Supplementary Informations for
Mobility Census for monitoring rapid urban development

Gezhi Xiu*, Jianying Wang, Thilo Gross, Mei-Po Kwan, Xia Peng, Yu Liu

CONTENTS

I. Data Description and Construction 2
A. Mobile Phone Data 2
B. Mobility Variables: Human Mobility Attributes 3
1. Human mobility attributes at the individual level 3
2. The collective-level mobility features 4
3. Local inequality metrics 7

II. Details in Statistical Analysis 8
A. Spearman Rank Correlation 8
B. Parameters of GMMs and key factors for different clusters 8

III. Diffusion Maps 10

IV. Analysis on the Inverse Participation Ratios 10

V. Dominant eigenfeatures of 2018 mobility data 13

VI. Explanatory mobility variables for 2021 Mobility Census 13

References 20

I. DATA DESCRIPTION AND CONSTRUCTION

A. Mobile Phone Data

We used Beijing as the study area, characterized by mobile phone signal (MP) records. The MP record dataset is collected by Unicom, one of the three major mobile network operators in China, which covers about 6.87 million subscribers. As a background, Beijing's approximate population is 21 million. The subscribers are usually regarded as a large and unbiased subsample of the population in terms of demography and socioeconomic properties. In addition, the densely deployed cell towers frequently collect the location of the subscribers when they are calling or accessing online resources. Thus, a location's activity can be characterized by the usage logs which include the anonymous user location and checkin time of the subscribers nearby.

The spatial resolution of our study is determined at grid cells of 500 meters. Within the Sixth Ring region of Beijing, the cell-tower coverage is sufficiently dense to provide a 250-meter grid-cell spatial resolution. However, as we aim to study the spatial evolution of Beijing's population landscape, especially the polycentric transitions, it is essential to include the surrounding areas, where cell-tower coverage is sparser but sufficient to support a 500-meter study resolution.

The temporal accessibility of the dataset is between 2018 and 2021. To balance the relatively complete characterization of the city and computational efficiency, we choose the study period of five-continuous weeks, one in 2018 and another in 2021. Throughout 2018-2021, significant changes took place in Beijing. Parts of the Beijing municipal government moved from the capital's downtown center to the Tongzhou district in 2019, and the rest is also moving; In the meantime, Beijing Daxing International Airport was put into use in 2019. The significant events in the city's periphery and the gradual development of the subcenters and industrial developments have stretched Beijing's job-housing mismatching structures, leading to significant changes in human mobility patterns. On the contrary, finding the invariants of the cities in mobility and identifying the critical transitions of the mobility landscape contribute to the designing further development of the cities.

B. Mobility Variables: Human Mobility Attributes

To quantitatively represent the mobility patterns of the different censuses, we use mobility metrics concerning radius, frequency, in/outflow ratio, etc., both at the individual and collective level. Notably, the individual-level metrics indicate those metrics assessed from individual trajectories (e.g., length of displacements, the entropy of visitation, etc.), and collective-level metrics indicate those assessed from the origin-destination matrix between different censuses. The features are categorized into the group at the individual level and that at the collective level specified in the following subsections. To conclude, we adopted 1,106 features, (442 individual-level mobility variables and 111 collective-level mobility variables, $\times 2$, for each cell as the origin and destination of flows) to reflect the local inequality.

1. *Human mobility attributes at the individual level*

From the raw human mobility dataset, we first identify the stay points as the pass-by points i where individuals stay for more than 30 min, which is commonly used for trajectory analysis [1, 2]. Accordingly, we capture the number of stays points N_u, the number of distinct places an individual visit N_P, the average duration an individual spend at a stay point T_s, average commuting time per trip $T_{avg} = \sum T_i/N_u$, total commuting time $T = \sum T_{N_u}$, and average Euclidean distance per trip $\sum L_{N_u}/N$. We further decompose the individual trajectories by the mobility features of 500m grids as the origins and destinations of trajectories. In general, we get the features at the individual level with a dimensionality of 442 (17 features $\times$13 statistical attributes $\times$2 for working places or residential places).

Then, we estimate the travel distance L_{ij}^T, commuting speed V_{ij} and detour ratio R_{ij}^l between L_{ij} and L_{ij}^T (to quantify the detour extent of a trip) as deterministic mobility features. We use entropy to measure the complexity of an area's visitation, which is defined as:

$$E_u = -\sum_{j=1}^{N_u} P_u \log P_u(j), \tag{1}$$

where N_u is the number of distinct places visited by an individual u, and $P_u(j)$ is the probability for a trip to point j. We also define stay-duration-weighted entropy E_u^t by assigning P_u^j as the probability that finding an individual appears in the area j.

Meanwhile, the radius of gyration (ROG) represents the characteristic distance of individuals' travels (3). It is defined as:

$$\mathrm{ROG}_u = \sqrt{\frac{1}{N}\sum_{i=1}^{N_u} w_j \cdot \mathrm{dist}\,(r_u(i) - r_u(cm))}, \tag{2}$$

where $r_u(i)$ is the position of point i; $r_u(cm)$ is the centre of mass of trajectories; dist is the Euclidean distance; w_j is the visit frequency. When w_j is the probability of stay duration in location j of an individual, we get a stay-duration-weighted ROG ROG_u^t. Meanwhile, most individuals' daily movements are characterized by a few locations, such as homes, workplaces, supermarkets, and schools. Therefore, considering the k areas with the largest visitation volume, the radius of gyration of these areas is defined as:

$$\mathrm{ROG}_u^k = \sqrt{\frac{1}{k}\sum_{i=1}^{k} w_j \cdot \mathrm{dist}\,(r_u(i) - r_u^k(cm))}, \tag{3}$$

where $r_u^k(cm)$ is the weighted center of the population. We also assess the stay-duration-weighted k-ROG ROG_u^{kt}. Further, we propose using the ratio R_r^k (R_r^{kt}) between ROG_u^k (ROG_u^{kt}) and ROG_u (ROG_u^t) to quantify how much an individual's daily movements are characterized by their familiar places.

To comprehensively describe the spatial distributions of individuals with different mobility metrics, we use the average value, standard deviation, skewness, kurtosis, and deciles (from 10% to 90%) of these individual-level mobility features aggregated based on their home and workplace location respectively. The distributions of some reprehensible mobility attributes are shown in Fig 1.

2. *The collective-level mobility features*

The collective-level mobility features are mainly extracted from the origin-destination matrix (OD matrix) to reveal the temporal variations of the spatial intersections. The temporal variations can roughly be divided into commuting and non-commuting hours. Such categorization is essential because of the dominance of commuting mobility in urban studies and its impact on urban development. The component a_{ij} of non-commuting OD matrices represents the travels starting from area i and ending at area j at the time period around h. It does not consider the via points of trajectories. The component a_{ij}^c of commuting OD matrices represents the number of individuals who live in area i and work in area j.

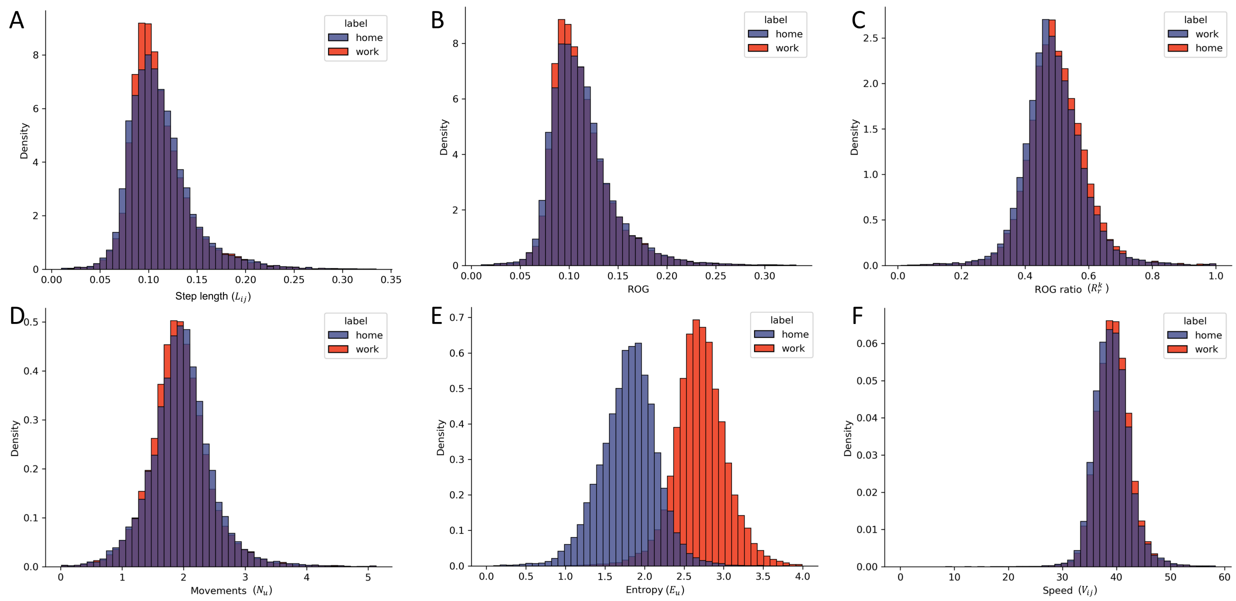

FIG. 1. The aggregated mobility index at the individual level. Step length (home) is the average distance aggregated based on individuals' home location. ROG is the average of the ROG at the individual level. ROG ratio stands for the quotient between ROG2 and ROG. Movements stand for an average of the number of daily trips individuals make. Entropy is the average temporal-uncorrelated entropy of a set of destinations at the individual level. Speed stands for the average speed for trips.

Home and work locations are easy to recognize according to individual activity rhythms. A workplace is where an individual has their daily activities (e.g., workplace or school), and a home place is where individuals spend their evenings. To conclude, the collective mobility features have a dimensionality of 111: first, 4×24 for hourly sampled inflow, outflow, and their h-indices; secondly, the daily sums of inflow, outflow, and their h-indices; lastly, 11 collected features of commuting flows, identified by trajectories destinated at the locations of residential and workplaces: inflow volume, outflow volume, inflow divided by outflow, the entropies of inflow and outflow, the weighted direction of all the flows by the volumes, and the inflow and outflow's h-index, respectively.

For OD matrices F_{ij}^h, we first reveal the patterns of in-flows and out-flows, demonstrating the total straightness of a grid's interactions with others. The in-flow and out-flow of grid i

at hour h are defined as:

$$F_{out}^{h} = \sum_{j=0}^{N} F_{ij}^{h}, \tag{4}$$

and

$$F_{in}^{h} = \sum_{j=0}^{N} F_{ji}^{h}. \tag{5}$$

Meanwhile, the ratio and difference between in- and out-flow is $R_f = F_{out}/F_{in}$ and $D_f = F_{out} - F_{in}$. A cell with $R_f > 1$ indicates its role as a local sink. We also sample the distributional features of OD flows beyond point estimations of mobility features. We use the entropy of F_{ij}^{h} to quantify the randomness of the in-flows. Finally, we calculate each cell's total in-flow and out-flow to represent their connection strength or attractiveness. We also use the H-index, the number of trips larger than H being at least H to evaluate cells' cumulative attractiveness [3]. H_i^o and H_i^d are the H-index of area i as source and sink of traveling, respectively. In general, we get the features with a dimensionality of 100 $(4 \times 24 + 4)$. The distributions of some typical parameters are shown in Fig. 2.

Particularly, for commuting OD matrices, in-flow and out-flows are always referred as residential population (P_r) and work populations (P_w). We further quantify the ratio between the residential population and work populations, and the randomness of the sources of the residential population and work populations. We also use the H-index to evaluate cells' cumulative attractiveness considering the commuting matrix. Besides, since all the flows of a cell can be defined as a vector centered in this cell i, $F_{ij}\vec{\mu}_{ij}$, where $\vec{\mu}_{ij}$ is the unit vector origins from the location i to j. Then the net flows of this cell could be defined as

$$F_{net}\vec{\mu}_{net} = \sum_{j=0}^{N} F_{ij}\vec{\mu}_{ij}, \tag{6}$$

where F_{net} is the volume of the net flow and $\vec{\mu}_{net}$ demonstrate its directions. Further, we turn the vector $\vec{\mu}_{net}$ into the angle A_{ij} to represent the direction. In general, we get the features with a dimensionality of 11 to reflect the patterns of the commuting OD matrices.

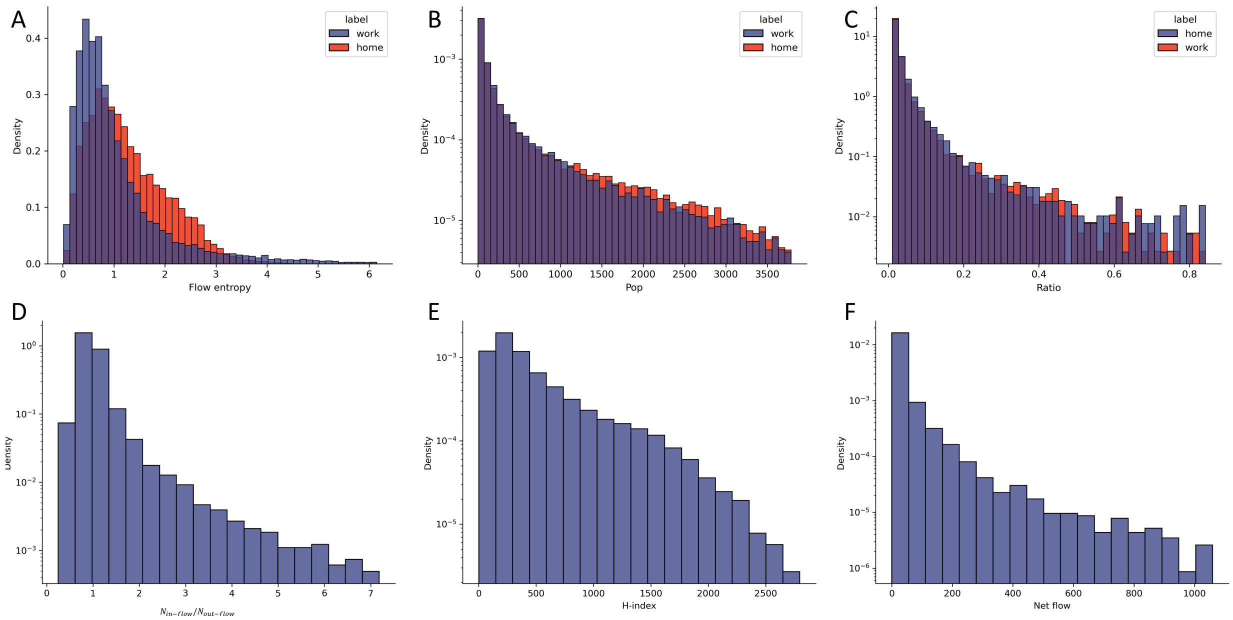

FIG. 2. The mobility index is quantified by the OD matrices. Flow entropy (work) quantifies the randomness considering the distribution of home locations (origin of the commuting flow). Pop stands for the residential and working population of each spatial unit. The ratio is the quotient between the flows and the residential/working population. $N_{in-flow}/N_{out-flow}$ indicates the quotient between the in-flow and out-flow. The h-index represents the relative importance of a spatial unit in OD matrices.

3. Local inequality metrics

We use the small-area level (SAL) Gini for radii of $r = 0.02$ degrees around a cell i to quantify the local heterogeneity of the mobility patterns. It is defined as:

$$\mathrm{Gini}_i = \frac{1}{2N^2 u} \cdot \sum_{j=1}^{N_{\mathrm{local}}} \sum_{k=1}^{N_{\mathrm{local}}} |a_k - a_j|.$$

We use the SAL location ratio with the same radius to quantify the relative value of mobility metrics of cell i compared with its neighborhoods. It is defined as:

$$LR_i = a_i/u_i. \tag{7}$$

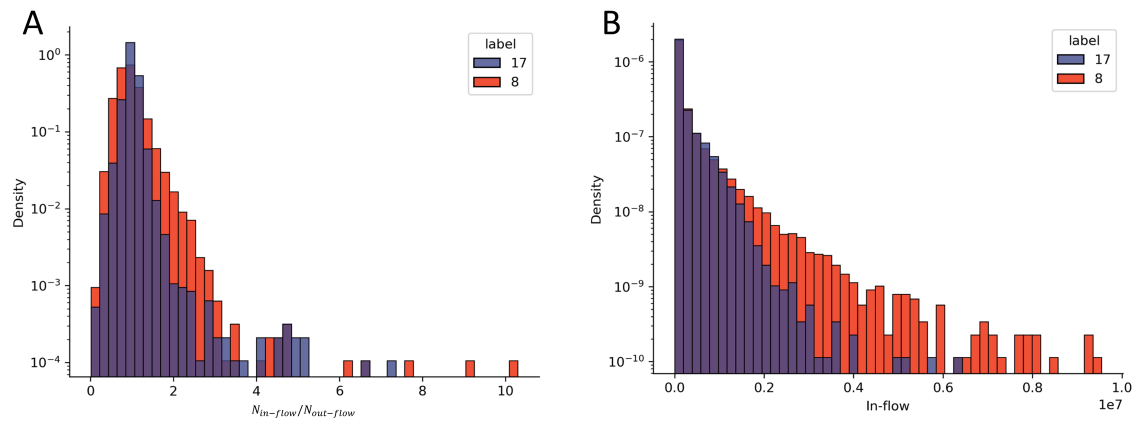

FIG. 3. The temporal difference for the in- and out-flow ratio and the distribution of the in-flow. 17 stands for the mobility index from 17:00 pm to 18:00 pm.

II. DETAILS IN STATISTICAL ANALYSIS

A. Spearman Rank Correlation

Here, cell xx's mobility variables are denoted by XX. We adopt the Spearman Rank Correlation from the definition

$$r_s = \rho_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{\mathrm{cov}(\mathrm{R}(X), \mathrm{R}(Y))}{\sigma_{\mathrm{R}(X)}\sigma_{\mathrm{R}(Y)}}, \tag{8}$$

where ρ denotes the Pearson correlation coefficient applied to the ranks $R(X)$ and $R(Y)$, $R(X)$ is the ranks of X's elements out of the corresponding mobility variables, and σ denotes the standard deviation of X.

B. Parameters of GMMs and key factors for different clusters

To identify the low-dimensional urban structure captured by the mobility census, we apply Gaussian mixture model (GMM) clustering. This model assumes the hidden structures, namely the hidden sub-populations of data, could be represented by a combination of various Gaussian distributions. We used the Bayesian information criterion (BIC), which is a penalized form of the log-likelihood, to find the optimal number of clusters. BIC considers both the accuracy (based on log-likelihood) and the complexity (based on the penalty term

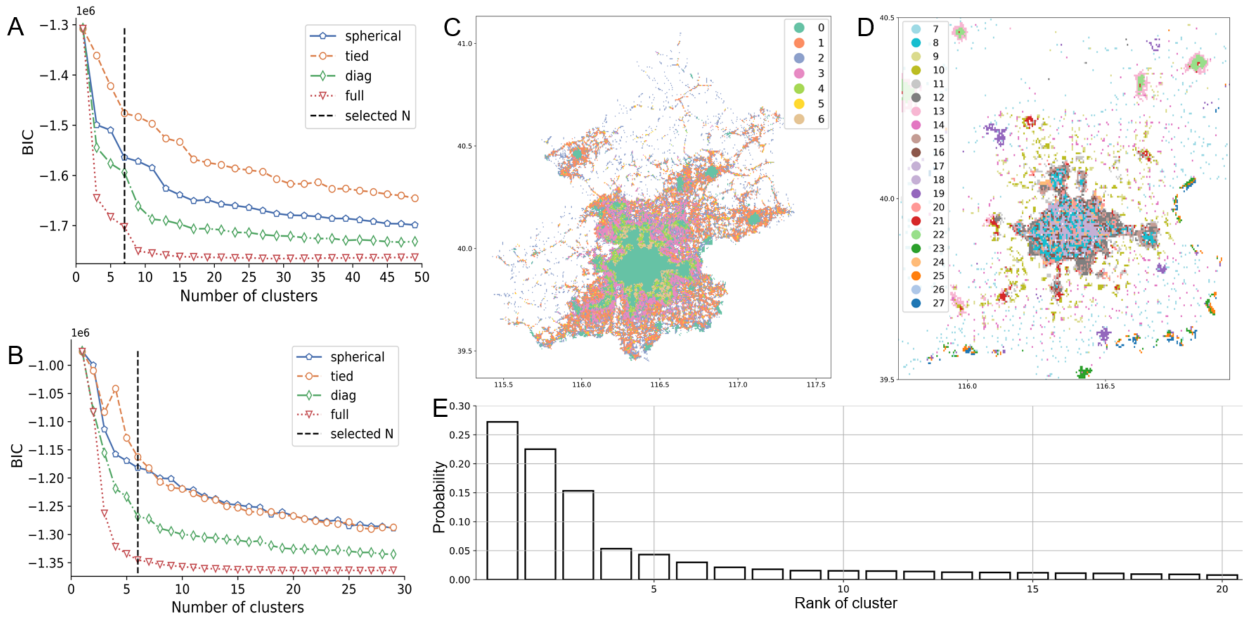

FIG. 4. The GMM clustering parameter selection. **A**. BIC changes with the number of clusters in 2021. **B**. BIC changes with the number of clusters in 2018. **C**.The spatial distribution of top 6 clusters with the largest spatial coverage. Grids sorted into other clusters are labeled as cluster 0. **D**. The spatial distribution of clusters except for the top 6 clusters. These clusters further identify the land-use distribution in the center and sub-centers of Beijing. **E**. The probability distribution of the spatial coverage of different clusters.

for the number of parameters). As Fig. 4A and Fig. 4B shows, GMM models with full covariance parameters and 6 clusters could obtain a suitable trade-off between the complexity and BIC (based on knee-point detection). Fig. 4C-Fig. 4E demonstrates the characteristics of clusters predicted by GMM with the highest BIC. As Fig. 4E indicates, the top 66 clusters contain most of the spatial units, and each of the other clusters only accounts for no more than 3% spatial units. We further illustrated the spatial distribution of the first 6 clusters (denoted as 1-61-6) and other clusters (denoted as 00) in Fig. 4C. The top 66 underlines the significant difference in land use between rural and urban areas in Beijing. Meanwhile, clusters focus on further dividing urban areas into different land-use types (Fig. 4C). Therefore, we focus on the top six clusters to highlight the most significant changes.

III. DIFFUSION MAPS

Diffusion mapping was applied to extract the low-dimensional structure beneath the Mobility Census data table [4]. The diffusion mapping procedures were regularized by Barter and Gross [5], and optimized by the authors of this article. Briefly, the method involves (i) for each grid cell, calculating its Pearson rank correlation coefficient with each of the rest grid cells, and storing the id of k other grids of the highest coefficients as the grid cell's nearest neighbors; (ii) iterating the above procedure by increasing k from 1 to 1000, until the graph is connected; (iii) converting the graph to its weighted adjacency matrix, and compute the adjacency matrix's corresponding row-normalized Laplacian matrix; and finally, (iv) perform eigendecomposition to the Laplacian matrix, and sort the eigenvalues from the least positive to the most positive. The eigenvectors corresponding to the smallest few positive eigenvalues represent the diffusion variables of the most important variations in the dataset. The importance of the diffusion variables is determined by corresponding eigenvalues, which also correspond to the characteristic time scale of the data's diffusive modes.

IV. ANALYSIS ON THE INVERSE PARTICIPATION RATIOS

As we decomposed the mobility landscapes in the previous section, the areas are quantified by categories such as workplace quality, housing, entertainment, and new company hubs. If it is surrounded by various compact mobility labels and concentrated demographic features, an area can experience traffic jams and exhibit unsatisfactory working/living conditions at peak times. We take the inflow volumes and the outflow volumes as mobility labels attached on each grid cell, to examine how different areas are pressured as origins and destinations. In this section, we introduce the eigenvectors' inverse participation ratio (IPR)

$$\mathrm{IPR}^i = \frac{\sum_{j=1}^{N} w_j^h (f_i^j)^4}{\frac{1}{N}\sum_{j=1}^{N} w_j^h}, \tag{9}$$

a measure from statistical physics, to distinguish between one eigenfeature with a coincidence of feature emergence on the source and sink of flows. Here, h is the index of the weights, e.g., in-flow, out-flow, population, etc., and w is the weight. The physical meaning of IPR is understood by two limiting cases: identical components, and thus no emergence of phenomena, $1/\sqrt{N}$ leads to IPR=$1/N$ which approaches 0; and whereas totally localized components $(1, 0, \dots, 0)$, has sIPR=1. For example, if an eigenfeature highlights the tourist attractions, but when weighted by the inflows, this eigenfeature gives a near-zero value, predictably, no traffic congestion would emerge for the tourist sites in this city.

To specify how IPRs identify the grid cells' heterogeneous response to inflow and outflow, we measure three sets of weight: $w_j \equiv 1$ defines standard IPR (sIPR); w_j = the residential population (rIPR) and work population (wIPR) corresponds to the OD gap. From the physical meanings, we interpret that sIPR as well as wIPR and rIPR is related to the number of weighted eigenfeature components that are significantly different from zero. If the highlighted locations in an eigenfeature are relatively more important as work locations than residential locations, the rIPR should be smaller than the eigenfeature's wIPR.

Table I ranks the IPR, wIPR, and rIPR of the 10 most dominant eigenfeatures. Generally, localization corresponds to a large weighted IPR for an eigenfeature, which is related to the high coincidence of the eigenfeature with its weight. Specifically, wIPR is maximized by f_5, indicating the extreme components of f_5 are actually the core working locations. In fact, in f_5 we find the financial street, International trade center, and Xierqi, the most recognized functional zones of finance and high-tech industries in Beijing. f_4 and f_1 are similar in terms of localization, but f_4 ranks higher in rIPR while f_1 ranks higher in wIPR. This indicates f_4 is highlighted by residential population, and by looking at the extreme values in $f_4 - f_1$ we have almost all the large communities in Beijing.

sIPR	ID1	rIPR	ID2	wIPR	ID3
0.008982	2	0.005576	2	0.018148	5
0.001172	6	0.005157	6	0.005671	2
0.000890	5	0.004847	5	0.004659	6
0.000626	8	0.002844	7	0.002688	7
0.000503	4	0.002065	4	0.002295	1
0.000496	7	0.001746	1	0.001477	4
0.000269	9	0.000660	3	0.001389	3
0.000236	1	0.000410	8	0.000369	8
0.000152	3	0.000241	9	0.000313	9

TABLE I. The Spatial IPR, residential IPR and work IPR. Rank represents the ascending order of the first 1010 EFs sorted by the IPRs. **ID1**, **ID2**, and **ID3** indicate the indices of EFs sorted by **sIPR**, **rIPR**, and **wIPR**, respectively.

V. DOMINANT EIGENFEATURES OF 2018 MOBILITY DATA

VI. EXPLANATORY MOBILITY VARIABLES FOR 2021 MOBILITY CENSUS

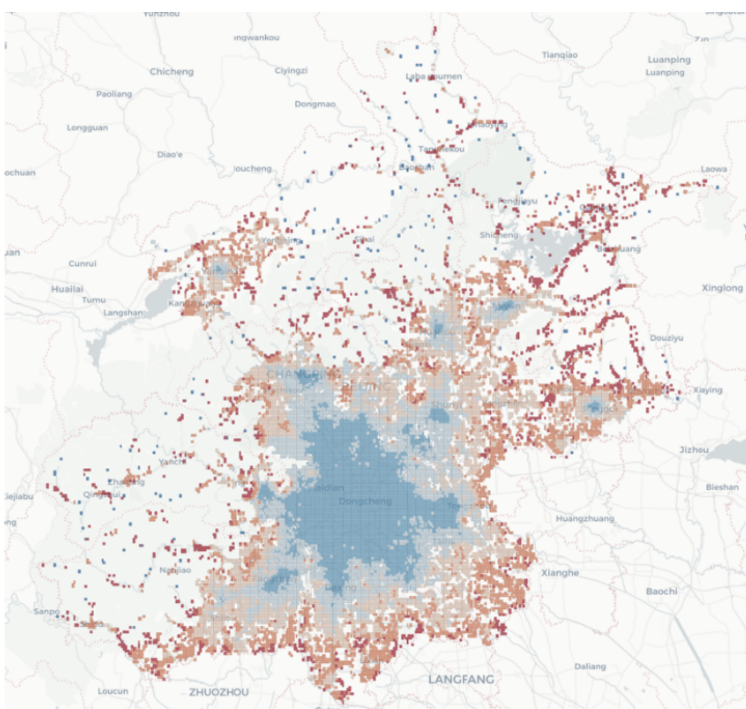

FIG. 5. The MC eigenfeature 1 of Beijing in 2018.

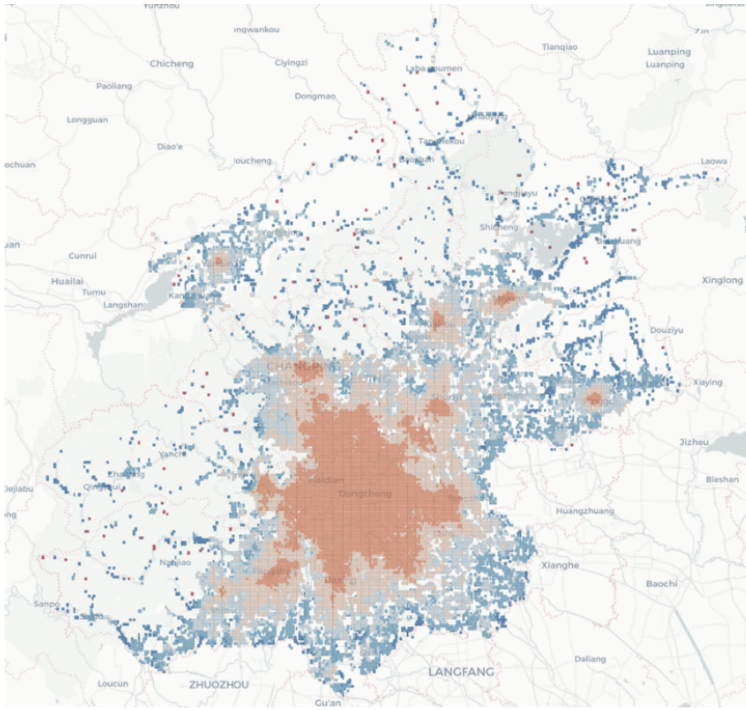

FIG. 6. The MC eigenfeature 2 of Beijing in 2018.

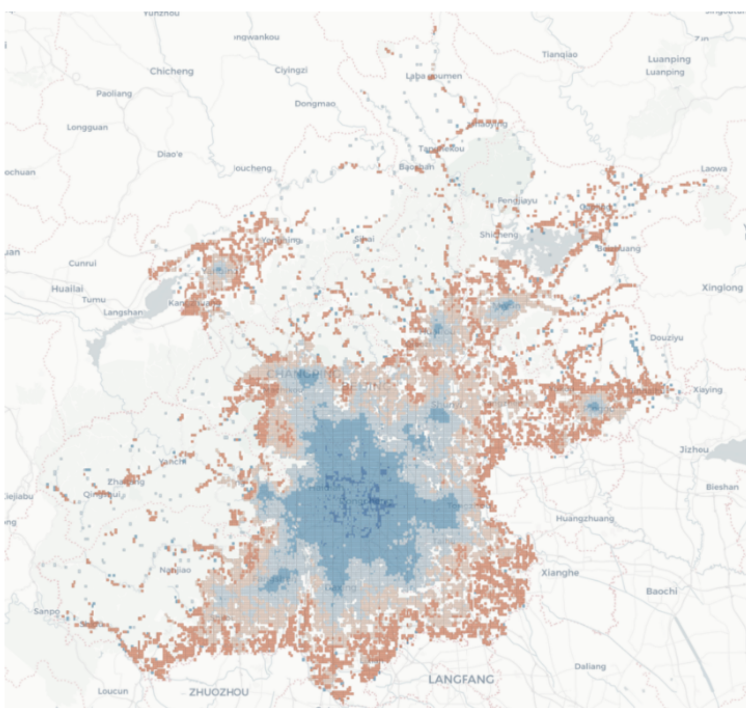

FIG. 7. The MC eigenfeature 3 of Beijing in 2018.

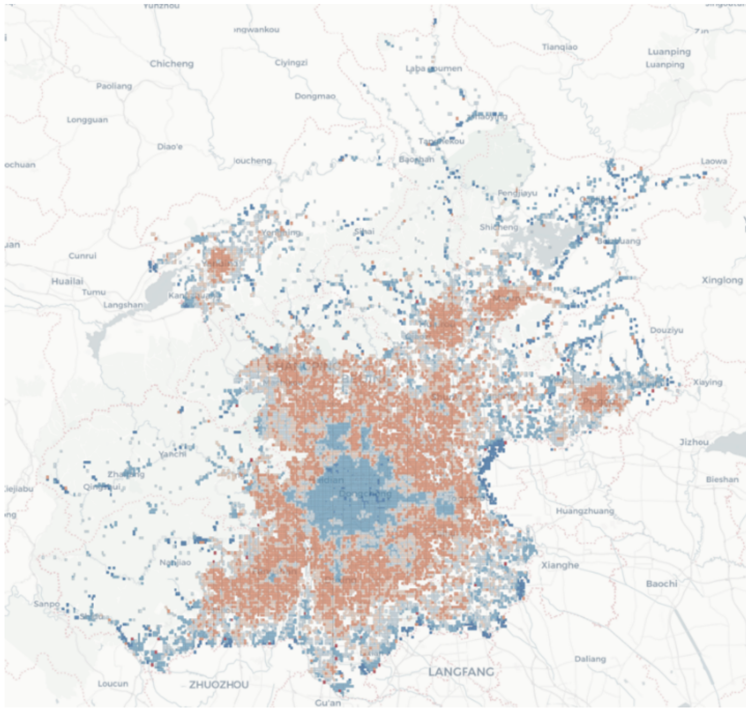

FIG. 8. The MC eigenfeature 4 of Beijing in 2018.

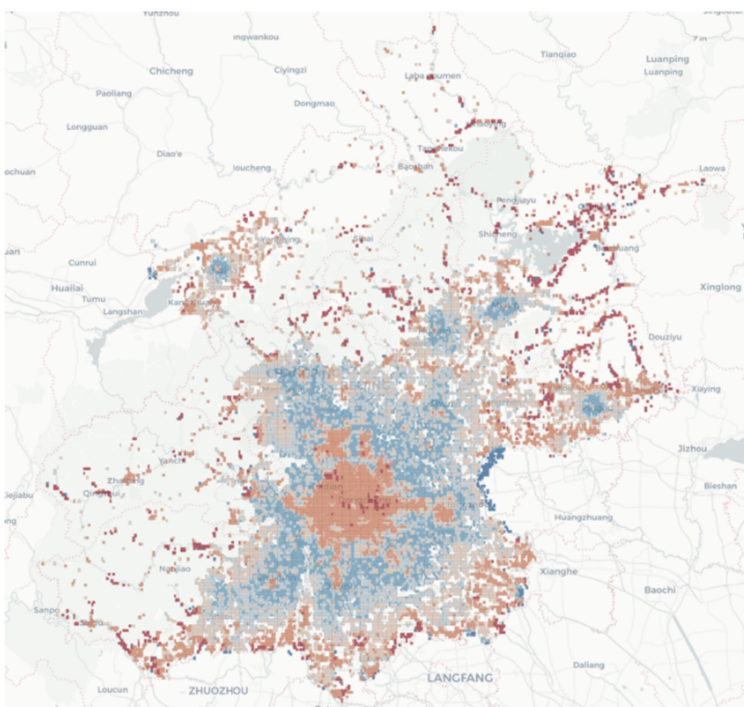

FIG. 9. The MC eigenfeature 5 of Beijing in 2018.

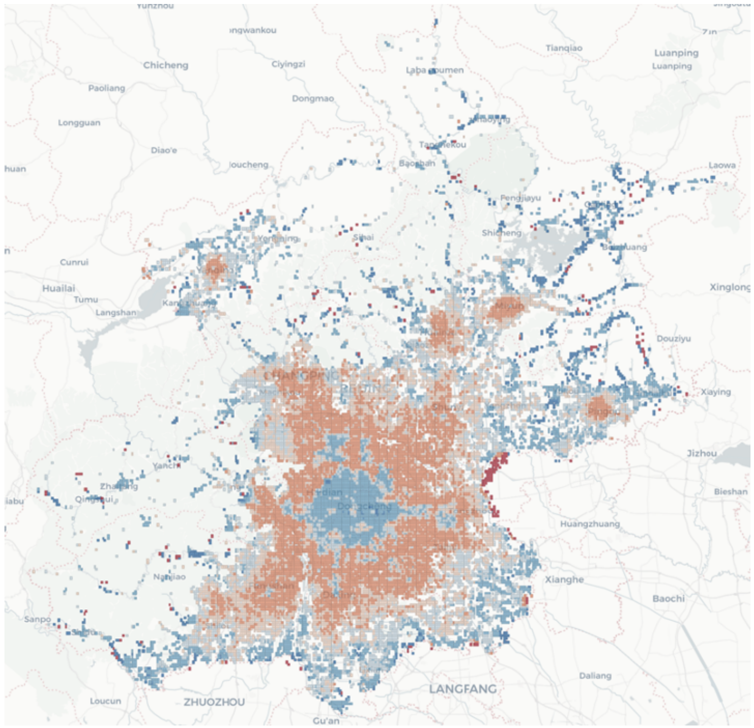

FIG. 10. The MC eigenfeature 6 of Beijing in 2018.

Rank	f_1	**Corr**	f_2	**Corr**	f_3	**Corr**	f_4	**Corr**	f_5	**Corr**
1	H-index	-0.8807	Gini (Flow ratio (11))	0.6804	Gini (detour ratio work (p2))	0.7142	Gini (H-index)	-0.3679	Net flow	0.3203
2	out-flow (17)	-0.8489	Gini (Flow ratio (16))	0.6773	Gini (detour ratio work (p1))	0.7140	Rog2 ratio work (p2)	0.3399	in-flows (8)	0.3203
3	out-flow (12)	-0.8484	Gini (Flow ratio (17))	0.6714	Gini (detour ratio home (p2))	0.7136	Rog2 ratio work (p3)	0.3367	in-flows (9)	0.3028
4	out-flow (11)	-0.8483	Gini (Flow ratio (14))	0.6612	Gini (detour ratio home (p1))	0.7134	Rog2 work (p5)	0.3333	Gini (H-index)	0.3026
5	out-flow (16)	-0.8481	Gini (Flow ratio (13))	0.6566	Gini (detour ratio home (p3))	0.7128	Rog2 work (p4)	0.3316	in-flows (10)	-0.2757

TABLE II. Correlations between mobility variables and diffusion map eigenvectors. Gini stands for SAL Gini, LR stands for local ratio, and K stands for the kurtosis of the distribution; (P1) stands for the first percentiles of the value, and (1) stands for the mobility metric at 1am.

1	Bainaohui Computer Shops	2	Xiushuijie Mall
3	Wanda Plaza	4	The Place (Shimaotianjie)
5	Wangfujing Avenue	6	Beijing TV station
7	Parkview Green *Fangcaodi*, Beijing	8	Jianwai SOHO
9	Financial Street Plaza	10	Financial Street Shopping Centre
11	Fuxing Commercial Town	12	Yitai Plaza
13	Hopson One *Beijing*	14	Guijie Street
15	Xidan Joycity	16	Sanlitun
17	Taikoo Li Sanlitun	18	Wudaokou Farmer's Market
19	Lanjinglijia Dazhongsi Home Plaza Store	20	Yuelin Baixing Plaza
21	Beijing South Railway Station Shopping	22	Kirin Place (Hesheng Qilin Xintiandi)
23	Changan Shopping Mall	24	Wudaoko Mall
25	Sanlitun SOHO	26	Uptown Mall
27	Fortune Shopping Centre	28	China World Mall
29	Donghua Jinjie Shopping Mall, Beijing	30	New World Department Store
31	Chunping Plaza	32	Huaxi Live, Wukesong
33	Capitalland, Wangjing	34	Jiacheng Plaza

TABLE III. Geographical place labels for Figure 3 in the Main Text (Part I).

c1	The teacher dormitory of Tsinghua University
c2	The Zijing student dormitory of Tsinghua University
c3	The student dormitory of Tsinghua University
c4	The west student dormitory of Tsinghua University
d1	Nanhu Compound
d2	Huajiadixili compound
d3	Nanhuxili Compound
d4	Daxiyang compound
d5	Huajiadi compound
d6	West Wangjing compound
d7	Wangjing compound
d8	Guofengshangguan compound
d9	Dahshanzixili compound
d10	Rongke compound
d11	baoqinghuajue
e1	beijing west railway station
f1	beijing north railway station
g1	Xinfadi market Agricultural Products Wholesale Market (Farm product)
g2	Xinfadi market Agricultural Products Wholesale Market (seafood)
g3	Xinfadi market Agricultural Products Wholesale Market (organic product & updated area)
g4	Xinfadi market Agricultural Products Wholesale Market (fruit)
a1	Chuangchunyuan of Peking University
a2	Chuangchunxinyuan of Peking University
a3	Furongli compound
a4	Shaoyuan of Peking university
a5	Yannanyuan of Peking university
a6	Haidianlu subdistrict

TABLE IV. Geographical place labels for Figure 3 in the Main Text (Part II).

[1] J. Wang, M.-P. Kwan, W. Cao, Y. Gong, L. Guo, and Y. Liu, Assessing changes in job accessibility and commuting time under bike-sharing scenarios, Transportmetrica A: Transport Science , 1 (2022).

[2] S. Jiang, J. Ferreira, and M. C. Gonzalez, Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore, IEEE Transactions on Big Data **3**, 208 (2017).

[3] J. E. Hirsch, An index to quantify an individual's scientific research output, Proceedings of the National academy of Sciences **102**, 16569 (2005).

[4] R. R. Coifman and S. Lafon, Diffusion maps, Applied and computational harmonic analysis **21**, 5 (2006).

[5] E. Barter and T. Gross, Manifold cities: social variables of urban areas in the uk, Proceedings of the Royal Society A **475**, 20180615 (2019).