SCIENTIFIC REP⚙RTS

OPEN

Influence of sociodemographics on human mobility

Maxime Lenormand[1], Thomas Louail[2,3], Oliva G. Cantú-Ros[4], Miguel Picornell[4], Ricardo Herranz[4], Juan Murillo Arias[5], Marc Barthelemy[2,6], Maxi San Miguel[1] & José J. Ramasco[1]

Human mobility has been traditionally studied using surveys that deliver snapshots of population displacement patterns. The growing accessibility to ICT information from portable digital media has recently opened the possibility of exploring human behavior at high spatio-temporal resolutions. Mobile phone records, geolocated tweets, check-ins from Foursquare or geotagged photos, have contributed to this purpose at different scales, from cities to countries, in different world areas. Many previous works lacked, however, details on the individuals' attributes such as age or gender. In this work, we analyze credit-card records from Barcelona and Madrid and by examining the geolocated credit-card transactions of individuals living in the two provinces, we find that the mobility patterns vary according to gender, age and occupation. Differences in distance traveled and travel purpose are observed between younger and older people, but, curiously, either between males and females of similar age. While mobility displays some generic features, here we show that sociodemographic characteristics play a relevant role and must be taken into account for mobility and epidemiological modelization.

Everyday, billions of individuals generate a large volume of geolocated data by using their mobile phone, GPS, public transport cards or credit cards. Such a vast amount of data is bringing new opportunities for the research in socio-technical systems[1–3]. Indeed, geolocated data allow the identification of when and where people interact with or through ICT tools. Each time someone makes a phone call or pays with a credit card the event gets registered contributing to massive databases with potential to provide useful insights on human behavior and mobility[4–9]. For example, the authors of Refs. 6, 7 used credit card and mobile phone datasets to study statistical characteristics of mobility patterns and showed that the distribution of displacement of all users can be approximated by a Levy law. Recently, geolocated data has been also employed to study the spatial structure of cities by detecting hotspots[10] or to characterize land use patterns in urban areas[11–15] with mobile phone records, Twitter data[16] or both together[17]. On a larger scale, comparisons and relations between different cities[18] or even between countries[19,20] have also been also investigated.

Beyond mere location, some datasets offer the opportunity to gather extra information about the type and duration of the interaction or the operation through ICT tools. For instance, it is possible to know from mobile phone records where and when an individual makes a call, but sometimes information such as the ID of the callee and the call duration are also available. This information enables researchers to move further on the study of human behavior by analyzing the structure, intensity and spatial properties of social interactions. Some examples include the analysis of the structure of social networks[21–27],

[1]IFISC, Instituto de Física Interdisciplinar y Sistemas Complejos (CSIC-UIB), Campus Universitat de les Illes Balears, E-07122 Palma de Mallorca, Spain. [2]Institut de Physique Théorique, CEA-CNRS (URA 2306), F-91191, Gif-sur-Yvette, France. [3]Géographie-Cités, CNRS-Paris 1-Paris 7 (UMR 8504), 13 rue du four, FR-75006, Paris, France. [4]Nommon Solutions and Technologies, Plaza del Marqués de Salamanca 11, 28006, Madrid, Spain. [5]BBVA Data & Analytics, Avenida de Burgos 16D, 28036 Madrid, Spain. [6]Centre d'Analyse et de Mathématique Sociales, EHESS-CNRS (UMR 8557), 190-198 avenue de France, FR-75013 Paris, France. Correspondence and requests for materials should be addressed to M.L. (email: maxime@ifisc.uib-csic.es)

the correlation between mobility and social network[28–30], information diffusion[31] and the role played by social groups[26,32].

However, many previous studies lack sociodemographic resolution on the characteristics of the individuals. Except for some features such as language or place of work and/or residence identified in Refs. 19, 33, information about gender, age or occupation are typically missing from studies based on ICT data. This information is of great relevance to characerize the city structure, to estimate population needs in urban planning, transport demand and also for public health. For example, regarding age, knowing the areas of concentration of younger and older population helps to optimize infrastructure such as location of schools, care facilities, etc. Another aspect for which this information is relevant is the modeling of infectious diseases spreading. The models rely on the interplay among hosts, which is related to their location and mobility. Recent epidemic modeling has incorporated mobility information as a way to get closer to real disease spreading[4,34–47]. Additionally, demographic factors such as age or gender can also play an important role in disease transmission and, therefore, must be taken into account when modeling certain infections[48–54]. Furthermore, in a sort of feedback loop, these sociodemographic factors influence mobility as well.

Some works based on smaller-scale surveys point out towards a number of significant differences between men and women in terms of their travel purposes and the activities they pursue[55–57]. More recently, quantitative studies of social networks dynamics have also shown that people behave differently according to the gender and age[58,59]. In this paper, we go beyond by analyzing a credit card use database containing over 40 million card transactions in order to explore consumption and mobility patterns of bank customers in the two most populated provinces of Spain according to three sociodemographic characteristics: gender, age and occupation.

Methods

Dataset description. Our dataset comes from an extraction of the Banco Bilbao Vizcaya Argentaria (BBVA) database on credit card transactions. Different extractions of this data have been used in open data challenges[60] and other scientific works[61]. The data contains information about 40 million bank card transactions made in the provinces of Madrid and Barcelona in 2011. Each transaction is characterized by its amount (in euro currency) and the time when the transaction has occurred. Each transaction is also linked to a customer and a business using anonymized customer and business IDs. Customers are identified with an anonymized customer ID connected with sociodemographic characteristics (gender, age and occupation) and the postcode of his/her place of residence. For convenience sake, we consider five age groups (]15, 30],]30, 45],]45, 60],]60, 75], >75) and five types of occupations (student, unemployed, employed, homemaker, and retired). In the same way, businesses are identified with an anonymized business ID, a business category (accommodation, automotive industry, bars and restaurants, etc.) and the geographical coordinates of the credit card terminal.

The geographical extent of our data is restricted to the provinces of Barcelona and Madrid. For both case studies, we only consider the credit card payments made in the province by individuals living in the province (Fig. 1). Table 1 presents some basic statistics on the data collected. Both provinces have similar features in terms of population size, area and number of businesses, but the number of users and transactions are higher in Madrid than in Barcelona. The number of users represents about 8% of the total census population in Madrid and 5% of that of Barcelona.

Results

The statistical features of the data for Barcelona and Madrid are very similar. Therefore, the data is aggregated for analyzing general properties in the next two sections and segregated later in the third one to study mobility patterns. The aggregation provides higher statistical power, while the disaggregation is needed due to the different geographical shapes of both provinces. Due to the optimization of space, only figures obtained for Madrid are displayed in the third section on mobility. Still equivalent results for Barcelona are found and can be seen in the Supporting Information (Figs. S9–S15).

General features. In order to have a first look at the data, we plot in Fig. 2 some descriptive statistics about individuals according to their sociodemographic characteristics. Figure 2 shows the proportion of individuals according to gender, age and occupation in the dataset and the corresponding fractions as observed in the census[62]. We note an over-representation of men and middle-aged individuals (30-60) in the dataset compared to census data. Moreover, employed people represent about 80% of the individuals, which is two times higher than the proportion of employed people in Spain. Therefore, since the data are not representative of the population, in the rest of the manuscript only indicators and measures normalized by the total number of individuals in each groups will be considered. It is also important to note that the three distributions are not independent, for example, the proportion of individuals according to the age is not the same for student and retired individuals. In the same way, the proportion of individuals according to the occupation is different for men and women. For example, there are more female homemakers than male homemakers. For more details, histograms of the three joint distributions are available in Supporting Information (Figs. S1–S3).
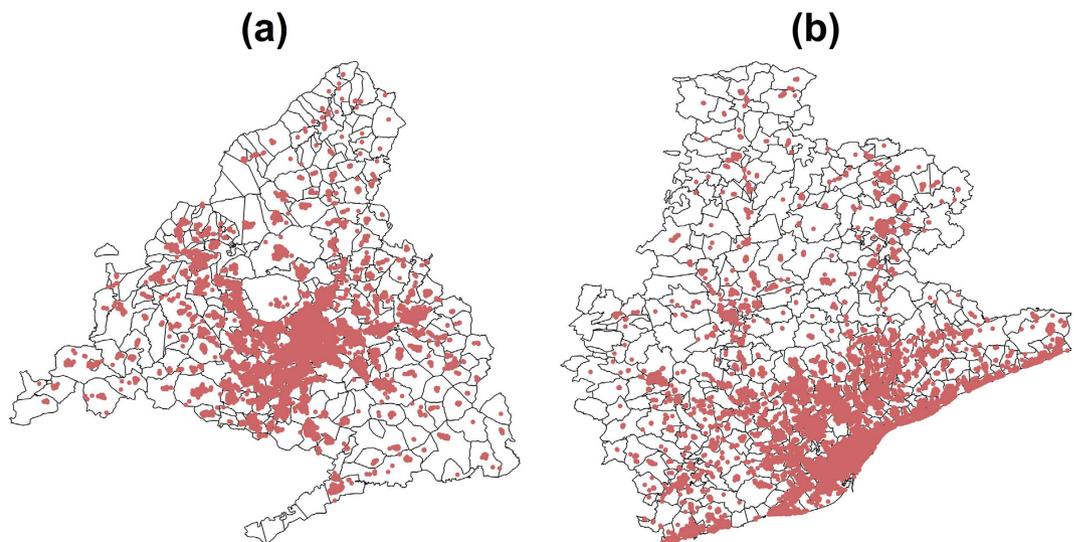
(a)   **(b)**

Figure 1. Maps of the transactions. The red dots represent the locations of the transactions on a map of the province of Madrid (**a**) and Barcelona (**b**) The small areas correspond to postcodes. This map was generated using standard packages of the R statistical software for handling spatial data. The vector layer of the Spanish postcode boundaries is available under free license on multiple websites.

Statistics	Barcelona	Madrid
Number of postcodes	368	271
Number of inhabitants	5,540,925	6,489,680
Area (km^2)	7,733	8,022
Number of customers	270,205	531,818
Number of transactions	13,077,178	24,920,896
Number of businesses	111,956	109,707

Table 1. Summary statistics of the two provinces.

To highlight differences between individuals having different sociodemographic characteristics, we also plot on Fig. 2 the median number of transactions per user, the median amount of money spent per user and the median average amount of money spent per transaction per user. We used the median instead of the average because the distributions exhibits a large number of outliers (see Figs. S4–S6 in Supporting Information for more details). It can be observed that individuals do not spend their money in the same way according to whether they are men or women, young or old and active or inactive. For instance, the number of transactions and the amount of money spent is higher for women than for men and decreases with age. Furthermore, they are also higher for employed persons and homemakers than for unemployed individuals, students and retired people (which is probably related to the age). Inversely, the average amount of money spent per transaction is higher for men than women and increases with age.

To investigate the influence of sociodemographics on the way people spend their money, we plot on Fig. 3 the average fraction of money spent by an individual according to the business category and his/her sociodemographic characteristics. Since the total amount of money spent in 2011 is different from one individual to another, the distribution has been normalized for each user by the total amount of money he/she spent during the year. Note that the distribution is very different for men and women. Indeed, women spend more money than men in Fashion, Food/Hypermarkets, Health and Wellness/Beauty whereas men spend more money than women in Automotive Industry, Bar/Restaurants, Technology and Transport. We also find that the proportion of money spent in Fashion, Food/Hypermarkets, Sports/Toys, Technology and Transport globally decreases with age. Inversely, the amount of money spent in Automotive Industry, Health, Travel Agencies and Wellness/Beauty increases with age. Finally, the differences between people having different occupation are explored. For instance, students spend more money in Bar/Restaurant, Fashion, Sports/Toys and Technology than others types of occupation.
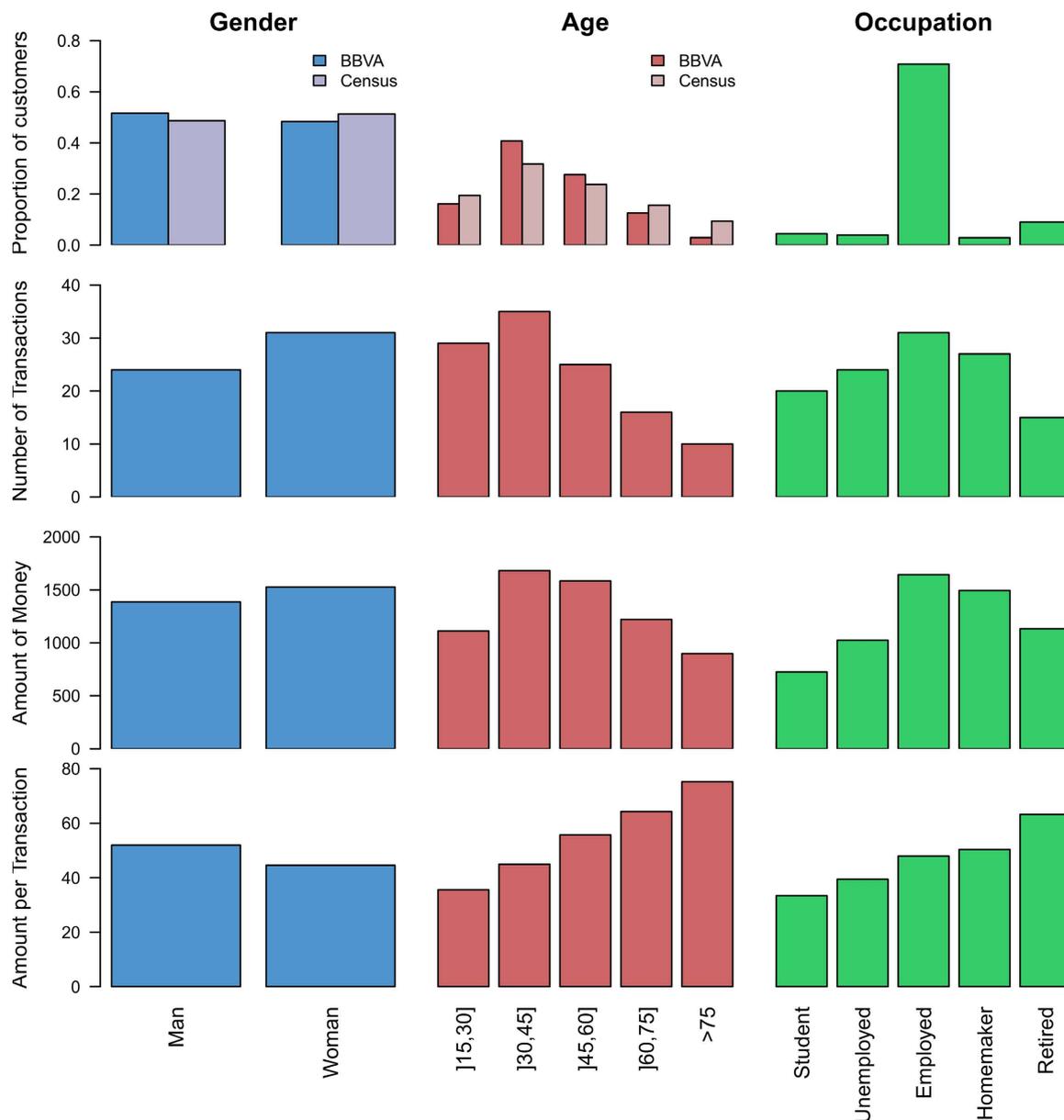
Figure 2. Descriptive statistics according to the individual sociodemographic characteristics. From top to bottom, proportion of individuals, median number of transactions per user and per year, median amount of money spent per user and per year (in euro) and median of the average amount of money spent per transaction (in euro) according to, from left to right, the gender, the age and the occupation.

Since the proportion of individuals according to the occupation is different for men and women, and in order to take away potential bias, we have studied the average fraction of money spent by an individual according to the business category and his/her sociodemographic characteristics but only for employed individuals. We reach the same conclusions as for the overall sample, see Fig. S7 in Supporting Information.

Time evolution of the amount of money spent. To study how the amount of money spent by BBVA customers changes over time during an average week, the days of the week have been divided into four groups: one, from Monday to Thursday representing a normal working day (hereafter called *WD*) and three more for Friday, Saturday and Sunday (hereafter called *Fri*, *Sat* and *Sun*). The average amount of money spent per day as a function of the hour of the day is displayed in Fig. 4a (gray curve). Globally, the amount of money spent is significantly higher during the week days, Friday and Saturday than on Sunday. This can be explained by the fact that most of the business were closed on Sunday in Spain in
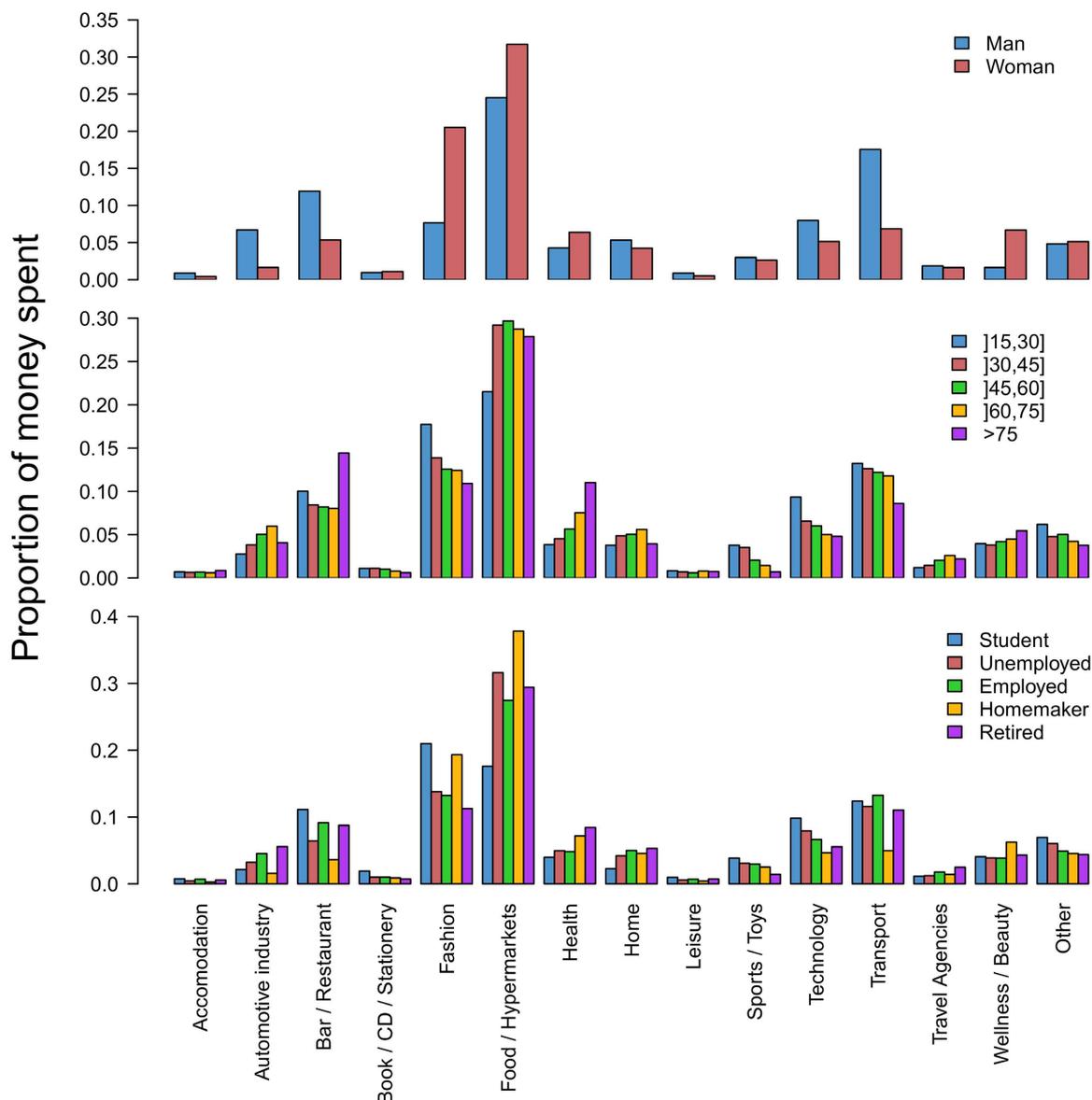
Figure 3. Average fraction of money spent by an individual according to the business category and his/her sociodemographic characteristics. From the top to the bottom: gender, age and occupation.

the time that the data was collected. The activity on Sunday takes place between 10*am* and 7*pm* with a small peak around 4*pm*. During the week days, Friday and Saturday money is spent between 8*am* and 10*pm*. For these days the curves show two peaks, one around noon and another one around 7*pm*. It is interesting to note that for the week days and Friday the second peak is higher than the first one whereas the opposite behavior is observed on Saturday. A small peak around 11*pm* corresponding to the nightlife activity is also observed for the three first days.

To go further in the analysis, a k-means clustering algorithm with Euclidean distance[63] is applied in order to identify clusters naturally present in the data. The purpose is to cluster together individuals exhibiting temporal distribution of money spent. The total amount of money spent in 2011 is different from one individual to another so we have normalized the temporal distribution of money spent for each user by the total amount of money he/she spent in 2011. To choose the number of clusters, we use the pseudo-F statistics which describes the ratio of between-cluster variance to within cluster variance[64]. The optimal number of clusters is the one for which the highest pseudo-F value is obtained, in our case we found two opposite clusters (see Fig. S8 in Supporting Information for more details). Figure 4a displays the results of the clustering analysis, we observe an opposition between active and inactive individuals. The first cluster represents one third of the individuals and is characterized by a higher activity during the morning and during weekdays in opposition with the second cluster in which individuals tend to
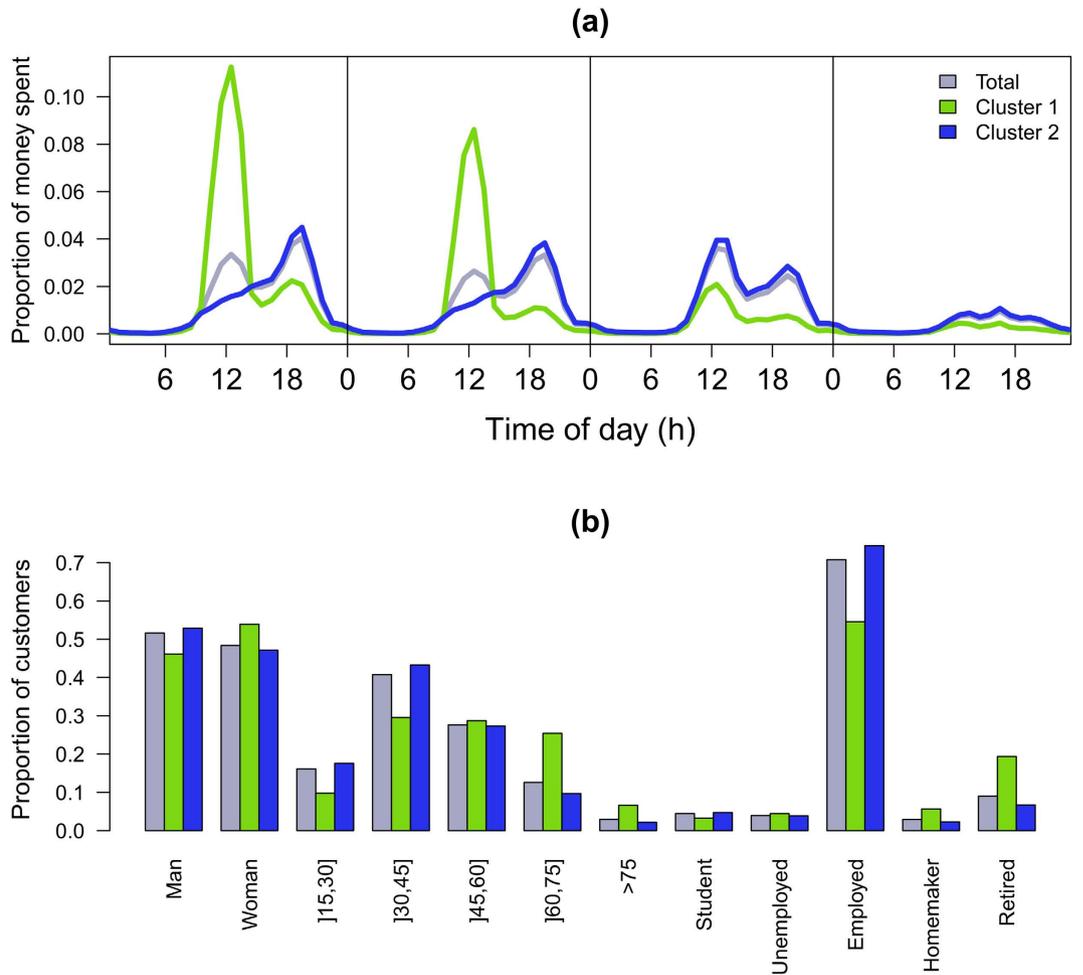
Figure 4. Time evolution of the amount of money spent. (**a**) Average amount spent per day as a function of the hour of the day in total and according to the cluster. From left to right: weekdays (aggregation from Monday to Thursday), Friday, Saturday and Sunday. (**b**) Proportion of individuals in total and in each cluster according to, from left to right, the gender, the age and the occupation.

spend more money after 6*pm* and during week end days. It is interesting to note that the first cluster is over-represented by women, old people and homemaker and retired individuals compared to the whole population (Fig. 4b).

Mobility patterns. In order to characterize mobility patterns of each user, we have considered three variables: Δ_t, the time elapsed between two consecutive transactions, Δ_r, the distance traveled between two consecutive transactions, and r_g, the radius of gyration[7]. The radius of gyration is defined as

$$r_g = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\overrightarrow{p}_k - \overrightarrow{p}_c\right)^2},$$

(1)

where $\overrightarrow{p}_k$ represents the kth position of the user displacements in 2011 and $\overrightarrow{p}_c = \frac{1}{n}\sum_{k=1}^{n}\overrightarrow{p}_k$ is the center of mass of his/her motions. It is important to note that r_g is defined per user whereas Δ_t and Δ_r are computed for each displacement. Although Δ_r and r_g are related, Δ_r informs us on the distance traveled by users, which might depend on the frequency at which each person uses its credit card, whereas r_g gives us a more holistic view of how people moves around their centers of mass. To avoid the introduction of bias in the mobility patterns analysis, all the consecutive user's positions geo-located in the province and the distances between them are considered whatever the elapsed time between consecutive transactions.

Figures 5a, 6a and 7a display the probability density function of the three variables. The distribution of Δ_t is a decreasing density function exhibiting circadian rhythms. The average and median time between two transaction are, respectively, around 5 days and 2 days. The distribution of Δ_r show two
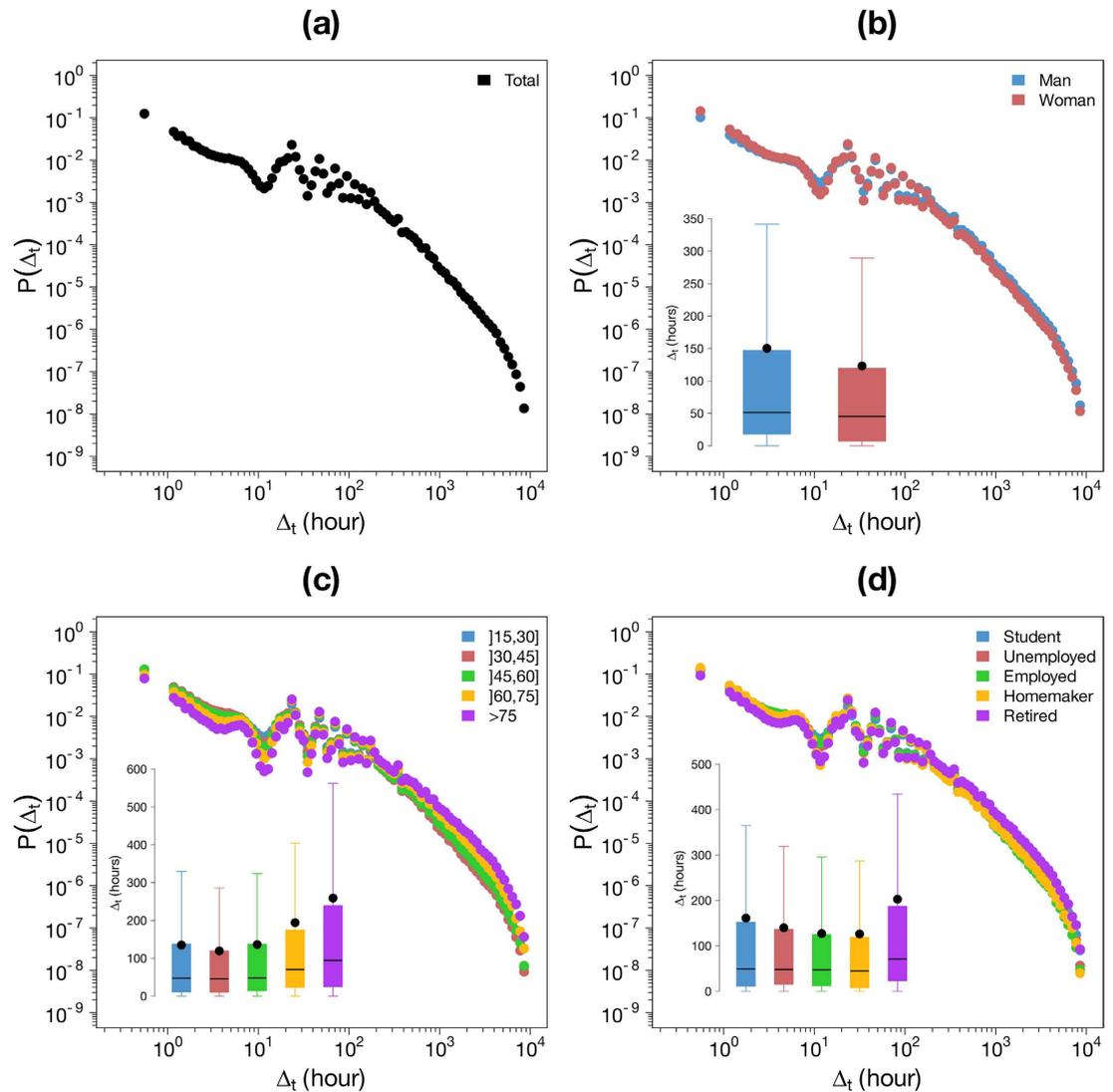
Figure 5. Inter-event time distribution $P(\Delta_t)$. (**a**) Probability density function of Δ_t. (**b**) - (**d**) Probability density function of Δ_t according to the gender (**b**) the age (**c**) and the occupation (**d**). The insets show the Tukey boxplot of the distributions, the black points represent the average.

different regimes. First the distribution exhibits a slow decay, and then, beyond 40 kilometers the distribution is characterized by a rapid decay. This cutoff is introduced by the limited geographical scale of the provinces. The probability density function $P(r_g)$ increases very slowly until reaching a maximum around 6 kilometers and then the distribution is characterized by a rapid decay.

In this work we have also assessed the influence of sociodemographic characteristics on the individual mobility patterns. The results obtained are plotted on the Figs. 5, 6 and 7. For each sociodemographic characteristic and each variable, we performed two non-parametric tests to assess the statistical significance of the differences between the different type of users' mobility using the Mann "Whitney U test[65] to compare the distributions and the Mood's median test[66] to compare the medians. For both case studies the differences between distributions and medians are always significant (*p-values* lower than 10^{-4}) except for the difference between radius of gyration of individuals of age between 15 and 30 and those between 30 and 45 in Barcelona.

Figure 5 displays the inter-event time distribution according to the gender (Fig. 5b), the age (Fig. 5c) and the occupation (Fig. 5d). The average and median inter-event time are higher for men than women and increases with age. They are also higher for unemployed individuals, students and retired people than for employed persons and homemakers. We observe an negative correlation between the time elapsed between two consecutive transactions and the number of transactions per individual described in the first section.

The results obtained for Δ_r and r_g are plotted in Figs. 6 and 7, respectively. Based on these results, one can understand that, depending on his/her sociodemographic characteristics, an individual can
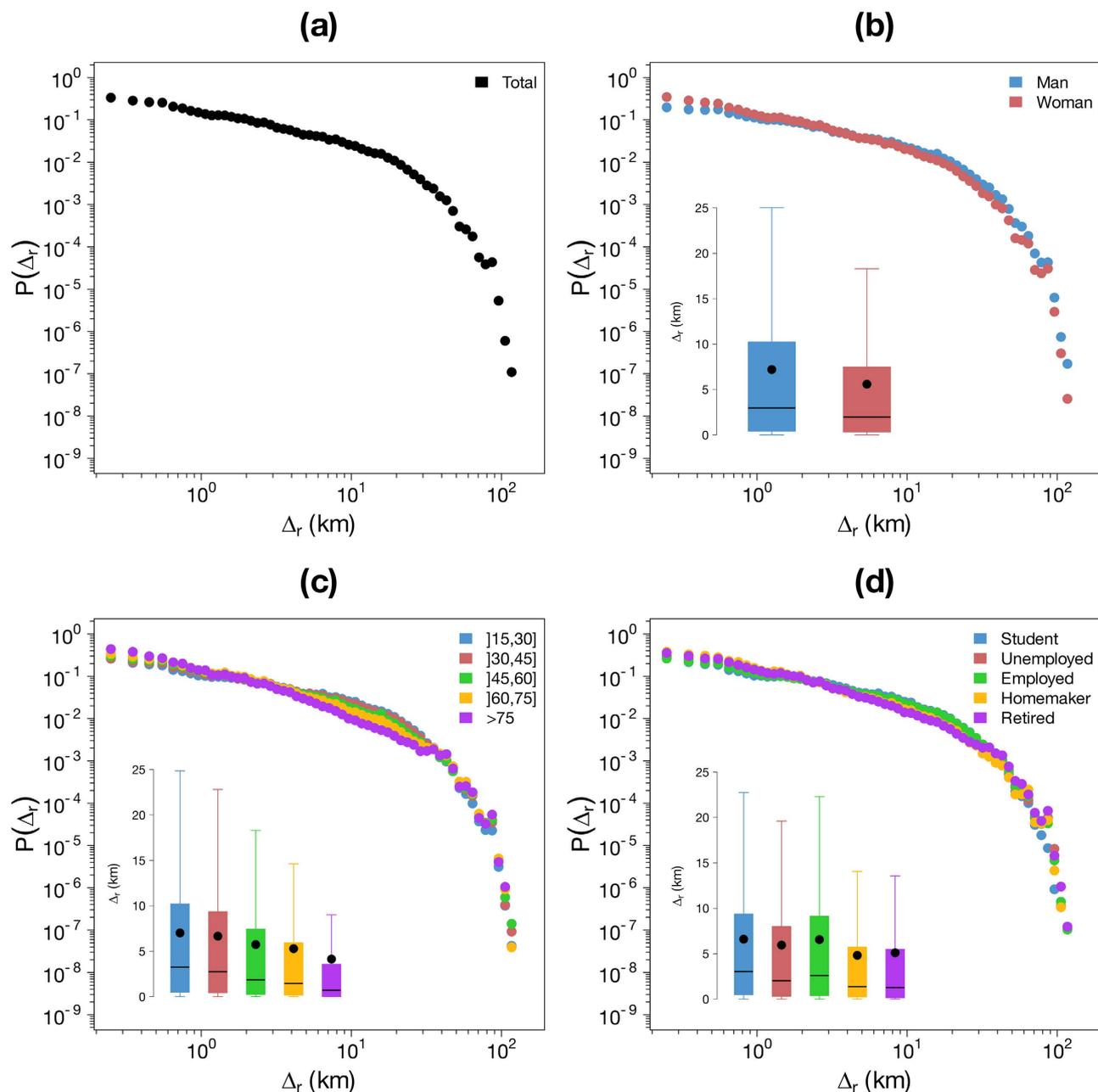
Figure 6. Distribution of the distance traveled by an individual between two consecutive transactions $P(\Delta_r)$. (**a**) Probability density function of Δr. (**b**) – (**d**) Probability density function of Δr according to the gender (**b**) the age (**c**) and the occupation (**d**). The insets show the Tukey boxplot of the distributions, the black points represent the average.

travel short or long distances and stays more or less close to his/her center of mass. Three main differences are observed. First, women travel shorter distances than men and their trajectory stays closer to their center of mass. Second, the average distance traveled between two consecutive positions and the radius of gyration decrease with age. Finally, an opposition between active and inactive individual is highlighted. Indeed, retired, homemaker and, to a lesser extent, unemployed individuals travel shorter distances and stay closer to the center of mass than other people.

As previously mentioned, the distance traveled by an individual between two consecutive transactions might depend on the frequency at which an individual uses his/her credit card, and therefore, the differences between people observed for Δ_r could be a consequence of the differences observed for Δ_t. Although the same conclusion are reached for the radius of gyration, which does not depend on the frequency at which someone uses his/her credit card, it could be interesting to study how the average value of Δ_r evolves as a function of Δ_t according to the individual's sociodemographic characteristics.
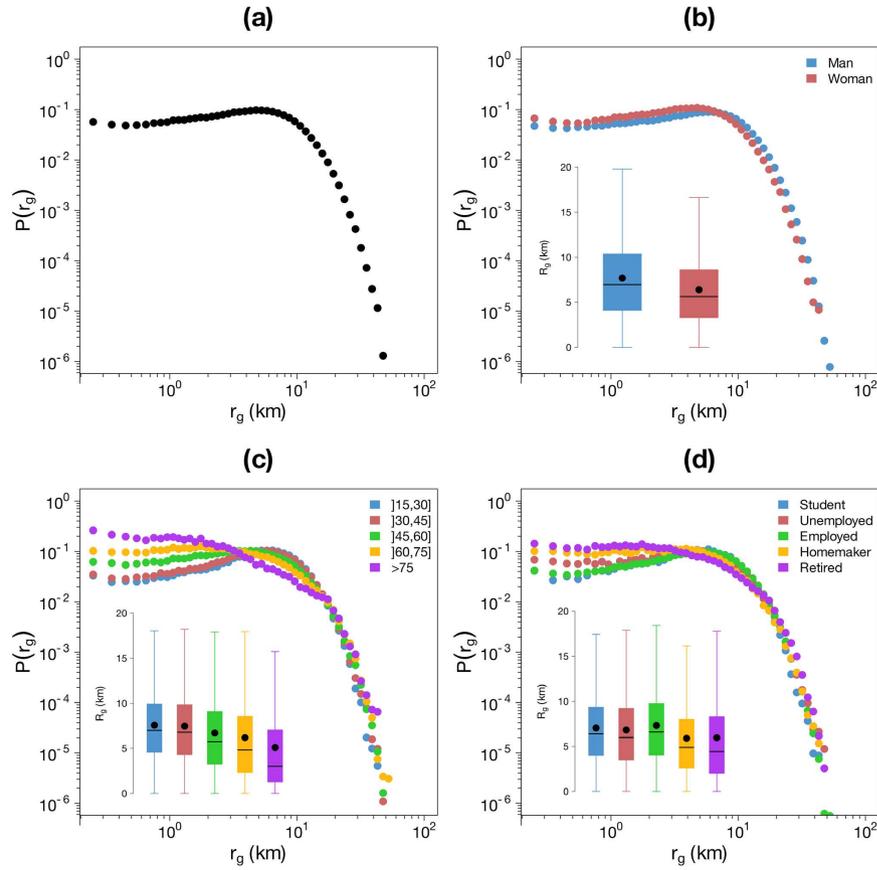
Figure 7. **Distribution of the radius of gyration** $P(r_g)$. (**a**) Probability density function of r_g. (**b**) – (**d**) Probability density function of r_g according to the gender (**b**) the age (**c**) and the occupation (**d**) The insets show the Tukey boxplot of the distributions, the black points represent the average.
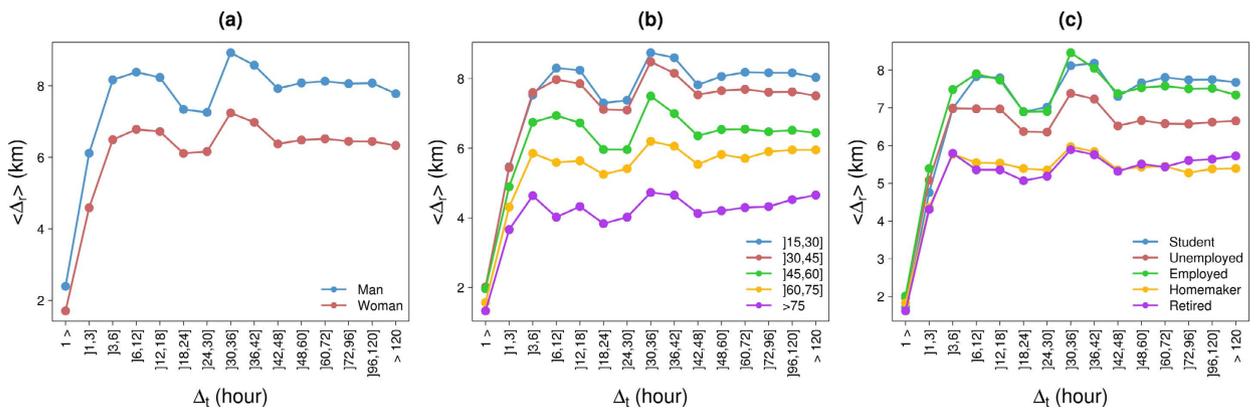

Figure 8. **Average** $\langle \Delta_r \rangle$ **value as a function of** Δ_t **according to the gender (a) the age (b) and the occupation (c).**

We can observe in Fig. 8 that the differences between the different types of individuals in terms of distances traveled always exist whatever the time elapsed between two consecutive transactions. It is also worth noting that the value of $<\Delta_r>$ is not completely independent of Δ_t. Obviously, for small values of Δ_t ($\Delta_t < 6$) the value of $<\Delta_r>$ increases with the value of Δ_t due to physical constraints but we can also note a valley for $\Delta_t \in]18, 30]$ followed by a peak for $\Delta_t \in]30, 42]$. This phenomenon seems to be more pronounced for active people than for inactive people, possibly reflecting the home-to-work/school commuting.
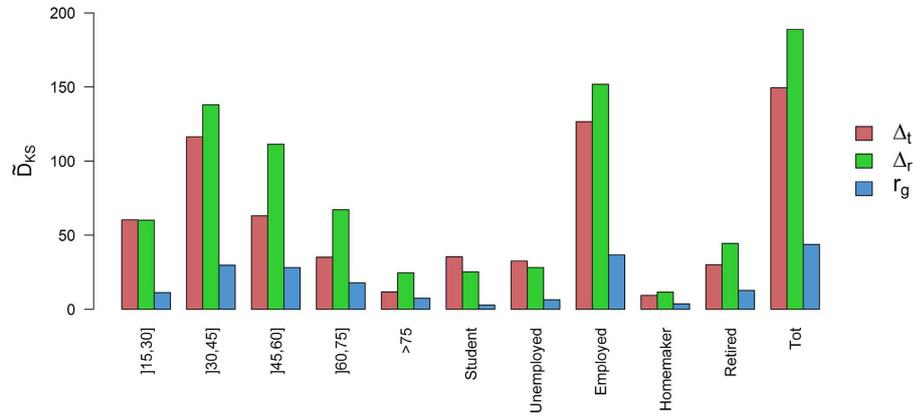
Figure 9. Kolmogorov-Smirnov distance between men and women's Δ_t distributions (in red), Δ_r distributions (in green) and r_g distributions (in blue) according to their sociodemographic characteristics.
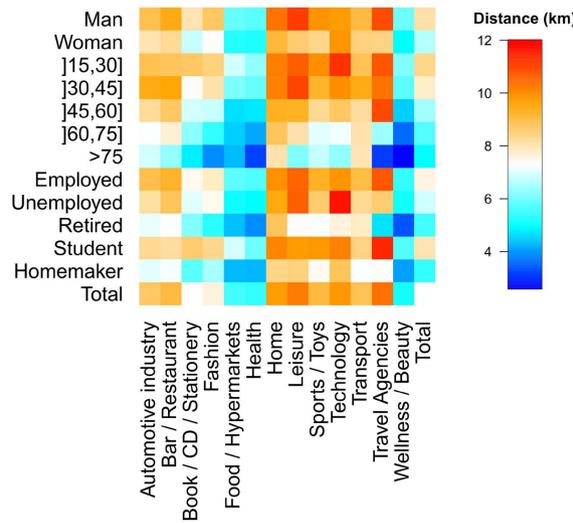

Figure 10. Average distance between individuals residence and business according to sociodemographics and business category. Distances are expressed in kilometer and are computed using the Haversine distance between the latitude and longitude coordinate of the centroid of the postcode of residence and the business' latitude and longitude coordinates for each transaction.

Among all these comparisons, discrepancy in mobility between men and women is the most challenging. In order to verify that this difference is significant and it is not related to other sociodemographic variables, the Kolmogorov-Smirnov (KS) distance between men and women's Δ_t, Δ_r and r_g distributions are computed (Fig. 9). The Kolmogorov-Smirnov (KS) distance between two empirical probability distributions X and Y is defined as

$$D_{KS} = \sup_x \left| F_X(x) - F_Y(x) \right|, \tag{2}$$

where F_X and F_Y are the empirical cumulative distribution function of X and Y respectively. Since, the sample size of both distributions may vary from one sociodemographic variable to another we need to normalize D_{KS} according to the sample sizes,

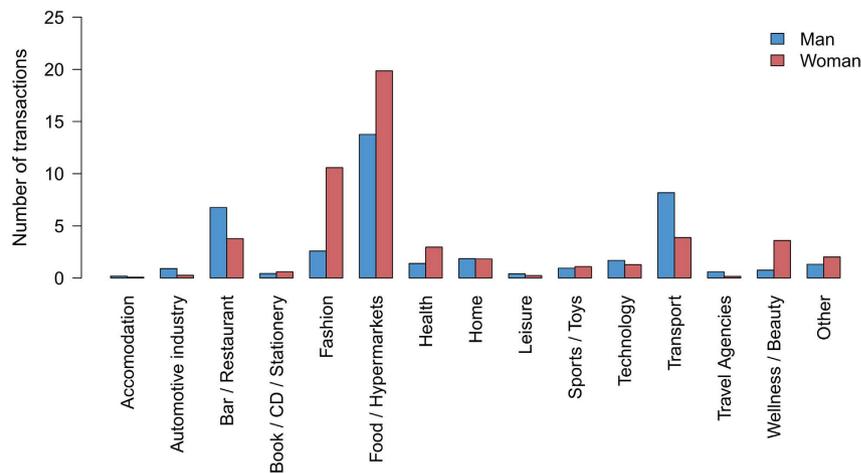$$\widetilde{D}_{KS} = \sqrt{\frac{n_X n_Y}{n_X + n_Y}} \, \sup_x \left| F_X(x) - F_Y(x) \right|, \tag{3}$$

Figure 11. Number of transactions according to the gender and the business category.

where n_X and n_Y represent the sample sizes of X and Y, respectively. This allows for a direct comparisons of the Kolmogorov-Smirnov distances. Moreover, using this normalization, the null hypothesis that the two data samples come from the same distribution is rejected at level 0.001 if $\widetilde{D}_{KS} > 1.95$.

First, we observe that a significant difference between men and women appears whatever the socio-demographic characteristic of the population is filtered out (i.e. $\widetilde{D}_{KS}$ is always higher than 1.95), which means that on average, women have an inter-event time lower than men and men do longer journeys than women. Second, one can observe that this gendered difference is more important for middle age individual than for young and old people, but also that is more pronounced for employed people.

To go further, we have studied too the influence of the individuals sociodemographic characteristics and the business category on the distance traveled between home and business. To do so, we computed for each transaction the distance between the individual's place of residence and the business. As residence location, we use the centroid of the individual's postcode of residence. Finally, these distances were averaged according to individual and business type. These average distances can be observed in Fig. 10. First, we observe that the same differences between type of individuals as the ones highlighted previously are obtained whatever the business category. For each business category, the distance between home and business is globally higher for men than women, it decreases with age and it is higher for employed and student than for the other occupation categories. Although, the average distance between home and business changes according to the category of business. Indeed, distances between home and businesses belonging to the categories Food/Hypermarkets, Health, Wellness/Beauty and Book/CD/Stationery are lower than for the other categories. It is interesting to note that these business category are also the type of business in which the number of transactions is higher for women than for men (Fig. 11). This partially explains why women travel shorter distance than men to go shopping.

Discussion

In summary, we have shown in this study that it is possible to use information provided by credit card data to assess the influence of sociodemographic characteristics on the way people move and spend their money. We highlighted differences in consumption habits and mobility patterns of bank customers according to their gender, age and occupation. First, we shown that according to the business type the fraction of money spent can be very different from one individual to another. In particular, women tend to spend more money in Fashion, Food/Hypermarkets, Health and Wellness/Beauty than men whereas men spend more money than women in Automotive Industry, Bar/Restaurants, Technology and Transport. We have also studied the time evolution of the amount of money spent along the week according to the individual's sociodemographic characteristics. An opposition between two types of individuals has been identified. The temporal distribution of money spent by the first type of individuals which is over-represented by inactive people is characterized by a higher activity during the morning and during weekdays in opposition with the second type of individuals more active after working hours and during week end days. Then, we investigated the properties of people mobility patterns using three variables: the time elapsed between two consecutive transactions, the distance traveled by an individual between two consecutive transactions and the radius of gyration. Three main differences between groups of people were identified: differences between men and women, young and old people and active and inactive individuals. In the three cases, people of the first group (men, young people and active people) travel shorter distances and their trajectory stays closer to their center of mass than individuals of the second groups (women, old individual and inactive people).

Among all the differences emphasized in this paper the one between men and women is the most difficult to explain. In all the comparisons we have carefully checked that this difference was not related to other sociodemographic variables and it was not the case. It could be interesting to verify whether this difference is related to other social characteristics such as the number of children for example. Indeed, the fact that the difference in terms of mobility patterns between men and women is less pronounced for old people and students may reflect that women with children move differently than women without children. While further data is required to assess whether these differences between individuals are universal, i.e., to which extend they are specific or not to urban areas or the cities of the country analyzed, our results point toward the possibility that mobility may display significant differences for different types of individuals.

References

1. Watts, D. J. A twenty-first century science. *Nature* **445,** 489 (2007).
2. Lazer, D. *et al.* Computational Social Science. *Science* **323,** 721–723 (2009).
3. Vespignani, A. Predicting the Behavior of Techno-Social Systems. *Science* **325,** 425–428 (2009).
4. Chowell, G., Hyman, J., Eubank, S. & Castillo-Chavez, C. Scaling laws for the movement of people between locations in a large city. *Phys. Rev. E.* **68,** 066102 (2003).
5. Barrat, A., Barthélemy, M., Pastor-Satorras, R. & Vespignani, A. The architecture of complex weighted networks. *Proc. Nat. Acad. Sci. USA* **101,** 3747–3752 (2004).
6. Brockmann, D., Hufnagel, L. & Geisel, T. The scaling laws of human travel. *Nature* **439,** 462–465 (2006).
7. Gonzalez, M. C., Hidalgo, C. A. & Barabàsi, A.-L. Understanding individual human mobility patterns. *Nature* **453,** 779–782 (2008).
8. Song, C., Qu, Z., Blumm, N. & Barabàsi, A.-L. Limits of Predictability in Human Mobility. *Science* **327,** 1018–1021 (2010).
9. Bagrow, J. P. & Lin, Y.-R. Mesoscopic Structure and Social Aspects of Human Mobility. *PLoS ONE* **7,** e37676 (2012).
10. Louail, T. *et al.* From mobile phone data to the spatial structure of cities. *Sci. Rep.* **4,** 5276 (2014).
11. Ratti, C., Pulselli, R. M., Williams, S. & Frenchman, D. Mobile Landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design* **33,** 727–748 (2006).
12. Reades, J., Calabrese, F., Sevtsuk, A. & Ratti, C. Cellular Census: Explorations in Urban Data Collection. Pervasive Computing, *IEEE* **6,** 30–38 (2007).
13. Soto, V. & Frías-Martínez, E. Automated land use identification using cell-phone records. In *Proceedings of the 3rd ACM international workshop on MobiArch*, HotPlanet 11, 17–22 (ACM, New York, NY, USA, 2011).
14. Pei, T., Sobolevsky, S., Ratti, C., Shaw, S. L. & Zhou, C. A new insight into land use classification based on aggregated mobile phone data. *ArXiv e-print* **arxiv: 1310.6129** (2013).
15. Toole, J., Ulm, M., González, M. & Bauer, D. Inferring land use from mobile phone activity. Paper presented at: *The ACM SIGKDD International Workshop on Urban Computing, Beijing.* Place of publication: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing, doi:10.1145/2346496.2346498* (2012).
16. Frías-Martínez, V., Soto, V., Hohwald, H. & Frías-Martínez, E. Characterizing urban landscapes using geolocated tweets. Paper presented at: *SOCIALCOM-PASSAT '12, Privacy, Security, 2012 International Conference on Risk and Trust (PASSAT) and 2012 International Confernece on Social Computing (SocialCom), Amsterdam.* Place of publication: *Proceedings of the Privacy, Security, 2012 International Conference on Risk and Trust (PASSAT) and 2012 International Confernece on Social Computing (SocialCom), doi:10.1109/SocialCom-PASSAT.2012.19* (2012).
17. Lenormand, M. *et al.* Cross-checking different source of mobility information. *PLoS ONE* **9,** e105184 (2014).
18. Noulas, A., Scellato, S., Lambiotte, R., Pontil, M. & Mascolo, C. A tale of many cities: Universal patterns in human urban mobility. *PLoS ONE* **7,** e37027 (2012).
19. Hawelka, B. *et al.* Geo-located Twitter as a proxy for global mobility patterns. *Cartography and Geographic Information Science* **41,** 260–271 (2013).
20. Lenormand, M., Tugores, A., Colet, P. & Ramasco, J. J. Tweets on the road. *PLoS ONE* **9,** e105407 (2014).
21. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P. & Tomkins, A. Geographic routing in social networks. *Proc. Natl. Acad. Sci. USA* **102,** 11623–11628 (2005).
22. Onnela, J. *et al.* Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. USA* **104,** 7332–7336 (2007).
23. Java, A., Song, X., Finin, T. & Tseng, B. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, 56–65 (ACM, 2007).
24. Huberman, B. A., Romero, D. M. & Wu, F. Social networks that matter: Twitter under the microscope. *First Monday* **14,** 1–2 (2008).
25. Eagle, N., Pentland, A. S. & Lazer, D. From the Cover: Inferring friendship network structure by using mobile phone data. *Proc. Natl. Acad. Sci. USA* **106,** 15274–15278 (2009).
26. Ferrara, E. A large-scale community structure analysis in Facebook. *EPJ Data Science* **1,** 9 (2012).
27. Grabowicz, P. A., Ramasco, J. J., Goncalves, B. & Eguiluz, V. M. Entangling mobility and interactions in social media. *PLoS ONE* **9,** e92196 (2014).
28. Backstrom, L., Sun, E. & Marlow, C. Find Me if You Can: Improving Geographical Prediction with Social and Spatial Proximity. In *Proceedings of the 19th International Conference on World Wide Web*, 61–70 (ACM, 2010).
29. Calabrese, F., Smoreda, Z., Blondel, V. D. & Ratti, C. Interplay between Telecommunications and Face-to-Face Interactions: A Study Using Mobile Phone Data. *PLoS ONE* **6,** e20814 (2011).
30. Phithakkitnukoon, S., Smoreda, Z. & Olivier, P. Socio-Geography of Human Mobility: A Study Using Longitudinal Mobile Phone Data. *PLoS ONE* **7,** e39253 (2012).
31. Ferrara, E., Varol, O., Menczer, F. & Flammini, A. Traveling Trends: Social Butterflies or Frequent Fliers? In *Proc. 1st ACM Conf. on Online Social Networks (COSN)*, 213–222 (2013).
32. Grabowicz, P. A., Ramasco, J. J., Moro, E., Pujol, J. M. & Eguiluz, V. M. Social Features of Online Networks: The Strength of Intermediary Ties in Online Social Media. *PLoS ONE* **7,** e29358 (2012).
33. Mocanu, D. *et al.* The Twitter of Babel: Mapping world languages through microblogging platforms. *PLoS ONE* **8,** e61981 (2013).
34. Rvachev, L. A. & Longini, I. M. A mathematical model for the global spread of influenza. *Mathematical Biosciences* **75,** 3–22 (1985).
35. Grais, R. F., Hugh Ellis, J. & Glass, G. E. Assessing the impact of airline travel on the geographic spread of pandemic influenza. *Eur. J. Empidemiol.* **18,** 1065–1072 (2003).
36. Eubank, S. *et al.* Modelling disease outbreaks in realistic urban social networks. *Nature* **429,** 180–184 (2004).

37. Hufnagel, L., Brockmann, D. & Geisel, T. Forecast and control of epidemics in a globalized world. *Proc. Natl. Acad. Sci. (USA)* **101,** 15124–15129 (2004).
38. Longini, I. M. *et al.* Containing pandemic influenza at the source, *Science* **309,** 1083–1087 (2005).
39. Ferguson, N. M. *et al.* Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* **437,** 209–214 (2005).
40. Riley, S. Large-Scale Spatial-Transmission Models of Infectious Disease, *Science* **316,** 1298–1301 (2007).
41. Colizza, V., Barrat, A., Barthélemy. M., Valleron, A. J. & Vespignani, A. Modeling the Worldwide spread of pandemic influenza: baseline case and containment interventions. *PloS Medicine* **4,** e13 (2007).
42. Ciofi degli Atti, M. L. *et al.* Mitigation measures for pandemic influenza in Italy: An individual based model considering different scenarios, *PLoS ONE* **3,** e1790 (2008).
43. Balcan, D. *et al.* Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci. USA* **106,** 21484–21489 (2009).
44. Bajardi, P. *et al.* Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic. *PLoS ONE* **6,** e16591 (2011).
45. Meloni, S. *et al.* Modeling human mobility responses to the large-scale spreading of infectious diseases. *Sci. Rep.* **1,** 62 (2011).
46. Tizzoni, M. *et al.* Real-time numerical forecast of global epidemic spreading: Case study of 2009 A/H1N1pdm. *BMC Medicine* **10,** 165 (2012).
47. Poletto, C., Tizzoni, M. & Colizza, V. Heterogeneous length of stay of hosts' movements and spatial epidemic spread. *Sci. Rep.* **2,** 476 (2012).
48. Wallinga, J., Teunis, P. & Kretzschmar, M. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *Am. J. Epidemiol.* **164,** 936–944 (2006).
49. Brauer, F. Epidemic models with heterogeneous mixing and treatment. *Bull. Math. Biol.* **70,** 1869–1885 (2008).
50. Nishiura, H. Travel and age of influenza a (h1n1) 2009 virus infection. *J. Trav. Med.* **17,** 269–270 (2010).
51. Nishiura, H., Cook, A. R. & Cowling, B. J. Assortativity and the probability of epidemic extinction: A case study of pandemic influenza a (H1N1-2009). *Interdisciplinary Perspectives on Infectious Diseases* **2011,** 194507 (2011).
52. Rocha, L., Liljeros, F. & Holme, P. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Comput. Biol.* **7,** 1001109 (2011).
53. Apolloni, A., Poletto, C. & Colizza, V. Age-specific contacts and travel patterns in the spatial spread of 2009 H1N1 influenza pandemic. *BMC Infectious Diseases* **13,** 1–18 (2013)
54. Apolloni, A., Poletto, C., Ramasco, J. J., Jensen, P. & Colizza, V. Metapopulation epidemic models with heterogeneous mixing and travel behaviour. *Theoretical Biology and Medical Modelling* **11,** 3 (2014).
55. Golob, T. F. & McNally, M. G. A model of activity participation and travel interactions between household heads. *Transportation Research Part B: Methodological* **31,** 177–194 (1997).
56. Hamed, M. M. & Mannering, F. L. Modeling Travelers' Postwork Activity Involvement: Toward a New Methodology. *Transportation Science* **27,** 381–394 (1993).
57. Bianco, M. & Lawson, C. Trip chaining, childcare and personal safety: critical issues in women's travel behavior. In Proceedings from the second national conference on women's travel issues. Washington DC: US Department of Transportation, Federal Highway Administration (1996).
58. McPherson, M., Lovin, L. S. & Cook, J. M. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* **27,** 415–444 (2001).
59. Stehle, J., Charbonnier, F., Picard, T., Cattuto, C. & Barrat, A. Gender homophily from spatial behavior in a primary school: A sociometric study. *Social Networks* **35,** 604–613 (2013).
60. Innova Challenge http://www.centrodeinnovacionbbva.com/en/innovachallenge/what-innova-challenge. Date of access 03/12/2014.
61. Sobolevsky, S. *et al.* Mining Urban Performance: Scale-Independent Classification of Cities Based on Individual Economic Transactions. In *Proceedings of ASE BigDataScience 2014 conference* (2014).
62. Spanish Census 2011 (Instituto Nacional de Estadstica): http://www.ine.es/censos2011/tablas/Inicio.do. Date of access 12/03/2015 2011.
63. Hartigan, J. A. & Wong, M. A. A K-Means Clustering Algorithm. *Applied Statistics* **28,** 100–108 (1979).
64. Caliński, T. & Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics-Simulation and Computation* **3,** 1–27 (1974).
65. Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* **18,** 50–60 (1947).
66. Brown, G. W. & Mood, A. M. On Median Tests for Linear Hypotheses. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 159–166 (University of California Press, Berkeley, Calif., 1951).

Acknowledgements

Author Contributions

M.L. coordinated the study, designed the study, analysed the data and wrote the manuscript; T.L. and M.S.M. coordinated the study and wrote the manuscript; O.G.C., M.P., R.H. and M.B. coordinated the study; J.M.A. obtained the data and coordinated the study; J.J.R. coordinated the study, designed the study and wrote the manuscript. All authors read, comments and approved the final version of the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Lenormand, M. *et al.* Influence of sociodemographics on human mobility. *Sci. Rep.* **5,** 10075; doi: 10.1038/srep10075 (2015).