Supplementary Materials for *Gender gaps in urban mobility*

Laetitia Gauvin[1,+,*], Michele Tizzoni[1,+], Simone Piaggesi[1,2], Andrew Young[3], Natalia Adler[4], Stefaan Verhulst[3], Leo Ferres[5,6], and Ciro Cattuto[1,7]

[1]ISI Foundation, Turin, 10144, Italy
[2]Doctoral School in Data Science and Computation, University of Bologna, Bologna, Italy
[3]The Governance Lab, New York University, New York, NY, United States
[4]United Nations International Children's Emergency Fund (UNICEF), New York, NY, United States
[5]Data Science Institute, Universidad del Desarrollo, Santiago, Chile
[6]Telefónica R&D, Santiago, Chile
[7]University of Torino, Italy
[*]laetitia.gauvin@isi.it

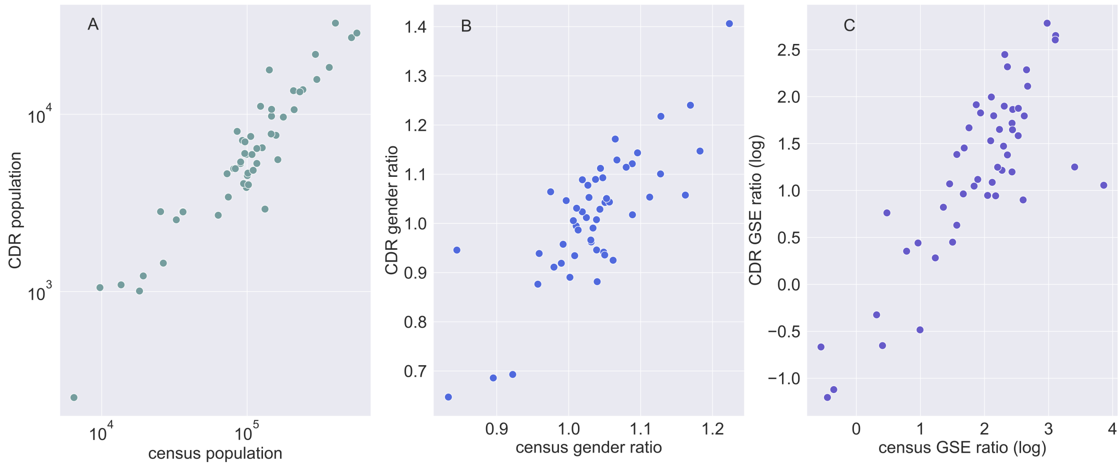

Figure S1: a) Scatter plot of the gender ratio measured on the CDR data by municipality with respect to the census gender ratio. b) Scatter plot of the ratio of the poorest population (according to the GSE ratio) over the weathiest population by municipality measured computed from the survey with respect to the one measured using the metadata associated to the CDR data. The CDR population sample is highly representative of the population distribution as confirmed by the high correlations of 0.93, 0.79, 0.88 measured between the CDR data and the baseline (census or survey) respectively for the population, the gender ratio and the GSE ratio. We remind that the GSE population ratio corresponds to the ratio of the population having an income below a reference income $P(I < I^*)$ over the population having an income below this reference income $P(I > I^*)$. The repartition of the socio-economic groups in each *comunas* has been extracted using the survey *Encuesta Origen Destino de Viajes* (EOD) realized by the *Secretaría de Planificación de Transporte* (SECTRA) between July 2012 and November 2013, by assignment of Chilean Ministry of Transport and Telecommunications. The dataset contains the income of 60,054 individuals (47% males, 53% females). Socio-economic groups were assigned to each household of the survey based on their income using the definition of the socio-economic groups as given by http://www.emol.com/noticias/Economia/2016/04/02/796036/Como-se-clasifican-los-grupos-socioeconomicos-en-Chile.html

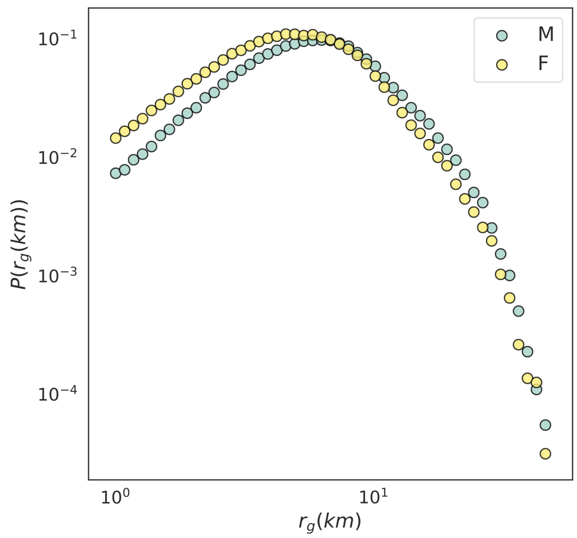

Figure S2: Distribution of the radius of gyration (r_g) disaggregated by gender. The r_g distribution spans a similar range of distances for women and men, however women tend to have a smaller radius of gyration. Average values are: $\langle r_g \rangle_M = 6.68$ Km, $\langle r_g \rangle_F = 5.59$ Km.

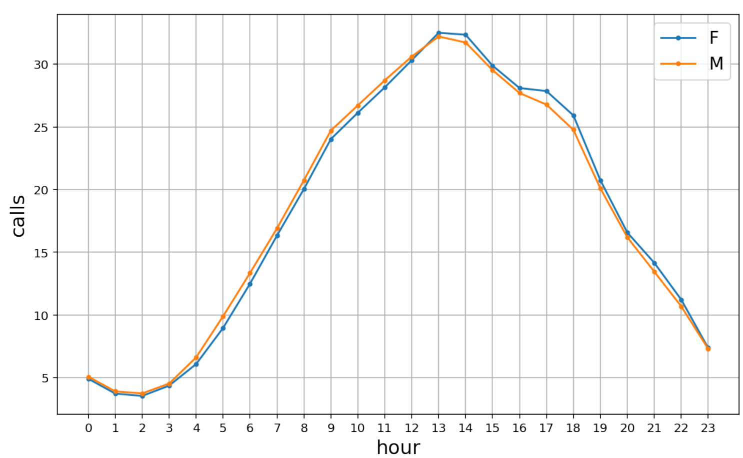

Figure S3: Distribution of the average number of calls made or received by the users during the day, disaggregated by gender. The average is computed hourly over the three month period of our study, and then normalized by the total average calling activity for males and females respectively.

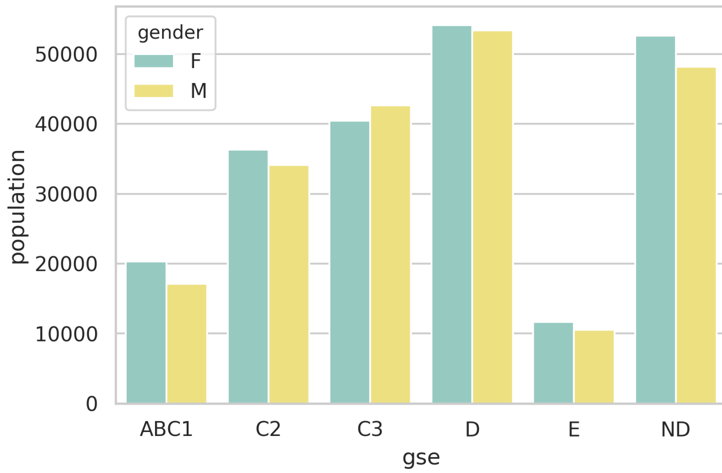

Figure S4: Distribution of the CDR users by GSE and by gender. There are 5 GSE classes from the highest to the lowest level: ABC1, C2, C3, D, and E. Users in the class ND have no GSE assigned. Overall, we observe that both genders are generally equally represented across socio-economic segments.

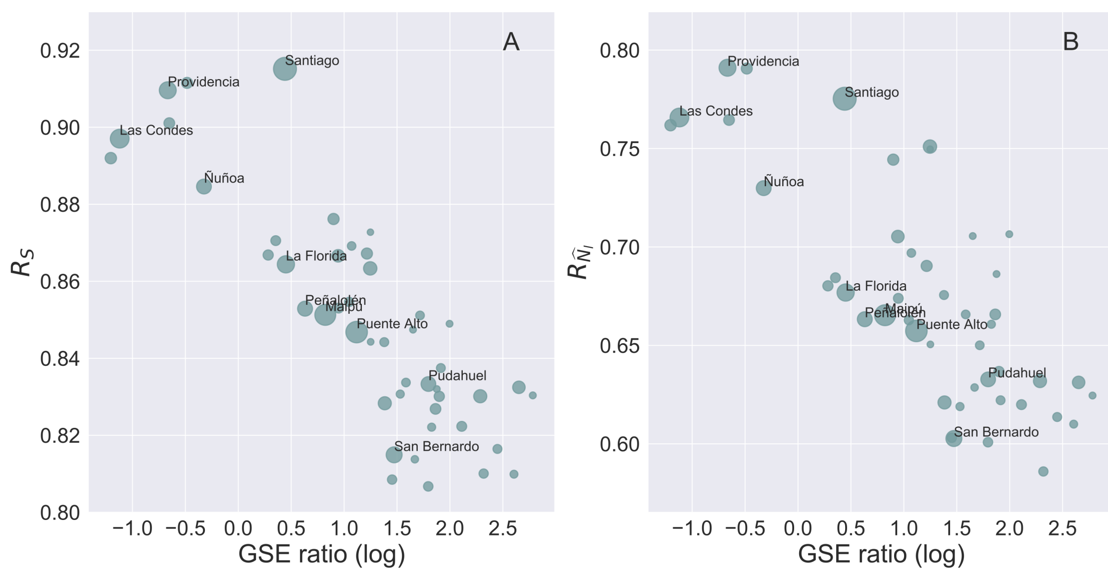

Figure S5: Scatterplots of the mobility gap against the GSE ratio by Comuna of the Santiago Metropolitan area. The mobility gap is measured by the entropy ratio (**A**) and the women to men ratio of the number of visited locations that correspond to at least 80% of the activity of each user (**B**).The size of the points is proportional to the population of the *comunas*.

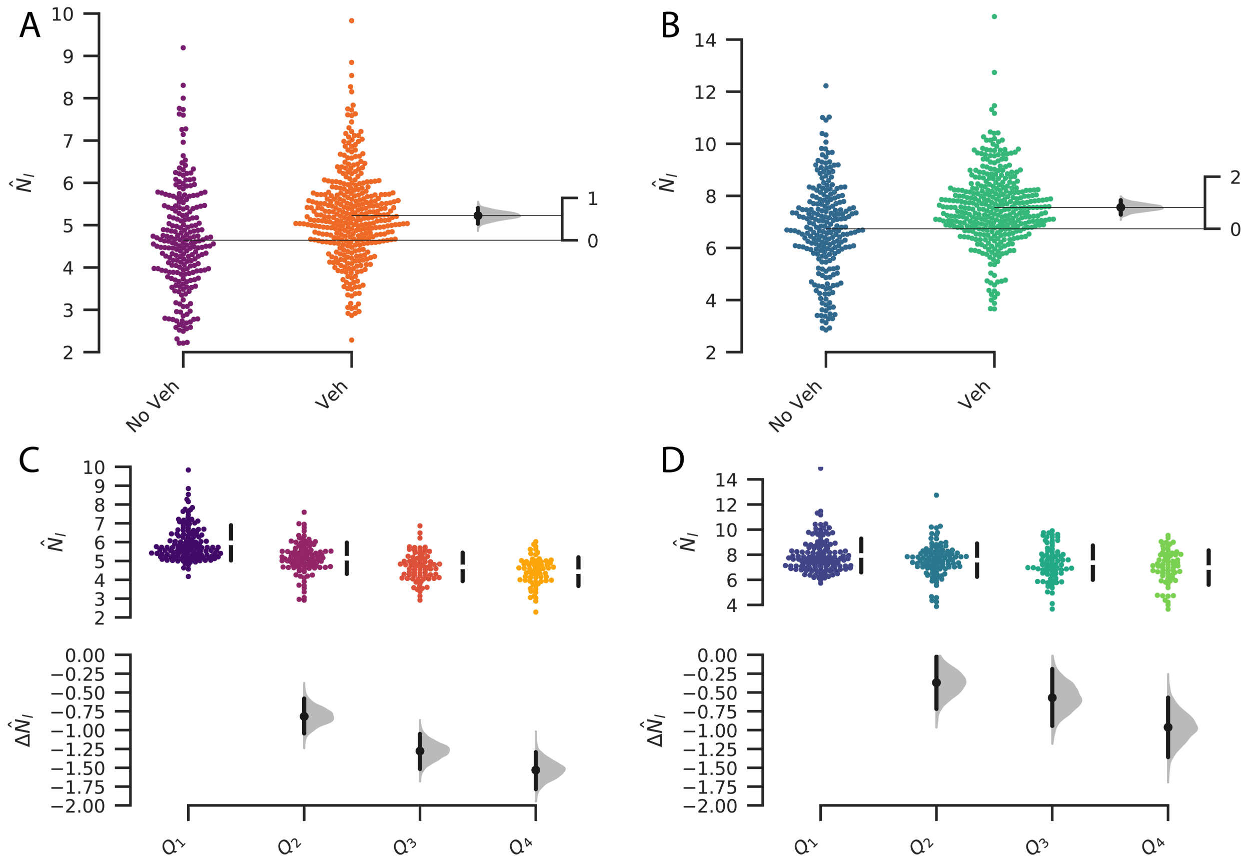

Figure S6: Estimation plots of the average difference in number of locations visited by women and men by cells classified according to access to private transportation and socio-economic groups. Estimation plots of the difference in number of locations visited by women (**A**) and men (**B**) under the conditions of living in cells where the average number of cars is below (No Veh) or above (Veh) 1 over 4 per person**.** Estimation plots of the difference in number of locations visited by women (**C**) and men (**D**) cells where the average number of cars isabove1 per 4 person by quartile of GSE ratio. Each group represents cells in decreasing quartiles of income, from the top 25% to the bottom 25%. As for public transportation, we observe an increase in mobility both for women and mean when their probability of having access to a private vehicle is higher.If we focus on cells where having access to car is more likely (i.e. where the average number of cars is above the threshold we introduced), we notice a smaller difference in the number of locations visited by the residents living in cells belonging to the different quartiles of GSE ratio for men than women. Overall the access to transportation either public or private is linked to higher mobility but it does not fill the gender gap observed for the poorest areas of Santiago.

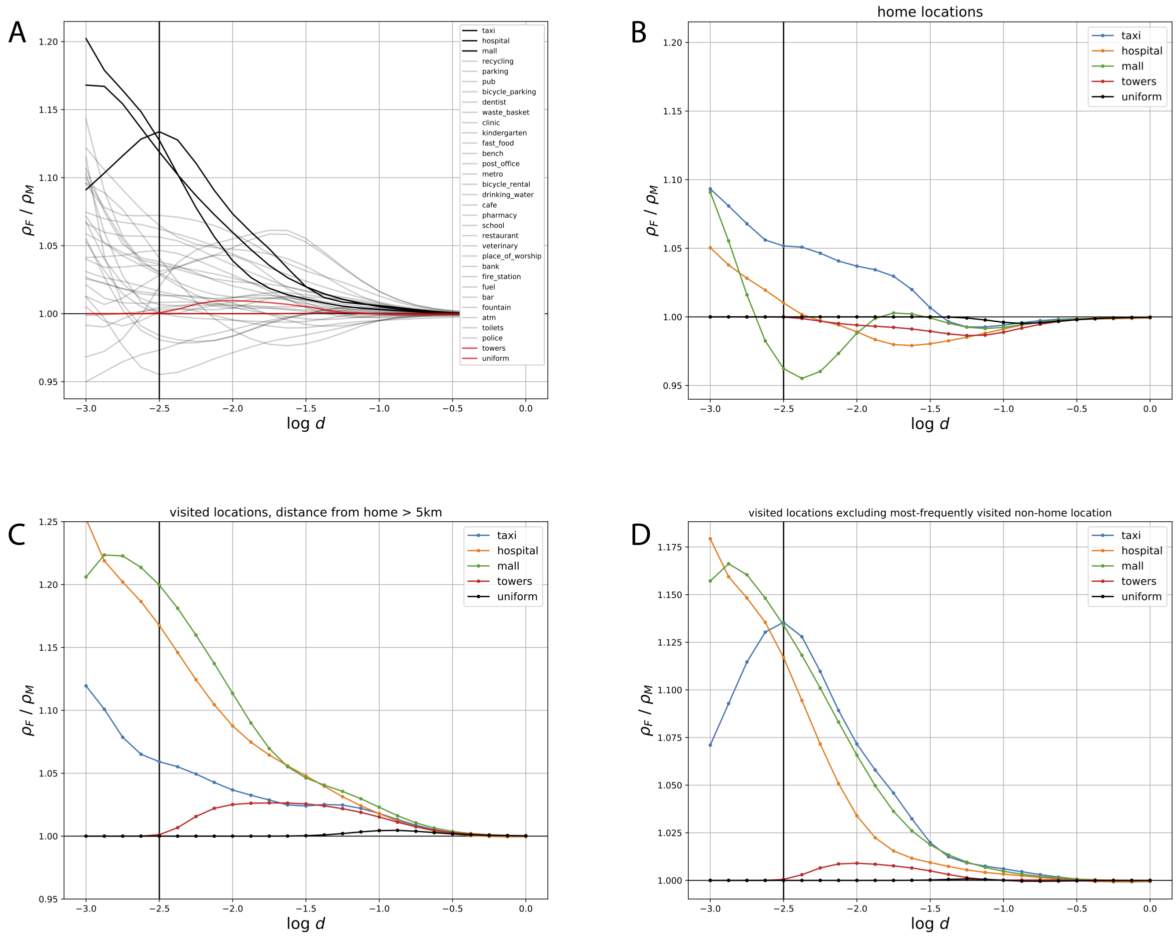

Figure S7: Gender differences in visit patterns. Gender ratio $r_k = \rho_F/\rho_M$ as a function of the bandwidth d (decimal logarithm of the value in decimal degree). The vertical line corresponds to a kernel bandwidth comparable to the spatial resolution afforded by the CDR data we use. (**A**) r_k is shown for all POI types considered, including the "towers" and "uniform" reference layers. The ratio for the three POI types showing the largest gender imbalance at $log(d) = -2.5$ ("taxi", "hospital", and "mall") is highlighted in black. (**B**) The gender ratio r_k is computed by considering only home locations as the set of visited locations of each user. The gender ratio drops significantly close to or below 1 for all the three POIs ("taxi", "hospital", and "mall"). This means that the gender imbalance observed in the main analysis for all non-home locations is larger than we could expect if only based on the differences in home locations by gender. (**C**) The gender ratio is computed by excluding from the set of visited locations of each user all locations that fall within a 5 km distance from home. (**D**) The gender ratio is computed by excluding from the set of visited locations of each user both the first the second most frequently visited locations (home and work).

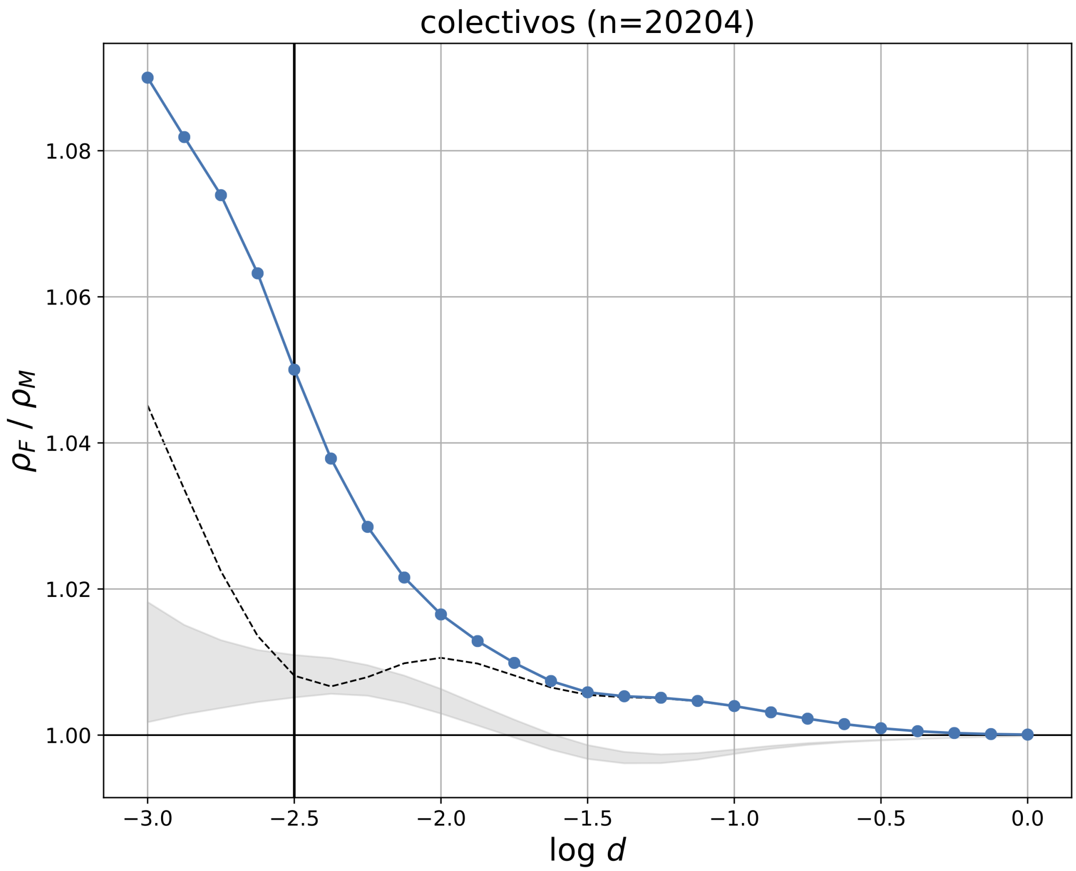

Figure S8: Gender differences in visit patterns: collective taxis. Gender ratio ρ_F/ρ_M as a function of the bandwidth d (decimal logarithm of the value in decimal degree), for the "colectivos" (collective taxis) POI type. This POI layer includes locations along the routes of shared taxis in Santiago ("colectivos"), obtained by scraping the geographical information used to draw taxi routes on the Web site (www.ubicatucolectivo.cl/cliente_final/all_lines/vercion_1.php?id=14). The vertical line corresponds to a kernel bandwidth comparable to the spatial resolution afforded by the CDR data we use. The shaded area indicates, for each value of the bandwidth d, the 95% confidence interval of the values ρ_F/ρ_M computed on the realizations of the spatial sampling null model. The dashed line indicates the gender ratio computed on the spatially perturbed null model.

Table S1: Descriptive statistics of mobility metrics, S, r_g, N_l, $\hat{N}_l$ by gender. Each column reports the average, $\langle\dot{}\rangle$, and the 95% reference range (2.5 and 97.5 percentile points in square brackets) of the distributions.

sex	S	r_g	N_l	$\hat{N}_l$
F	1.647 [1.644, 1.65]	5.59 [5.58, 5.61]	28.48 [28.37, 28.59]	4.79 [4.77, 4.81]
M	1.908 [1.905, 1,911]	6.68 [6.66, 6.70]	37.05 [36.92, 37.18]	6.92 [6.89, 6.95]

Table S2: Pearson correlation values between R_S and $R_{\hat{N}_l}$ and the sociodemographic features of 51 municipalities in the SMR. Statistical significance is assessed by comparing the observed correlation value against 1,000 realizations of a randomized null model where each user is assigned every time to a random municipality, preserving the original population ratio females/males in each municipality. P-values are then computed as one-tailed tests on the probability distribution of correlation values generated by the null model. We report statistical significance using the following notation: $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$.

	R_S	$R_{\hat{N}_l}$
GSE ratio (log)	-0.77***	-0.70***
HDI	0.61***	0.50***
higher education gender ratio	-0.14	0.01
secondary education gender ratio	0.27	0.44*
primary education gender ratio	0.24	0.09
employment gender ratio	0.60***	0.37***
general fertility rate	-0.59***	-0.40***
couples household	0.63***	0.60***
extended household	-0.71***	-0.71***
family household	-0.20	-0.03
single parent household	-0.24	-0.35
single person household	0.55***	0.47***

Table S3: Spatial correlations between POI types. For each pair of POI layers, A (rows) and B (columns), the spatial correlation is obtained by first computing the kernel density estimator (Eq. 2 of the main text) on layer A with a bandwidth $d = 10^{-2.5}$ kms and then using it to estimate the average likelihood of any given point of layer B, $\langle \mathcal{L}(B) \rangle$. In other terms, we measure the spatial correlation by measuring how much layer A is predictive of any given point of layer B. Values reported in the table correspond to the ratio of the average likelihoods $\langle \mathcal{L}(B) \rangle / \langle \mathcal{L}(A) \rangle$.

POI type	hospital	mall	taxi	towers	uniform
hospital	**1**	0.23	0.15	0.06	0.02
mall	0.07	**1**	0.10	0.03	0.01
taxi	0.03	0.06	**1**	0.02	0.0004
towers	0.35	0.60	0.46	**1**	0.27
uniform	0.42	0.62	0.47	**1**	**1**

Text A: Controlling for differences in call activity by gender

We investigated the relationship between socio-demographic features and the gender differences in mobility by computing the Pearson semi-partial correlation coefficient between any two types of variables, controlling both for call activity and population size. For completeness, we also performed a sensitivity analysis of our results with respect to two types of sampling of the users' call activity distribution. First, we down sampled the call activity of men by randomly removing 25% of calls (resp. 50%) for each male user. Then, we computed the average Nemenman-Shafee-Bialek (NSB) entropy estimator [44] for women, $NSB_F = 1.7$ and men $NSB_M = 1.87$ (resp $NSB_F = 1.7$ and $NSB_M = 1.94$). We see that even after a drastic down-sampling of men's call activity, the entropy of women is still smaller on average.

Second, we selected only users who made at least 200 calls during the three month time period of our study. This corresponds to restricting our sample to 72% of the total number of users. Then, we selected 200 calls at random for each user and recompute the discrete entropy estimator so to compare users with the same total activity. In this case, the measured entropy estimators are $NSB_F = 1.73$ and $NSB_M = 2.01$, with women still displaying a smaller average entropy. In both cases, we observe a difference between the male and female sample that is statistically significant according to a Kruskal-Wallis test ($p < 0.001$).

Text B: Temporal stability of the entropy

To test the robustness of our results with respect to the time period under consideration, we measured the temporal stability of users' entropy S with a $k-$fold approach. To this aim, we divided the time window of our study into $k = 10$ folds. For each user, we computed the entropy averaged over the $k - 1$ folds obtained by excluding fold k. Thus, we obtain a vector V_1 containing the entropy of each user averaged over $k - 1$ folds. Then, we create a second vector V_2 with users' entropy measured on the $k - th$ fold. The average Pearson correlation between V_1 and V_2, for all users, is $r = 0.81 \pm 0.01$ indicating that entropy is a stable metric over time.

Text C: Sensitivity to the spatial resolution

In our study, the highest spatial resolution at which we can compute users' mobility metrics is the level of cells (towers clusters). To test how our results are impacted by the spatial resolution considered, we computed the semi-partial correlation between the GSE ratio (as defined in the Materials and Methods) and the gender ratio of the mobility metrics (the entropy, S, and the set of locations accounting for 80% of a user's activity $\hat{N}_l$) at the cell level. The correlation between socio-economic variables and the gender mobility gap, observed for the *comunas*, is still present at the level of cells. Indeed the Pearson correlation between the GSE ratio and R_S is $r = -0.45$ ($p < 10^{-25}$) and the correlation between the GSE ratio and $R_{\hat{N}_l}$ is $r = -0.40$ ($p < 10^{-25}$).

Text D: Robustness with respect to call activity

In our study, we selected our users' sample by considering only those who made or received at least one call per day, on average, over the three month study period. To test the sensitivity of our results to the chosen activity threshold, we also selected a smaller sample of users who made or received at least 3 calls per day, on average, during the study period, resulting in a total of $254,586$ users. Overall, the differences in mobility between men and women are also observed in the restricted users' sample: women display a lower mobility entropy and visit a smaller number of locations. On average, women visit less distinct locations than men. Considering the complete set of locations visited by a user, we find women visit about 12 locations (95% CI $[11.57, 12.01]$) less than men. If we only look at distinct locations that characterize the core of users' daily activity, ($\hat{N}_l$), the difference between men and women becomes $\Delta\hat{N}_l = 2.73$ $[2.68 - 2.79]$. The average radius of gyration of women, $\langle r_G \rangle_F$, is 1.14 Km (95% CI $[1.10 - 1.17]$) shorter than $\langle r_G \rangle_M$. Also, women's mobility patterns are characterized by a smaller Shannon entropy compared to men, $\Delta S = 0.30$ (95% CI $[0.29 - 0.30]$).

Text E: Sensitivity analysis of gender differences in visit patterns
Here, we discuss several aspects of the analysis of gender differences in visit patterns presented in the main text. In particular, we describe the sensitivity analysis done to check the robustness of our results and some limitations related to it.

First, the use of CDR data, and the spatial resolution at which we work, only allow us to speak of proximity to a given POI, hence our use of POI densities, estimated via kernel methods, over spatial scales of 1km or more. We make no claims about gendered visits to places or services that can not be properly described at a finer spatial scale.

Second, the observed gender differences might be ascribed to gender biases in the *residence locations* of the individuals under study. To rule this out, we repeat the POI analysis by using, for each user, a single location corresponding to the inferred home tower, and we find (Fig. S7B) that the gender imbalance vanishes for malls and hospitals, and is drastically reduced for taxi stands (we remind that the inferred home location for each user was excluded in the POI analysis reported in Figure 4 of the main text). To further investigate this point, we also repeat the POI analysis of Figure 4 by restricting it to *distant locations* for each user: that is, for each user we remove all visited locations within 5km of the inferred home location. The results (Fig. S7C) show that visits to distant places near hospitals and malls are still strongly gendered (in fact, the effect grows slightly stronger), while the gender imbalance in visits associated to taxi POIs is reduced, as expected if we assume that females use taxi stands in the vicinity of their home more than males.

Third, the observed gender imbalance of Figure 4 might be ascribed to gender differences in *employment* at, e.g., malls and hospitals. To investigate this, we check for the robustness of the POI analysis on removal of the most-frequently-visited non-home location for each user, which we will assume indicates the likely work location of employed individuals. On doing so (Fig. S7D), we observe no significant change in our results. This robustness holds on removing the top-k most-frequently-visited non-home locations for each user, for several values of $k > 1$ (not shown), confirming that the gender differences we report are not associated to users' frequently visited locations (e.g., work place or other relevant place), but rather correspond to places in the long tail of visited locations.

Fourth, limitations in the quality of the POI data we use, especially the OpenStreetMap POIs, might in principle influence our results. As far as the mall and *hospital* POIs are concerned, data quality is actually not an issue: the list of *mall* POIs is generated by looking up towers within the boundary of malls, and the location of malls has been manually verified. The OpenStreetMap *hospital* POIs were successfully checked against official databases of healthcare facilities in Santiago. It is also important to remark that, for both malls and hospitals, the spatial extent of the facilities ($\sim 1\text{km}^2$) is comparable to the spatial resolution of the CDR data, hence a CDR record associated to mall or hospital cell tower corresponds with high probability to a user actually visiting those facilities. As far as *taxi* POIs are concerned, it is challenging to assess spatial biases in their reporting in OpenStreetMap. However, we observe similar gender differences for other type of non-public-transport POIs, such as the *collectivos* shared taxis data (see Fig. S8).

Finally, to check that the gender differences observed in the visited locations are not due to differences in call activity we computed the average call activity by gender in the cells where the different POIs are present. We observe that the POI types showing the strongest gender imbalances, i.e. the largest gender density ratio, are not associated to a higher call activity for women. Indeed, the average call activity by female users is always lower than the one by male users for the *hospital* (F average activity 38.5 vs M average activity 40.4), mall (F average activity 34.1 vs M average activity 36.3) and *taxi* (F average activity 28.1 vs M average activity 28.9) layers.

References

[44] Ilya Nemenman, Fariel Shafee, and William Bialek. Entropy and inference, revisited. In *Advances in neural information processing systems*, pages 471–478, 2002.