nature cities

Article

https://doi.org/10.1038/s44284-024-00051-7

# Infrequent activities predict economic outcomes in major American cities

Received: 8 July 2023

Accepted: 16 February 2024

Published online: 15 March 2024

Check for updates

Shenhao Wang[1,2,8,9] ✉, Yunhan Zheng[3,8], Guang Wang[4,5], Takahiro Yabe[4,6], Esteban Moro[2,4,7,9] ✉ & Alex 'Sandy' Pentland[2,4]

Many studies have revealed the predictive power of the most frequent, regular and habitual mobility patterns. However, it remains unclear which components of the mobility patterns contain the most informative signals for predicting disparate economic development across urban areas. Here we use machine learning to predict economic outcomes by analyzing the heterogeneous mobility networks of 687 activities from more than 560,000 anonymized users in Boston, Chicago and Miami. We find that mobility patterns are highly predictive of the current and future economic development in major American cities but, surprisingly, the high predictive power is concentrated on infrequent, irregular and exploratory activities. These predictive activities account for only less than 2% of total visits but successfully explain more than 50% of variation in economic outcomes. Future research should shift more attention from regular visits to irregular activities, and policymakers could leverage these infrequent yet informative activities to manage urban economic development.

Mobility patterns can effectively predict economic outcomes in cities[1–3]. Economic outcomes—typically measured by wealth, GDP and income—are associated with the regular mobility patterns at the macroscale because the volume of mobility and economy both grows proportionally with urban population[4–8]. Economic outcomes could also be associated with the irregular mobility patterns at the microscale[9–11], which facilitate the local governments to proactively monitor and develop the local economy through mobility-related interventions[4]. This relationship between mobility and economy has been framed with various theoretical underpinnings, including neighborhood effects[12,13], weak ties[14,15], economic complexity[16] and structural diversity[17–19].

However, it is still controversial why mobility networks can effectively predict urban economic outcomes. Although the socioeconomic status is correlated with the diversity in mobility networks, simply improving diversity fails to enhance economic productivity in field experiments[17,20–23]. In social network literature, diversity is mainly attributed to the heterogeneous and, particularly, weak social ties rather than the massive volume of strong ones[14,15]. However, heterogeneous ties facilitate only specific (but not all) aspects of job searches[24–26], leading to a 'paradox' of their effects[24]. In cities, this paradox is not only a theoretical issue but also poses practical challenges. Policymakers could be bewildered by a large number of mobility diversity definitions using travel modes or displacements. Those debates indicate that the fundamental link between mobility and economy is still missing.

This missing link is partially caused by the insufficient focus on human activities. It's important to recognize that urban mobility serves as a means to an end, and these ends are the various activities conducted by individuals within the urban environment. Most studies treated human movements as only homogeneous edges in mobility networks, overlooking the diversity and nuances of underlying activities[4–7,17,20]. The simplified homogeneity assumption could dilute the

[1]Department of Urban and Regional Planning, University of Florida, Gainesville, Florida, USA. [2]Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, USA. [5]Department of Computer Science, Florida State University, Tallahassee, USA. [6]Tandon School of Engineering, New York University, Brooklyn, NY, USA. [7]Network Science Institute, Northeastern University, Boston, MA, USA. [8]These authors contributed equally: Shenhao Wang, Yunhan Zheng. [9]These authors jointly supervised this work: Shenhao Wang, Esteban Moro. ✉e-mail: shenhaowang@ufl.edu; emoro@mit.edu

https://doi.org/10.1038/s44284-024-00051-7

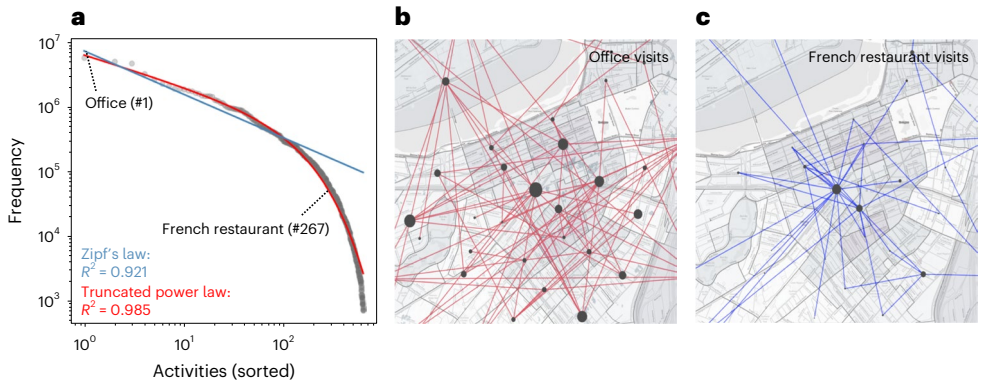


**Fig. 1 | Diverse activities in mobility networks. a**, Power-law distribution of human activities in cities. The activity frequency follows the truncated power-law distribution $\mathbb{P}(c) \propto r_c^{-\beta} \exp(-r_c/K)$ (ref. 8; see Methods for details). **b**, Office visits as an example of frequent and regular activities with a scattered spatial distribution. **c**, French restaurant visits as an example of infrequent and irregular activities with a concentrated spatial distribution.

predictive power in mobility networks because only certain activities contain economic signals. Although several studies have examined urban activities[27,28], they typically apply data mining methods without pinpointing the signals in the mobility networks for predicting economic outcomes.

Here we seek to predict economic outcomes using mobility patterns by characterizing the diverse human activities. We first examine the activity frequency by fitting it to the truncated power law and Zipf's law[8]. We then employ two machine learning methods to predict economic outcomes, utilizing activity-specific visits collected from phone trajectories and points-of-interest (POIs) in three prominent metropolitan areas in the United States. After identifying the predictive activities, we examine their frequency and regularity. We find that the economic signals are concentrated on mobility patterns characterized by infrequent and irregular activities, although the frequent and regular ones also contain moderate predictive power. We further demonstrate the robustness of our findings by investigating model sensitivity to noises, missing values and outliers, testing the model transferability across six major US cities, and examining the impacts of changing the activity grouping strategy. Our findings have broad scientific implications because they refine the theories of social diversity, weak ties and economic complexity; they also have practical implications for policymakers, who can use the infrequent yet informative mobility signals to monitor future economic development.

## Results

### Data and methods

The mobility network for each activity is constructed by aggregating all trips from users' census tract home $i$ to POIs of a given category $c$ in census tract $j$ (Fig. 1 and Methods). Each node in the mobility network is a census tract that includes various node features, and edges are weighted by the number of trips $w_{ij,c}$ in activity category $c$ from census tracts $i$ to $j$. For simplicity, we represent edges by an activity indicator $a_{ij,c} = \mathbb{1}\{w_{ij,c} \geq \delta\}$, with $\delta$ representing a volume threshold (Methods and Supplementary Information). After processing, the data include 687 fine-grained activity categories that are mapped to 15 parent activity categories. This activity categorization has been widely adopted and is highly consistent with the default categorization from the data provider (Foursquare)[29,30]. The node features include two economic outcomes (median household income and property values) as the model outputs and around ten sociodemographic variables as the model inputs; these were extracted from the American Community Survey (ACS). Although they were collected through phone trajectories, the mobility network data are representative of the population and socioeconomic distributions from the census population, indicating that smart phones have been widely adopted across social classes (Methods and Supplementary Section 2.3). The mobility networks leverage the unique power of the phone trajectories, which provide fine-grained activity dynamics and refine economists' traditional conclusions based on the survey data[31–33]. In fact, these mobility networks, which include most of the human activities, have been widely used to analyze segregation[34], consumer behavior[35,36], poverty[37], wealth[1], city structures[38], economic growth[20] and epidemic spreading[39], generating insights that are unavailable from the survey-based data sources.

Activities are quite heterogeneous. As shown in Fig. 1a, the activity frequency, which is measured by the total number of visits for each activity, can be accurately fitted by the truncated power law[8], suggesting a substantial variation ranging from the most frequent visits to the office (Fig. 1b) to places visited relatively infrequently (for example, French restaurants; Fig. 1c). Despite this significant variation, most studies on human mobility and its universal laws focus solely on aggregate and frequently occurring activities. Here we study all activity networks independently with a multiscale perspective to consider both frequent and infrequent human activities.

We predict economic outcomes $y_{i,t'}$ through machine learning by evaluating three groups of input variables: using exclusively activity-specific visits, $\alpha_{ij,c,t}$; using exclusively sociodemographic variables, $x_{i,t}$; and combining activity and sociodemographic variables (Methods). The predictive power is examined by predicting both the present and future economic outcomes ($t' = t$ and $t' = t + \delta t$), thus identifying the sources of the economic signals that consistently exist in two time periods. The current case refers to a cross-sectional analysis in which both the mobility inputs and economic outputs were collected in 2016, whereas the future case refers to the economic outcomes in 2018 as the outputs. Two machine learning methods are used. The first is a linear regression with elastic net regularization to tackle overfitting[37,40] caused by the high dimension of the activities ($|C| = 687$). The elastic net regressions can automatically learn predictive linear relationships among activities (Methods). An example of elastic net regression that uses only mobility patterns as inputs is:

$$y_{i,t'} = \sum_{c \in C} \beta_c \sum_j a_{ij,c,t} + \epsilon_{i,t} \quad (1)$$

where $y_{i,t'}$ represents the economic outcomes, and $\sum_j \alpha_{ij,c,t}$ represents the node degree for activity $c$, which is the sum of the neighboring edges around node $i$. $\beta_c$ represents the coefficient of the node degree for activity $c$, and $\epsilon_{i,t}$ denotes the error term. The second machine learning method is gradient boosting regressions (GBR), which ensembles decision trees and identifies the non-linearly predictive activities using their importance score. In short, our experiments use two machine learning methods to examine two output variables (median income

https://doi.org/10.1038/s44284-024-00051-7

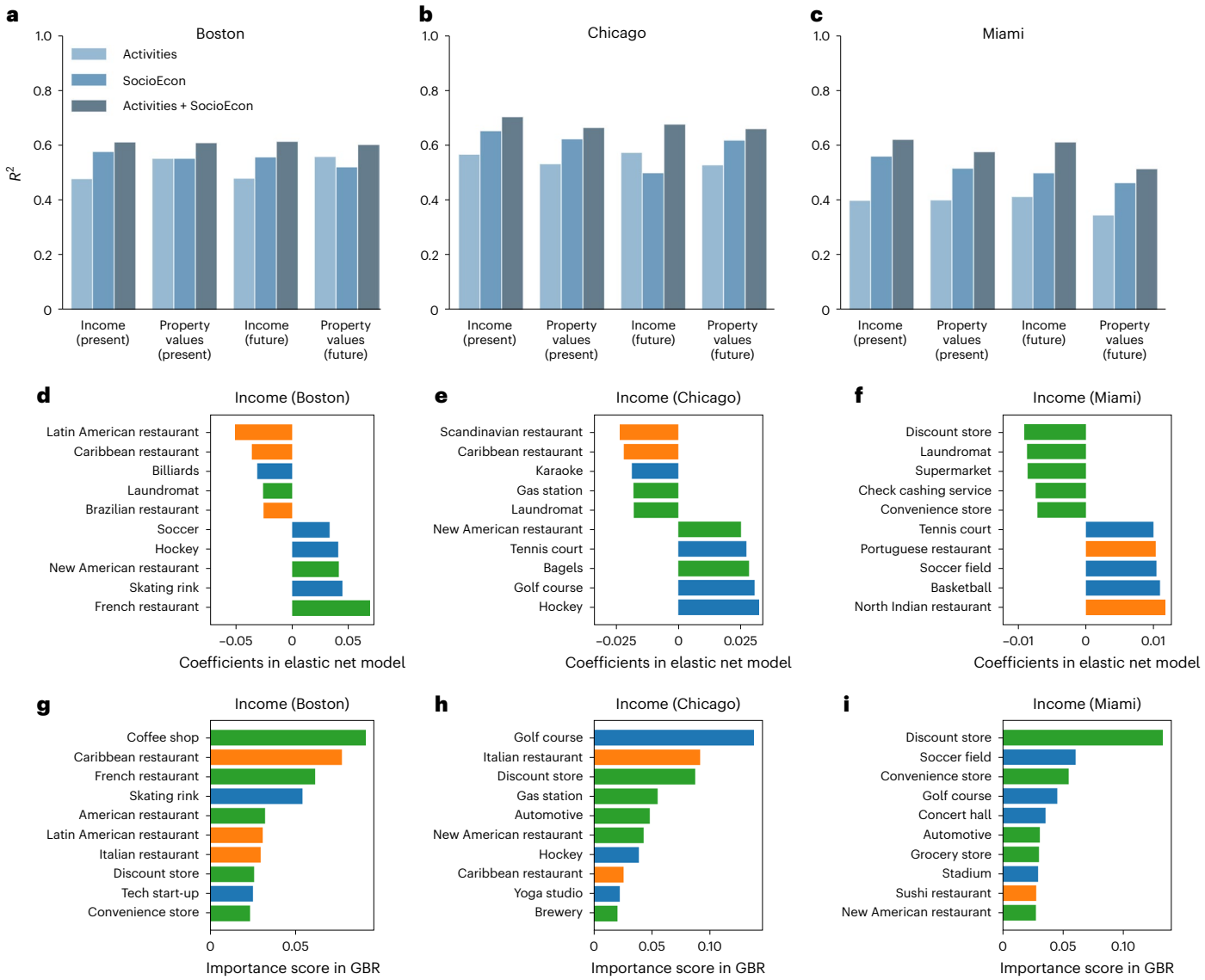


**Fig. 2 | Mobility activities predict economic outcomes in Boston, Chicago and Miami. a–c**, Predictive performance of elastic net regressions using only activity data or sociodemographic variables, and the integration of both, in Boston (**a**), Chicago (**b**) and Miami (**c**). The performance of GBRs is included in Supplementary Section 4, demonstrating similar results. **d–f**, The coefficients of predictive activities in elastic net regressions for Boston (**d**), Chicago (**e**) and Miami (**f**). **g–i**, The importance scores in GBRs for Boston (**g**), Chicago (**h**) and Miami (**i**). In both machine learning methods, the predictive activities include aspirational behaviors (for example, tennis, hockey; blue); non-essential consumption (for example, French restaurants; green); and latent socioeconomic cues (for example, visits to Latin American or Caribbean restaurants; orange).

and property values) in three US metropolitan areas at two time periods, leading to a total of 24 experimental scenarios.

## Mobility activities predict economic outcomes

The mobility activities can predict more than 50% of variation in income and property values using elastic net and gradient boosting regressions for the majority of the experimental scenarios. As shown in Fig. 2a–c, the cross-validated $R^2$ of elastic nets for income and property values using only activities reach around the 40–60% range. The GBRs can further advance the $R^2$ performance to the 50–70% range (Supplementary Section 4.2). The predictive performance of our model is higher than 50% for most of the scenarios, except for predicting the economic outcomes with simple linear models for Miami. The predictive performance is higher in Boston and Chicago than Miami, indicating a higher predictability of economic outcomes in the two dense urban areas than the sprawled urban environment. Our results improve upon the state-of-the-art methods by at least 20%. In fact, past studies often find that it is quite challenging to predict income and wealth because the state-of-the-art performance of $R^2$ was only around[37] 0.40–0.46, and it sometimes[41] approaches as low as 0.25. Furthermore, the high performance is consistent across Boston, Chicago and Miami without much variation, indicating that the predictive power embedded in mobility activities is consistent.

The mobility data are as predictive as the sociodemographic variables from the census, and the two data sources are highly complementary. In Fig. 2a–c, the mobility and the sociodemographic variables achieve quite comparable predictive performance. For example, in Boston, the $R^2$ is around 55.8% using the sociodemographic variables—slightly higher than 48.1% using the mobility activities data to predict income. Meanwhile, for predicting property values, the $R^2$ value using mobility data stands at around 55.9%, outperforming the 52.2% achieved when solely relying on sociodemographic variables. In fact, the prediction power of the two data sources is consistently comparable across the three cities because both data sources achieve higher

https://doi.org/10.1038/s44284-024-00051-7

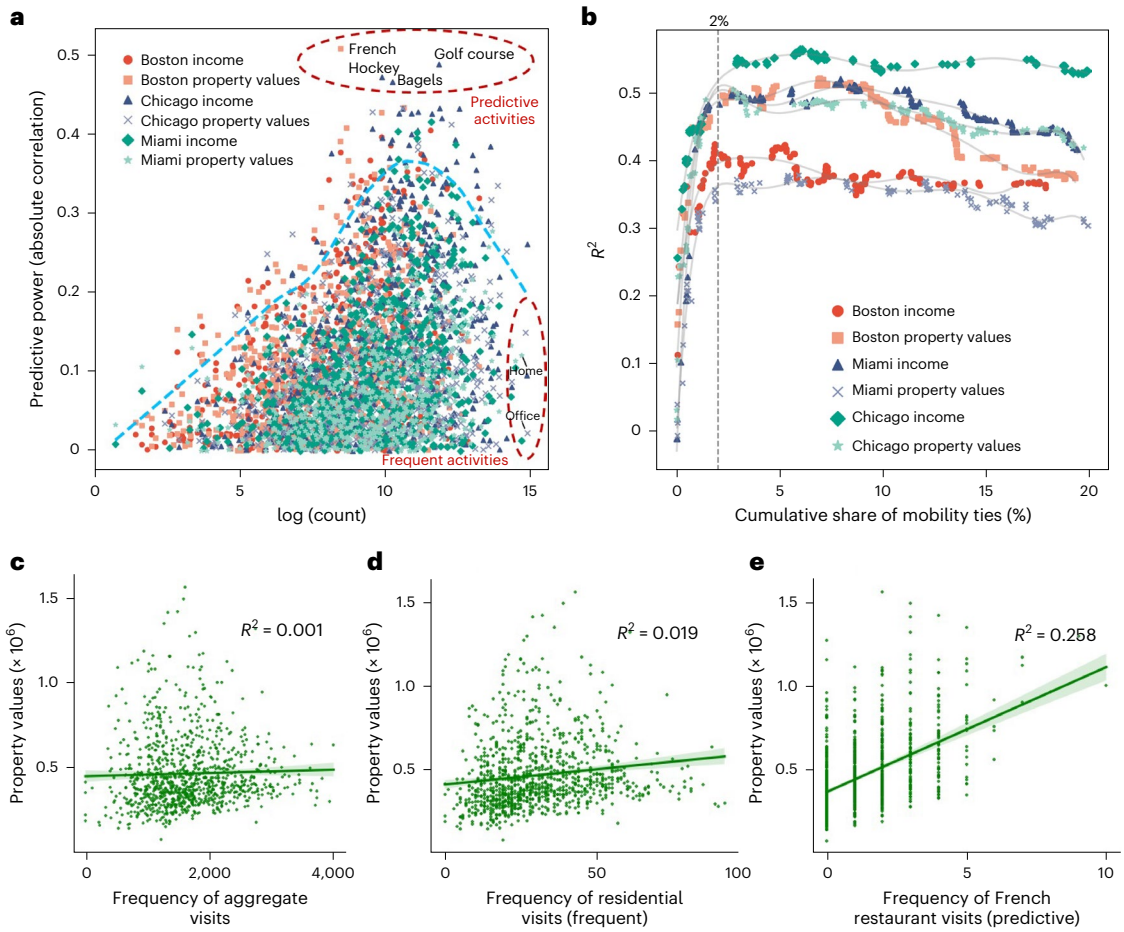


**Fig. 3 | Predictive activities are associated with infrequent mobility patterns.** **a**, Predictive power and frequency of activities. The predictive power is measured by the absolute value of the Pearson correlation between individual activities and economic outcomes. The red dashed ovals indicate the predictive and frequent activities, whereas the blue curve connects the upper five percentiles of the activities with a smoothing function. The most predictive activities are not the most frequent ones, but rather the relatively infrequent ones. **b**, Lorenz curve of the cumulative activity frequency and $R^2$ in elastic net regressions. Around 2% of activities are sufficient to achieve the highest predictive performance for income and property values. **c**–**e**, An example about the correlation of property values with aggregate (**c**), frequent (**d**) and predictive (**e**) activities in Boston. Only the infrequent vists to French restaurants are predictive for property values. The regression line and 95% confidence bands of the best fit line (shaded areas) are shown.

performance in Boston and Chicago, and relatively lower performance in Miami. The integrated regressions always achieve higher predictive performance when the two data sources are combined. For example, in Boston, the $R^2$ of the integrated regressions achieves 61.5% and 60.4% in predicting income and property values—higher than the results using the two data sources separately, demonstrating strong complementarity between the two data sources. The $R^2$ of GBRs using the integrated data sources is even higher, approaching around 70–80% in Boston and Chicago, demonstrating similar complementarity effects as elastic nets (Supplementary Section 4). This strong complementarity effect triggers the further question of which economic signals are captured by the mobility data but elude the traditional public surveys.

To understand the sources of predictive power, we identify the top predictive activities using the coefficients in the elastic nets and the importance scores in the GBRs. The visits to tennis courts, soccer fields, science museums and theaters are highly predictive. Such activities are the aspirational behaviors that take physical or mental effort. We also find that the visits to New American or French restaurants—those high-end non-essential consumptions[42], as well as to discount stores and pawn shops—can also predict economic outcomes. Finally, the visits to Brazilian, Caribbean, Spanish and North Indian restaurants are also predictive, potentially because these latent socioeconomic cues effectively differentiate people from sub-ethnic groups; such cues typically do not exist in the public census data. Overall, aspirational activities, non-essential consumption and latent socioeconomic cues are found as critical economic signals in both machine learning methods, as represented by the three colors in Fig. 2d–i. The three labels are designed to facilitate result interpretation but they are neither collectively exhaustive nor mutually exclusive categories. For example, New American and French restaurants are labeled as non-essential consumptions as they are more expensive than other restaurants, but they could also reveal ethnic information and thus could be labeled as latent socioeconomic cues.

## Predictive activities are infrequent

Figure 3a illustrates that the predictive and most frequent activities do not overlap. The most frequent activities, which include visits to offices and residential areas, are weakly associated with economic outcomes, with correlation coefficients smaller than 0.2. However, the correlation coefficients of the most predictive activities are larger than 0.4, such as the visits to French restaurants and golf courses. As Fig. 3a uses a logarithmic scale for the $x$-axis, the predictive activities are generally around four magnitudes smaller than the most frequent ones, indicating that the frequency of the predictive activities is around 10,000 smaller than the most frequent activities. However, although the predictive activities are relatively infrequent, the reversed statement

https://doi.org/10.1038/s44284-024-00051-7

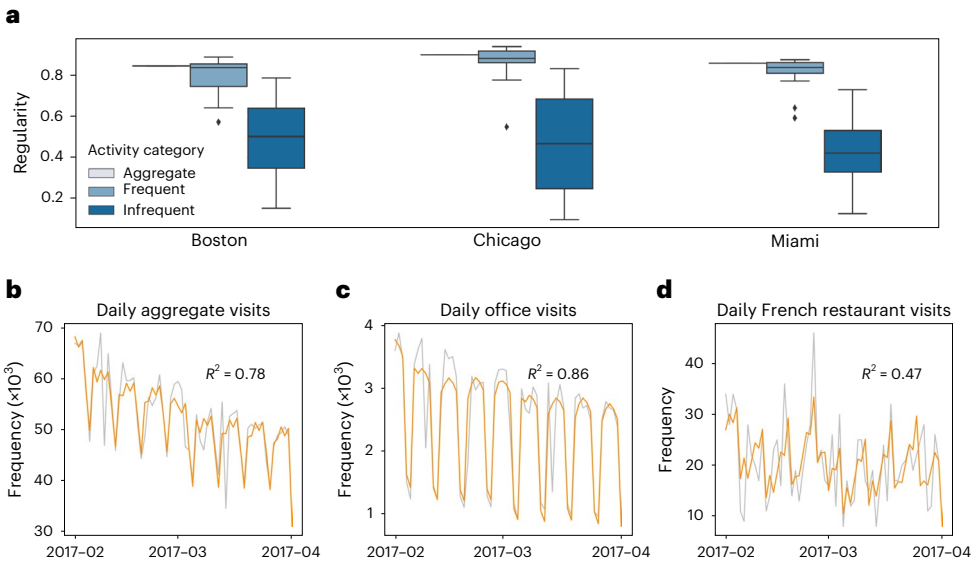


**Fig. 4 | Predictive activities are associated with irregular mobility patterns.** **a**, Regularity metrics for the aggregate, frequent and predictive activities in three cities. The box plots follow the standard format with the center line representing the median, the box limits representing the upper and lower quantiles, and whiskers representing 1.5 × the interquartile range; $N = 20$ (frequent, Boston), 20 (frequent, Chicago), 20 (frequent, Miami), 26 (infrequent, Boston), 21 (infrequent, Chicago), 23 (infrequent, Miami). The predictive activities are highly irregular in the behavioral patterns (see Methods about temporal regularity). **b**–**d**, The daily aggregate (**b**), office (**c**) and French restaurant (**d**) visits in Boston, with the orange and grey lines indicating predicted and true values, respectively. The predictive activity exhibits a highly irregular behavioral pattern, as measured by the low $R^2$.

is not necessarily true. The most infrequent activities are not highly predictive because the most infrequent activities cannot cover the large spatial scale, thus limiting their capacity in economic prediction. As summarized by the blue trend line in Fig. 3a, the concave shape of predictability in activities successfully demonstrates that the predictive activities are moderately infrequent—much less frequent than the regular daily activities, but with relatively large spatial coverage to provide signals for all the census tracts in a metropolitan area.

Figure 3a cannot examine the interaction effects among the predictive activities because it visualizes the Pearson correlation between individual activities and economic outcomes. We therefore cumulatively add the most predictive activities into regressions and visualize the Lorenz curves of $R^2$ regarding the cumulative shares of activities in Fig. 3b. The six Lorenz curves rise quite sharply and peak around the region of only 2% of the cumulative mobility patterns, predicting about 40–60% of variation in income and property values, indicating that only a small fraction of infrequent activities is needed to achieve relatively high performance. In fact, incorporating more frequent activities could even lower the predictive performance, as the Lorenz curves start to decrease when more than 2% of activities are used.

To provide further intuition, we visualize the correlation of aggregate, frequent and predictive activities with Boston property values in Fig. 3c–e. The figures show stark contrasts in their activity frequency ($x$-axis): the aggregate mobility patterns reach about 4,000 counts, the residential visits reach 100 counts, whereas French restaurant visits have only ten counts per day. However, their predictive capabilities are reversed: the mobility patterns of aggregate and frequent activities have nearly zero predictive power for property values, whereas a single infrequent activity—visiting a French restaurant—can effectively predict 25.8% of variation in property values.

## Predictive activities are irregular

Besides frequency, we compare the temporal regularity of the aggregate, frequent and predictive activities in the three cities, and the temporal regularity metric is quantified using the $R^2$ from time-series regressions that account for both temporal trend and cyclical patterns. We compare the linear, quadratic and cubic function specifications for the temporal trend, and include day of the week, week of the month, and month fixed effects for the cyclical patterns. Empirical results demonstrate that the temporal regularity does not vary considerably with the specifications of temporal trends, but is mainly captured by the cyclical patterns (Methods and Supplementary Section 4.2). Figure 4a compares the temporal regularity across the three groups of activities in the three cities. Figure 4b–d presents the temporal patterns of the aggregate, office and French restaurant visits, with the last two serving as examples of frequent and predictive activities.

Figure 4a illustrates that the predictive activities are associated with temporally irregular mobility patterns. The predictive activities have much lower temporal regularity than the aggregate and frequently occuring activities in Boston, Chicago and Miami. The aggregate and frequently occuring activities are quite regular because the time-series regressions can easily capture around 80% of variation by solely considering the cyclical patterns and the temporal trends (Fig. 4a). On the contrary, the predictive activities are much more irregular: only less than 50% of variation can be explained by temporal trends and cyclical patterns. The results suggest that the most frequent activities are the most habitual visits with clear temporal cycles, whereas the predictive activities are often the irregular activities with much weaker cycling and temporal trends.

Figure 4b–d demonstrates why the predictive activities are irregular by contrasting the temporal patterns of the aggregate, frequent and predictive activities. The aggregate mobility patterns are highly predictable because the time-series regression can capture the weekly fluctuation and the overall declining trend with $R^2 = 78\%$ (Fig. 4b). As an example of frequent activities, office visits exhibit even higher predictability than aggregate visits. This is attributed to the much higher office visits during weekdays than weekends, leading to a distinct weekly recurring pattern with a $R^2 = 86\%$ for explained variation. However, the predictive activity is much more irregular. Although the French restaurant visits do exhibit some underlying trends and seasonality, only around 47% of the temporal dynamics can be explained by these factors (Fig. 4d). This irregular temporal dynamic of the predictive activities is associated with their low frequency: visits to French restaurants happen so rarely, and thus it is highly challenging to form a regular visitation pattern.

### Robustness check 1 (model sensitivity)

We examine the model robustness by conducting four robustness checks. In the first, the sensitivity to noise is tested by introducing Gaussian random noises with variance equivalent to 10% of the variance in the economic outcomes. The sensitivity to missing values is tested by randomly omitting 10% of the phone trajectory data. The sensitivity to outliers is tested by dropping the data points below the permissible lower bound and above the upper bound. In all three tests, the $R^2$ and the predictive activities remain highly stable (Supplementary Section 5.1).

### Robustness check 2 (transferability test)

Besides in-city model performance, we also test model transferability across cities. In addition to Boston, Chicago and Miami, we added Washington DC, Detroit and Philadelphia to the transferability test using the top-20 and top-10 predictive activities. The results of the transferability analysis demonstrate that, on average, 55.6% of the variance can be explained by the top-20 predictive activities identified from other cities, and 44.2% of the variance can be explained by the top-10 activities. This result indicates strong cross-context robustness of our empirical findings (Supplementary Section 5.2).

### Robustness check 3 (benchmarking frequent activities)

The predictive power of the infrequent activities is compared with the most frequent ones as a benchmark. This comparison uses the likelihood ratio test, in which log-likelihood is computed by the model fitting with an ordinary least square after the feature selection of elastic net[43]. The likelihood ratio tests consistently reject the null hypotheses, indicating a statistically significant higher performance in the infrequent activities over the frequent ones (Supplementary Section 5.3).

### Robustness check 4 (grouping activities)

We examine whether our findings are sensitive to the aggregation of activity categories. We find that the 15 parent activity categories still retain moderate predictive power, achieving around 25–50% $R^2$ for economic outcomes; this range is notably lower than the predictive performance from the infrequent activities. This moderate predictive power is a benchmark representing a conservative estimate in mobility networks for predicting economic outcomes. Meanwhile, the coefficients of the 15 parent activity categories demonstrate similar patterns, because the relatively infrequent parent activity categories (that is, sports and arts/museums) are more predictive. Therefore, the predictive signals of economic outcomes concentrate on the infrequent activities, although the frequent activities also maintain certain predictive power (Supplementary Section 5.4).

## Discussion

Past studies demonstrate that mobility networks contain economic signals[1,5–7]; however, none of these studies have detected the exact sources of the signals. This study elaborates that the diverse human activities—which follow the power-law distribution in visit frequency—can effectively predict socioeconomic status. The predictive activities are disproportionately concentrated in the infrequent and irregular ones, including aspirational behavior, non-essential consumption and latent socioeconomic cues, but excluding the most frequent and regular displacements such as visits to jobs or residential areas. In other words, it is the infrequent, irregular and exploratory activities—rather than the frequent, regular and habitual ones—that are the most important socioeconomic signals in mobility patterns.

The findings have broad scientific and practical implications. They refine the existing theories of structural and social diversity through the lens of human activities[15,24,25,44]. Although structural diversity still matters[17,20,21], we find that only the long-tailed activities, such as the infrequent aspirational behaviors, are contributing positively to economic complexity and development[16]. Although social diversity still matters[12,13], we find that it might be a concept rooted in the local context. For example, the subgroups within the Asians and the Hispanics are more critical socioeconomic differentiators than the general ethnic groups in Boston and Miami, as shown by their significant coefficients in the predictive models.

This study also demonstrates the outstanding explanatory power by combining machine learning and mobility network data, which can complement the public surveys that typically do not collect the infrequent activity signals. Public surveys might easily miss out these infrequent but informative visits because the informative activities account for only a small portion of the total activities. Unlike the public surveys, the mobility network data can provide real-time insights into economic cycles or time-dependent policy changes through dynamic human activities, which are critical for policymakers to take action, particularly when the economy experiences fast and systematic disruptions (for example, COVID-19). The infrequent activities could be used to monitor economic inequality as well. When the non-essential consumption is concentrated in certain urban regions, this concentration could imply spatial segregation and inequality. Therefore, policymakers can detect sudden changes in socioeconomic indices by combining these predictive activities and machine learning models, and proactively design policies to manage economic development and mitigate economic inequality.

This study generates causal hypotheses by analyzing the predictive power of the heterogeneous mobility networks, but it cannot pinpoint the exact causal mechanisms. The aspirational behaviors could lead to higher socioeconomic status because they represent the entrepreneurship and risk-seeking preferences of individuals, although a reversed causality cannot be excluded[45]. High-end consumption (for example, New American and French restaurants) is predictive potentially because it can reveal the expected growth in household wealth. But such high-end consumption appears implausible to be the cause of high socioeconomic status, but more likely the consequence. In fact, both elastic net and GBR models could suffer from omitting variable or reversed causality issues, thus biasing the coefficient estimates. Therefore, our causal hypotheses need further validation by combining experimental data and causal models in future studies.

This study is limited by its analytical contexts. We examine relatively large American metropolitan areas without investigating the transferability of our findings to the relatively poor and rural areas, which potentially limits generalizability[46]. The predictive activities with high signal-to-noise ratios, such as Pilates and Yoga studios, tend to exist in only the relatively wealthy areas. Even the visits to the pawn shops, which indicate a low socioeconomic status, suggest the existence of a basic financial institute, precluding the exceedingly poor areas without any financial infrastructure. Unfortunately, it is challenging to obtain the highly diverse long-tail activity data in less developed areas, which could exhibit an economy–mobility link that is different to our findings. Many unpredictable factors, such as pandemic, economic downturns or policy changes, might simultaneously influence both mobility and economy. Some of the unpredictable factors could be revealed indirectly through the mobility data, although they are not explicitly integrated into this study.

## Methods

### Ethics

We obtained Institutional Review Boards (IRB) exemption to use the Cuebiq mobility data from the MIT IRB office (COUHES protocol no. 1812635935 and its extension, no. E-2962).

### Data sources

Two major data sources are the mobility data from a geospatial data provider, Cuebiq, and the socioeconomic data from the census. The privacy-enhanced mobility data are highly granular with a broad coverage of anonymized users, venues and activities. To become part of Cuebiq's panel, users provide informed consent to data collection

for research purposes, under an opt-in process that is compliant with both the General Data Protection Regulation and California Consumer Privacy Act. The dataset contains 126 million visits to 386,000 venues from 560,000 anonymized users in Boston, Chicago and Miami, covering a wide span of 687 activities, inlcuding visits to offices, supermarkets, coffee shops, among many others. Its temporal resolution is at seconds, and spatially, the longitude and latitude of activity locations are collected. The mobility data start on 30 September 2016 and ends on 1 April 2017, covering the six months during the 2016–2017 period. Although individual behaviors are observed in our raw mobility data, the individuals are aggregated to census tracts, thus protecting privacy and matching the standard spatial resolution of the census. Our socioeconomic data are collected from ACS at the census tract level[47]. As the mobility data reflect the mobility patterns in 2016–2017, we examine whether the mobility data could predict the future using the economic outcomes in 2018. Please refer to Supplementary Section 2 for descriptions of the data and details on data processing).

### Data representativeness

Our dataset represents approximately 2.61% of the total census population across the three cities (1.54% for Boston, 3.82% for Chicago and 4.68% for Miami). To assess data representativeness, we compared our sample population estimates with 2012–2016 five-year ACS data for each census tract in Boston, Chicago and Miami (see Supplementary Section 2.3 for details). Our analysis revealed a moderately high correlation between the sample and census tract populations, with correlation coefficients of ~0.7 for Boston, ~0.66 for Chicago and ~0.58 for Miami. We also evaluated the representativeness of income and property value distributions in our sample across the three cities through quantile–quantile plots. Our analysis indicated a substantial degree of alignment between these distributions across the three cities overall (see Supplementary Section 2.3 for details).

### Graph representation

The mobility network is built using census tracts as nodes and visits between homes and activities as edges. The mobility networks are defined as

$$G_c = \{V, E_c\} \tag{2}$$

We use $G$ to represent the graph, $V$ for nodes and $E$ for edges, following the standard notations. We also introduce the symbol $c$ as an activity index to account for network heterogeneity in edges. In the network, nodes correspond to census tracts, and we incorporate socioeconomic features from the ACS as node attributes. Specifically, we use two economic indicators (namely, median household income and property values) as model outputs denoted by $y_i$. The sociodemographic variables are denoted as $x_i$, which include education, gender, racial composition and other demographic factors. Although this study mainly uses the mobility edges to predict economic variables $y_i$, we also examined the complementarity of the mobility edges and sociodemographic variables $x_i$. Edges are represented by the home-to-activity visitation, which is in turn represented by a high-dimensional activity indicator:

$$a_{ij,c} = \mathbb{1}\{w_{ij,c} \geq \delta\} \tag{3}$$

where $w_{ij,c}$ is the observed counts of visits from node $i$ to $j$, $\delta$ is the volume threshold, and $a_{ij,c}$ is an indicator for the existence of an edge. Activities are collected by matching the phone trajectories and the POIs, which incorporate 687 activities ($|C| = 687$) and 126 million trip records from 560,000 users, collected from September 2016 to April 2017 in Boston, Chicago and Miami. To align with the temporal resolution of the economic variables $y_i$, we add the mobility ties over the six-month research period to create a weighted adjacency matrix $W_c$, in which each individual edge weight $w_{ij,c} = \sum_{t=1,\ldots T} w^t_{ij,c}$. With the weighted adjacency matrix $W^t_c$, we could create the unweighted adjacency matrices $A_c$, in which each element in equation (3) is $a_{ij,c}$.

### Power-law distribution for activities

The activity frequency follows the truncated power-law distribution[8], indicating a substantial frequency variation in human activities. The frequency of activity $c$ is computed as:

$$f_c = \sum_{i,j \in V} w_{ij,c} \tag{4}$$

The distribution of activity frequency is $\mathbb{P}(c) = f_c / \sum_{c \in C} f_c$, which can be more accurately fitted by the truncated power law[8] than Zipf's law (Fig. 1a). The activity frequency varies considerably. For example, in the six-month period, office visits have 5,594,655 counts and rank first among all 687 activities, whereas the visits to French restaurants have only 65,534 visits in total, ranking only 267th. The truncated power-law distribution is:

$$\mathbb{P}(c) \propto r_c^{-\beta} \exp(-r_c/K) \tag{5}$$

in which $r_c$ represents the rank of activity $c$. Note that this formula combines the polynomial and the exponential decay in the tails, with $r_c^{-\beta}$ and $\exp(-r_c/K)$ representing polynomial and exponential decay, respectively. The Zipf's law is based on only a polynomial decay $\mathbb{P}(c) = K r_c^{-\alpha}$. With logarithmic transformation, the equation becomes

$$\log \mathbb{P}(c) = -\alpha \log(r_c) + \log K \tag{6}$$

which serves as a benchmark model in our analysis of activity frequency.

### Elastic net regressions

The elastic net regressions target to predict the economic outcomes $y_i$ by comparing the mobility networks, sociodemographics, and integrated sociodemographic and network data. The first regression aggregates the mobility edges around each node $i$ regarding an activity $c$ with the following formula:

$$y_{i,t+\Delta t} = \sum_{c \in C} \beta_c \sum_{j \in N(i)} a_{ij,c,t} + \epsilon_{i,t} \tag{7}$$

in which $N(i)$ represents the neighborhoods of node $i$, and $\sum_{j \in N(i)} a_{ij,c,t}$ adds the neighboring edges around each node $i$. This regression examines whether the economic outcomes can be predicted by the activities visits. It is designed as a prediction task since the outputs are from 2018, whereas the inputs are from 2016. The second regression uses only the sociodemographics from the ACS:

$$y_{i,t+\Delta t} = \sum_k \beta_k x_{ik,t} + \epsilon_{i,t} \tag{8}$$

in which $x_{ik,t}$ represents the sociodemographic variables. The third regression combines the two data sources with the following equation:

$$y_{i,t+\Delta t} = \sum_k \beta_k x_{ik,t} + \sum_{c \in C} \beta_c \sum_{j \in N(i)} a_{ij,c,t} + \epsilon_{i,t} \tag{9}$$

The results from the three regressions are compared to demonstrate the complementarity of the mobility and survey data.

The use of mobility networks as predictors in regressions can potentially result in overfitting, primarily due to the high dimensionality of activity categories ($|C| = 687$). To address this issue, we employ elastic net regularization, which helps mitigate overfitting while enabling the automatic identification of significant activities. The elastic net regression optimizes its objective function using mean-squared error as the primary target, complemented by L1 and L2 regularization

terms. This optimization process is illustrated in equation (10), where the goal is to minimize the loss function:

$$\min_{\beta_c} L(y_i, \hat{y}_i) = \min_{\beta_c} \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda_1||\beta_c||_1 + \lambda_2||\beta_c||_2^2 \quad (10)$$

Here, $\hat{y}_i$ represents predicted economic outcomes, and $\lambda_1$ and $\lambda_2$ control L1 and L2 regularization, respectively. We systematically apply this regression framework to 12 experimental scenarios, involving the prediction of income and property values in Boston, Chicago and Miami for both 2016 and 2018.

In the process of selecting $\lambda_1$ and $\lambda_2$, we start by splitting the data into training and testing sets. To optimize the elastic net model's hyperparameters efficiently, we follow a two-step approach. In the first step, we perform 20 randomized searches over a predefined hyperparameter grid, considering different values for $\alpha$ (where $\alpha = \lambda_1 + \lambda_2$) and $r$ (where $r = \lambda_1/(\lambda_1 + \lambda_2)$). During this step, we employ fivefold cross-validation to minimize mean-squared error on the training set. Subsequently, we conduct a grid search centered around the best hyperparameters from the randomized search ($\alpha^*$ and $r^*$), exploring neighboring values. The final model is trained using the best $\alpha$ and $r$ values from this grid search, and its performance is evaluated on the testing data. This approach ensures efficient hyperparameter selection while maintaining computational efficiency, with an average training time of approximately 75 s for each economic outcome and city (see Supplementary Section 4.1 for details).

### Gradient boosting regression

In addition to elastic net, we also assess the predictive capacity of mobility networks for forecasting economic outcomes using GBR, which—implemented through Python's sklearn library with GradientBoostingRegressor—is adept at capturing intricate non-linear relationships among variables. However, GBR does not provide coefficients that directly indicate the magnitude and direction of effects, as seen in linear models. We instead report feature importance when using GBR. In training the GBR model, we also employ a fivefold cross-validation strategy to fine-tune hyperparameters. This involves a grid search over learning rates, maximum depth of individual regression estimators, and the number of boosting stages. After identifying the best model, we calculate feature importances and assess its performance on the testing dataset. Notably, the GBR model exhibits an average training time of approximately 88 s for a single city and economic outcome, comparable with the efficiency achieved with the elastic net model, emphasizing the computational effectiveness of our approach in both modeling techniques (see Supplementary Section 4.1 for details).

Due to the unique model designs, both elastic net and GBRs are robust to real-world data challenges such as noises, missing values and outliers. GBRs are the ensemble model with iterative algorithms fitting the residuals. As a result of the model ensemble, GBRs tend to generate a smooth function space, which is relatively insensitive to the random noises and outliers. Elastic net regressions can address missing values by simply ignoring them in the computation of the loss function. Elastic net regressions can also address outliers and noises because the L1 and L2 regularizations stablize the coefficients, which create models being less sensitive to outliers and noises. But meanwhile, such shrinkage and sparsity effects lead to potential biases when the individual coefficients are used for model interpretation. The biases in individual coefficients exert limited impacts on this study, because we mainly focus on the generalizability and robustness of model prediction.

### Temporal regressions

To investigate temporal regularity, we employed time-series regression methods to model the trends and cyclic patterns within mobility data. The regression equation is defined as follows:

$$y_t = f(t) + c(t) + \epsilon_t \quad (11)$$

Here, $f(t)$ and $c(t)$ capture the underlying trend and cyclical patterns in the time-series, with $y_t$ representing the daily activity counts. The trend function $f(t)$ can take on the form of a linear ($\beta_0 + \beta_1 t$), quadratic ($\beta_0 + \beta_1 t + \beta_2 t^2$) or cubic ($\beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$) equation. In our primary model, we employ the cubic equation, whereas the outcomes related to the linear and quadratic equations are detailed in Supplementary Section 4.2.4. The cyclical pattern is captured by the seasonality function $c(t) = \alpha_1 D(t) + \alpha_2 W(t) + \alpha_3 M(t)$, in which $D(t)$, $W(t)$, and $M(t)$ denote the day of the week, week of the month and month fixed effects, collectively accounting for cyclic patterns within the mobility network. An example of the cubic trend function with the seasonality function is denoted as:

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \alpha_1 D(t) + \alpha_2 W(t) + \alpha_3 M(t) + \epsilon_t \quad (12)$$

In fact, the three forms of the temporal trend function yield quite similar empirical performance (see Supplementary Section 4.2 for details), suggesting that the cyclical patterns are more important than the overall trends. This temporal regression is trained using the mean-squared error as the objective function, and its $R^2$ value quantifies the proportion of variance explained by the trend and cyclical components:

$$R_c = \frac{\sum_t \left(y_{t,c} - \hat{y}_{t,c}\right)^2}{\sum_t \left(y_{t,c} - \bar{y}_c\right)^2} \quad (13)$$

Here, $R_c \in [0, 1]$ represents the temporal regularity score for a specific activity $c$.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The data that support the findings of this study are available from Cuebiq through their Data for Good program, but restrictions apply to the availability of these data, which were used under the licence for the current study and are therefore not publicly available. Information on how to request access to the data, and its conditions and limitations, can be found at https://www.cuebiq.com/about/data-for-good/. The locations and activity categories of visits were obtained via Foursquare using their Public Search API. The public data source of this study (for example, ACS) is available in the Github repository at https://github.com/cjsyzwsh/economic_growth_usa.git.

## Code availability

The analysis was conducted using Python. The scripts that support the findings of this study are also available via the Github repository at https://github.com/cjsyzwsh/economic_growth_usa.git.

## References

1. Aiken, E., Bellue, S., Karlan, D., Udry, C. & Blumenstock, J. E. Machine learning and phone data can improve targeting of humanitarian aid. *Nature* **603**, 864–870 (2022).
2. Smythe, I. S. & Blumenstock, J. E. Geographic microtargeting of social assistance with high-resolution poverty maps. *Proc. Natl Acad. Sci. USA* **119**, e2120025119 (2022).
3. Chi, G., Fang, H., Chatterjee, S. & Blumenstock, J. E. Microestimates of wealth for all low-and middle-income countries. *Proc. Natl Acad. Sci. USA* **119**, e2113658119 (2022).
4. Bettencourt, L. M. A., Lobo, J., Helbing, D., Kuhnert, C. & West, G. B. Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl Acad. Sci. USA* **104**, 7301–7306 (2007).
5. Simini, F., Gonzalez, M. C., Maritan, A. & Barabasi, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96–100 (2012).

6. Schlapfer, M. et al. The universal visitation law of human mobility. *Nature* **593**, 522–527 (2021).
7. Alessandretti, L., Aslak, U. & Lehmann, S. The scales of human mobility. *Nature* **587**, 402–407 (2020).
8. Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).
9. Bettencourt, L. M. A., Lobo, J., Strumsky, D. & West, G. B. Urban scaling and its deviations: revealing the structure of wealth, innovation and crime across cities. *PLoS ONE* **5**, e13541 (2010).
10. Song, C., Koren, T., Wang, P. & Barabasi, A.-L. Modelling the scaling properties of human mobility. *Nat. Phys.* **6**, 818–823 (2010).
11. Pappalardo, L. et al. Returners and explorers dichotomy in human mobility. *Nat. Commun.* **6**, 8166 (2015).
12. Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R. & Porter, S. R. *The Opportunity Atlas: Mapping The Childhood Roots of Social Mobility* (National Bureau of Economic Research, 2018).
13. Bell, A., Chetty, R., Jaravel, X., Petkova, N. & Van Reenen, J. Who becomes an inventor in America? The importance of exposure to innovation. *Q. J. Econ.* **134**, 647–713 (2019).
14. Granovetter, M. S. The strength of weak ties. *Am. J. Sociol.* **78**, 1360–1380 (1973).
15. Granovetter, M. The impact of social structure on economic outcomes. *J. Econ. Perspect.* **19**, 33–50 (2005).
16. Hidalgo, C. A. & Hausmann, R. The building blocks of economic complexity. *Proc. Natl Acad. Sci. USA* **106**, 10570–10575 (2009).
17. Eagle, N., Macy, M. & Claxton, R. Network diversity and economic development. *Science* **328**, 1029–1031 (2010).
18. Gomez-Lievano, A., Patterson-Lomba, O. & Hausmann, R. Explaining the prevalence, scaling and variance of urban phenomena. *Nat. Human Behav.* **1**, 0012 (2016).
19. Bettencourt, L. M. A., Samaniego, H. & Youn, H. Professional diversity and the productivity of cities. *Sci. Rep.* **4**, 5393 (2014).
20. Chong, S. K. et al. Economic outcomes predicted by diversity in cities. *EPJ Data Sci.* **9**, 17 (2020).
21. Pentland, A. Diversity of idea flows and economic growth. *J. Social Comput.* **1**, 71–81 (2020).
22. Llorente, A., Garcia-Herranz, M., Cebrian, M. & Moro, E. Social Media Fingerprints of Unemployment. *PLoS ONE* **10**, e0128692 (2015).
23. Su, J., Kamath, K., Sharma, A., Ugander, J. & Goel, S. An experimental study of structural diversity in social networks. In *Proc. International AAAI Conference on Web and Social Media* Vol. 14, 661–670 (AAAI, 2020).
24. Gee, L. K., Jones, J. J., Fariss, C. J., Burke, M. & Fowler, J. H. The paradox of weak ties in 55 countries. *J. Econ. Behav. Organization* **133**, 362–372 (2017).
25. Jahani, E., Fraiberger, S., Bailey, M. & Eckles, D. Long ties, disruptive life events, and economic prosperity. *Proc. Natl Acad. Sci. USA* **120**, e2211062120 (2022).
26. Centola, D. & Macy, M. Complex contagions and the weakness of long ties. *Am. J. Sociol.* **113**, 702–734 (2007).
27. Jiang, S., Ferreira, J. & Gonzalez, M. C. Clustering daily patterns of human activities in the city. *Data Min. Knowl. Discov.* **25**, 478–510 (2012).
28. Jiang, S., Ferreira, J. & Gonzalez, M. C. Activity-based human mobility patterns inferred from mobile phone data: a case study of singapore. *IEEE Trans. Big Data* **3**, 208–219 (2017).
29. Hunter, R. F. et al. Effect of COVID-19 response policies on walking behavior in us cities. *Nat. Commun.* **12**, 3652 (2021).
30. Yang, Y., Pentland, A. & Moro, E. Identifying latent activity behaviors and lifestyles using mobility data to describe urban dynamics. *EPJ Data Sci.* **12**, 15 (2023).
31. Solow, R. M. A contribution to the theory of economic growth. *Q J. Econ.* **70**, 65–94 (1956).
32. Barro, R. J. Economic growth in a cross section of countries. *Q. J. Econ.* **106**, 407–443 (1991).
33. Glaeser, E. L., Scheinkman, J. A. & Shleifer, A. Economic growth in a cross-section of cities. *J. Monetary Econ.* **36**, 117–143 (1995).
34. Moro, E., Calacci, D., Dong, X. & Pentland, A. Mobility patterns are associated with experienced income segregation in large us cities. *Nat. Commun.* **12**, 4633 (2021).
35. Dong, X. et al. Social bridges in urban purchase behavior. *ACM Trans. Intell. Syst. Techno.* **9**, 1–29 (2017).
36. Singh, V. K., Bozkaya, B. & Pentland, A. Money walks: implicit mobility behavior and financial well-being. *PLoS ONE* **10**, e0136628 (2015).
37. Blumenstock, J., Cadamuro, G. & On, R. Predicting poverty and wealth from mobile phone metadata. *Science* **350**, 1073–1076 (2015).
38. Barbosa, H. et al. Uncovering the socioeconomic facets of human mobility. *Sci. Rep.* **11**, 1–13 (2021).
39. Aleta, A. et al. Quantifying the importance and location of SARS-CoV-2 transmission events in large metropolitan areas. *Proc. Natl Acad. Sci. USA* **119**, e2112182119 (2022).
40. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
41. Kreindler, G. E. & Miyauchi, Y. Measuring commuting and economic activity inside cities with cell phone records. *Rev. Econ. Stat.* **105**, 899–909 (2019).
42. Deaton, A. & Muellbauer, J. *Economics and Consumer Behavior* (Cambridge Univ. Press, 1980).
43. Belloni, A., Chernozhukov, V. & Hansen, C. Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Studies* **81**, 608–650 (2014).
44. Luptakova, I. D., Simon, M. & Pospichal, J. Weak ties and how to find them. In *23rd International Conference on Soft Computing* 16–26 (Springer, 2019).
45. Bilbao-Osorio, B. & Rodriguez-Pose, A. From R&D to innovation and economic growth in the EU. *Growth and Change* **35**, 434–455 (2004).
46. Blumenstock, J. E. Estimating economic characteristics with phone data. In *AEA papers and Proceedings* Vol. 108, 72–76 (AEA, 2018).
47. *2015–2019 American Community Survey 5-Year Estimates* (United States Census Bureau, 2019); https://www.census.gov/programs-surveys/acs


## Acknowledgements

We thank Cuebiq, who kindly provided us with the mobility dataset for this research through their Data for Good program. S.W. acknowledges partial support from a University of Florida ROSF-2023 grant. G.W. is partially supported by the NSF (grant no. 1952096). E.M. acknowledges support by Ministerio de Ciencia e Innovación/Agencia Española de Investigación (MCIN/AEI/10.13039/501100011033) through grant no. PID2019-106811GB-C32, and the NSF under grant no. 2218748.

## Author contributions

S.W., G.W., T.Y., E.M. and A.S.P. conceptualized the work. S.W. and Y.Z. performed the methodology, and designed and implemented the experiments. S.W. wrote the original draft, which was reviewed and edited by S.W., Y.Z., G.W. and T.Y. S.W., E.M. and Y.Z. curated the data, which were visualized by S.W. and Y.Z. S.W. and A.S.P. supervised and administered the project.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s44284-024-00051-7.

**Correspondence and requests for materials** should be addressed to Shenhao Wang or Esteban Moro.

**Peer review information** *Nature Cities* thanks Yang Yue and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Shenhao Wang

Last updated by author(s): Feb 13, 2024

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed |
|---|---|
| ☐ | ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☒ | ☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ A description of all covariates tested |
| ☒ | ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☒ | ☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | No specific software was used to collect the data. |
| Data analysis | Data analysis was done using Python. Please refer to the supplementary information for further details. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Mobility data are available from Cuebiq, available upon request submitted to https://www.cuebiq.com/about/data-for-good/. Other data used come from the American Community Survey (5y) from the Census, which is publicly available on their websites. See the supplementary materials for a description of them

## Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

| | |
|---|---|
| Reporting on sex and gender | We do not report the sex and gender of the participants. |
| Population characteristics | See below |
| Recruitment | See below |
| Ethics oversight | The privacy-enhanced mobility data was collected by the company Cuebiq using anonymized records of GPS locations from users that opted-in to share the data anonymously through a General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA) compliant framework. Additionally, we obtained IRB exemption to use the mobility data from the MIT IRB office. (COUHES protocol #1812635935 and its extension #E-2962) |

Note that full information on the approval of the study protocol must also be provided in the manuscript.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Study description | We perform a quantitative study on observational data using statistical methods to detect the relationship between mobility network and economic outcomes. The elastic net regularization is used to automatically choose the most important activities. |
| Research sample | Data used are geo-locations from anonymous opted-in devices collected by the company Cuebiq in 3 metro areas in the US. Data has been aggregated at the level of census areas where a number of devices are present to prevent de-anonymization. |
| Sampling strategy | Post-stratification techniques has been used to correct for potential biases in the sample of users and to ensure population representation. |
| Data collection | Data collection was done by the company Cuebiq |
| Timing | Data was collected from October 2016 through March 2017 |
| Data exclusions | No data was excluded. |
| Non-participation | Only anonymous opted-in devices where used in the analysis |
| Randomization | NA |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |