What Banking and Phone Data tell us about the Socioeconomic Groups and their Consumption Patterns?

Ángel F. Agudo-Peregrina
Department of Industrial Management, Business Admin. and Statistics
Universidad Politécnica de Madrid
af.agudo@upm.es

Diego Pérez,
Luis C. Reyes
Research
Clarity AI
diego.perez@clarity.ai;
luis.reyes@clarity.ai

Martin Langberg,
Martin Minnoni
Research Lab
Grandata
mlangberg@grandata.com
martin@grandata.com

***Abstract*— This paper makes use of a large dataset of anonymized banking transactions and phone calls to classify individuals into socioeconomic groups (SEGs) and social networks, determine their consumption patterns, and compare the latter with equivalent information available from household surveys. The results obtained demonstrate that classification into SEGs by aggregated bank income provides a robust breakdown of the population that is validated by a social network analysis of the phone data. In addition, the spending profile obtained for each SEG shows that individuals behave similarly according to their income and their spending can be accurately categorized. Furthermore, the consumption patterns obtained from this novel approach can be contrasted directly with those obtained from national household surveys, potentially overcoming some of the limitations of traditional approaches in terms of coverage and inclusiveness. The work presented here shows the feasibility and capabilities of Big Data sources and tools to understand the consumption behavior of the population.**

I. Introduction

Understanding the consumption behavior of the population has been important in different dimensions: to determine their preferences from the analysis of the budget breakdown, to measure the welfare of the population through the analysis of the products and services included in their consumption profiles, and to understand the consumption differences among different groups of the population, such as those arousing from social inequality **[1,2].** Historically, household surveys have been the main source of information to understand consumption at the household level. Thus, a lot of efforts have focused on improving, extending, standardizing and analyzing these surveys to get the most out of them **[3,4,5]**. However, household surveys suffer from several endemic concerns: the process of getting the information is very slow and expensive, the coverage of the population is limited (richer people are less likely to participate in the surveys), there are issues in the connection of the information coming from the household surveys and the national accounts, and the frequency of updated information varies a lot from one country to another, primarily in developing countries **[6]**. In addition to consumption, household surveys have also been a tool to determine household income, even though they were limited: in poor countries it has been easy to find information about the consumption of the population and difficult to know their salaries (most of the population are self-employed), while in developed countries the problem has been the opposite: it is relatively easy to get information about the income (most of the population are waged employees), but people are reluctant to spend their time answering time consuming surveys **[7]**. The fast dissemination of electronic payment alternatives (debit cards, credit cards, digital payment methods like PayPal, Stripe or similar; mobile payments, etc.) opens an interesting alternative to household surveys to understand the consumption of the population. Thus, the availability of transactional and personal data together with the emergence of Big Data tools and techniques opens a window into data-first, bottom-up approaches that overcome some of the serious limitations of traditional methods to quantify people's consumption. Along these lines, this paper makes use of anonymized bank transactions and mobile-phone call history for a period of one year for millions of individuals to answer questions such as: Can individual banking and mobile phone data be used to measure the income of those individuals and with that, classify individuals into representative socioeconomic groups? And can it be used to determine the consumption profile of the individuals and their communities?

II. Data Description and analysis

Banking transactions dataset (BDS): The anonymized banking dataset consists of millions of debit card transactions collected in a 12-month span. It contains the purchase history of 8,757,080 users who made at least one transaction during this period. In terms of merchants, the sample includes 1,412,719 different merchants, out of which 1,073,066 have a merchant category code (MCC) assigned. MCC is the standard to classify the merchants in card transactions and consists of 28 different groups **[8]**. More than 80% of debit card transactions are located in the service providers group, and inside this group, the vast majority of the transactions are considered cash withdrawals. In addition to debit card transactions, BDS also includes income transactions from which the individual's total income is calculated. Users must have at least two months with transactions in more than three different MCCs. Regarding the income, they must have a regular income every month. After the cleaning process, BDS contains a total of 950,543 unique users.

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 © 2018 IEEE

Phone interactions dataset (PDS): This dataset includes phone calls covering the same timespan as BDS. Phone data is also anonymized, in a way that allows the matching of anonymized users between both datasets. For each call detail record (CDR), the dataset contains the anonymized (encrypted) phone number associated to the contract, geo-location of the cell tower used, anonymized issuer and receiver of the call, duration of the call, and the carrier. The combined dataset must contain users included in both datasets (BDS and PDS); that is, those banking active users that have calling activity. Thus, users from the phone dataset who are not active users in the bank dataset are excluded. This results in 41,753 individuals considered active in the bank and that have at least one CDR.

Income estimated from banking data: Based on the information available in BDS, individuals can be classified according to their total spending (as seen in **[9]**) or through their income. Less than 50% of the population in the studied country have access to a bank account, and of those that have it, the majority are in the richest classes. Thus, in order to preserve the complete vision of all individuals living in the country, the boundaries for defining socioeconomic levels that properly accounts for all levels of wealth are taken from the census, which is based on quarterly income collected on national surveys. Once the income levels from the census are applied, the raw number of people that deposit their income and spend money with their debit card, are mainly wealthy. As explained in section 1, this is precisely a segment of the population that is less likely to answer surveys and thus, both approaches can be used to validate and/or complement each other. Thus, special attention must be put into understanding the limitations of each available data source, like the banking dataset used in this analysis, which after careful cleaning is mostly representative of wealthy individuals. Therein, lies the value of searching for alternate and complementary data sources such as mobile phone data that can help complete and validate the vision of the population as a whole.

***Social Network Analysis and Results*:** According to **[10]**, friends in the same social group often belong to the same SEG. This assumption is checked with the available data using a social network analysis, by identifying the friends of each individual and comparing their socioeconomic levels as determined in the previous section. Two metrics are used to quantify friendship between individuals from their phone calls: i) the number of times that they were in touch, and ii) the total length of phone calls between those individuals measured in seconds **[11]**. Each metric is normalized with respect to their maximum value (per individual) and then divided by a factor of 2, so that each normalized metric takes a value up to 0.5 and the sum of the two metrics is less or equal than 1. This is our combined metric for friendship. Next, the 75th percentile (on an individual basis) is used as a threshold to end up with a list of friends for each user. The SEG of each friend is available from the combined data and Fig. 1 shows the distribution of friends according to SEG. It's clear from the figure that individuals mainly make friends with people in their own and neighboring groups. This robust distribution of friends among SEG groups indicates that social network analysis of phone data and other social networks could be used to determine an individual's SEG from that of his/her contacts. This could be especially useful to extend data about consumption and income when the coverage is for only a fraction of the population.

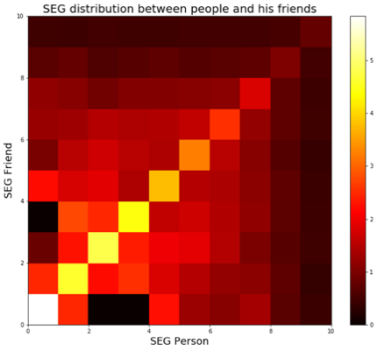

Fig. 1. **Friends' concentration by SEG.** For the SEG of each individual (horizontal axis), the ratio *r* of number of his/her friends per SEG group to what would be expected from a random distribution is plotted using a logarithmic scale. Ratios are overwhelmingly higher in the diagonal, indicating that friends are typically concentrated in the same SEG as the individual.

III. Consumption Patterns and Comparison with Surveys

Spending transactions in the BDS are identified by their MCC. Nevertheless, national household surveys of consumption behavior normally make use of the Classification of Individual Consumption by Purpose (COICOP) categories **[5]**. COICOP is often used by institutions like United Nations or local governments to measure the consumption patterns of the population. As a consequence, it was necessary to map the MCC of each transaction into the corresponding COICOP category **[12]**. The individual consumption vector is defined as the fraction of spending in each of the 12 COICOP categories and the average consumption vector per SEG group is calculated from its individual consumption vectors. The spending profile of each SEG in terms of the 12 consumption categories is shown in Fig. 2. The data shows that the 10th SEG, which consist of the wealthiest individuals, spend a smaller fraction of their money in food and non-alcoholic beverages (1st category) and in transport (7th category), in comparison with the other groups; but spend more in the "miscellaneous" category (12th category, which includes financial services), and restaurants (11th category). Furthermore, for SEGs in the low and middle part of the spectrum, spending in department stores and wholesale clubs represent the bulk of transactions in "miscellaneous". The distance between normalized consumption vectors is calculated according to **[9]** with identical findings in regards to significant differentiation among SEGs and low intragroup dispersion, thus confirming that consumption vectors are a robust proxy to understand the behavior of each socioeconomic group.

Finally, the consumption vectors calculated in this analysis from BDS are compared to equivalent consumption vectors from national surveys. The values were calculated subtracting the spending percentage of a certain category according to the banking data with respect to the result from the national surveys. Thus, positive values are indicative of people spending less money in this category than the surveys say. Fig. 3 shows overall differences of 15% or less across all SEGs. Outstanding

differences in housing expenses and miscellaneous goods and services (which are in the same direction for all groups) can be attributed to bias in the use of cash with respect to debit/credit cards as a form of payment. Presently, consumption data from cash payments is being added to the current analysis in order to obtain a full picture of individual spending through their different facets.

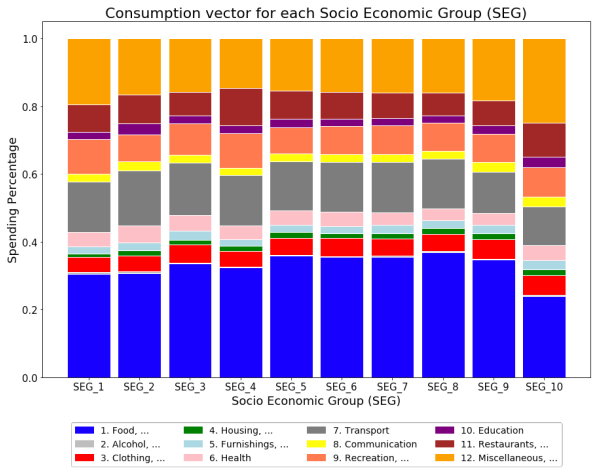

Fig. 2. **Spending profile for each SEG** in terms of COICOP categories.

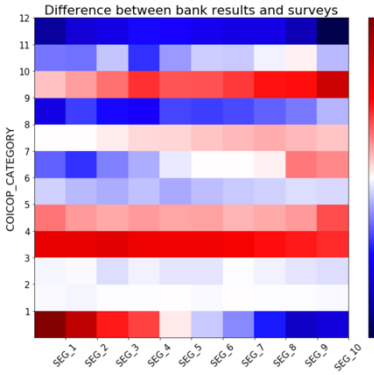

Fig. 3. **Differences in the spending profiles obtained from national surveys and those calculated from BDS**. Rows correspond to the COICOP categories shown in Fig.2. A positive difference indicates that the spending according to national surveys is greater than the value calculated from the banking dataset.

IV. Conclusions

This paper makes use of anonymized bank transactions of millions of individuals to classify them into socioeconomic groups according to their income. This was accomplished after a careful and comprehensive cleaning of the data to minimize bias and provide robustness to the results. Spending and income are correlated, thus providing a two-pronged approach to classify individuals: based on total spending **[9]** and based on income as shown here. Banking data like the one used in this analysis is dominated in raw numbers by wealthy individuals, which was expected since poor individuals are less likely to use banking services. In terms of determining the consumption behavior of individuals this is both a challenge, since lower income individuals are poorly represented in the sample, and an opportunity, since wealthy individuals are less likely to answer surveys designed with this goal. Phone call data from the same individuals are used to validate the classification found from banking data through a social network analysis and to explore possible ways of using geographical information from the mobile phone antennas to characterize those individuals. The social network analysis corroborates the socioeconomic classification based on income and shows that this type of new data sources could be used as a proxy to extrapolate income level (and consumption behavior as discussed below) from data-rich individuals to other individuals in their networks for which not enough data is available.

The consumption patterns of each socioeconomic group were constructed as an average spending profile based on properly vetted bank transactions instead of the self-reporting typically used in national surveys. This required the mapping between MCC and COICOP categories, which is in itself a challenging and important task, since the lack of universal standards obscures the correct determination of spending breakdown for each one of the SEGs. The similarity analysis for consumption vectors and cash retrieval confirms that individuals within the same SEG behave similarly, while individuals from socioeconomic groups from opposite ends of the spectrum have very different consumption patterns.

V. References

[1] Deaton, A. (1993) Understanding Consumption. Oxford University Press.

[2] Deaton, A. (2001). Counting the world's poor: problems and possible solutions. *The World Bank Research Observer*, *16*(2), 125-147.

[3] Deaton, A. (1997). *The analysis of household surveys: a microeconometric approach to development policy*. World Bank Publications.

[4] Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public opinion quarterly*, *70*(5), 646-675.

[5] United Nations Statistics Division (1999) Classification of Individual Consumption According to Purpose (COICOP). https://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=5

[6] Deaton, A. (2005). Measuring poverty in a growing world (or measuring growth in a poor world). Review of Economics and statistics, 87(1), 1-19.

[7] Gasparini, L., & Glüzmann, P. (2012). Estimating income poverty and inequality from the GallupWorld Poll: The case of Latin America and the Caribbean. *Journal of Income Distribution*, *21*(1), 3-27

[8] American Express (2008) Merchant Category Codes and Groups Directory. Available from: https://amex.co/2JHbLHH.

[9] Leo Y, Karsai M, Sarraute C, Fleury E (2016) *Correlations of consumption patterns in social-economic networks.* In: Proceedings IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

[10] Soto V., Frias-Martinez V., Virseda J., Frias-Martinez E. (2011) *Prediction of Socioeconomic Levels Using Cell Phone Records*. In: Konstan J.A., Conejo R., Marzo J.L., Oliver N. (eds) User Modeling, Adaptation and Personalization. UMAP 2011. Lecture Notes in Computer Science, Vol. 6787. Springer, Berlin, Heidelberg.

[11] Eagle N, Pentland A, Lazer D. (2009) *Inferring friendship network structure by using mobile phone data.* PNAS September 8, 2009. 106 (36) 15274-15278; https://doi.org/10.1073/pnas.0900282106

[12] Ernst & Young (2016). *Reducing the Shadow Economy through Electronic Payments, Technical Appendices.* Available from: https://go.ey.com/2GWjDH