SUPPLEMENTARY INFORMATION
Cities through the prism of people spending behavior

Stanislav Sobolevsky[1,*], Izabela Sitko[2], Remi Tachet[1], Bartosz Hawelka[2], Juan Murillo Arias[3], Carlo Ratti[2]

1 Senseable City Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

2 Department of Geoinformatics - Z_GIS, University of Salzburg, Salzburg, Austria

3 New Technologies, BBVA, Madrid, Spain

∗ E-mail: Corresponding stanly@mit.edu

Demographic normalization

Let us describe the normalization procedure used to account for demographic discrepancies between cities. The idea is to compare the observed value of a given parameter with its theoretical expected one (computed using the city demographic profile). Let $(p_c)_{c\in C}$ be the measured parameter where C denotes the entire set of customers, C_X the subset containing only customers from city X and $C_{g,a}$ the customers of gender g and age a. The average quantity for a given gender g and age a and for a given city X are

$$Q_{g,a} = \frac{\sum\limits_{c\in C_{g,a}} p_c}{|C_{g,a}|} \qquad Q_X = \frac{\sum\limits_{c\in C_X} p_c}{|C_X|}$$

The expected value of the parameter based on the demography of city X is

$$E_X = \frac{\sum\limits_{g,a} |C_X \cap C_{g,a}| Q_{g,a}}{|C_X|}$$

In the end, the normalized value used as a measure of city X economic behavior is $\frac{Q_X}{E_X}$

Scaling of all five characteristics of individual spending behavior with city size

Parameter	City definition	Exponent	Confidence intervals	p-value	R^2
Activity	LUZ	1.052	$[1.0, 1.1]$	4.5%	17%
	FUA	1.044	$[1.0, 1.08]$	3.2%	11.5%
	CON	1.048	$[1.03, 1.06]$	$3e$-9%	15.6%
Avg. amount	LUZ	−0.007	$[-0.05, 0.03]$	71.2%	0.6%
	FUA	0.002	$[-0.03, 0.04]$	87.9%	$6e$-4%
	CON	0.008	$[0.07, 0.02]$	28.6%	0.5%
Diversity	LUZ	0.033	$[0.0, 0.064]$	4.2%	17.5%
	FUA	0.035	$[0.01, 0.06]$	0.49%	19.0%
	CON	0.051	$[0.04, 0.06]$	$2e$-15%	26.1%
Distant mob.	LUZ	−0.06	$[-0.24, 0.11]$	45.8%	2.5%
	FUA	0.035	$[-0.1, 0.16]$	60%	0.7%
	CON	0.158	$[0.11, 0.20]$	$6e$-11%	18.65%
Local mob.	LUZ	−0.10	$[-0.24, 0.04]$	15.2%	9.1%
	FUA	−0.073	$[-0.17, 0.03]$	16.1%	5.1%
	CON	−0.031	$[-0.07, 0.01]$	15.4%	0.7%

Table 1. Scaling of customers spending behavior characteristics with city size for different city definitions.

As mentioned in the main text, the spending diversity exhibits a small but consistent scaling with city size for the three definitions considered: the graphs can be found in Figure 1.

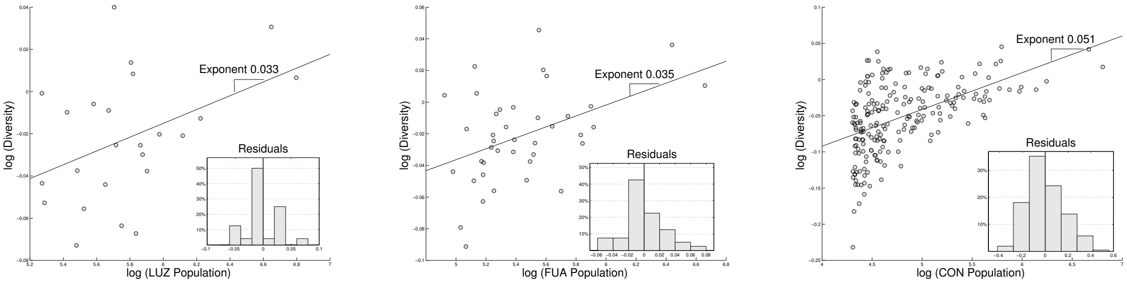

Figure 1. Scaling of diversity with LUZ, FUA and CON (left to right) size. LUZ Exponent: 3.3%, CI: [0.0,0.064], p-value: 4.2%, $R^2 = 17.5\%$. FUA Exponent: 3.45%, CI: [0.01,0.06], p-value: 0.49%, $R^2 = 19.0\%$. CON Exponent: 5.1%, CI: [0.04,0.06], p-value: $2e$-15%, $R^2 = 26.1\%$.

Total amount of transactions

As a supplement to the graphs from the main text, Figure 2 shows the total amount of money spent during 2011 for both genders against age.

Pairwise visualization of correlations in urban characteristics feature space

Figure 3 shows the existing correlations in the urban characteristics feature space for all 3 levels of city definition. Most of the characteristics appear to be independent from one another. As expected relatively strong correlation is observed between spending activity and diversity and also weaker ones between activity and average purchase as well as between the average purchase and distant mobility bot only in case of LUZs and FUAs.

Validation of the clustering schemes

Similarity between the signatures of cities within each definition level, i.e. CONs, FUAs and LUZs, was assessed with the k-means clustering algorithm [1]. In order to select the optimal number of clusters, we validated different approaches with the silhouette metric *s(i)* [2], with $k \in \{2{:}10\}$. Silhouette aims to reflect how well each object fits to its cluster based on the comparison of an object dissimilarity (in our case Euclidean distance) to the points grouped in the same cluster and to the points grouped within the next best fitting cluster. It is computed according to the equation:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

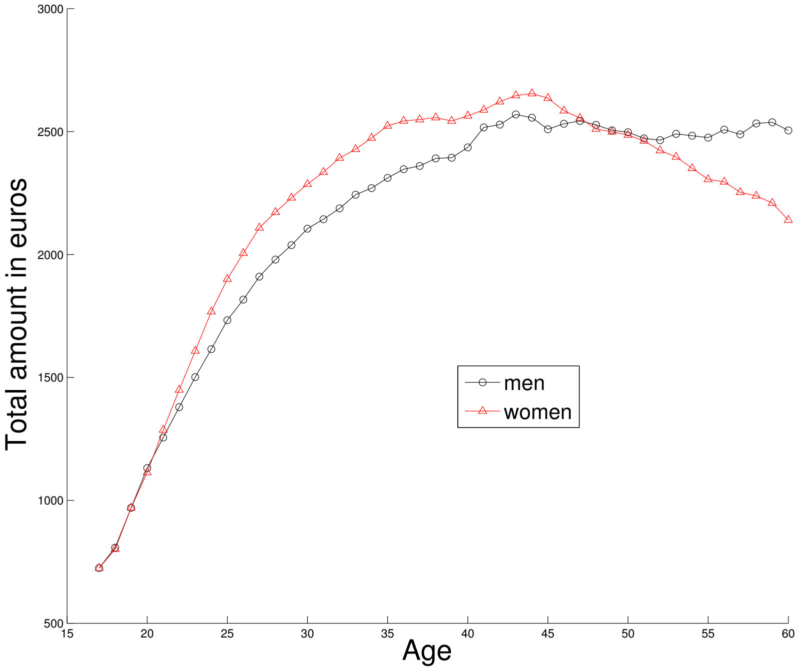

Figure 2. Impact of age and gender on the total amount of money spent every year

where *a(i)* is the average dissimilarity of a city *i* to the other cities assigned to the same cluster, and *b(i)* is the average dissimilarity of a city *i* to its next best fitting cluster. Silhouette varies in the range of {-1:1}. Positive values indicate a good match with the own cluster (small *a(i)*) and a bad match with the neighboring cluster (high *b(i)*). On the contrary, negative values indicate that a data points is more similar to the neighboring cluster, while values around 0 imply that a point is on the edge of two clusters. In our case, we compared the values of *s(i)* across the clustering schemes with different *k*, assuming that the highest values indicate the optimal split of cities. Received values are presented in the Figure 4. We observed that the silhouette metric peaked at k = 3 for the levels of CONs and FUAs, and k = 2 for LUZs. For the sake of consistency, we selected three-clusters approach for all the levels. Additionally, we validated selected algorithm, i.e. k-means, against its common variation, that is k-medoids [3]. As the latter one resulted in lower silhouette values for all tested k, we retained the k-means approach.

References

1. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1: Statistics: 281–297.

2. Rousseeuw P (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20: 53–65.

3. Kaufman L, Rousseeuw P (1987) Clustering by means of medoids. In: Y D, editor, Statistical Data Analysis Based on the L1 Norm and Related Methods, North-Holland. pp. 405–416.

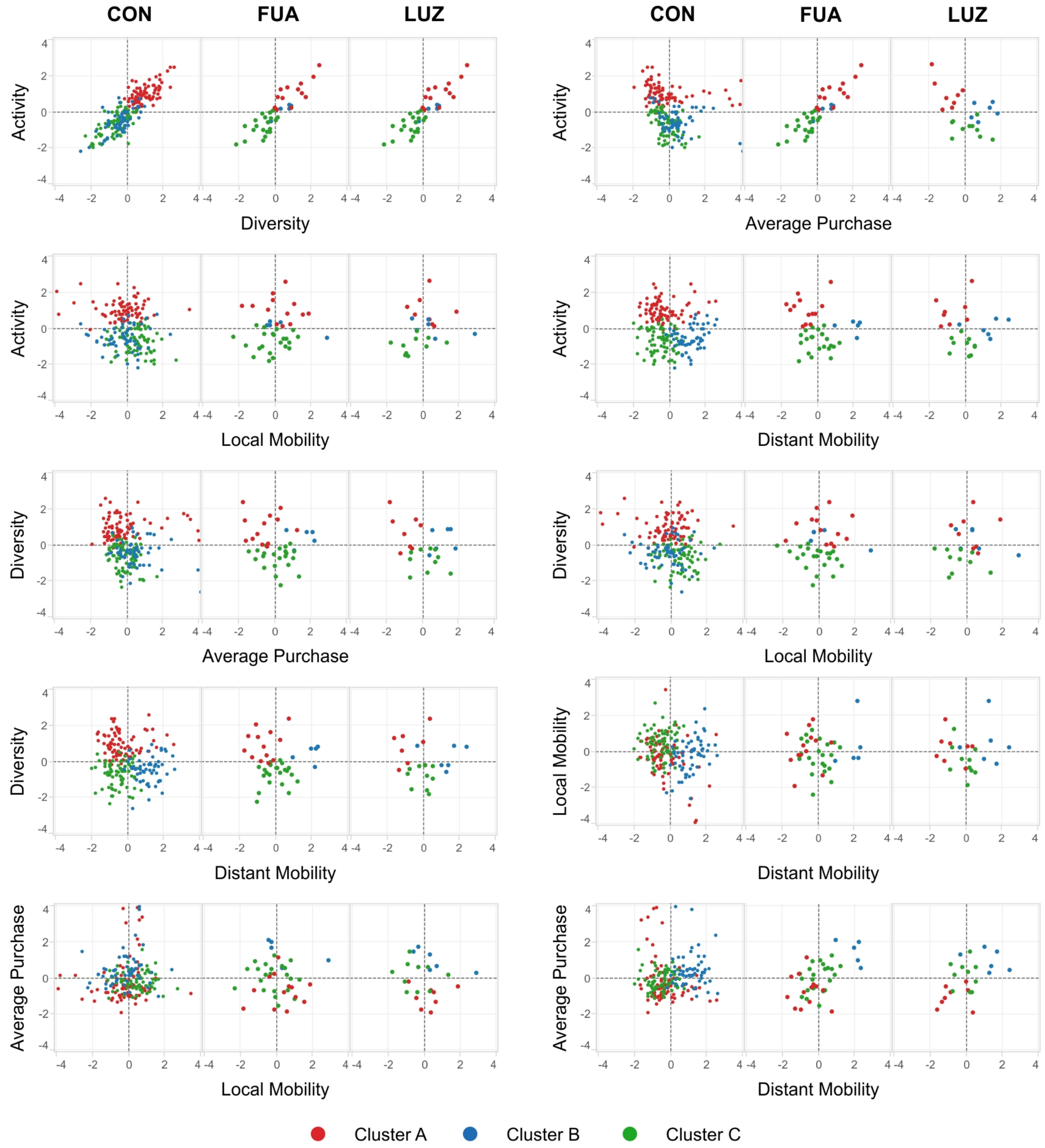

Figure 3. Pairwise correlations in the feature space of five urban characteristics for the three levels of city definitions.

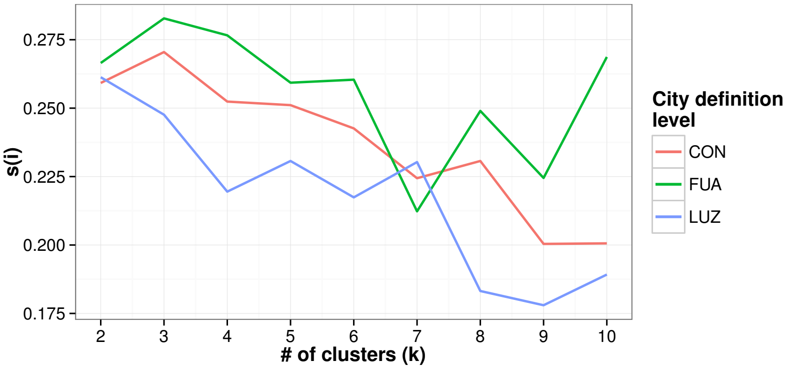

Figure 4. Values of the silhouette metric for clustering schemes with k varying from 2 to 10. Higher values indicate a better fit of data points to the clusters they were assigned (more appropriate clustering approach).