Supporting materials

In this supporting information, we present the essential extended details of the results and analysis.

Calibration results

This section gives the results of calibrating home and work identification algorithm with the ground truth data, given the optimal objective. Besides, we also evaluate the calibrated results based on mean distance error, standard deviation of distance error and identification coverage (that is, the share of users with home/work identified in those with home/work labeled from the ground truth) (*c.f. Supporting materials*, Table S3).

After 20000 times of iteration, the objective functions under seven settings of clustering radius all converge to some single value. It means that the pre-set algorithmic control parameters of CRS (including, iteration stages N and initial points in storage p) are sufficiently moderate to generate a set of parameters with which most individuals' resulting identified home and workplaces give the best fittings to the labeled ones.

We further calculate the 25th, 50th, 75th, 97.5th percentile error between the home and workplaces as identified by our algorithm and the labeled ones as in ground truth. Both algorithms perform well, achieving median errors of 0.08 km and 0.64 km, respectively. Moving out to the 75th percentile, the home algorithm continues to work well, with 2.90 km of error, whereas the work algorithm's error increases to 6.58 km. A further exploitation into the few cases with larger distance errors reveals that the poor performance usually happens to users who do not use Weibo frequently at workplace, or users who would post a microblog only when they are working overtime, which time period is not covered by the peak work event hours defined in our algorithm. In general, our home algorithm performs better than work algorithm.

Sample reconstruction

It has to be noted that, in this study, due to the limited size of non-commuters identified with check-in data set, population reconstruction for those has to sample from the whole microdata. Considering the data set we use is only a subset of all geo-tagged check-in records available on Weibo, when a sufficiently large data set is utilized, this process would improve further.

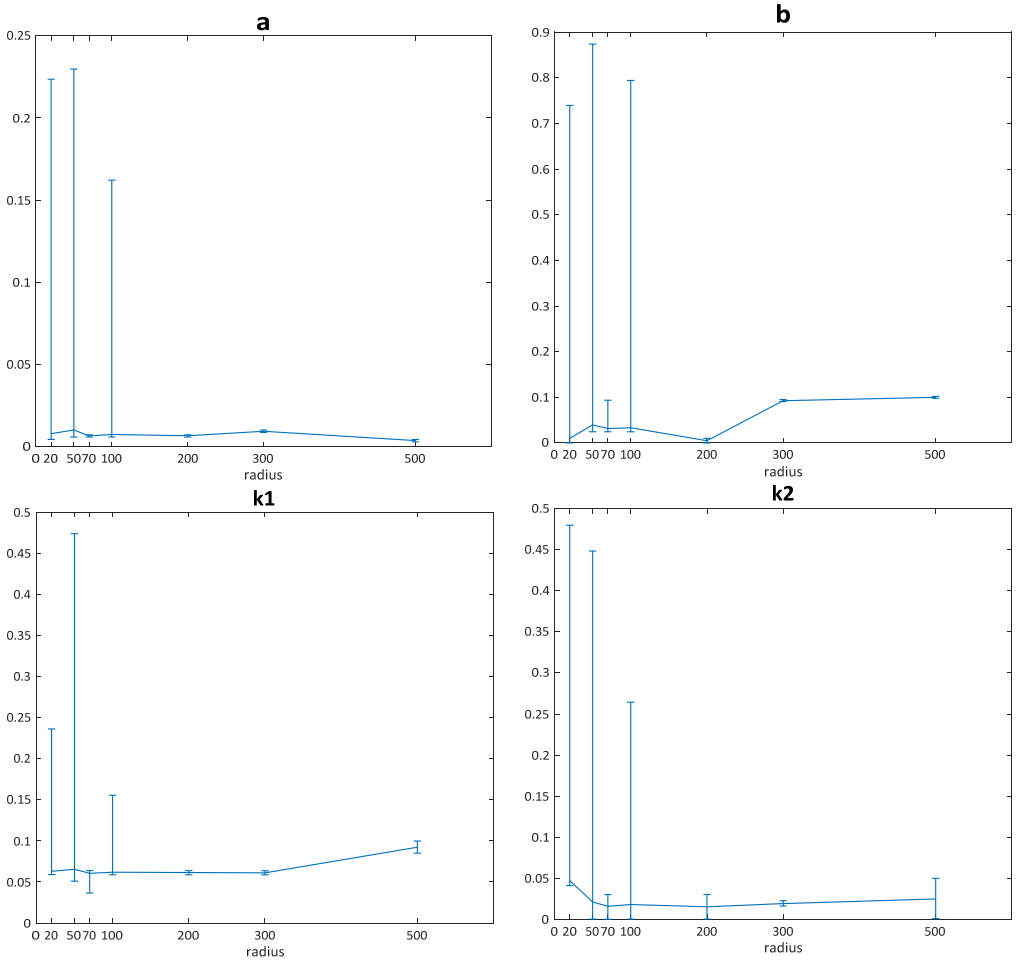

Fig. S1 The first 200 best parameter estimates of CRS algorithm under different settings of clustering radius (m). The upper bond: indicates the maximum value, lower bond, the minimum value and center, the best estimate of objective function. Comparatively, all of the four parameters give the best converging performance at the radius of 300 m.

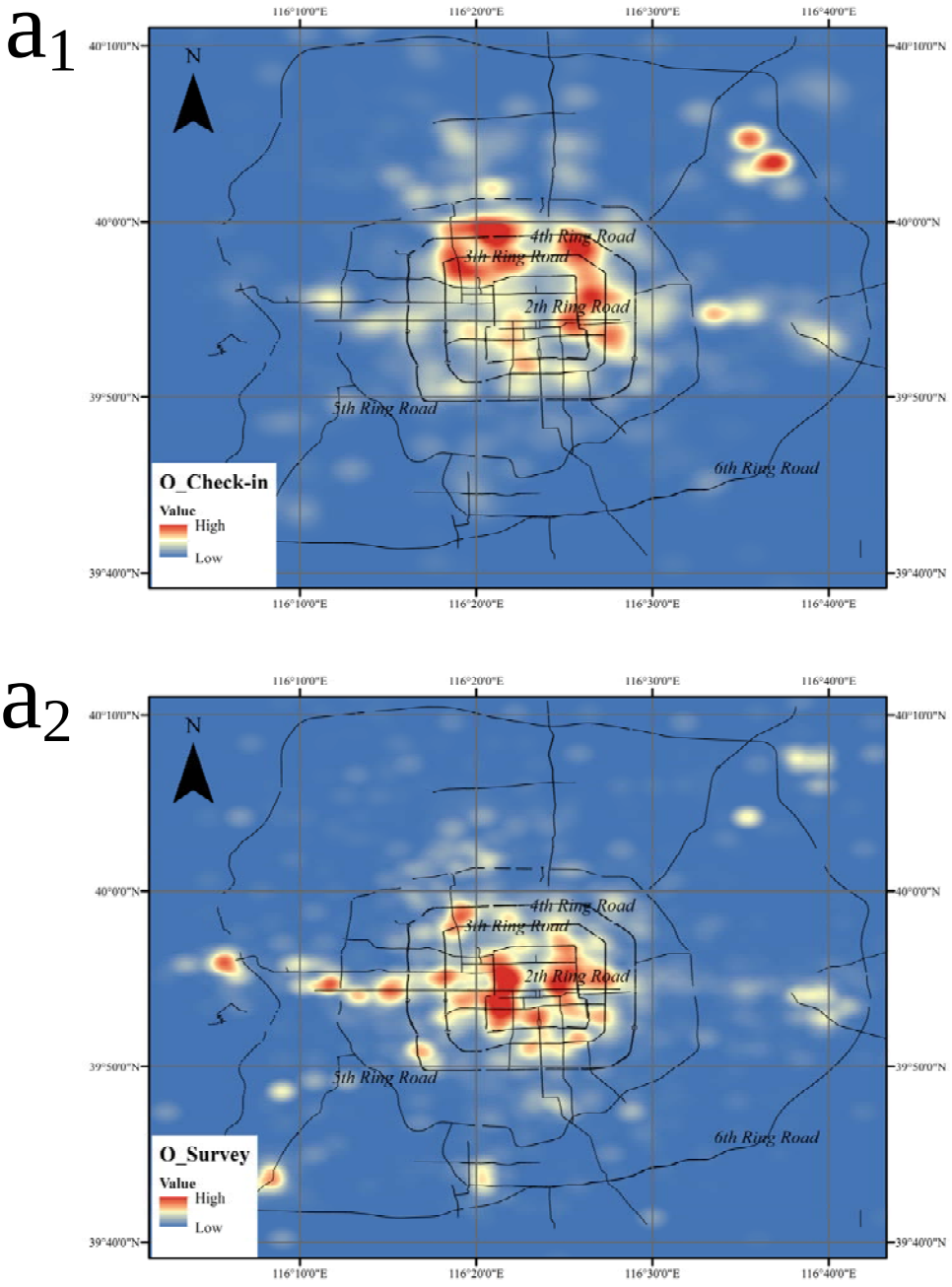

Fig. S2 Other activity density maps. The identified (a_1) and recorded (a_2) other activity density maps for the major metropolitan area in Beijing, with a *Cosine Similarity* index of 0.35. Among all the non-commuting locations, we calculate the number of days each unique venue is visited by each person (*Days*). Then, the popularity of a certain place as a candidate location for other activity can be calculated as the sum of the corresponding *Days* of all the persons.

Table S1 Bags of words used when building ground truth data

Activity	Work/School	Home
Words	work/school, company, office room, working station, classroom, work overtime, on duty	Home, bed, wake up, sleep, morning, good night, dormitory, go out

Table S2 Mean values of commuting distances by attributes

Attributes	Categories	Distance (km)	Attributes	Categories	Distance (km)
Gender	female	2.20	Education	primary	2.01
	male	2.02		secondary	2.10
				tertiary	2.13
	Chi-sq (P)	**0.01(0.91)**		**Chi-sq (P)**	**0.00(0.99)**
Age	15–24	1.81	District	Dongcheng	1.89
	25–26	1.67		Xicheng	1.53
	27–28	2.12		Chaoyang	2.07
	29–30	2.53		Haidian	1.39
	31–33	2.66		Shijingshan	3.27
	34–39	2.58		Fengtai	2.82
	40–49	1.86		Changping	3.03
	50 +	0.93		Shunyi	3.25
				Tongzhou	3.49
				Daxing	2.50
	Chi-sq (P)	**335.92(0.00)**		**Chi-sq(P)**	**103.64(0.00)**

Notes: Kruskal–Wallis test (or, KW test), a non-parametric method for testing whether samples originate from distributions with the same median, is used to select attributes strongly related to travel behavior. A significant KW test, with a P value close to zero, indicates that at least one sample is from a distribution with a different median.

Table S3 Calibration results of home and workplace identification

R (m)	Home			Work			Objective Function
	Mean	Standard Deviation	Coverage	Mean	Standard Deviation	Coverage	
20	3.32	6.83	91%	4.60	7.97	87%	4.28E-04
50	3.24	6.76	88%	4.42	7.87	86%	4.37E-04
70	3.20	6.63	95%	4.69	7.76	93%	4.38E-04
100	3.18	6.64	92%	4.66	7.90	92%	4.34E-04
200	3.43	6.89	97%	4.66	7.29	96%	5.05E-04
300	3.18	6.54	92%	4.48	7.25	91%	5.08E-04
500	3.67	7.32	97%	5.21	7.87	99%	5.03E-04

Table S4 Representation of the model constraint at the TAZs level

Constraints	Commuting		Non-commuting	
	TAE	CPE (%)	TAE	CPE (%)
District	0	0	0	0
Age	0	0	0	0
Gender	0	0	0	0
Age by Gender	0	0	0	0

Notes: Total Absolute Error (TAE) and Cell Percentage Error (CPE) are used to evaluate the goodness-of-fit of the synthetic population. TAE is calculated with Eq. (6) in the manuscript, while CPE is derived by TAE/N*100, where N is the population of the relevant cell, zone or attribute.

Table S5 Sub-district level *Coincidence Ratio* (CS) of commuting trip-length distribution before and after sample reconstruction

Districts	Sample CS	Synthesis CS	Sample size
Dongcheng	0.24	0.59	1589
Xicheng	0.20	0.58	2137
Chaoyang	0.25	0.55	6979
Haidian	0.23	0.52	5820
Fengtai	0.24	0.60	2671
Shijingshan	0.20	0.49	714
Changping	0.28	0.43	1698
Shunyi	0.26	0.25	977
Tongzhou	0.22	0.41	925
Daxing	0.20	0.42	1008

Notes: Sample CS represents the coincidence ratio between survey data and Weibo data before sample bias modification; Synthesis CS represents the coincidence ratio between survey data and Weibo data after sample bias modification. There are 24,518 records in total. The sample CS and synthesis CS for the whole dataset is 0.23 and 0.63, respectively.