$	$0.394 \pm <0.001$	$0.379 \pm <0.001$	< 0.001
Mexico City	0.486	$0.457 \pm <0.001$	$0.461 \pm <0.001$	$0.457 \pm <0.001$	< 0.001
Rio de Janeiro	0.614	$0.572 \pm <0.001$	$0.578 \pm <0.001$	$0.572 \pm <0.001$	< 0.001

S3.4 Prediction Error Correlation Analysis

To understand whether the four models capture distinct or overlapping aspects of commuting behavior, we computed pairwise Pearson correlations between prediction residuals (observed minus predicted flows) for all models and cities. Figure S21 displays the residual correlation matrices.

The most striking pattern is the near-perfect residual correlation between Gravity and BMS Plausible ($r = 0.96$–0.99 across all cities), consistent with the structural similarity of their formulations [2]. Both models rely on origin and destination populations weighted by distance decay, so they succeed and fail on essentially the same OD pairs. WorkReach's residual correlations with the benchmarks vary across cities: highest in Los Angeles ($r = 0.92$–0.95) and lowest in the Bay Area ($r = 0.86$–0.92) and Mexico City ($r = 0.66$–0.86), indicating that the inclusion of $\text{ECI}^{\text{employment}}$ and informality captures distinct predictive information in these cities. The Extended Radiation model diverges most from all other models, particularly in Mexico City ($r = 0.66$ with WorkReach, $r = 0.77$ with Gravity) and Los Angeles ($r = 0.73$ with WorkReach, $r = 0.80$ with Gravity), likely reflecting the radiation model's reliance on intervening opportunities rather than parametric distance decay.

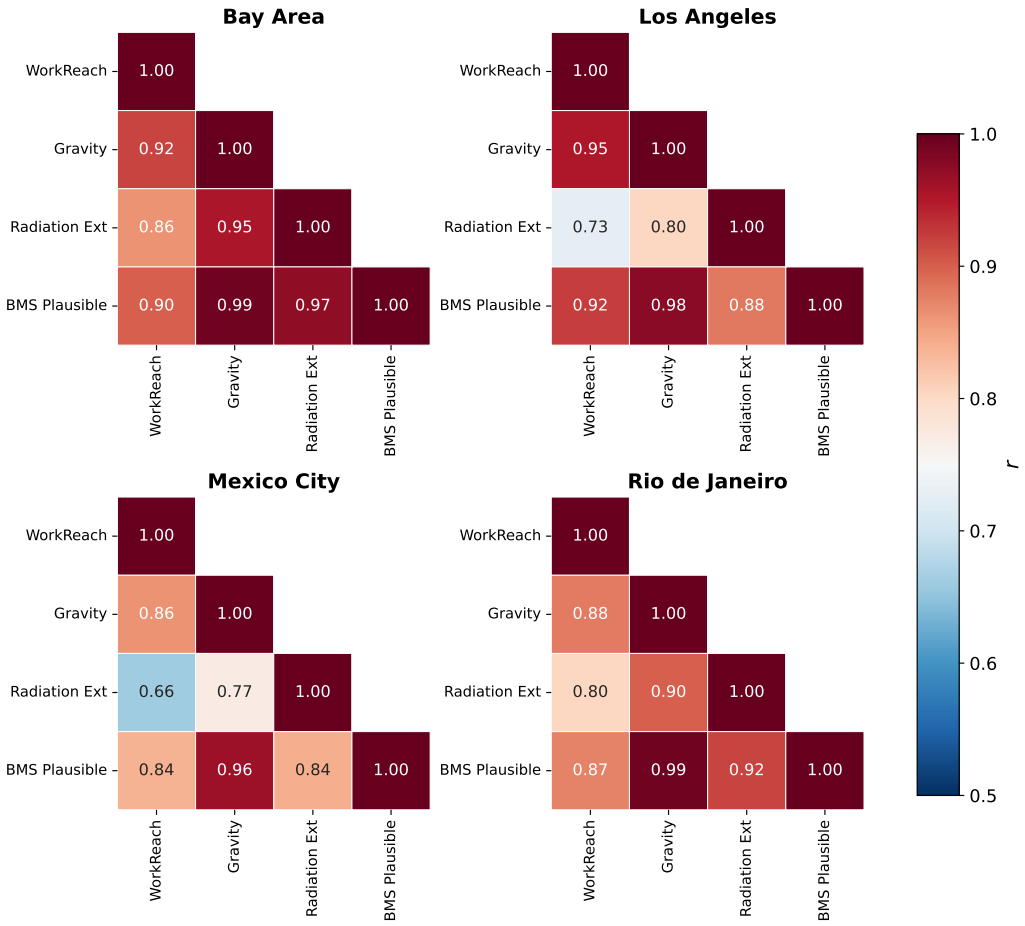

Fig. S21 Pairwise Pearson correlations between prediction residuals (observed − predicted) for all four models, by city. Gravity and BMS Plausible share near-identical residual structure ($r > 0.96$). The Extended Radiation model is the most distinct, particularly in Mexico City and Los Angeles.

S3.5 Out-of-Sample Cross-Validation

To assess whether the reported performance metrics reflect genuine predictive ability rather than overfitting, we conducted a stratified 5-fold cross-validation on origin zones. In each fold, 80% of origin zones are used for training and the remaining 20% are held out for evaluation. The split is stratified by quartile bins of zone population, $\text{ECI}^{\text{employment}}$, and informality rate to ensure that both train and test sets span the full range of zone characteristics. Crucially, all destination zones remain in the choice set for both train and test origins: we withhold only the observed outflow rows of test origins during parameter estimation, preserving the multinomial logit structure

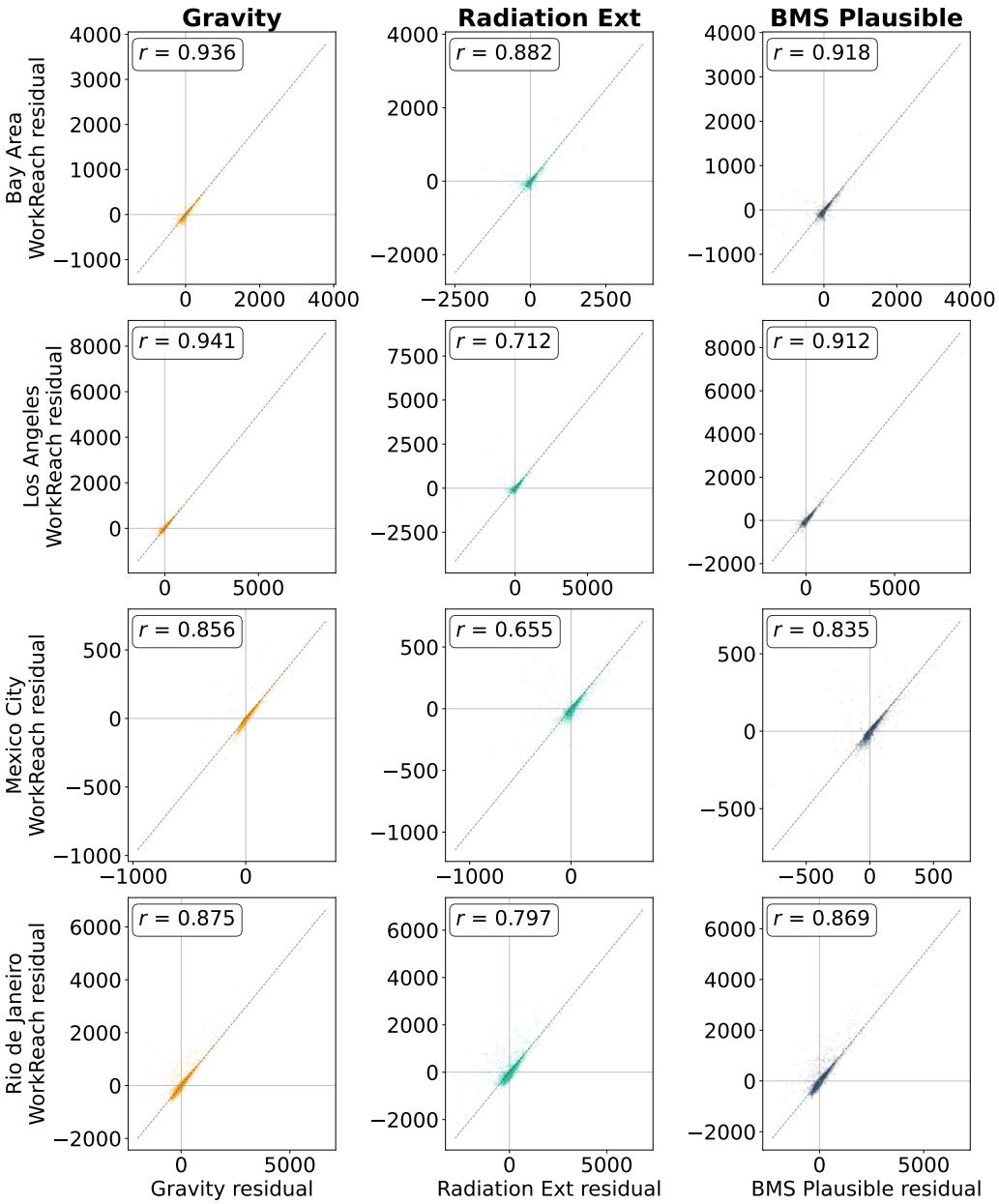

Fig. S22 WorkReach prediction residuals versus benchmark model residuals for each city (rows) and benchmark (columns). Points near the diagonal indicate shared prediction errors; deviations reveal OD pairs where WorkReach predicts differently from the benchmark. Pearson correlations are annotated in each panel.

in which every origin faces the same set of destination alternatives. Each model is re-estimated from scratch on the training origins; the fitted parameters are then used to predict destination shares for the held-out origins.

Because each origin zone appears in the test set in exactly one fold, stitching the test predictions across all five folds yields a complete out-of-sample prediction for every origin–destination pair. Table S5 reports the mean train and test CPC and Pearson r across folds; Figure S23 visualizes the comparison.

The train–test performance gap is negligible for all models and cities (ΔCPC $\leq$ 0.001, $\Delta r \leq 0.02$). This confirms that the parsimonious parameterizations of all four models (1–6 free parameters on 40,000–630,000 OD pairs) are not susceptible to overfitting, and that the in-sample performance comparison reported in the main text accurately reflects out-of-sample predictive ability.

Table S5 Mean train and test performance across 5 stratified folds ($\pm$ standard deviation). The negligible train–test gap confirms the absence of overfitting across all models.

		CPC		Pearson r	
City	Model	Train	Test	Train	Test
Bay Area	WorkReach	0.389 ± 0.001	0.389 ± 0.007	0.524 ± 0.013	0.524 ± 0.029
	Gravity	0.347 ± 0.001	0.347 ± 0.002	0.461 ± 0.014	0.454 ± 0.045
	Ext. Radiation	0.335 ± 0.001	0.335 ± 0.003	0.413 ± 0.003	0.415 ± 0.013
	BMS Plausible	0.342 ± 0.001	0.342 ± 0.003	0.468 ± 0.017	0.460 ± 0.050
Los Angeles	WorkReach	0.413 ± 0.002	0.413 ± 0.006	0.580 ± 0.015	0.580 ± 0.044
	Gravity	0.418 ± 0.002	0.417 ± 0.006	0.595 ± 0.015	0.581 ± 0.046
	Ext. Radiation	0.406 ± 0.001	0.406 ± 0.006	0.530 ± 0.011	0.522 ± 0.043
	BMS Plausible	0.414 ± 0.002	0.413 ± 0.007	0.606 ± 0.018	0.587 ± 0.056
Mexico City	WorkReach	0.486 ± 0.002	0.486 ± 0.005	0.658 ± 0.003	0.658 ± 0.009
	Gravity	0.488 ± 0.002	0.488 ± 0.006	0.661 ± 0.010	0.664 ± 0.035
	Ext. Radiation	0.462 ± 0.002	0.461 ± 0.006	0.613 ± 0.003	0.613 ± 0.013
	BMS Plausible	0.474 ± 0.002	0.473 ± 0.007	0.646 ± 0.010	0.649 ± 0.038
Rio de Janeiro	WorkReach	0.614 ± 0.002	0.614 ± 0.003	0.677 ± 0.005	0.677 ± 0.015
	Gravity	0.639 ± 0.002	0.639 ± 0.005	0.694 ± 0.008	0.694 ± 0.033
	Ext. Radiation	0.644 ± 0.001	0.644 ± 0.005	0.707 ± 0.005	0.707 ± 0.018
	BMS Plausible	0.641 ± 0.001	0.641 ± 0.004	0.697 ± 0.008	0.697 ± 0.033

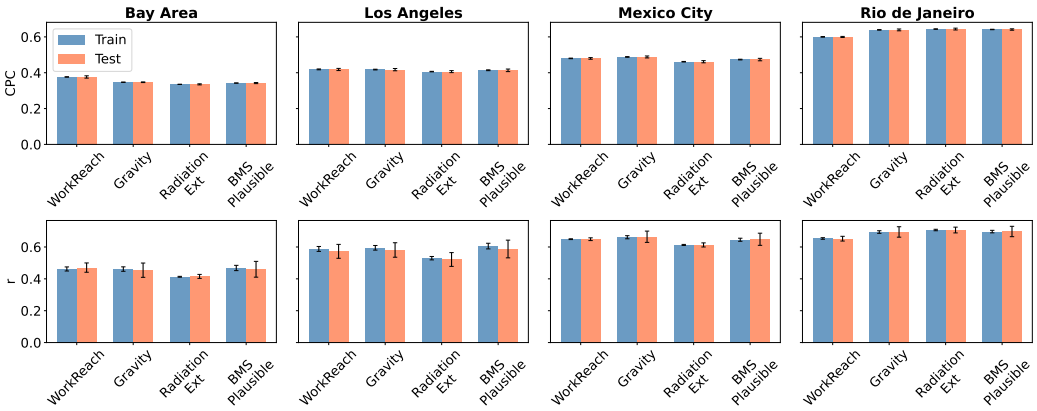

Fig. S23 Train (blue) versus test (coral) predictive performance across 5 stratified folds for all four models and cities. Top row: CPC; bottom row: Pearson r. Error bars show ± 1 standard deviation across folds. The negligible gap between train and test confirms that none of the models overfit.

S4 Robustness of the Economic Complexity Index

A central claim of WorkReach is that $\text{ECI}^{\text{employment}}$ captures a meaningful dimension of destination quality not reducible to simpler alternatives. This section examines that claim from five complementary angles. We first verify that $\text{ECI}^{\text{employment}}$ and the model's other covariate, residential informality, are not proxies for the same underlying variable (Section S4.1). We then confirm that the zone–sector binary matrices underlying the $\text{ECI}^{\text{employment}}$ computation exhibit the nested structure required for reliable complexity rankings (Section S4.2). We also characterize the product-space network that underlies the $\text{PCI}^{\text{employment}}$ and proximity calculations, documenting meaningful cross-city differences in co-location density (Section S4.3). Next, we show that $\text{ECI}^{\text{employment}}$ retains predictive power even after removing all variation explained by its two raw ingredients, diversity and mean ubiquity (Section S4.4). Finally, we compare $\text{ECI}^{\text{employment}}$ against income and wealth proxies across all four cities and find that both carry complementary information about destination attractiveness (Section S4.5).

S4.1 $\text{ECI}^{\text{employment}}$ and Informality as Complementary Dimensions

The WorkReach model enters $\text{ECI}^{\text{employment}}$ as a destination attribute and informality as an origin attribute, assigning them structurally different roles in the utility function. To verify that these two variables are not redundant, we examine their cross-zone correlation, which quantifies the degree to which they capture overlapping spatial variation. For each city, every H3 zone (or AEP zone for Rio de Janeiro) carries both an $\text{ECI}^{\text{employment}}$ value derived from its employment structure and a residential informality rate derived from census data. Figure S25 and Table S6 report this correlation across all four cities.

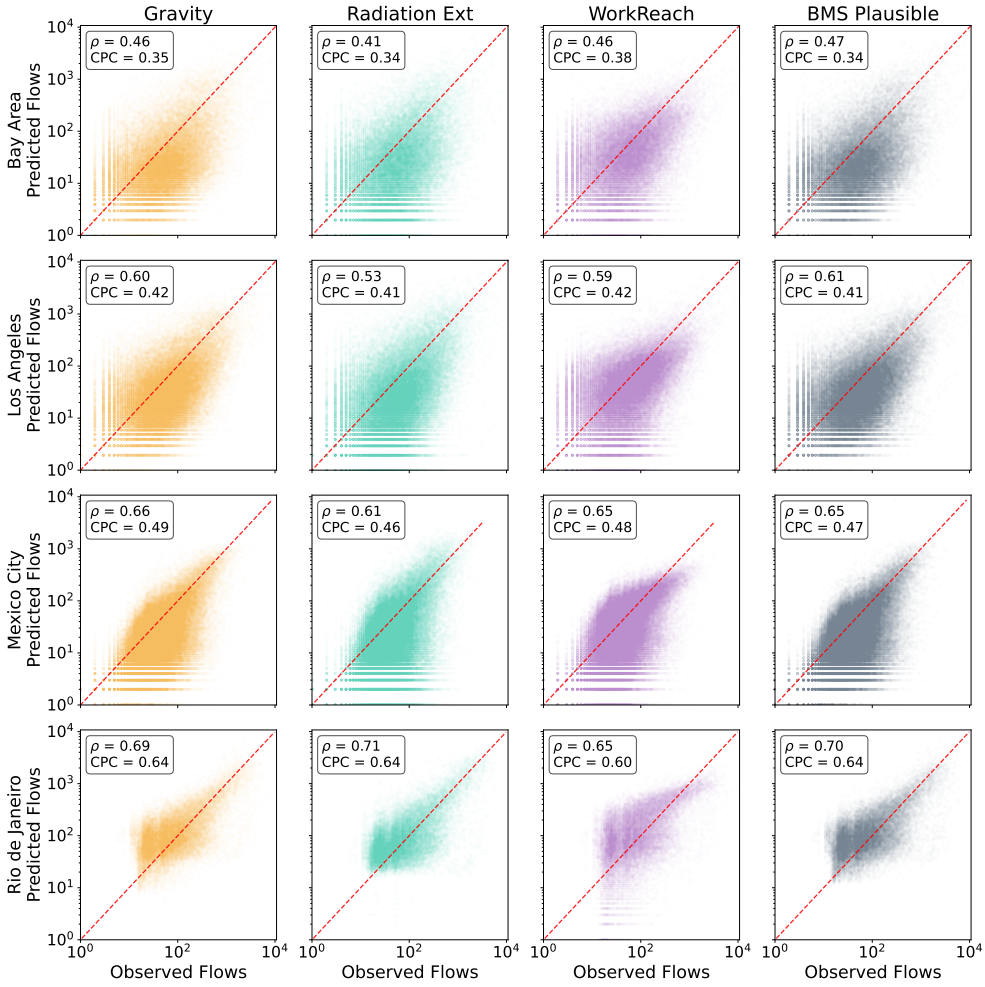

Fig. S24 Out-of-sample observed versus predicted flows (log–log scale) for all four models and cities. Each prediction is fully out of sample: origin zones were stitched across the five cross-validation folds so that every origin appears in the test set exactly once. CPC and Pearson r values are comparable to the full-data estimates reported in the main text (Figure 4), confirming that the in-sample results are not driven by overfitting.

At the zone level, the Pearson correlations for the Bay Area and Los Angeles are $r = 0.030$ and $r = 0.016$ respectively (both $p > 0.4$), indicating that $\text{ECI}^{\text{employment}}$ and residential informality are essentially uncorrelated in these labor markets. High-$\text{ECI}^{\text{employment}}$ employment zones are found across the full informality spectrum of residential zones; no systematic spatial alignment exists between productive complexity and informality in these cities. In Mexico City and Rio de Janeiro, the zone-level correlations are $r = -0.401$ ($p < 0.001$) and $r = -0.394$ ($p < 0.001$) respectively. The negative sign is consistent with a spatial structure in which central zones have

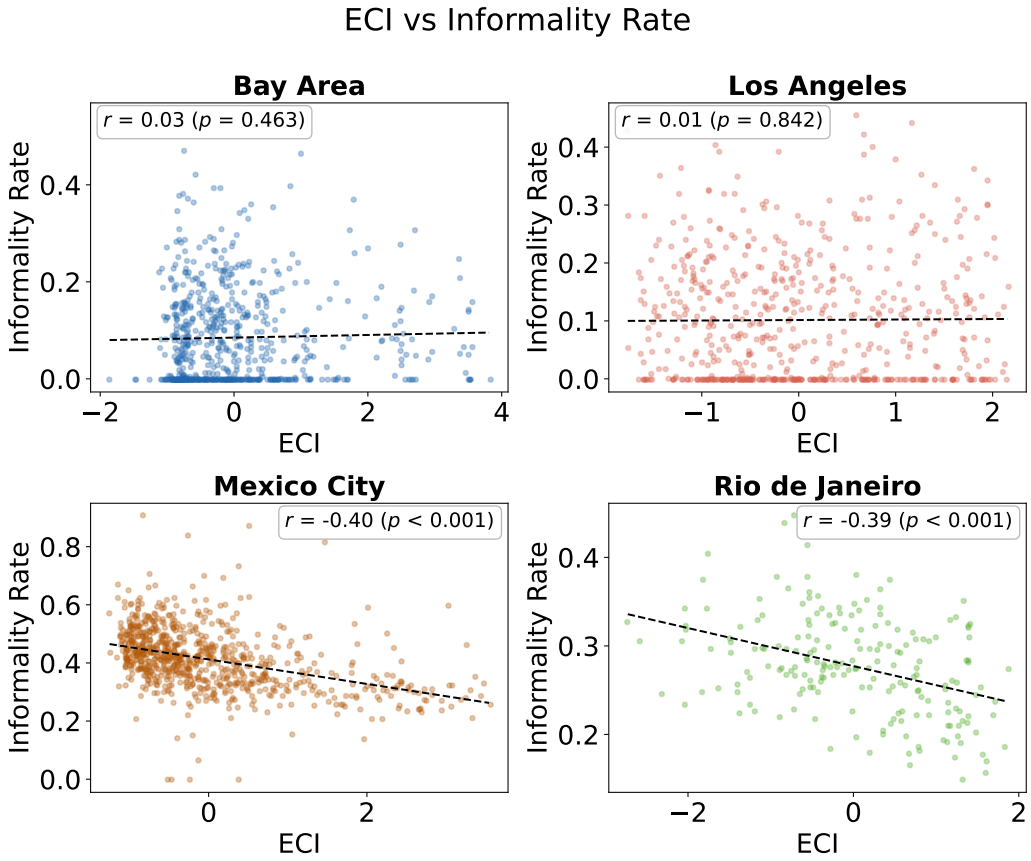

Fig. S25 Zone-level $\text{ECI}^{\text{employment}}$ versus residential informality rate for each city. Dashed lines show OLS fits; Pearson r and p-values are reported. The correlation is near-zero in the U.S. cities and moderately negative in the Latin American cities.

both higher employment complexity and lower residential informality, while peripheral zones have lower complexity and higher informality. However, the moderate magnitude ($|r| < 0.45$, $R^2 \leq 0.16$) indicates that each variable accounts for at most 16% of the variance in the other.

Because the model pairs origin informality with destination $\text{ECI}^{\text{employment}}$, we also compute the flow-weighted Pearson correlation between home-zone informality and work-zone $\text{ECI}^{\text{employment}}$ across all observed origin-destination pairs (Table S6). In the U.S. cities the flow-weighted correlations are near-zero and not statistically significant ($r = 0.022$, $p = 0.124$ for the Bay Area; $r \approx 0.000$, $p = 0.995$ for Los Angeles). In Mexico City and Rio de Janeiro they are $r = -0.153$ and $r = -0.164$ respectively (both $p < 0.001$, permutation test), substantially weaker than the zone-level values. The two variables as they actually enter the model are thus only weakly associated, with $R^2 < 0.03$ even in the Latin American cities.

Table S6 Cross-correlation between $\text{ECI}^{\text{employment}}$ and residential informality rate. The first two columns report the zone-level Pearson correlation (each zone carries both an $\text{ECI}^{\text{employment}}$ value and an informality rate). The last column reports the flow-weighted Pearson correlation computed across all observed origin-destination pairs, where origin informality and destination $\text{ECI}^{\text{employment}}$ are the two variables and commuting flows are the weights.

City	Zones (n)	Pearson r (zone)	p-value	Pearson r (flow-wtd)	p-value
Bay Area	669	0.030	0.430	0.022	0.124
Los Angeles	603	0.016	0.698	-0.000	0.995
Mexico City	792	-0.401	< 0.001	-0.153	< 0.001
Rio de Janeiro	199	-0.394	< 0.001	-0.164	< 0.001

Several mechanisms, which are not mutually exclusive, can account for why high-$\text{ECI}^{\text{employment}}$ destinations attract workers across the formality spectrum. High-$\text{ECI}^{\text{employment}}$ zones concentrate a broader and more diverse set of employment opportunities, so workers at all skill levels are more likely to find a suitable match, an effect akin to the labor-market pooling described in the agglomeration literature [3]. In addition, the employment data used to compute $\text{ECI}^{\text{employment}}$ in Latin American cities records formally registered establishments (DENUE in Mexico City; RAIS in Rio de Janeiro); employment in these zones therefore tends to carry statutory protections, including social security affiliation and health coverage, regardless of the worker's role. A worker employed in a non-specialized capacity at a formally registered firm, such as building maintenance, administrative support, or supply services, gains institutional protections that would be unavailable through informal employment near their residence. Finally, workers whose skills are not matched by local demand must cross greater distances to reach zones where that demand exists. Our aggregate flow data do not permit us to distinguish these channels, and all three are consistent with a positive β_{ECI} and a positive $\beta_{\text{informality}}$ in Latin American cities.

S4.2 Nested Structure of Zone–Sector Binary Matrices

A binary matrix is nested if the sector portfolio of less-diversified zones is a proper subset of the portfolio of more-diversified zones, a property that supports reliable $\text{ECI}^{\text{employment}}$ estimation [4]. For each city we construct the zone–sector binary matrix $M_{zs} = \mathbf{1}[\text{RCA}_{zs} \geq 1]$ and sort rows by diversity (descending) and columns by ubiquity (descending). Figure S26 shows the resulting staircase pattern across all four cities.

We quantify nestedness using the NODF metric (Nestedness based on Overlap and Decreasing Fill), which measures the extent to which the sectoral composition of less diversified zones forms subsets of more diversified ones. The observed score is compared against a null distribution obtained by randomizing the zone–sector matrix while preserving row and column totals (Table S7). In all four cities the observed NODF substantially exceeds the null expectation ($p < 0.001$), confirming significant nestedness.

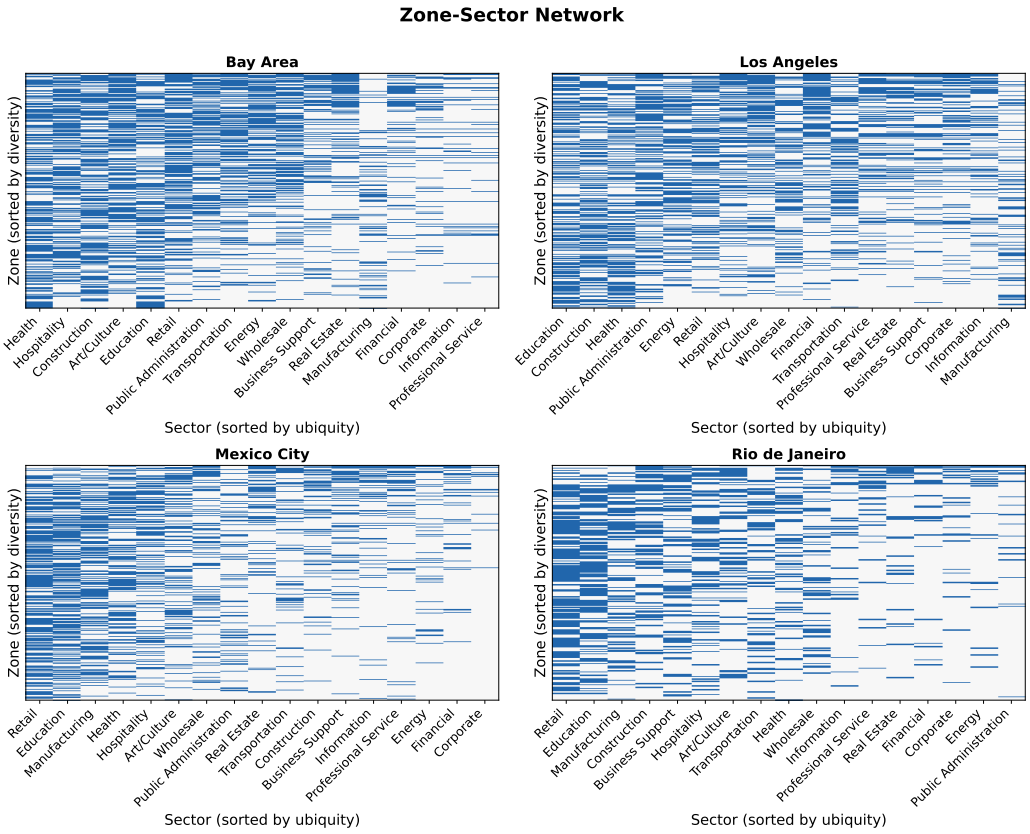

Fig. S26 Zone–sector binary RCA matrices for the four cities, sorted by zone diversity (rows, high to low) and sector ubiquity (columns, high to low). Blue cells indicate RCA ≥ 1; white cells RCA < 1. The staircase pattern is the signature of a nested structure.

Table S7 NODF nestedness scores for the four zone–sector binary matrices compared to a random null (999 permutations). All observed values significantly exceed the null expectation.

City	NODF (observed)	Null mean $\pm$ SD	p-value
Bay Area	47.7	38.1 ± 0.4	< 0.001
Los Angeles	44.0	40.0 ± 0.3	< 0.001
Mexico City	42.6	31.8 ± 0.4	< 0.001
Rio de Janeiro	42.1	32.2 ± 0.9	< 0.001

S4.3 Product-Space Network Structure

The product-space proximity matrix $\mathbf{\Phi}$ quantifies the tendency of sectors to co-locate within the same zones: $\phi_{ss'} = \Pr(\mathrm{RCA}_{zs} \geq 1 \mid \mathrm{RCA}_{zs'} \geq 1)$, taken as the minimum of the two conditional probabilities. Because $\mathrm{PCI}^{\mathrm{employment}}$ and the skill-relatedness

score are both derived from this matrix, it is useful to characterize its structure across the four cities.

With 17 NAICS sectors the complete graph contains $\binom{17}{2} = 136$ possible edges. Following the approach of Hidalgo et al. [5], who used a maximum spanning tree (MST) supplemented by all edges above a proximity threshold to visualize the product space, we construct each city's network as the union of the MST and all edges with $\phi > \phi^*$, where $\phi^* = 0.30$ for the U.S. cities and $\phi^* = 0.20$ for the Latin American cities. The MST guarantees a single connected component, while the threshold controls network density. We chose a lower cutoff for Mexico City and Rio de Janeiro because their proximity matrices are sparser, reflecting sharper sectoral specialization (Table S8).

Table S8 Product-space network edge counts (and mean degree) under different filtering strategies. "MST only" retains the 16 edges of the maximum spanning tree. Each subsequent column shows the result of supplementing the MST with all edges above the indicated proximity threshold. The final column ($\phi > 0.20$ only) applies a fixed threshold without the MST. The maximum possible number of edges is 136 ($= \binom{17}{2}$). Bold values indicate the configuration used in the manuscript figures.

City	MST	$\phi > 0.30$	$\phi > 0.40$	$\phi > 0.50$	$\phi > 0.60$	$\phi > 0.20$ only
Bay Area	16	**62 (7.3)**	48 (5.6)	26 (3.1)	18 (2.1)	86 (10.1)
Los Angeles	16	**73 (8.6)**	42 (4.9)	27 (3.2)	20 (2.4)	106 (12.5)
Mexico City	16	30 (3.5)	20 (2.4)	16 (1.9)	16 (1.9)	**57 (6.7)**
Rio de Janeiro	16	33 (3.9)	17 (2.0)	16 (1.9)	16 (1.9)	**60 (7.1)**

The U.S. cities exhibit denser co-location structure: many sectors co-occur at comparable rates, yielding numerous edges above even a moderate threshold. The Latin American cities display sparser networks with fewer strong ties, indicating sharper sectoral specialization in which zones tend to concentrate employment in a smaller set of sectors. This asymmetry is consistent with the higher nestedness scores observed for the U.S. cities (Table S7). Figure S27 visualizes the product-space network for each city, with nodes colored by $\text{PCI}^{\text{employment}}$ and sized by total sector employment.

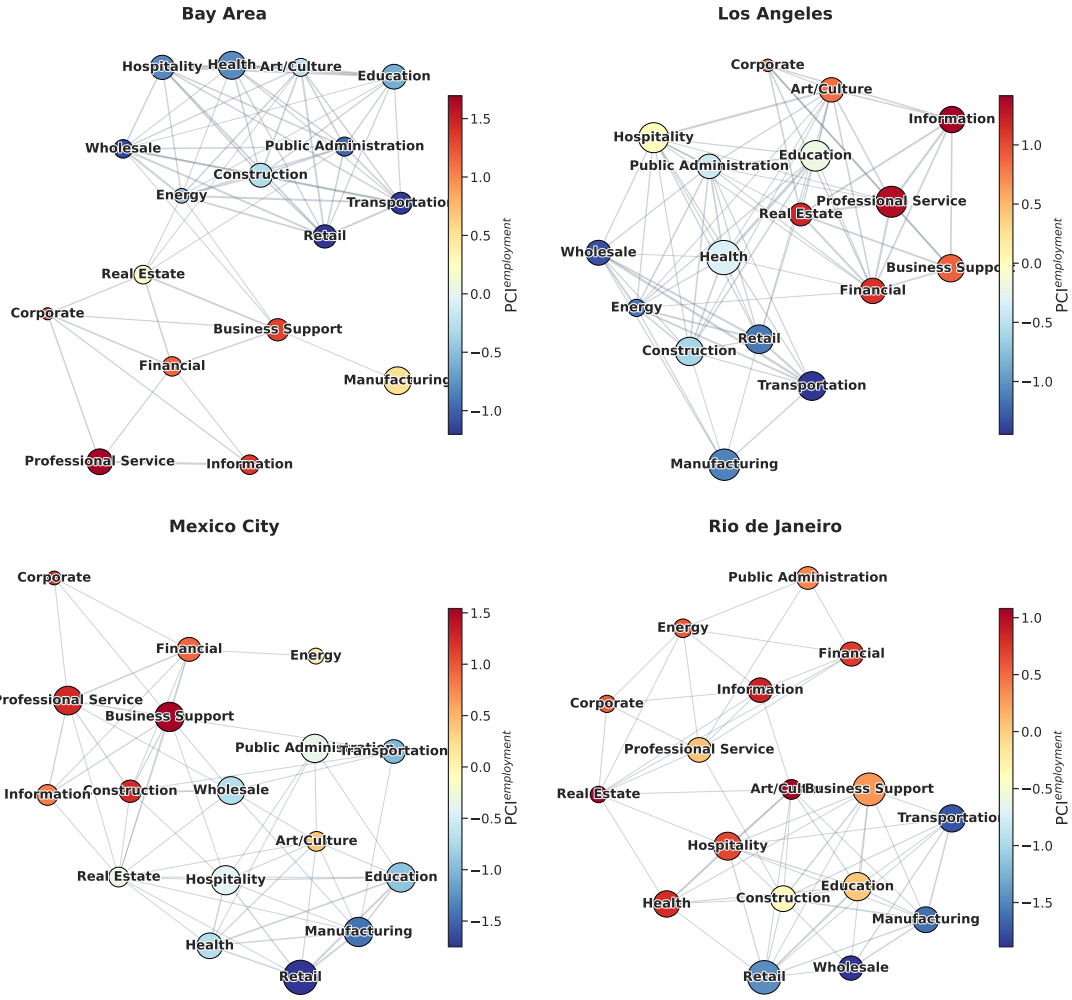

Fig. S27 Product-space networks for the four cities. Nodes represent NAICS sectors, colored by $\text{PCI}^{\text{employment}}$ and sized by total sector employment. Edge width is proportional to proximity ϕ. U.S. cities use MST + $\phi > 0.30$; Latin American cities use MST + $\phi > 0.20$. The denser networks in the U.S. cities reflect more homogeneous sector co-location patterns, while the sparser Latin American networks indicate sharper sectoral specialization.

S4.4 $\mathrm{ECI}^{\mathrm{employment}}$ versus Diversity and Ubiquity

$\mathrm{ECI}^{\mathrm{employment}}$ is constructed from two zone-level quantities: industrial diversity (the number of sectors in which a zone holds a revealed comparative advantage) and sector ubiquity (how widespread those sectors are across zones). To test whether $\mathrm{ECI}^{\mathrm{employment}}$ carries information about destination attractiveness beyond what diversity and ubiquity alone provide, we decompose it into a component explained by these two inputs and an orthogonal residual, then evaluate whether the residual retains predictive power in the WorkReach model.

Definitions and Orthogonal Decomposition

Given the binary RCA matrix M_{zs}, we define for each zone z:

$$\mathrm{Diversity}_z = \sum_s M_{zs}\,, \tag{1}$$

$$\overline{\mathrm{Ubiquity}}_z = \frac{1}{\mathrm{Diversity}_z} \sum_s M_{zs} \cdot \left(\sum_{z'} M_{z's} \right), \tag{2}$$

where $\overline{\mathrm{Ubiquity}}_z$ is the mean ubiquity of the sectors in which zone z has a comparative advantage. We then regress $\mathrm{ECI}^{\mathrm{employment}}$ on diversity and mean ubiquity via ordinary least squares:

$$\mathrm{ECI}_z^{\mathrm{employment}} = \alpha + \beta_D\,\mathrm{Diversity}_z + \beta_U\,\overline{\mathrm{Ubiquity}}_z + \varepsilon_z\,. \tag{3}$$

The residual ε_z, which we denote $\mathrm{ECI}_z^{\mathrm{employment},\perp}$, is orthogonal to both regressors by construction and isolates the information in $\mathrm{ECI}^{\mathrm{employment}}$ that cannot be recovered from diversity or mean ubiquity alone.

Table S9 reports the R^2 of this regression for each city. The fraction of $\mathrm{ECI}^{\mathrm{employment}}$ variance explained by its two ingredients varies widely: from 90% in the Bay Area to 9% in Rio de Janeiro. In all cities, a non-trivial share of $\mathrm{ECI}^{\mathrm{employment}}$ variation is orthogonal to both components, confirming that the iterative reflections extract a joint signal that neither diversity nor mean ubiquity fully captures.

Correlation between $\mathrm{ECI}^{\mathrm{employment}}$ and Its Components

Figure S28 shows $\mathrm{ECI}^{\mathrm{employment}}$ plotted against diversity (top row) and against negative mean ubiquity (bottom row) for all four cities. $\mathrm{ECI}^{\mathrm{employment}}$ is positively correlated with diversity (r between 0.18 and 0.49) and negatively correlated with mean ubiquity (r between -0.29 and -0.95), but the strength varies across cities. In the Bay Area and Mexico City, mean ubiquity is a strong predictor of $\mathrm{ECI}^{\mathrm{employment}}$ ($r = -0.95$ and -0.87), whereas in Los Angeles and Rio de Janeiro the relationship is weaker ($r = -0.46$ and -0.29). Diversity shows moderate to weak correlations in all cities.

Table S9 Variance in $\text{ECI}^{\text{employment}}$ explained by diversity and mean ubiquity (R^2 of OLS regression). The residual $\text{ECI}^{\text{employment},\perp}$ captures the remaining variation, which is orthogonal to both regressors by construction.

City	R^2
Bay Area	0.900
Los Angeles	0.219
Mexico City	0.757
Rio de Janeiro	0.090

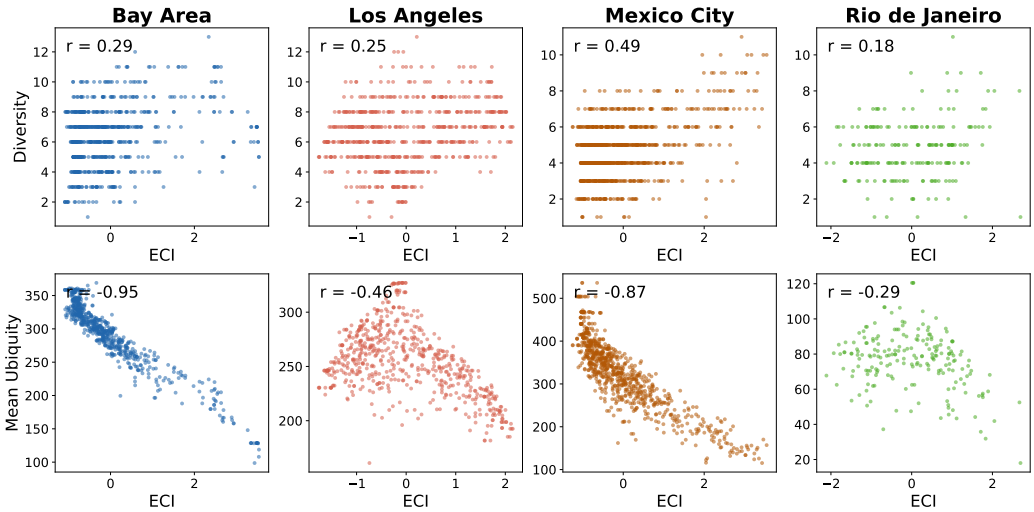

Fig. S28 Zone-level $\text{ECI}^{\text{employment}}$ versus diversity (top row) and mean ubiquity (bottom row, sign inverted for visual clarity) for the four cities. Pearson correlation coefficients are shown for each city.

Predictive Performance of $\text{ECI}^{\text{employment}}$ and Its Components

To test whether diversity or ubiquity alone can substitute for $\text{ECI}^{\text{employment}}$ in the WorkReach model, we estimate four specifications per city: the original WorkReach with $\text{ECI}^{\text{employment}}$, a diversity-only variant, a ubiquity-only variant, and a variant using the orthogonal residual $\text{ECI}^{\text{employment},\perp}$. Table S10 reports CPC and r with bootstrap 95% confidence intervals.

Ubiquity alone yields higher CPC and r than $\text{ECI}^{\text{employment}}$ in all four cities, and diversity outperforms $\text{ECI}^{\text{employment}}$ in Mexico City and Rio de Janeiro. Because $\text{ECI}^{\text{employment}}$ combines diversity and ubiquity into a single index, a model that uses

Table S10 WorkReach model performance using $\text{ECI}^{\text{employment}}$, diversity alone, mean ubiquity alone, or the orthogonal residual $\text{ECI}^{\text{employment},\perp}$ as the destination attribute. Bootstrap 95% confidence intervals (1,000 replicates) are shown in brackets.

City	Model	CPC [95% CI]	r [95% CI]
Bay Area	WorkReach	0.389 [0.381, 0.397]	0.524 [0.487, 0.560]
	Diversity	0.363 [0.355, 0.370]	0.470 [0.454, 0.493]
	Ubiquity	0.392 [0.384, 0.400]	0.514 [0.493, 0.536]
	$\text{ECI}^{\text{employment},\perp}$	0.367 [0.359, 0.375]	0.486 [0.459, 0.509]
Los Angeles	WorkReach	0.413 [0.405, 0.420]	0.580 [0.539, 0.612]
	Diversity	0.412 [0.404, 0.420]	0.574 [0.546, 0.593]
	Ubiquity	0.430 [0.422, 0.437]	0.601 [0.573, 0.619]
	$\text{ECI}^{\text{employment},\perp}$	0.412 [0.404, 0.419]	0.569 [0.539, 0.590]
Mexico City	WorkReach	0.486 [0.480, 0.493]	0.658 [0.648, 0.667]
	Diversity	0.486 [0.480, 0.492]	0.661 [0.651, 0.670]
	Ubiquity	0.496 [0.489, 0.502]	0.682 [0.671, 0.693]
	$\text{ECI}^{\text{employment},\perp}$	0.483 [0.477, 0.490]	0.661 [0.651, 0.670]
Rio de Janeiro	WorkReach	0.614 [0.606, 0.621]	0.677 [0.659, 0.693]
	Diversity	0.623 [0.616, 0.629]	0.690 [0.673, 0.706]
	Ubiquity	0.654 [0.646, 0.661]	0.735 [0.719, 0.749]
	$\text{ECI}^{\text{employment},\perp}$	0.609 [0.601, 0.616]	0.670 [0.652, 0.687]

one component directly has more flexibility to fit city-specific patterns. Notably, even the orthogonal residual $\text{ECI}^{\text{employment},\perp}$, which by construction contains no information from diversity or mean ubiquity, substantially outperforms the randomized-$\text{ECI}^{\text{employment}}$ null from the permutation analysis (Table S4): for example, CPC of 0.367 versus 0.330 in the Bay Area and 0.609 versus 0.572 in Rio de Janeiro. This confirms that $\text{ECI}^{\text{employment}}$ encodes information about destination attractiveness beyond what its two constituent ingredients provide. The advantage of $\text{ECI}^{\text{employment}}$ is not strictly predictive but structural: it is embedded in the broader economic complexity framework, which provides the feasibility score needed to identify which new sectors a zone could realistically develop and to simulate the resulting change in commuting accessibility (Section S8). Neither diversity nor ubiquity alone supports this type of counterfactual analysis.

S4.5 $\text{ECI}^{\text{employment}}$ versus Income and Wealth Proxies

$\text{ECI}^{\text{employment}}$ captures the sophistication of a zone's employment structure, which may overlap with local income or wealth levels. To test whether it adds predictive value beyond a direct economic proxy, we estimated three model variants for each city: the original WorkReach specification ($\text{ECI}^{\text{employment}}$ only), a wage-only specification replacing $\text{ECI}^{\text{employment}}$ with a measure of destination economic status, and a combined model including both. Because direct wage data at the work location are unavailable for the Latin American cities, we construct city-specific proxies described below.

Income and Wealth Data Sources

Bay Area and Los Angeles. Wage data come from the Replica synthetic employment dataset. We aggregate individual incomes at the work location to zone-level means.

Rio de Janeiro. Individual-level wage records are not available. Instead, we use residential household income from the 2020 Brazilian Census at the zone level (in BRL per month). Because census income is measured at the *residential* location, direct assignment to destination zones would mix residential and employment characteristics. We therefore aggregate income to the employment location using the observed commuting flows as weights:

$$\bar{y}_j = \frac{\sum_i T_{ij}\, y_i}{\sum_i T_{ij}}\,, \tag{4}$$

where y_i is the residential-zone mean household income and T_{ij} is the commuting flow from zone i to zone j. This yields a flow-weighted income proxy for each destination zone, reflecting the income distribution of workers who actually commute there.

Mexico City. No equivalent income variable with broad spatial coverage is available for Mexico City. We construct a composite wealth index from the 2020 Mexican Population and Housing Census at the AGEB (census tract) level, following previous work [6]. The index is based on a principal component analysis (PCA) of sixteen household-level indicators: access to internet, motor vehicle, computer, microwave, washing machine, cable television, streaming services, videogame consoles, and mobile phone; share of dwellings with three or more rooms; access to a water storage tank; absence of dirt floors; absence of overcrowding; and an overall access-to-information-devices composite. The first principal component explains the largest share of variance across these dimensions and is converted to an index (0–1). The resulting residential wealth index is then aggregated to destination zones using the same flow-weighting formula as for Rio de Janeiro.

$\mathrm{ECI}^{\mathrm{employment}}$ and Income/Wealth Correlation

Figure S29 shows the zone-level relationship between $\mathrm{ECI}^{\mathrm{employment}}$ and the respective income or wealth proxy across all four cities. $\mathrm{ECI}^{\mathrm{employment}}$ and destination economic status are positively correlated in every city (Pearson $r = 0.53$ for the Bay Area, $r = 0.61$ for Los Angeles, $r = 0.59$ for Mexico City, and $r = 0.54$ for Rio de Janeiro; all $p < 0.001$). The correlation is moderate rather than near-perfect, indicating that $\mathrm{ECI}^{\mathrm{employment}}$ and wages or wealth capture overlapping but distinct dimensions of destination quality.

Model Comparison

Table S11 reports predictive performance with block-bootstrap 95% confidence intervals for all four cities using mean income or wealth as the destination proxy. For Rio de Janeiro and Mexico City the proxy is the flow-weighted mean of the residential income or wealth variable defined above.

Likelihood-ratio (LR) tests confirm that each variable contributes independently when added to the other (Table S12). In every city and both test directions, the

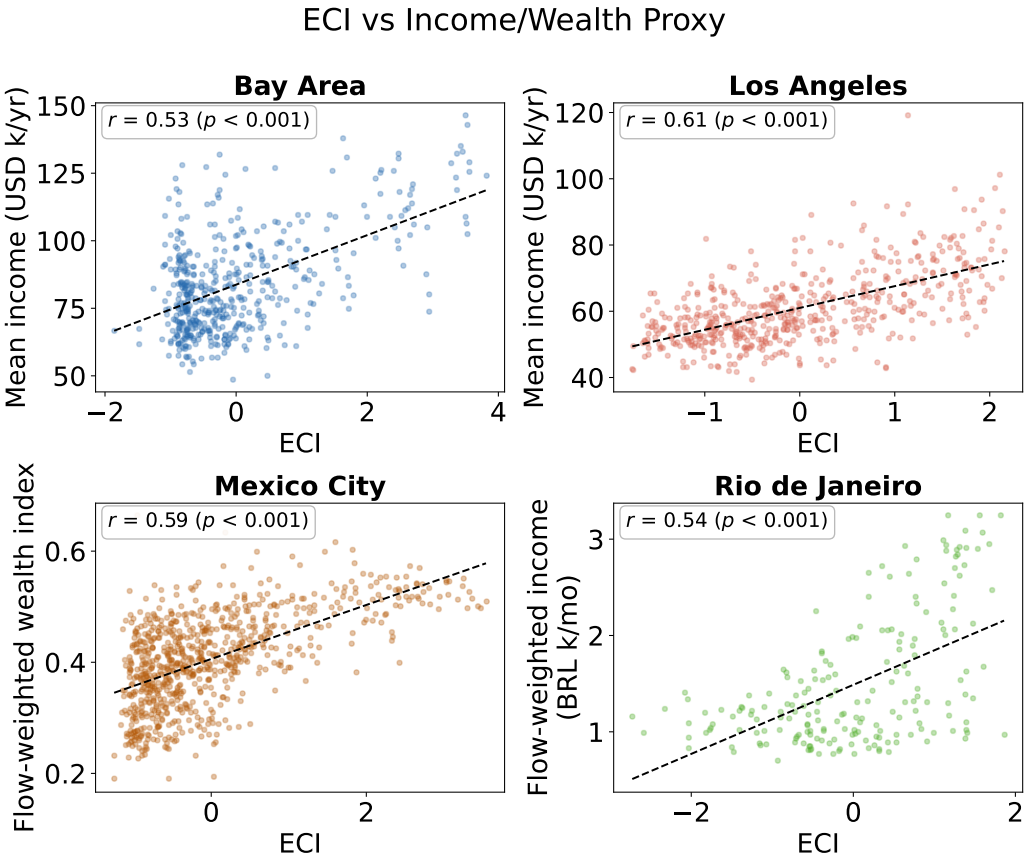

Fig. S29 Zone-level $\text{ECI}^{\text{employment}}$ versus employment-location income or wealth proxy across all four cities. U.S. cities use mean annual income from the Replica dataset (USD); Rio de Janeiro uses flow-weighted mean household income from the 2020 Census (BRL/month); Mexico City uses a flow-weighted mean of a PCA-based wealth percentile (0–1) derived from 16 census indicators. Dashed lines show OLS fits; correlation statistics are shown for each city.

improvement in log-likelihood is highly significant ($p < 0.001$), indicating that $\text{ECI}^{\text{employment}}$ and the income or wealth proxy carry complementary information about destination attractiveness.

In the Bay Area, $\text{ECI}^{\text{employment}}$ outperforms the wage proxy ($\Delta\text{CPC} = +0.012$; $\Delta r = +0.040$), while in Los Angeles the two are essentially tied ($\Delta\text{CPC} = +0.001$; $\Delta r = +0.008$). In the Latin American cities, the flow-weighted income or wealth proxy outperforms $\text{ECI}^{\text{employment}}$ alone ($\Delta\text{CPC} = -0.011$ for Mexico City, -0.022 for Rio de Janeiro). An important caveat applies to these two cities: because no direct work-location wage data are available, we construct the proxy by aggregating residential income to destination zones using the observed commuting flows themselves. This flow-weighting procedure introduces a circularity, as the same flow matrix enters both the

Table S11 Predictive performance across all four cities for three model variants: WorkReach, Proxy-only, and combined ($ECI^{employment}$ + Proxy). The proxy is mean annual income for U.S. cities (Replica), flow-weighted mean household income for Rio de Janeiro (2020 Census, BRL/month), and flow-weighted mean PCA wealth percentile for Mexico City (0–1).

City	Model	CPC [95% CI]	r [95% CI]
Bay Area	WorkReach	0.389 [0.381, 0.397]	0.524 [0.487, 0.560]
	Proxy-only	0.377 [0.368, 0.386]	0.484 [0.450, 0.520]
	WorkReach + Proxy	0.390 [0.382, 0.399]	0.537 [0.501, 0.573]
Los Angeles	WorkReach	0.413 [0.405, 0.420]	0.580 [0.539, 0.612]
	Proxy-only	0.412 [0.404, 0.419]	0.572 [0.544, 0.590]
	WorkReach + Proxy	0.413 [0.406, 0.421]	0.580 [0.553, 0.597]
Mexico City	WorkReach	0.486 [0.480, 0.493]	0.658 [0.648, 0.667]
	Proxy-only	0.497 [0.491, 0.503]	0.675 [0.661, 0.688]
	WorkReach + Proxy	0.499 [0.493, 0.506]	0.677 [0.663, 0.690]
Rio de Janeiro	WorkReach	0.614 [0.606, 0.621]	0.677 [0.659, 0.693]
	Proxy-only	0.636 [0.628, 0.644]	0.680 [0.660, 0.700]
	WorkReach + Proxy	0.635 [0.628, 0.643]	0.679 [0.660, 0.698]

Table S12 Likelihood-ratio tests for the independent significance of $ECI^{employment}$ and the income or wealth proxy across all four cities.

City	Test	$\Delta(2\ell)$	p-value
Bay Area	WorkReach → $ECI^{employment}$ + Proxy (proxy adds to $ECI^{employment}$)	263,136	< 0.001
	Proxy-only → $ECI^{employment}$ + Proxy ($ECI^{employment}$ adds to proxy)	831,813	< 0.001
Los Angeles	WorkReach → $ECI^{employment}$ + Proxy (proxy adds to $ECI^{employment}$)	2,596	< 0.001
	Proxy-only → $ECI^{employment}$ + Proxy ($ECI^{employment}$ adds to proxy)	118,590	< 0.001
Mexico City	WorkReach → $ECI^{employment}$ + Proxy (proxy adds to $ECI^{employment}$)	839,904	< 0.001
	Proxy-only → $ECI^{employment}$ + Proxy ($ECI^{employment}$ adds to proxy)	279,335	< 0.001
Rio de Janeiro	WorkReach → $ECI^{employment}$ + Proxy (proxy adds to $ECI^{employment}$)	604,977	< 0.001
	Proxy-only → $ECI^{employment}$ + Proxy ($ECI^{employment}$ adds to proxy)	194,580	< 0.001

dependent variable and the construction of the covariate, which may inflate the apparent predictive advantage of the proxy in Mexico City and Rio de Janeiro. In all cases the absolute CPC differences remain small (≤ 0.022), and both covariates contribute independently to the combined model (all LR tests $p < 0.001$). Importantly, wages capture only one dimension of destination quality: the economic complexity framework additionally encodes inter-industry relatedness, productive knowledge agglomeration, and the feasibility of diversification into new activities, dimensions that we explore further in the skill-relatedness analysis (Section S5).

Bootstrap Significance of $ECI^{employment}$ vs. Proxy Differences

We test whether these CPC and r differences are significant using a block bootstrap (1,000 replicates), resampling origin zones with replacement to preserve within-origin

correlation. Figure S30 shows the resulting 95% confidence intervals. In the Bay Area, the CPC difference between $\text{ECI}^{\text{employment}}$ and the wage proxy is not significant ($p = 0.38$), although the Pearson r difference is ($p = 0.001$). In Los Angeles, $\text{ECI}^{\text{employment}}$ significantly outperforms the proxy on both CPC ($p < 0.001$) and r ($p < 0.001$). In Mexico City and Rio de Janeiro, the proxy outperforms $\text{ECI}^{\text{employment}}$ on both metrics ($p \leq 0.004$), though the circularity noted above applies to this comparison. Taken together, the results confirm that $\text{ECI}^{\text{employment}}$ and the income proxy are not interchangeable predictors across the four cities.

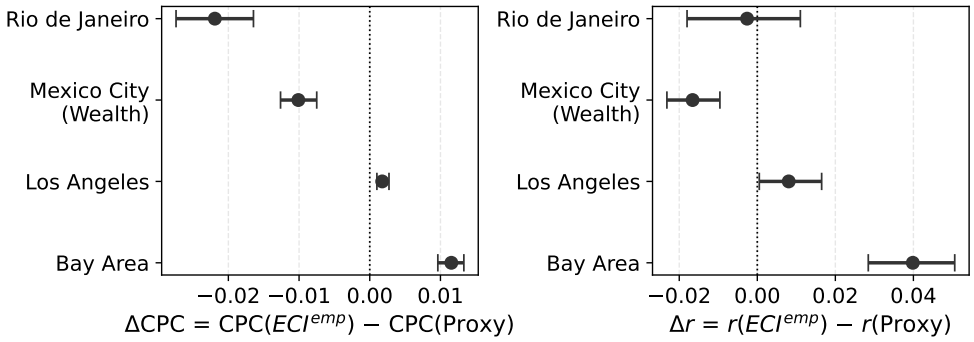

Fig. S30 Bootstrap 95% confidence intervals for the difference in CPC (left) and Pearson r (right) between the $\text{ECI}^{\text{employment}}$-only model and the income/wealth proxy-only model across all four cities. Values above zero indicate that $\text{ECI}^{\text{employment}}$ outperforms the proxy; values below zero indicate the reverse.

S5 Skill Relatedness Analysis

The economic complexity framework encodes not only the sophistication of individual zones ($\text{ECI}^{\text{employment}}$) but also the relatedness between economic sectors. Two sectors are considered skill-related when they tend to co-locate in the same zones, which reflects shared labor requirements, technologies, or institutional conditions. This co-location structure is captured by the product-space proximity matrix $\mathbf{\Phi}$ [5], whose entries are

$$\phi_{ss'} = \frac{\sum_z M_{zs} M_{zs'}}{\max(K_s, K_{s'})},$$

where M_{zs} is the Revealed Comparative Advantage indicator and $K_s = \sum_z M_{zs}$ is the ubiquity of sector s. The matrix $\mathbf{\Phi} \in \mathbb{R}^{S \times S}$ has entries $[\mathbf{\Phi}]_{ss'} = \phi_{ss'}$, so that each element measures how frequently sectors s and s' appear together across the urban economy, normalized by the more common of the two.

For each origin–destination pair (i, j), we aggregate these pairwise sector proximities into a single skill-relatedness score

$$\text{SR}_{ij} = \mathbf{c}_i \cdot \mathbf{\Phi} \cdot \mathbf{c}_j ,$$

where $\mathbf{c}_i = (c_{i1}, \ldots, c_{iS})$ is the employment-share vector of zone i across $S = 17$ harmonized NAICS sectors. Intuitively, SR_{ij} is high when both zones specialize in sectors that are closely related in the product space, and low when the destination offers employment that is structurally different from what the origin provides.

Following the same comparison strategy as for the income and wealth proxies (Section S4.5), we estimated three model variants per city: the original WorkReach, a model replacing $\text{ECI}^{\text{employment}}$ with SR, and a combined model including both. The SR and combined models are estimated on the subset of origin–destination pairs for which both zones have sector-level data; WorkReach is re-estimated on the same subset for a valid likelihood-ratio comparison.

Table S14 reports parameter estimates and performance metrics for all variants across the four cities. The SR coefficient is negative in all four cities, indicating that workers systematically commute to destinations whose sector composition is less related to their home zone, consistent with a necessity or skill-mismatch hypothesis rather than a skill-matching one.

Likelihood-ratio tests confirm that SR significantly improves model fit beyond $\text{ECI}^{\text{employment}}$ alone in all four cities, and that $\text{ECI}^{\text{employment}}$ contributes independently beyond SR in all but Rio de Janeiro (Table S15).

$\text{ECI}^{\text{employment}}$ remains independently significant after controlling for SR in all four cities, as seen by the combined model retaining a positive β_{ECI} and a significant LR test statistic in every case. The marginal contribution of $\text{ECI}^{\text{employment}}$ beyond SR is largest in the Bay Area ($\Delta(2\ell) = 517{,}799$) and smallest in Rio de Janeiro ($\Delta(2\ell) = 140{,}059$). Pairwise Pearson correlations between destination $\text{ECI}^{\text{employment}}$ and SR across all off-diagonal origin–destination pairs are moderate in Mexico City ($r = -0.51$) and the Bay Area ($r = -0.48$), weak in Rio de Janeiro ($r = -0.24$), and negligible in Los Angeles ($r = -0.02$). The negative sign of β_{SR} across all cities supports the interpretation that long-distance commutes are associated with sector

Table S13 Predictive performance for WorkReach, SR-only, and combined ($ECI^{employment}$ + SR) models, with block-bootstrap 95% confidence intervals (2,000 replicates).

City	Model	CPC [95% CI]	r [95% CI]
Bay Area	WorkReach	0.389 [0.381, 0.395]	0.524 [0.485, 0.560]
	SR-only	0.391 [0.382, 0.398]	0.522 [0.499, 0.543]
	WorkReach+SR	0.399 [0.390, 0.405]	0.553 [0.530, 0.575]
Los Angeles	WorkReach	0.413 [0.405, 0.421]	0.580 [0.542, 0.606]
	SR-only	0.425 [0.415, 0.434]	0.583 [0.567, 0.602]
	WorkReach+SR	0.428 [0.419, 0.436]	0.597 [0.570, 0.613]
Mexico City	WorkReach	0.486 [0.480, 0.492]	0.658 [0.648, 0.666]
	SR-only	0.495 [0.488, 0.501]	0.684 [0.671, 0.694]
	WorkReach+SR	0.497 [0.491, 0.503]	0.681 [0.668, 0.691]
Rio de Janeiro	WorkReach	0.614 [0.607, 0.621]	0.677 [0.661, 0.694]
	SR-only	0.652 [0.646, 0.659]	0.734 [0.719, 0.747]
	WorkReach+SR	0.655 [0.649, 0.662]	0.735 [0.720, 0.748]

dissimilarity, meaning that workers travel to zones offering employment in sectors different from those available near home, consistent with an occupational-mismatch mechanism rather than skill matching.

$ECI^{employment}$ versus Skill Relatedness: Predictive and Policy Trade-offs

The SR-only and $ECI^{employment}$+SR models outperform WorkReach in CPC and r across all four cities, raising a natural question about why we center the WorkReach specification on $ECI^{employment}$ rather than SR. The answer lies in the distinct analytical roles of the two quantities and their suitability for policy applications.

$ECI^{employment}$ is an intrinsic property of a destination zone: it summarizes the diversity and sophistication of the zone's productive structure without reference to any particular origin. Skill relatedness, by contrast, is a pairwise quantity that measures the overlap in sector composition between a specific origin and a specific destination. It cannot be attributed to either zone in isolation; the same destination may have high SR with respect to one origin and low SR with respect to another. This asymmetry has three consequences for policy translation.

First, because $ECI^{employment}$ is a zone-level attribute, it can be paired with the feasibility score [5, 7] to design targeted economic development strategies. The feasibility score identifies which new sectors a zone could realistically develop given the relatedness of those sectors to the zone's existing economic activities. An increase in $ECI^{employment}$ resulting from the adoption of new sectors can be fed directly into WorkReach to simulate the change in commuting accessibility, as we demonstrate in Section S8. No comparable intervention exists for skill relatedness: because the metric depends jointly on the compositions of both origin and destination, there is no well-defined policy lever at a single location that would systematically shift SR across all origin–destination pairs.

Table S14 Parameter estimates with multinomial bootstrap 95% confidence intervals for WorkReach, SR-only, and combined ($\text{ECI}^{\text{employment}}$ + SR) models. Threshold τ is reported in km.

City	Model	β_d	β_{ECI}	β_{SR}	β_{inf}	τ [km]	k
Bay Area	WorkReach	−21.292 [−21.305, −21.278]	2.462 [2.459, 2.465]	–	−14.735 [−14.762, −14.707]	0.006 [0.006, 0.006]	128.764 [128.760, 128.766]
	SR-only	−13.140 [−13.152, −13.127]	–	−8.542 [−8.550, −8.533]	−2.427 [−2.457, −2.408]	0.006 [0.006, 0.006]	28.165 [28.159, 28.173]
	$\text{ECI}^{\text{employment}}$+SR	−17.683 [−17.701, −17.667]	1.562 [1.557, 1.566]	−8.542 [−8.552, −8.532]	−4.362 [−4.389, −4.340]	0.006 [0.006, 0.006]	37.972 [37.964, 37.979]
Los Angeles	WorkReach	−9.601 [−9.609, −9.593]	0.482 [0.479, 0.484]	–	−8.702 [−8.716, −8.687]	0.006 [0.006, 0.006]	32.413 [32.398, 32.439]
	SR-only	−5.778 [−5.791, −5.766]	–	−6.756 [−6.773, −6.742]	−1.794 [−1.808, −1.775]	2.189 [2.158, 2.213]	19.589 [19.541, 19.646]
	$\text{ECI}^{\text{employment}}$+SR	−5.858 [−5.872, −5.847]	0.698 [0.695, 0.701]	−6.714 [−6.729, −6.701]	−1.836 [−1.850, −1.818]	2.599 [2.571, 2.619]	20.438 [20.391, 20.489]
Mexico City	WorkReach	−21.881 [−21.898, −21.861]	4.220 [4.208, 4.230]	–	10.250 [10.207, 10.280]	16.201 [16.169, 16.222]	10.824 [10.782, 10.880]
	SR-only	−7.390 [−7.402, −7.381]	–	−4.595 [−4.604, −4.587]	−0.665 [−0.682, −0.650]	2.252 [2.246, 2.258]	46.397 [46.274, 46.512]
	$\text{ECI}^{\text{employment}}$+SR	−7.384 [−7.396, −7.375]	1.050 [1.047, 1.056]	−3.554 [−3.563, −3.543]	−4.620 [−4.644, −4.602]	2.254 [2.249, 2.260]	47.955 [47.951, 47.960]
Rio	WorkReach	−16.447 [−16.459, −16.424]	6.603 [6.593, 6.615]	–	4.267 [4.244, 4.298]	24.955 [24.904, 25.026]	9.654 [9.580, 9.720]
	SR-only	−3.692 [−3.701, −3.681]	–	−6.756 [−6.768, −6.743]	−0.792 [−0.814, −0.768]	2.126 [2.113, 2.139]	27.612 [27.524, 27.688]
	$\text{ECI}^{\text{employment}}$+SR	−4.027 [−4.036, −4.014]	0.935 [0.931, 0.939]	−6.756 [−6.767, −6.745]	−1.133 [−1.156, −1.112]	2.128 [2.114, 2.139]	27.617 [27.530, 27.684]

Table S15 Likelihood-ratio tests for the independent significance of $\text{ECI}^{\text{employment}}$ and skill relatedness across all four cities.

City	Test	$\Delta(2\ell)$	p-value
Bay Area	WR $\rightarrow$ $\text{ECI}^{\text{employment}}$ + SR (SR adds to $\text{ECI}^{\text{employment}}$)	552,839	< 0.001
	SR $\rightarrow$ $\text{ECI}^{\text{employment}}$ + SR ($\text{ECI}^{\text{employment}}$ adds to SR)	517,799	< 0.001
Los Angeles	WR $\rightarrow$ $\text{ECI}^{\text{employment}}$ + SR (SR adds to $\text{ECI}^{\text{employment}}$)	1,109,813	< 0.001
	SR $\rightarrow$ $\text{ECI}^{\text{employment}}$ + SR ($\text{ECI}^{\text{employment}}$ adds to SR)	231,432	< 0.001
Mexico City	WR $\rightarrow$ $\text{ECI}^{\text{employment}}$ + SR (SR adds to $\text{ECI}^{\text{employment}}$)	723,691	< 0.001
	SR $\rightarrow$ $\text{ECI}^{\text{employment}}$ + SR ($\text{ECI}^{\text{employment}}$ adds to SR)	220,929	< 0.001
Rio de Janeiro	WR $\rightarrow$ $\text{ECI}^{\text{employment}}$ + SR (SR adds to $\text{ECI}^{\text{employment}}$)	1,445,363	< 0.001
	SR $\rightarrow$ $\text{ECI}^{\text{employment}}$ + SR ($\text{ECI}^{\text{employment}}$ adds to SR)	140,059	< 0.001

Second, $\text{ECI}^{\text{employment}}$ captures information that neither SR nor wages provide. SR measures sector overlap between two zones but says nothing about the sophistication of those sectors; wages summarize current compensation but not the productive structure that sustains it. $\text{ECI}^{\text{employment}}$, by contrast, reflects which sectors co-locate and how that pattern compares across zones, encoding the agglomeration externalities and tacit knowledge that underpin long-run economic development [8, 9]. Wages, tested as a covariate in Section S4.5, summarize current compensation levels but provide no information about the structural conditions that sustain or could improve those levels. A policy objective framed as "increase wages at location j" lacks a mechanism; the complexity framework supplies that mechanism by identifying which productive activities are feasible and which would raise $\text{ECI}^{\text{employment}}$.

Third, adding SR as a sixth covariate to the existing WorkReach specification improves fit but complicates interpretation. The five-parameter model has a clear structure: each coefficient maps to a single behavioral quantity (distance penalty, attraction to complexity, effect of residential informality, regime threshold, transition steepness). An SR coefficient, by contrast, mixes labor-market mismatch with spatial sorting, making it harder to derive policy-relevant elasticities from the model.

In summary, SR and $\text{ECI}^{\text{employment}}$ are complementary quantities derived from the same product-space structure. We report the SR analysis in full to demonstrate that the economic complexity framework captures multiple drivers of commuting behavior, while retaining $\text{ECI}^{\text{employment}}$ in the main specification because it is the dimension most amenable to policy intervention.

To characterize how sector composition relates to commuting patterns, Figure S31 computes the within-sector commuting fraction: the share of workers in a given sector whose origin zone's dominant sector matches the destination sector. Because these are absolute shares across the full population, they reflect both how frequently a sector appears as the dominant activity across zones and how spatially concentrated its employment is. The fractions vary substantially across sectors and reveal clear distinctions between regions. Retail stands out in Latin American cities ($\approx$63% in Mexico City, $\approx$38% in Rio de Janeiro), far above the US cities where it remains below 10%, consistent with retail being the dominant sector in a large number of zones in those metropolitan areas. The pattern reverses for health care, professional services,

and manufacturing, all of which show markedly higher fractions in the Bay Area and Los Angeles than in Mexico City or Rio de Janeiro. These contrasts reflect differences in which sectors dominate the spatial landscape of employment in each region.

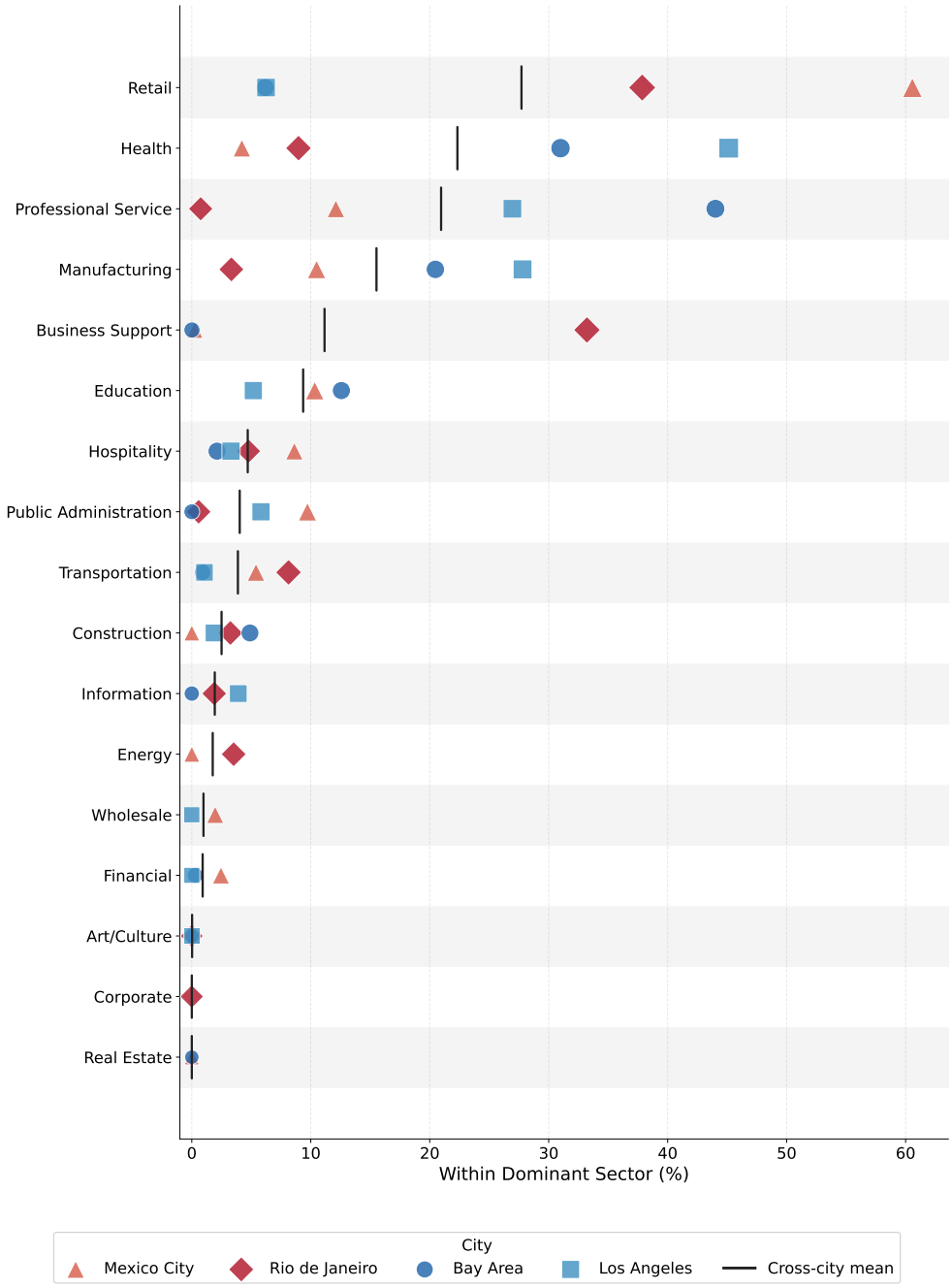

Fig. S31 Within-sector commuting fraction by sector and city. Each dot represents the fraction of commuters in a given sector whose origin zone's dominant sector matches the destination sector.

Figure S32 stratifies the SR–distance relationship by origin informality level using two measures: OD-pair SR (computed for each origin–destination pair) and per-origin mean SR (averaged across all destinations for a given origin). In Mexico City and Rio de Janeiro, the high-informality group exhibits a negative correlation between SR and commuting distance under both measures, while the low-informality group shows no significant relationship. In Bay Area and Los Angeles, the correlation is negative for all informality groups under both measures, with one exception: OD-pair SR is not significant for the low-informality group in the Bay Area. These patterns indicate that workers from high-informality zones are more likely to cross sector boundaries to reach distant employment, a relationship that is consistent across cities and robust to the choice of SR aggregation.

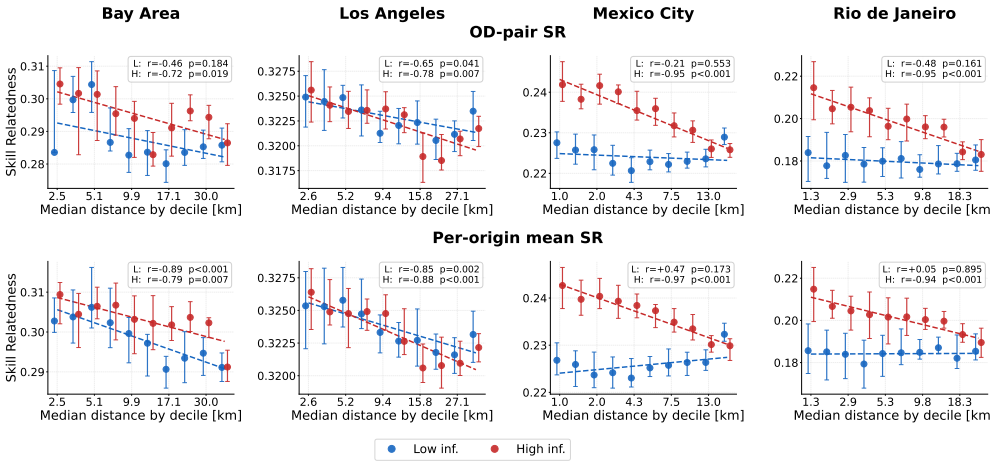

Fig. S32 Binned scatter plot of skill relatedness versus commuting distance decile, stratified by origin informality group. Left panels show OD-pair SR; right panels show per-origin mean SR. In Latin American cities, only the high-informality group exhibits a significant negative relationship; in US cities, the relationship is negative for all groups except the low-informality group in the Bay Area (OD-pair SR).

S6 Relationship between Accessibility and Informality

The WorkReach model allows to compute accessibility metrics based on utility (consumer surplus). Figure S33 shows the relationship between the Informality Rate of the home location and three different measures of accessibility (distance weighted, consumer surplus, and the 1st dimension of the PCA between the previous ones). All of the cities except for Mexico City exhibit medium to strong correlations for the three accessibility metrics.

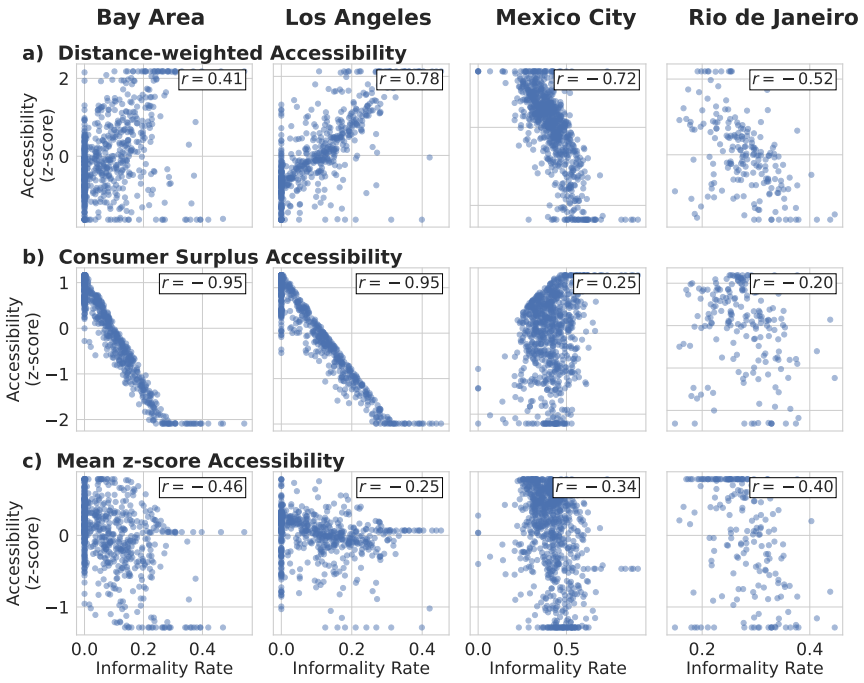

Fig. S33 Scatter plots of accessibility versus informality rate for each city. **a)** Distance-weighted accessibility. **b)** Consumer Surplus accessibility. **c)** Combined accessibility (mean of z-scored DW and CS measures). The plots show the Pearson correlation (r) for each relationship, highlighting how the relationship between accessibility and informality differs depending on the metric used.

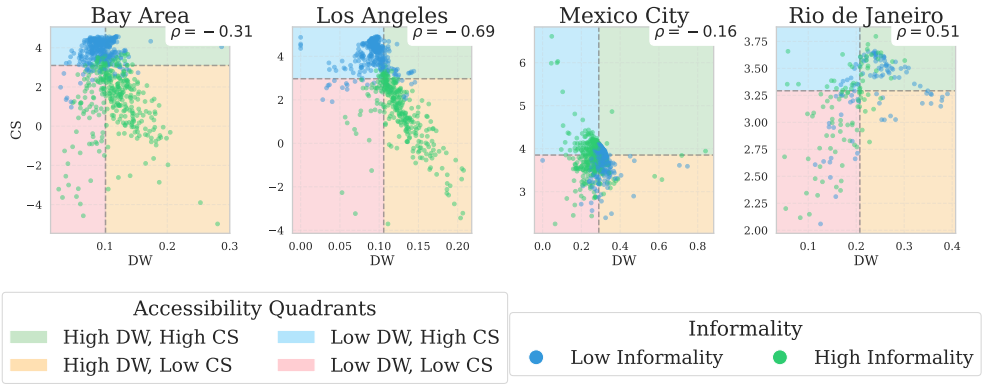

Fig. S34 Relationship between distance-weighted (DW) and consumer-surplus (CS) accessibility measures across cities. Scatter plots show the correlation between distance-weighted accessibility (x-axis) and consumer-surplus accessibility (y-axis) for each origin location, with points colored by informality level (blue = low informality, green = high informality). Background quadrants indicate accessibility combinations: high DW/high CS (green), high DW/low CS (orange), low DW/high CS (light blue), and low DW/low CS (pink). Pearson correlation coefficients (r) quantify the linear relationship between the two accessibility measures. Dashed lines indicate median values for each measure.

S7 Correlation to Census Variables

To contextualize the informality rate used in the WorkReach model, Table S16 reports the census variables most strongly correlated with the informality rate in each city.

Table S16 Top 5 Census Variables Correlated with Informality Rate

City	Variable Description	Correlation
Bay Area	Non-family households at/above poverty level (%)	0.479
	Family households at/above poverty level (%)	-0.427
	Female population 16+ didn't work. Household at/above poverty level (%)	-0.415
	Male population 16+ didn't work. Household at/above poverty level (%)	-0.413
	Female population 16+ not in labor force. Household at/above poverty level (%)	-0.410
Los Angeles	Foreign-born non-citizens at/above poverty line (%)	0.443
	Households in owner-occupied houses at/above poverty level (%)	-0.399
	Households with retirement income (%)	-0.395
	Foreign-born non-citizens below poverty line (%)	0.390
	Female population 16+ not in labor force. Household at/above poverty level (%)	-0.380
Mexico City	Occupants in overcrowded housing (%)	0.743
	Population 15+ without basic education (%)	0.742
	Dwellings without internet (%)	0.740
	Dwellings without refrigerator (%)	0.692
	Occupants in household without mobile phone (%)	0.558
Rio de Janeiro	Dwellings without computer at home (%)	0.783
	Housing inadequacy level	0.737
	Dwellings without washing machine (%)	0.728
	Household density (persons per room) (%)	0.699
	Household income (in minimum wages)	-0.681

S8 Counterfactual $\mathrm{ECI}^{\mathrm{employment}}$ Uplift and Worker Inflow

The WorkReach model can be used to evaluate how targeted changes in a zone's economic structure would reshape commuting flows. We combine the product-space framework with the fitted model to identify zones where gaining a single feasible sector would attract the most additional workers, and we decompose the predicted new inflow by the informality level of the sending origins.

Feasibility Scoring

For each zone z that does not yet hold a comparative advantage in sector s, we compute a feasibility score based on its relatedness density and the sector's product complexity:

$$F_{zs} = \omega_{zs} \cdot \mathrm{PCI}_s^{\mathrm{employment}} \cdot (1 - M_{zs}) \,, \tag{5}$$

where ω_{zs} is the relatedness density of zone z to sector s, defined as

$$\omega_{zs} = \frac{\sum_{s'} \phi_{ss'} \, M_{zs'}}{\sum_{s'} \phi_{ss'}} \,, \tag{6}$$

and $\phi_{ss'}$ is the product-space proximity between sectors s and s' (defined in Section S5). The feasibility score is high when a zone already hosts many sectors that are skill-related to s (high ω_{zs}) and s itself has high product complexity (high $\mathrm{PCI}_s^{\mathrm{employment}}$). For each zone we select the sector $s_z^* = \arg\max_s F_{zs}$ as the recommended diversification target.

Counterfactual $\mathrm{ECI}^{\mathrm{employment}}$ and Inflow Decomposition

We simulate the effect of each zone z gaining its recommended sector s_z^* by setting $M_{z,s_z^*} = 1$, recomputing $\mathrm{ECI}^{\mathrm{employment}}$ from the modified RCA matrix, and obtaining a counterfactual $\mathrm{ECI}^{\mathrm{employment}}$ gain $\Delta\mathrm{ECI}_z^{\mathrm{employment}}$. Zones with $\Delta\mathrm{ECI}_z^{\mathrm{employment}} \leq 0$ are discarded. We further restrict candidates to zones with above-median baseline inflow, so that the simulated changes affect a meaningful number of workers.

For each surviving candidate z, we replace its $\mathrm{ECI}^{\mathrm{employment}}$ value, recompute the full WorkReach probability matrix, and measure the change in predicted inflow. We further decompose the inflow change by whether the sending origin belongs to the high- or low-informality group, with percentage changes computed relative to each group's baseline inflow. The top ten zones per city are selected by total percentage inflow increase.

To illustrate where these opportunities concentrate, Figure S35 maps the selected zones for each city. The top row shows the ten highest-impact zones colored by $\Delta\mathrm{ECI}^{\mathrm{employment}}$, overlaid on the baseline $\mathrm{ECI}^{\mathrm{employment}}$ distribution; the bottom row colors all feasible zones by their recommended sector. In all four cities the selected zones are spatially dispersed and span a range of baseline complexity levels, indicating that the gains are not confined to already-complex cores.

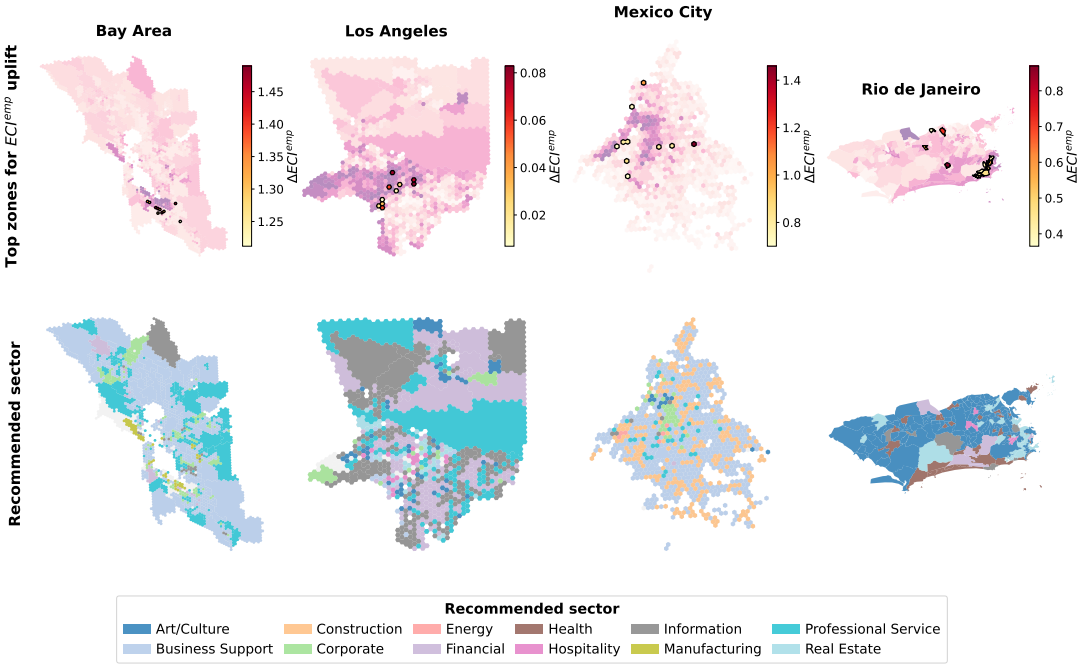

Fig. S35 Counterfactual $\text{ECI}^{\text{employment}}$ uplift analysis. Top row: the ten zones per city with the largest predicted percentage increase in worker inflow after gaining their most feasible sector, colored by $\Delta\text{ECI}^{\text{employment}}$ (background shading reflects baseline $\text{ECI}^{\text{employment}}$). Bottom row: all feasible zones colored by their recommended sector. The sector legend is shared across cities.

Figure S36 summarises the inflow response across cities: a) reports the median percentage change in worker inflow for all origins, high-informality origins, and low-informality origins, with interquartile ranges; b) shows the share of new workers arriving from high-informality zones, where values above 50% indicate that the uplift disproportionately benefits informal-sector workers; and c) reports the mean shift in each zone's high-informality workforce share (in percentage points), capturing whether the uplift changes the composition of the zone's labour force.

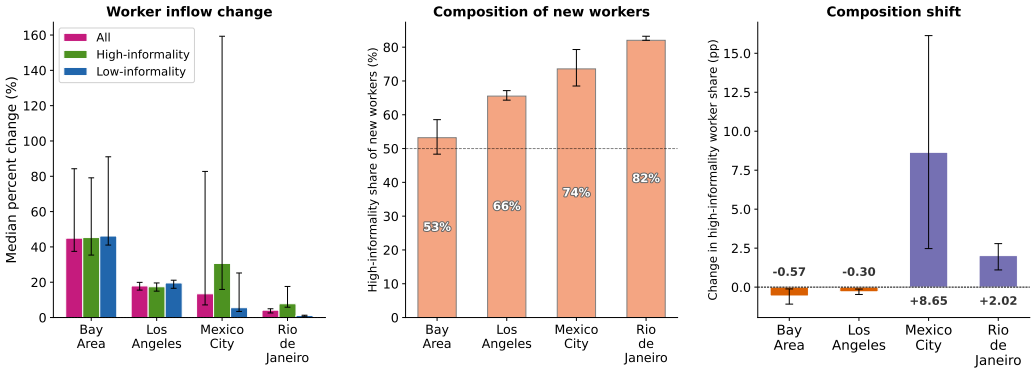

Fig. S36 Cross-city summary of counterfactual inflow changes for the ten highest-impact zones per city. **a)** Median percentage change in worker inflow, decomposed by origin informality group (error bars show interquartile range). **b)** Share of new workers originating from high-informality zones. **c)** Mean change in the high-informality share of each zone's workforce (percentage points); positive values indicate a compositional shift toward workers from high-informality origins.

References

1. INEGI. *Encuesta Origen-Destino en Hogares de la Zona Metropolitana del Valle de México (EOD) 2017* tech. rep. (Instituto Nacional de Estadística y Geografía, Aguascalientes, Mexico, 2017).
2. Cabanas-Tirapu, O., Danús, L., Moro, E., Sales-Pardo, M. & Guimerà, R. Human mobility is well described by closed-form gravity-like models learned automatically from data. *Nature Communications* **16.** Publisher: Nature Publishing Group, 1336. ISSN: 2041-1723. https://www.nature.com/articles/s41467-025-56495-5 (Feb. 2025).
3. Giuliano, G., Kang, S. & Yuan, Q. Agglomeration economies and evolving urban form. *The Annals of Regional Science* **63,** 377–398 (2019).
4. Bustos, S., Gomez, C., Hausmann, R. & Hidalgo, C. A. The Dynamics of Nestedness Predicts the Evolution of Industrial Ecosystems. *PLOS ONE* **7,** e49393 (2012).
5. Hidalgo, C. A., Klinger, B., Barabási, A.-L. & Hausmann, R. The Product Space Conditions the Development of Nations. *Science* **317.** Publisher: American Association for the Advancement of Science, 482–487. https://www.science.org/doi/10.1126/science.1144581 (July 2007).
6. Lucchini, L. *et al.* Socioeconomic disparities in mobility behavior during the COVID-19 pandemic in developing countries. *EPJ Data Science* **14.** Number: 1 Publisher: Springer Berlin Heidelberg, 25. ISSN: 2193-1127. https://epjds.epj.org/articles/epjdata/abs/2025/01/13688_2025_Article_532/13688_2025_Article_532.html (Dec. 2025).
7. Hidalgo, C. A. Economic complexity theory and applications. *Nature Reviews Physics* **3,** 92–113. ISSN: 2522-5820. https://www.nature.com/articles/s42254-020-00275-1 (Jan. 2021).
8. Hidalgo, C. A. & Hausmann, R. The building blocks of economic complexity. *Proceedings of the National Academy of Sciences* **106,** 10570–10575. ISSN: 0027-8424, 1091-6490. https://pnas.org/doi/full/10.1073/pnas.0900943106 (June 2009).
9. Balland, P.-A. *et al.* The new paradigm of economic complexity. *Research Policy* **51,** 104450. ISSN: 0048-7333 (2022).