JOURNAL OF THE ROYAL SOCIETY INTERFACE

Supplementary materials from

Quantifying segregation in an integrated urban physical-social space

Yang Xu,[1,*,+] Alexander Belyi,[2,3,+] Paolo Santi,[4,5] Carlo Ratti[4]

[1]Department of Land Surveying and Geo-Informatics,
The Hong Kong Polytechnic University, Hong Kong
[2]Singapore-MIT Alliance for Research and Technology, 1 Create Way, Singapore
[3]Faculty of Applied Mathematics and Computer Science,
Belarusian State University, Minsk, Belarus
[4]Senseable City Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA
[5]Istituto di Informatica e Telematica del CNR, Pisa, Italy
[*]Correspondence: yang.ls.xu@polyu.edu.hk; [+]Contributed equally

1. Distribution of housing units and prices

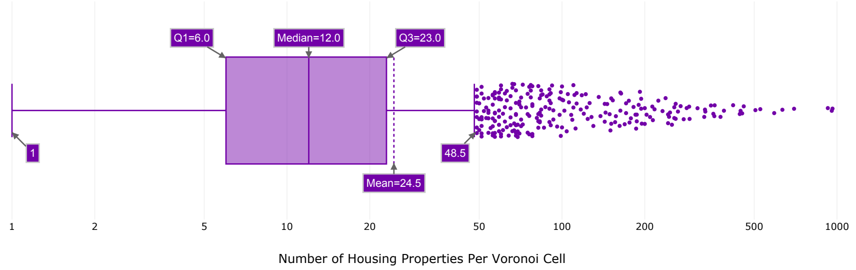

Figure S.1. Distribution of number of price values per tower service area.

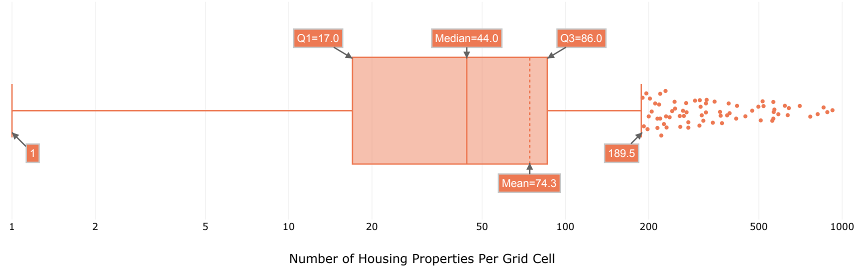

Figure S.2. Distribution of number of price values per 500 m grid cell.

In Figure S.1 and Figure S.2 we show the distribution of numbers of housing prices per Voronoi cell and per 500 m grid cell. Since most of the Voronoi cells in dense residential areas are much smaller than 500 m, by changing spatial resolution to the 500 m grid cells we increase the average number of prices per cell. This improves the quality of our estimate of the distribution of prices per cell and helps to assign more appropriate ranks to users.

2. Correlation between census and estimated populations

To evaluate whether our estimated home locations of each individual reflect the population distribution in Singapore we first calculate the total number of cellphone users with home location in each planning area and then compare these values with census data available from Department of Statistics Singapore, for the year 2010. As shown in Figure S.3, we find that the total number of cellphone users sampled in each planning area is strongly correlated with the population distribution recorded by the census data, with a Pearson's correlation of 97%.

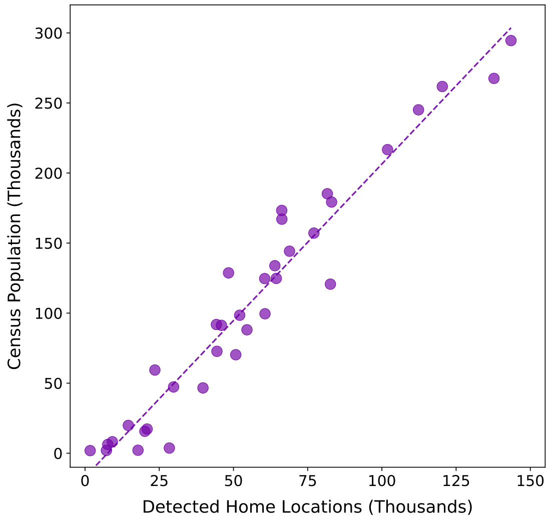

Figure S.3. Correlation between the number of detected home locations and census population (by planning area in Singapore) (Pearson's $R = 0.97$).

3. Correlation between housing price and income

To assess the feasibility of using housing price as an indicator of individuals' socioeconomic status (SES), we perform a correlation analysis at the level of Singapore's planning area by relating the housing price data and income data from the Household Interview Travel Survey (HITS). We first extract all the individuals from the HITS data who reported their monthly income (12,111 in total). We then aggregate individuals — based on the postal code of their reported residencies — by planning area, and calculate the average monthly income at each planning area. We then compute the average sale price of housing units in each planning area and correlate the two variables. We find that the average monthly income matches relatively well with the mean housing price (Figure S.4A) except for three outliers (Southern Islands, Sungei Kadut, and Novena). By further exploring the HITS dataset, we think this is partially caused by the sampling bias when individuals were selected for the travel survey. For example, only two individuals in Southern Islands were sampled from the 2011 HITS survey, and both them reported a monthly income of 500 SGD. However, Southern islands is well known as a planning area with many luxury housing communities. Housing price, in this sense, could be a more reasonable indicator of individual SES given the sparsity of HITS income data. Figure S.4B shows the relationship between the two variables after filtering these three outliers. The Pearson's R is 0.88, which suggests that it is reasonable to use housing price to approximate the SES of the underlying populations. Note, that in the main study we did not exclude residents of these tree outlying areas, as we believe that the bias comes mostly from the HITS data.

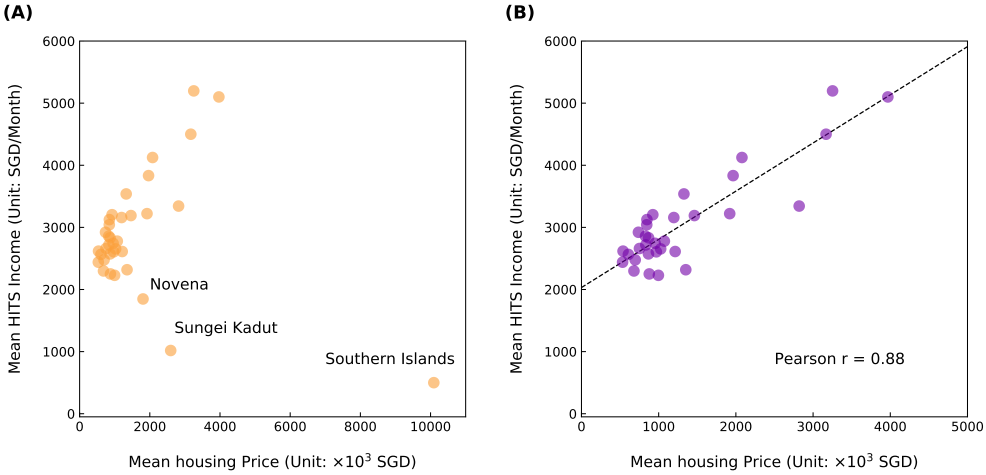

Figure S.4. The relationship between mean housing price and average monthly income **(A)** at the level of Singapore's planning area; **(B)** The correlation between the two variables after filtering the three outliers (Pearson's $R = 0.88$).

4. Random assignment of housing price

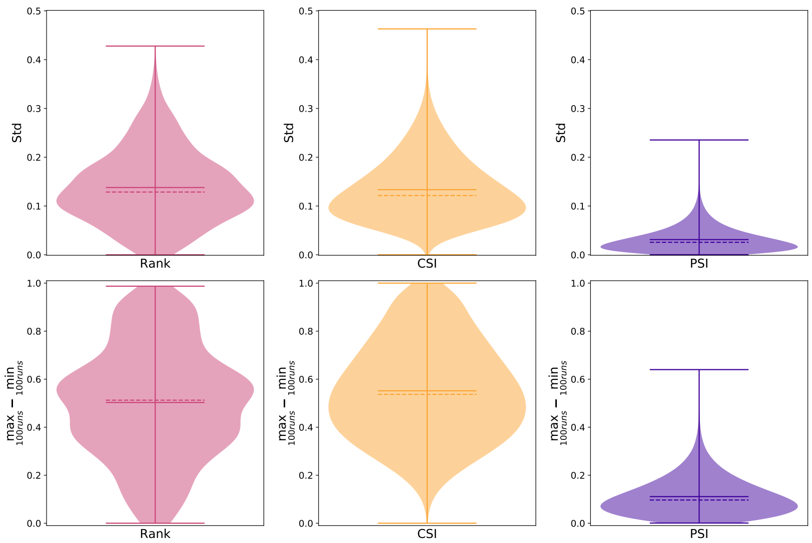

Figure S.5. Distribution of standard deviation and spread between min and max value of ranks, CSI and PSI of each user between 100 random assignments of housing prices.

In Figure S.5 we show how ranks, CSI and PSI change between assignments. One can see that for some users these values could change quite a lot. However, the overall standard deviations are relatively low, indicating that individuals tend to be assigned with similar SES values across the 100 assignments. Moreover, the overall distributions of CSI and PSI remain very similar from one assignment to another. This confirms that our conclusions about differences in segregation between classes are robust and do not depend on a particular way of assigning housing prices.

5. Segregation level of an individual under the assumption of a null model

Under the assumption of a *null model*, where each individual interacts equally with other individuals in a city (in the social or physical space), the segregation level of an individual can be quantified as the weighted sum of social similarity between him/her to all individuals, i.e., $\frac{1}{N}\cdot\sum_{j=1}^{N}s_{x\to j}$. We can prove that, in case of unique ranks $r_x = x$, this value is independent of the social rank x and is always 0.5. Given an individual's rank $x \leq \frac{N}{2}$, the weighted sum of social similarity can be calculated as:

$$\begin{aligned}
\frac{1}{N}\cdot\sum_{j=1}^{N}s_{x\to j} &= \frac{1}{N}\cdot\left(\sum_{j=1}^{x-1}s_{x\to j}+s_{x\to x}+\sum_{j=x+1}^{N}s_{x\to j}\right)\\
&= \frac{1}{N}\cdot\left(\sum_{j=1}^{x-1}s_{x\to j}+s_{x\to x}+\sum_{j=x+1}^{2x-1}s_{x\to j}+\sum_{j=2x}^{N}s_{x\to j}\right)\\
&= \frac{1}{N}\cdot\left(2*\sum_{j=1}^{x-1}s_{x\to j}+s_{x\to x}+\sum_{j=2x}^{N}s_{x\to j}\right)\\
&= \frac{1}{N}\cdot\left(\sum_{j=1}^{x-1}\frac{2N-4x-1}{N-1}+\sum_{j=1}^{x-1}\frac{4j}{N-1}+1+\sum_{j=2x}^{N}\frac{N}{N-1}-\sum_{j=2x}^{N}\frac{j}{N-1}\right)\\
&= \frac{1}{N}\cdot\frac{N^2-N}{2(N-1)}=\frac{1}{2}
\end{aligned}$$

We can prove the case similarly for $x > \frac{N}{2}$. The proof shows that this value is independent of social rank x.

So, if we assume that every person has equal chances to communicate with any other person, then:

$$\begin{aligned}
E(CSI_x) = E\frac{\sum_{j=1}^{N}f_j\cdot s_{x\to j}}{\sum_{j=1}^{N}f_j} &= \sum_{j=1}^{N}s_{x\to j}\cdot E\frac{f_j}{\sum_{i=1}^{N}f_i}\\
&= \sum_{j=1}^{N}s_{x\to j}\cdot\sum_{k=1}^{\sum_{i=1}^{N}f_i}\frac{1/N}{\sum_{i=1}^{N}f_i}=\frac{1}{N}\sum s_{x\to j}=\frac{1}{2}
\end{aligned}$$

And if at any place L and time period T all people have equal chances to be there, i.e., $prob_y(L,T) = prob(L,T)$, then:

$$PSI_x(L,T) = \frac{\sum_{y\in U(L,T)}prob_y(L,T)\cdot s_{x\to y}}{\sum_{y\in U(L,T)}prob_y(L,T)} = \frac{\sum_{y=1}^{N}prob(L,T)\cdot s_{x\to y}}{\sum_{y=1}^{N}prob(L,T)} = \frac{1}{N}\sum s_{x\to j} = \frac{1}{2}$$

and then

$$PSI_x(T) = \sum_{L\in Loc_x}prob_x(L,T)\cdot PSI_x(L,T) = \frac{1}{2}\sum_{L\in Loc_x}prob_x(L,T) = \frac{1}{2}$$

and

$$PSI(L,T) = \frac{\sum_{\{x|L\in Loc_x\}}prob_x(L,T)\cdot PSI_x(L,T)}{\sum_{\{x|L\in Loc_x\}}prob_x(L,T)} = \frac{1}{2}$$

6. CSI calculated including individuals living in the same cell

In Figure S.6 we present a distribution of CSI values calculated based on all users′ communications without filtering calls between users living in the same cell. One can see that distributions presented in Figure 2 (in the main text) and Figure S.6 are very similar. A slight and expected difference is that the histogram is shifted to the right with the mean equal to 0.576 (vs. 0.546 with filtering) while having the same standard deviation of 0.200. This confirms the expectation that a substantial proportion of all users communications consists of calls to the relatives and other people living together.

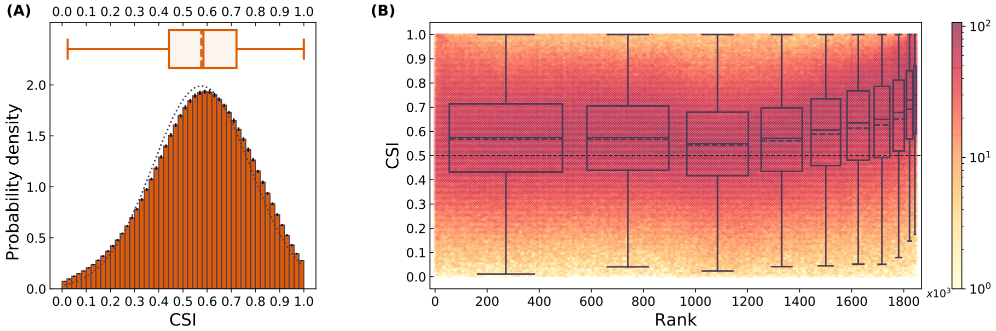

Figure S.6. Distribution of CSI values calculated without filtering calls between people living together.

7. CSI based on absolute rank difference

We understand that some readers might find our measure of social distance confusing at first. A more intuitive way of defining it could be just to take the absolute difference between people's ranks. In this case CSI could be defined by the same formula 2 from the main text, but using social distance defined as $d_{x\to j} = \frac{|r_x - r_j|}{N-1}$. This measure would totally make sense, but would lack a nice property of having the same baseline value for people from different classes. To see this, consider a person right in the middle of hierarchy, with rank $N/2$. For this person, her social distance to any other person range from 0 to $1/2$, i.e. social similarity range from $1/2$ to 1, and then average of these values will be $3/4$. While for the person with rank 1, her baseline CSI will stay 0.5.

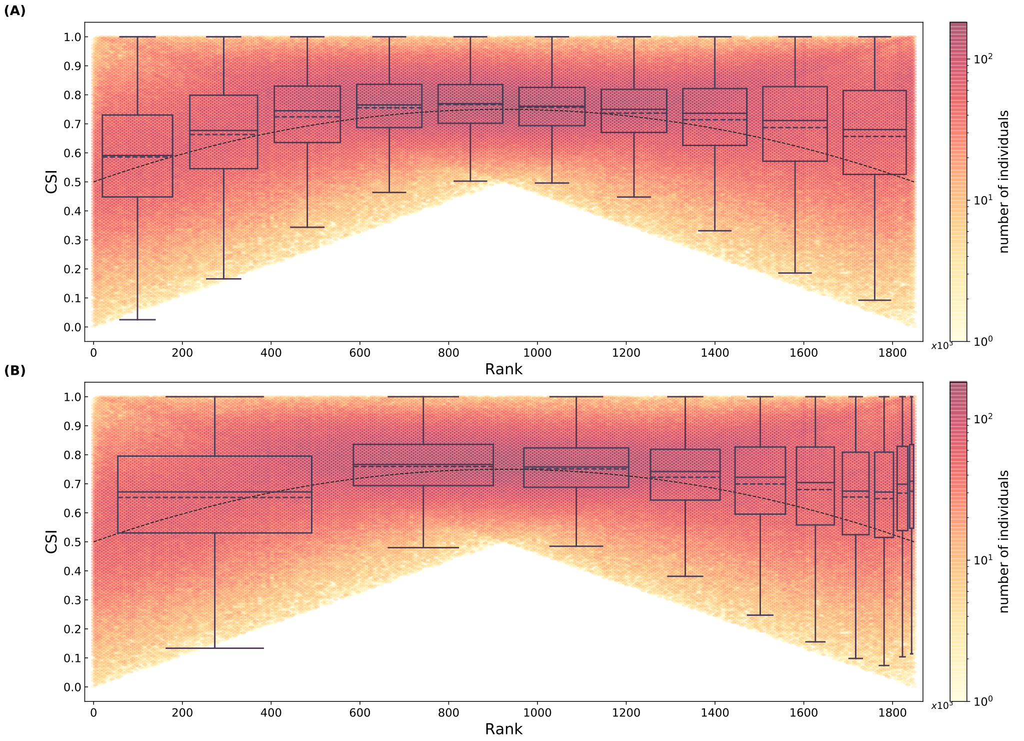

Figure S.7. Density of CSI based on absolute rank difference. We plot a line showing baseline CSI as well as box plots representing distribution of CSI values for **(A)** 10 equal-sized groups; **(B)** 10 groups with equal total housing price.

In Figure S.7, we show density plot similar to one from Figure 2B in the main text, but for this version of CSI. As it could be seen from the figure, just by raw CSI values it is hard to make any conclusions about the general level of social segregation and regarding different classes. To be able to make such conclusions one needs to compare these values with expected baseline values, outlining the importance of the social segregation metrics introduced in this paper.

In Figure S.8, we compare median values of original CSI, defined in the main text, with normalized CSI defined in this section. To normalize the values, we can either subtract baseline values from CSI values, or divide CSI values by baseline. We present both version in Figure S.8. From this figure, one could see that distribution of values for both versions of CSI are fairly similar. This indicates that these values represent qualitatively the same thing – level of segregation of a group of people. But CSI measure proposed in the main text does not require any normalization, and its values could be easily interpreted.

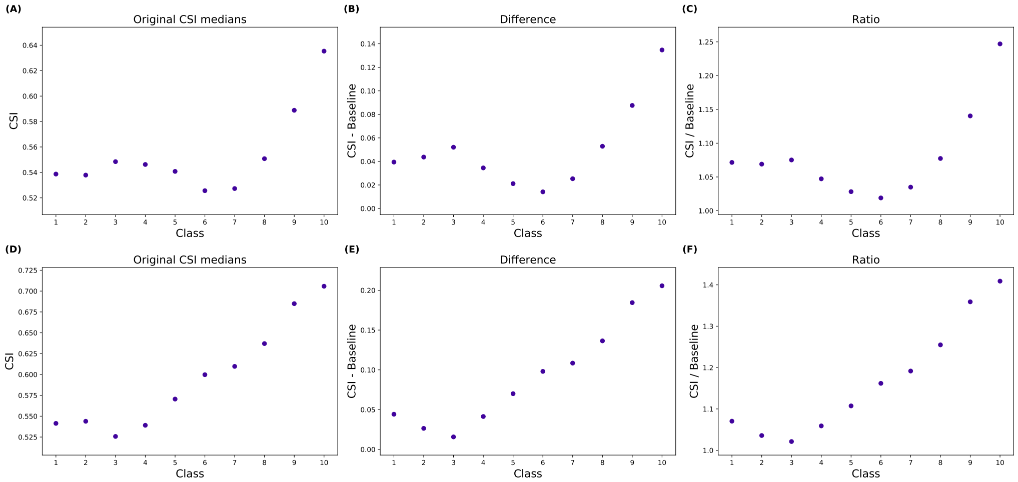

Figure S.8. Median CSI values of 10 classes. In the first row people split into equal-sized classes, in the second row – by equal cumulative housing price. **(A, D)** CSI measure proposed in this work; **(B, E)** Difference between median CSI based on absolute rank difference and median baseline; **(C, F)** Ratio between median CSI based on absolute rank difference and median baseline.

8. Measure segregation of the city using Theil's entropy index

The Theil's entropy index is frequently used in previous studies to quantify the residential unevenness of a city [1]. It measures the departure of the entropy of each spatial unit (e.g., census tract) — which is determined, for example, by the ethic/racial composition at that place — from the racial or ethnic entropy of the whole city. In this research, we apply the Theil's entropy index to quantify the unevenness of interactions across different social classes (e.g., income groups). The index is calculated as follows:

$$H = \sum_{i=1}^{n} \frac{S_i(E - E_i)}{EN}$$

where N denotes the total number of phone users in the city; n refers to the total number of spatial units; S_i and E_i stand for the expected number of phone users and the corresponding Shannon entropy of spatial unit i, respectively. E is the overall Shannon entropy of the city:

$$E = \sum_{m=1}^{M} p_m \cdot \log \frac{1}{p_m}$$

Here M denotes the total number of social classes that is predefined in the study, for example, income groups or classes derived from the housing price of phone users' residential locations. p_m stands for the proportion of class m users in the city. Similarly, E_i is calculated as:

$$E_i = \sum_{m=1}^{M} p_{i,m} \cdot \log \frac{1}{p_{i,m}}$$

where $p_{i,m}$ denotes the proportion of class m users in spatial unit i.

Unlike traditional measure of residential segregation which associates individuals to fixed locations (home), in this research, we aim to quantify the interactions of phone users across different classes, and examine how the level of segregation in a city changes over time. To take human movements into account, this research starts by first dividing a day into several time windows (e.g., 24 one-hour time windows). For a given time window T, we can estimate, for each phone user, the probability of stay at different spatial units (i.e., grid cells). Thus, the value of $p_{i,m}$, for a specific time window T, can be calculated as the proportion of class m (i.e., C_m) users at location i:

$$p_{i,m} = \frac{1}{S_i} \sum_{x \in C_m} prob_x(i)$$

where $prob_x(i)$ denotes the stay probability of phone user x' at location i. Note that:

$$S_i = \sum_{m=1}^{M} \sum_{x \in C_m} prob_x(i)$$

The Theil entropy index ranges from zero to one. A value of zero indicates that all the spatial units have the same entropy that is equal to the value of the whole city. A value of one indicates that each spatial unit only hosts one particular class, which results into an entropy value of zero. There are several important considerations when the phone user pool is divided into different social classes. First, we need to determine the number of classes (M). Second, we need to specify the criterion for phone user classification. Here, we use two ways to divide phone users into M classes:

- *Quantile*: each class includes the same number of mobile phone users.
- *Total Buying Power*: each class has the same amount of buying power (e.g., total housing price). In this research, the total buying power for each class is calculated as the sum of housing price value of all the phone users in that class.

Regarding the number of classes M, we test different values and compare the results.

Here we present results of analysis similar to the one for PSI presented in the main text, again just for one random assignment of housing prices. By distinguishing weekdays and weekends, we divide each type of day into 24 one-hour time windows. We then calculate Theil's entropy index for each time window. Figure S.9 shows the Theil's entropy indices of Singapore based on various combination of M and classification schemes (i.e., *Quantile* or *Total Buying Power*). We can see that different parameter sets produce very similar results on the overall level of segregation of the city as well as its diurnal patterns. During the day time, the city is more segregated on weekends than on weekdays.

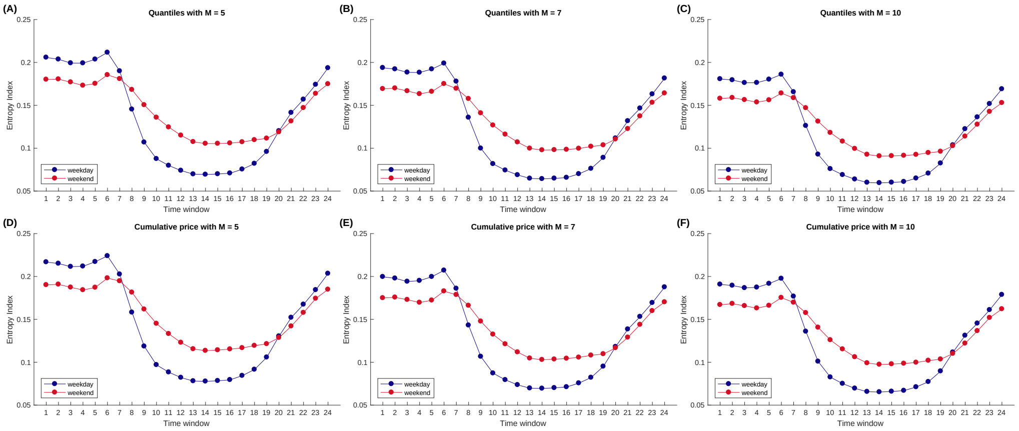

Figure S.9. Theil's entropy index and its temporal evolution over time. Results are generated based on: **(A)** Quantile classification with $M = 5$; **(B)** Quantile classification with $M = 7$; **(C)** Quantile classification with $M = 10$; **(D)** Buying Power classification with $M = 5$; **(E)** Buying Power classification with $M = 7$; **(F)** Buying Power classification with $M = 10$. The x-axis denote time windows, and y-axis denotes value of Theil's entropy index.

Figure S.10 shows values of 1.0 minus the normalized entropy value of grid cells for time windows *12AM – 1AM, 12PM – 1PM, 6PM – 7PM* on weekdays and weekends. Places with a high value of entropy (and respectively low value of 1 minus entropy) correspond to socially-mixed areas, while those with a low entropy (high bars) refer to those with higher levels of segregation. Hence, the method can be used to quantify: (1) the overall level of segregation in a city, and (2) the spatial heterogeneity.

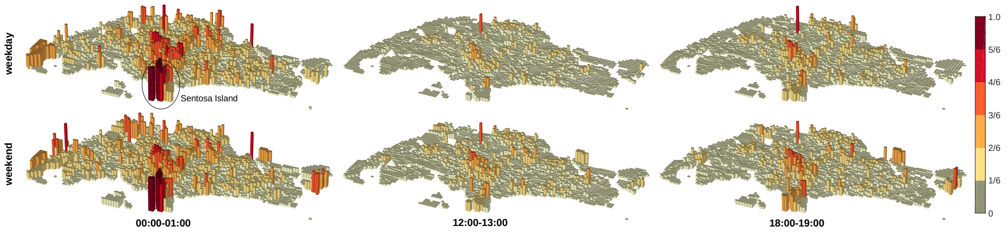

Figure S.10. $1 - \frac{E_i}{\log M}$ representing segregation of each grid cell at selected time windows (result based on quantile classification with $M = 10$)

References

1. Theil H. 1972 *Statistical decomposition analysis*. North-Holland Publishing Company Amsterdam.