News or Social Media? Socioeconomic divide of mobile service consumption

Iñaki Ucar, 1 Marco Gramaglia, 1,2 Marco Fiore, 3
Zbigniew Smoreda, 4 Esteban Moro 1,5,6*

1UC3M-Santander Big Data Institute,
Universidad Carlos III de Madrid, 28903 Getafe, Spain
2Department of Telematic Engineering,
Universidad Carlos III de Madrid, 28911 Leganés, Spain
3IMDEA Networks Institute, 28918 Leganés, Spain
4Sociology and Economics of Networks and Services Department,
Orange Labs, 92320 Châtillon, France
5Grupo Interdisciplinar de Sistemas Complejos, Department of Mathematics,
Universidad Carlos III de Madrid, 28911 Leganés, Spain
6Connection Science, Institute for Data Science and Society,
MIT, Cambridge, MA 02139

*To whom correspondence should be addressed; E-mail: emoro@math.uc3m.es.

Abstract

Reliable and timely information on socioeconomic status and divides is critical to social and economic research and policing. Novel data sources from mobile communication platforms have enabled new cost-effective approaches and models to investigate social disparity, but their lack of interpretability, accuracy or scale has limited their relevance to date. We investigate the divide in digital mobile service usage with a large dataset of 3.7 billion time-stamped and geo-referenced mobile traffic records in a major European country, and find profound geographical unevenness in mobile service usage – especially on news, e-mail, social media consumption, and audio/video streaming. We relate such diversity with income, educational attainment, and inequality, and reveal how low income or low education areas are more likely to engage in video streaming or social media, and less in news consumption, information searching, e-mail, or audio streaming. The digital usage
gap is so large that we can accurately infer socioeconomic status of a small area or even its Gini coefficient only from aggregated data traffic. Our results make the case for a cheap, privacy-preserving, real-time, and scalable way to understand the digital usage divide and, in turn, poverty, unemployment, or economic growth in our societies through mobile phone data.

Introduction

Inequality is a central societal problem, especially within rapidly expanding urban areas. While it is a crucial driver for economic growth [1], the progressive clusterization of workers, industries, companies, and services in cities has a tremendous cost in terms of segregation and discrimination. This cost is not only economic: in the same city, different areas can have a 10-to-15-year imbalance in life expectancy and highly divergent education levels, with little chances of social mobility [2]. The design and successful implementation of policies to alleviate those problems require fine-grained, frequently updated information about income, education, or inequality across metropolitan areas. However, most data sources employed today, like population censuses or surveys, suffer from sparsity in population coverage or infrequent updating, hence do not allow following the swift evolution that urban societies nowadays experience. Thus, the traditional ways of understanding cities tend to explain what happened five years earlier rather than nowcasting or even predicting urban transformations.

In recent years, digital data have been proposed as an alternative source for socioeconomic status (SES) inference [3, 4, 5]. The escalating use of mobile devices [6, 7, 8, 9], social media [10], or credit cards [11], and the growing availability of pervasive satellite imagery [12, 13] have allowed researchers to build SES models with unprecedented temporal and spatial resolutions. For example, income levels in urban areas were correlated with the unequal presence of trucks [14] or utilization of construction materials [15] extrapolated from imagery data. Similarly, the diversity in human mobility or social interactions observed in data from mobile phones
or social media was found to be correlated with higher income [8, 10]. However, while very successful in predicting SES in developing countries [12, 7, 9], these approaches are only moderately accurate in developed countries [8, 16], where variances in the penetration of mobile phones, in the use of credit cards, and in social segregation itself are more nuanced.

When considering economic and social inequality in wealthier countries, we argue that specific mobile services’ consumption may be a more suitable proxy for SES than other digital data considered to date. Indeed, a diffuse preference for particular mobile applications is a subtler indicator than the sheer adoption of mobile digital technologies, as it connects to finer-grained user traits like personal interests, digital skills, or accessibility to paying services [17, 18].

Several previous studies provide some evidence that corroborates our postulation. For instance, it has been hypothesized that mobile service usage can reveal the digital divide between different socioeconomic, gender or age groups [17]. There is qualitative confirmation that mobile digital usage might exacerbate socioeconomic inequalities given the impact that social media and online information resources have on the social, political, and economic aspects of our society [19, 18]. It is also known that time on some social platforms, watching videos or playing videogames [20] or news media consumption patterns [21] depends on user’s SES, or that students’ performance is related to different patterns in their Internet usage [22, 23]. All these studies suggest a significant disparity in how mobile services are consumed, even in developed economies where the technology access gap is not significant. Nevertheless, the limited scale and small granularity of existing studies do not allow forming a conclusive opinion on the magnitude of such a mobile application usage gap nor understand its repercussions on SES features.

In this paper, we present the first large-scale, quantitative study of the relationship between mobile service adoption and socioeconomic inequality. To that end, we analyze nationwide data traffic measurements collected by the leading mobile operator in a major European country (France), and find fundamental imbalances in the relative usage of specific mobile applications
by different income or education groups. The mobile service consumption gap is so profound that we can build fairly accurate models based on mobile traffic to estimate income, education level, and economic inequality at high spatial resolution.

Results

Our data consist of around 3.7 billion time-stamped and geo-referenced records of the mobile traffic generated by different applications, such as YouTube, Facebook or Netflix – including device-specific ones like Apple Store (run by iOS devices) or Google Play (run by Android devices). The data was collected between May and June 2017 over the whole France, and aggregated at the Base Station (BS) level. Due to their volume and scattered nature, some traffic from different applications were aggregated to common categories like mail, gaming, news consumption (mainly newspapers outlets) or audio streaming (see Supplementary Text and Table S1). We merge the per-service traffic volume recorded in each BS coverage area with socioeconomic indicators gathered from the 2014-2015 census, which include information about the income and population structure in each IRIS zone, i.e., French sub-municipal statistical unit (see Methods). The combination of the two data sets is performed via an areal interpolation that maps mobile traffic over BS coverage areas into IRIS zones, see Fig. 1.

Since our traffic data is collected by BS, it includes app usage by residents of that area and users from other areas that visit that BS throughout the day. To link traffic data to the residents of a particular statistical area, we implemented a temporal consolidation of our data in which we only consider the mobile service usage recorded during hours in which we can safely consider people to be at home, i.e., between 8pm and 7am during weekdays (see Methods and Supplementary Text).

The various mobile applications inherently entail very different traffic volumes: for instance, YouTube video streaming sessions consume much more data than Twitter messages. Therefore,
plain traffic byte counts per inhabitant are not comparable across services, and tend to conceal subtle differences in usage patterns as exemplified in Fig. 1 and S1. In order to bring patterns in the consumption of individual applications to the foreground, we use the Revealed Comparative Advantage (RCA) [24] to normalize the aggregated traffic by IRIS area and service. RCA measures the ratio between the share of traffic generated by an application in a certain IRIS area, and the same share computed in the whole country: it can thus reveal higher or lower relative adoptions of specific mobile services in a given area with respect to the national average.

The spatial, temporal and scale consolidation of the data outlined before allows revealing a structure of correlations in the usage of mobile service across geographical areas that was not recognized to date, and is displayed in Fig. 2. Specifically, previously observed strong correlations among different byte-count traffic flows [25] dissolve into a fabric of mild pairwise correlations and anti-correlations. We can clearly distinguish two groups of traffic flow RCAs which are loosely correlated within themselves and anticorrelated between them. We can easily recognize device-specific ones like the Apple Store, iCloud on one of them and their counterpart (Google Play) in the other. Beyond that, the former seems to be dominated by more information apps (Google, News, Mail), while the latter is composed by video streaming traffic or gaming. Social media traffic is different also across both groups: while Instagram and Twitter traffic flow seems to be more correlated with the News and Mail group, large Facebook or Snapchat usage co-occurs with generic video streaming and Google Play. Also gaming usage is different across groups, mainly concentrated in the group of high use of Facebook, Google Play and video streaming. The result highlights a pronounced spatial uniqueness in the consumption of each application, when relative usage is compared across different geographical units at a national scale.

In order to explore dependencies between such spatial diversity in mobile traffic and SES indicators, we gather three demographics variables in each IRIS area from census data: (i) the
median income, \((ii)\) the ratio of people with a professional activity that requires higher education (or higher education ratio, for short, hereafter), and \((iii)\) the Gini index of the income distribution, as a measure of local inequality. We model these three responses to try to explain them as a function of the relative traffic usage per category across areas. We use the population structure (i.e., population ratio by age ranges and immigrant ratio) as control variables in the framework of a generalized linear model weighted by the population counts in each area, with link functions specifically tailored to each response considered (Gamma regression with log link for median income; quasibinomial regression with logit link and fractional response for higher education ratio; and Beta regression with logit link for local inequality). All the regressors are standardized prior to model fitting. As for the spatial autocorrelation, the distribution of the response variables as well as the dimension of the problem (11k observations of 40 covariates) make traditional approaches (spatial lag/error models and eigenvector selection for semi-parametric spatial filtering) computationally unfeasible. Thus, we developed a hybrid approach between a spatial error model and spatial filtering, implemented in two stages: in a first stage, the model is fitted without taking into account the spatial dimension, which produces spatially autocorrelated residual deviances; then, these are spatially-lagged and re-introduced in a new fit as an additional auto-covariate. Our results show that this technique is not only much faster computationally, but also successfully filters the spatial autocorrelation in the final model (as measured by the Moran-I value), producing stable estimates (see Methods for further details).

Fig. 3a shows the quality of the regression on the three SES responses (i.e., median income, higher education ratio, and local inequality) using four sets of predictors: population (control) variables, normalized mobile service traffic, and both sets of variables without \((All)\) and with \((All+SF)\) spatial filtering. The left pane shows that the control variables alone explain a low ratio (in the 0.25–0.35 range) of the total variance, measured by an adjusted pseudo-\(R^2\), for
the three SES models. On the other hand, mobile application traffic features alone significantly predict SES responses (with up to 0.74 of variance explained for the higher education ratio). Jointly considering population and traffic variables, as well as adding spatial filtering, further improves the result: ultimately, 0.73, 0.84 and 0.87 of the variance can be predicted for local inequality, median income and higher education ratio, respectively. The right pane in Fig. 3a shows the mean absolute error (MAE), standardized by the mean response, so that the three models can be compared. Notably, the best model in terms of explained variance is the worst in terms of relative MAE, and vice versa. This can be explained by the much larger variability that the higher education ratio presents in comparison with the other two SES responses. As a consequence, this model, despite being very reliable when it comes to capture averages and general trends across spatial units (even for the traffic variables alone), is however less suitable for point estimates than the others. The overall predictive power for these models is depicted in Fig. 3b for the Paris metropolitan area. The three SES responses are bucketed in fine-grained categories. Predicted values, on the right, show an excellent agreement with the observed ones.

We compare the relative effect size of traffic variables, population control variables, and spatial filtering (SF) for the best models (All+SF) in Fig. 4, with 95% Confidence Intervals (CI). News and Facebook traffic stand out as key explanatory variables for all SES models, with especially high coefficients for the higher education ratio and median income. Their effect is however antithetical: a stronger usage of News applications positively correlates with income and education levels, whereas the increased usage of Facebook is associated with reduced income and education. Similar antagonistic behaviors are found in the specific groups of mobile services found in Fig. 2: for instance, a relatively higher consumption of WhatsApp, e-mail and audio streaming services are associated with higher income and education, but the increased use of Snapchat, video streaming, or adult services has an opposite effect. Gaming also has mixed relationship with SES: while Candy Crush is more used in areas with low income and
education ratio, the opposite happens for Clash of Clans. The large size effects identify for traffic variables suggest a deep usage gap of mobile phone data between different areas of income and educational attainment. On the other hand, traffic variables tend to have a different role when local inequality is concerned: as an example, audio and video streaming have reverse correlations (i.e., negative and positive, respectively) with this SES response. Especially good predictors of inequality are the adoption of iOS (i.e., Apple Store) devices in areas where income disparity is higher, and Android (i.e., Google Play) devices where the economic status of the population is instead more homogeneous. Putting together our results in Fig. 4, we can generally say that high income and education areas have relatively more traffic on information seeking tools (News, Google, Mail), e-commerce and audio-streaming, while areas with low SES indicators have higher relative traffic on social media activity and streaming (Facebook, Youtube, Snapchat).

Regarding population structure variables, most age groups show a similar positive effect on the responses, except for the higher education ratio, for which ages from 18 to 64 exhibit larger effects – as these are naturally the groups that had access to higher-level education. On the other hand, the ratio of immigrants is associated with lower levels of median income and education, as well as higher levels of inequality, which can be expected. Remarkably, this pattern is consistent with the estimates for mobile services like Facebook, generic messaging and Twitter. Finally, we quantify the relative importance of the spatial filtering in the same way as the rest of the variables in the model, although the method does not allow estimating an actual spatial correlation. In this sense, our estimates show that the spatial term is especially influential in the case of median income, but less important in the other responses.
Discussion

The data revolution has created an opportunity to scrutinize individual and collective behavior at an unprecedented scale, detail and speed. We have now the opportunity to measure, monitor and predict relevant aspects of socioeconomic status (SES) and growth in quasi real time by using satellite images, social media, or mobile phone data. More interestingly, some of those models relate socioeconomic development to meaningful measures of human behavior like diversity, expressed opinions, purchases, or the urban environment [13, 10, 8, 7]. Thus, not only they can be used to monitor human development, but also to understand the roots of socioeconomic status and their inequality. However, there seems to be a balance between predicting power and interpretability [13]. While machine learning models applied to satellite imagery and mobile phone data achieve typically high precision to explain SES [7], highly interpretable models based on diversity of mobility, purchases, content or other more interpretable metrics have less powerful explanatory power [8, 10, 13]. Our results show another dimension of human behavior obtained from mobile phone data, i.e., digital usage, can be used to achieve both high predicting power and interpretability of SES, even in developed countries. By just leveraging privacy-preserving aggregates of consumption of different services through mobile phones, we were able to have simple interpretable models for SES with high precision (≈ 80% of variance explained), larger than other models based on mobility diversity [8] or satellite imagery [13] for the same regions in France. Since our approach is complementary to those ones, there is a possibility that better precision can be obtained by combining our data with satellite imagery, for example. Finally, we found that usage gap is partly drawn around the already observed iOS/Android operating system divide [26], with a positive correlation for iPhone users and a negative one for Android devices. More importantly, we took a step further, and revealed that the gap goes beyond plain platforms, and roots deeply into the usage of different apps.
The success of our models is based on a dramatic difference of mobile phone usage behaviors across groups of different SES. The digital usage gap is so profound between low and high income or education areas, that it can be used clearly to distinguish between them or even identify the relative composition of those groups in a given area (Gini coefficient). High-income areas or those with larger education attainability show a more pronounced utilization of mobile devices to consume news, exchange e-mails, search for information, or listen to music. At the same time, they display a reduced use of some social media platforms or video streaming services. These results hold even when we control for age composition and other census variables like immigrant population. Although our models are equally accurate, the impact of the digital usage gap is more important on educational attainability. We can clearly see how regions that consume more Facebook content and less News have in general a lower fraction of population with higher education. This can be related to the two competing paradigms for online information consumption: the usage of traditional media vs. social media platforms. Social media has reshaped news by facilitating the involvement of audiences, and thus boosting engagement and dissemination [27]. Platforms such as Facebook and YouTube have been identified as the major pathways to the increasing habit of using social media as a news source [28, 29]. Even when perceived as unreliable, these platforms are used as “big outlets” for convenience, especially by young adults [30]. However, since we control for age composition, this is not strictly an effect of generational differences of social media and news usage. Rather, it might be related to how less-educated people consume news: for instance, U.S. adults who rely mostly on social media for news tend to have lower levels of education than those who mainly use several other platforms [31]. Another study in Chile found strong correlations of socio-demographics of users and online news media content [21]. Given that polarization and spreading of misinformation is more likely on social media [19], our results could be also used to identify those populations and areas which could be more susceptible to those problems.
Following the Bourdieusian framework [32], we can assume that the practices of individuals in the field of mobile Internet highlight interrelations between economic resources and social positioning—and, probably, internalized abilities. For example, in the analysis of the digital activities of Italian youth according to their social background, Micheli [33] finds that, while information seeking is positively correlated with the cultural capital of the students and the professional status of the parents, this is not the case for social media use. Adolescents from disadvantaged social backgrounds are more likely to actively participate in social media than adolescents from upper strata. Her interpretive analysis of qualitative data indicates that upper-middle class students replicate their parents’ attitudes toward the Internet as a tool for personal enrichment.

Finally, it is worth noting that our results are based on a fully privacy-preserving analysis of mobile phone data. While other metrics based on user mobility and communications need individual or high-resolution data, our variables are constructed using aggregates of traffic at network base stations. Such variables are fully compliant with the General Data Protection Regulation (GDPR), since they typically blend in a non-reversible way data generated by hundreds of users, hence do not incorporate any personal information and hinder the possibility to de-anonymize individual information. Also, they are compact enough to enable very large-scale analyses like the one we carried out, and they are relatively simple to collect for mobile network operators, easing the permanent availability of the statistics for longitudinal studies. More importantly, since our analysis is complementary to the ones using other dimensions of mobile phone data (mobility, diversity of communications), we believe our results will foster new analysis in the future about the relationships between different aspects of access to information, human communication and mobility and their impact on human development and SES.

Although our results are descriptive and do not imply causal relations, we believe that our findings could be used to point to important and previously overlooked factors of socioeconomic
inequality whose causal effect may be further tested through carefully designed experiments, interventions, or digital regulations. For example, the fact that low income or educational attainment is correlated with groups of services like social media, video streaming or messaging could be use to devise successful holistic interventions to minimize their use and promoting other mobile phone usages.

**Materials and Methods**

**Mobile service traffic data**

The network traffic dataset employed by our study comprises usage statistics of popular mobile applications. Data entries record as the uplink (data transmitted by the user device) and downlink (data flowing to the user device) byte counts per service, at a temporal granularity of 5 minutes and aggregated by Base Station (BS). The data was collected by Orange France within its own infrastructure during 1.5 months in May and June 2017. They describe the mobile behavior of the whole Orange subscriber base in France, i.e., approximately 15 million individuals distributed over more than 550,000 km², and served by over 25,000 BS. Usage statistics were collected by passive probes monitoring user sessions; the specific mobile service associated to each session was detected using Deep Packet Inspection (DPI) and fingerprinting techniques tailored to specific traffic types (see the SI Appendix for further details). The final dataset made available by the operator included the 40 services that generate the most traffic in the network, as detailed in Fig. S1.

**Geographical Data and Socioeconomic Indicators**

We used geographical information and census data from the French *Institut national de l’information géographique et forestière* (IGN), publicly available in their web pages. For the geographical description, we downloaded the *Contours IRIS édition 2016* dataset, which defines a polygon
in a Lambert-93 projection for each IRIS zone (i.e., aggregated unit for statistical information) in France, as well as an associated record containing the IRIS code, name and type among other information. For the population structure, we downloaded the *Population en 2015* dataset, which contains a description of the population structure by age group and other factors, such as socio-professional category and immigration. For the economic indicators, we downloaded the *Revenus, pauvreté et niveau de vie en 2014 (IRIS)* dataset, which contains a complete description of the income distribution deciles for residential IRIS zones. These are areas with more than 1,000 inhabitants, and their population generally falls between 1,800 and 5,000. Indicators for areas with less than 1,000 are not shared by the IGN due to privacy reasons.

**Areal Consolidation**

The coverage area of each BS in the Orange mobile network is modelled via a Voronoi tessellation that uses the BS location as the object positions on the geographical space. Such BS coverage areas have a different geometry than the IRIS zones for which income and population data are available; generally, coverage areas are much smaller than IRIS zones in urban centers, but the opposite occurs in the countryside and less-populated regions of the country. To spatially consolidate the data, we adopted an *areal-weighted interpolation* procedure to transfer BS-level traffic counts into IRIS zones. As exemplified in Fig. 1, the principle is computing the intersection between the two spatial bases, and then create a many-to-one mapping of BS coverage sub-areas to IRIS zones (i.e., determining which IRIS zones each BS coverage area intersects with) plus a set of associated areal weights (i.e., the surface fraction of original BS coverage area that falls in each BS sub-area). By assuming that mobile service traffic is evenly distributed within the BS coverage area, traffic counts for each BS sub-area are calculated as the areal weight multiplied by the total traffic recorded for the BS, for each service. Finally, the traffic counts for all relevant BS sub-areas are aggregated for each IRIS zone. After filtering out
IRIS zones without economic indicators, we have mobile service traffic data for 11,806 IRIS zones (out of 49,404), which encompass all main urban areas of France as depicted in Fig. S2. Classified by their degree of urbanization (according to Eurostat), we find that 78% of the IRIS zones in the final dataset correspond to urban areas, 19% are peri-urban areas and 3% are rural areas.

**Temporal Consolidation**

A mismatch between traffic and socioeconomic datasets exists also in the temporal dimension, due to the inherent *mobile* nature of the consumption of applications on portable devices opposed to the *static* character of census indicators. We resolve the discrepancy by only considering the mobile service usage that is most likely produced by users when they are at their locations of residence – which their socioeconomic indicators also refer to. More precisely, we filter out weekends and French holidays (May 25th and June 5th in the period considered), and we keep observations during home hours (from 8pm to 7am) on weekdays. As we show in the SI Appendix, our results are robust to that definition of home hours. Moreover, there is evidence that app usage peaks from 8pm [34], and that online consumption is more or less homogeneous throughout the day [35]. Thus, although there could be important differences of traffic during the day for individuals, we believe that our aggregate consumption data by area is highly representative of the daily online consumption of the population of the area.

**Scale Consolidation**

Different mobile applications generate heterogeneous volumes of network traffic depending on the nature of the data transferred (e.g., video streaming creates a much higher load per session than messaging) and popularity (with widely adopted services producing a much higher demand than niche ones). This results in diverse scales for traffic counts across services, which can span
several orders of magnitude as observed in Fig. S1. In addition, as shown in Fig. 1, raw byte counts are highly correlated across different mobile services, both spatially and temporally.

The scale mismatch and spatiotemporal correlation tend to hide differences in mobile service consumption. In order to give prominence to any such diversity, we aim at adopting a relative metric of the traffic with the property of being comparable across spatial zones and applications. Firstly, we consider the downlink byte counts for all services, which is aggregated on a hourly basis and normalized by census population. We then take the median values of the downlink bytes/inhabitant/hour during the whole 1.5-month observation period, for each IRIS zone and mobile service. Finally, we calculate the Revealed Comparative Advantage (RCA) [24] as follows:

\[
RCA_{ij} = \frac{T_{ij}}{T_i} \frac{T_i}{T_j} = \frac{T_{ij}}{T_j},
\]

where \(T_{ij}\) is the median hourly traffic per inhabitant in zone \(i\) for application \(j\); \(T_i\) is the median hourly traffic per inhabitant in zone \(i\) jointly generated by all considered applications; \(T_j\) is the median hourly traffic per inhabitant generated by service \(j\) in all zones at once; \(T\) is the median hourly traffic per inhabitant, aggregated over all zones and services. The index in Eq. (1) measures the proportion of traffic generated by a particular mobile application in a specific IRIS zone, normalized by the fraction of global (i.e., over all zones) traffic imputed to that same application. An advantage of service \(j\) is revealed in IRIS zone \(i\) if \(RCA_{ij} > 1\), implying a higher-than-ordinary usage of service \(i\) in area \(j\); conversely, if \(RCA_{ij} < 1\), application \(j\) presents a comparative disadvantage, i.e., a reduced adoption with respect to the national average, in zone \(i\). The metric allows measuring all traffic features in a common unit, and reveals a structure of mild correlations and anti-correlations, shown in Fig. 2, which is instead concealed by uniform strong interdependence when considering raw byte counts.
Multicollinearity handling

The RCA transformation is a relative measure of importance, and, as such, we have to drop at least one variable to avoid a perfect fit, i.e., that every RCA$_{ij}$ is an exact linear combination of remaining RCA$_{ik}$, $\forall k \neq i$. We dropped a No info service, which gathers traffic generated by unknown applications.

To diagnose the presence of multicollinearity in the remaining set of variables, we compute the Variance Inflation Factor (VIF) for each individual RCA$_{ij}$ using the median income as the dependent variable. This method reports a high level of multicollinearity, with an average VIF of $\sim 4$ and a median VIF of $\sim 27$ across variables.

Therefore, we proceed to manually remove a few residual traffic categories and uninformative ones that are of little interest from the behavioral perspective. Those are Pokemon Go, Other(s), Ads, Updates, Encrypted web and Generic web. After dropping these variables, 32 services remain, which are listed in Fig. 4; the same diagnostic run on the lasting variables reports average and median VIF of $\sim 2$, which corresponds to a low level of multicollinearity.

Regression Models

We consider two socioeconomic indicators in IRIS zones that are available in the public datasets, i.e., the median income and the ratio of people with a professional activity that requires higher education, or higher education ratio for short; in addition, we consider a third inequality indicator in the form of the Gini index computed from the income data (see the SI Appendix for further details). We model the dependency of the indicators on mobile service usage via a generalized linear model:

$$g\left(E[y_i]\right) = \alpha_0 + \sum_j \beta_j \cdot \text{RCA}_{ij} + \sum_k \gamma_k \cdot \text{POP}_k + \delta \cdot \text{SP}_{err},$$

(2)
where $y_i \equiv \{\text{income, education, inequality}\}$ for zone $i$ is modelled after the $RCA_{ij}$ values for each application $j$. $POP_k$ are control variables from the population structure, i.e., the ratio of inhabitants in the 11-17, 18-24, 25-39, 40-54, 55-64, 65-79, and 80+ ranges, plus the ratio of immigrant population. Both groups of regressors, $RCA_{ij}$ and $POP_k$, are standardized: scaled by the square root of the second raw sample moment of the whole group, so that the coefficient estimates and effect sizes across groups are comparable. The link function $g$ is tailored to the distribution of each response:

- Median income is a positive-definite continuous response that can be modelled after a Gamma function. Thus, we perform Gamma regression with log link, and estimates are interpreted as a means-ratio.

- Higher education ratio is the proportion of people with a professional activity that requires higher education, which is a counting process that may be overdispersed. Thus, we define a fractional model, i.e., a quasibinomial regression with logit link and fractional response, and estimates are interpreted as an odds-ratio.

- Local inequality is measured with the Gini coefficient, which can be modelled after a Beta distribution. Thus, we perform a Beta regression with logit link, and estimates are interpreted as an odds-ratio.

Finally, the high Moran-I value for each response (0.72, 0.81 and 0.74 respectively; see Table S2) justifies the use of a spatial model. Therefore, $SP_{err}$ is a variable created to filter the spatial correlation, and is defined as the spatially-lagged residual deviance of the rest of the model:

$$SP_{err} = W r_i$$

(3)

where $W$ is the row-standardized matrix of queen-contiguity spatial weights and $r_i$ is the deviance residuals for an initial fit with the rest of the variables involved (see the SI Appendix for
These models are thus fitted in four stages: (1) a reference fit with the population variables alone, which serves as a null model; (2) a second fit with traffic variables alone, to explore their explanatory power; (3) a complete model with both traffic and population variables; (4) a final model that performs spatial filtering by taking the deviance residuals $r_i$ from (3), which show spatial correlation, and incorporating them as $SP_{err}$ in a new fit. We checked that point estimates for $RCA_{ij}$ and $POP_k$ in (3) and (4) are very similar, but (4) succeeds in filtering out the spatial correlation ($p < 0.001$ for the Moran’s I test), thus producing better results and more precise and stable coefficients (see Tables S3-S6).

**Ethical Considerations**

The data from the Orange network probes used in this work were collected as part of the ABCD – Adaptive Behavior and Cloud Distribution collaborative research project founded by the French National Research Agency (ANR). The collection of this personal data has been authorized by the Data Protection Officer (DPO) of Orange according to article 89 of the European Union General Data Protection Regulation (GDPR), which provides an exemption for research, in particular for scientific and research purposes. The data were collected and processed exclusively on the Orange Labs secure Big Data platform. The data were processed in a server located in the operator premises, and accessible only to authorized researchers, and aggregated at the BS level so as to remove all privacy risk for individuals. All source data were deleted 12 months after the collection.

**Data Availability**

The data that support the findings of this study are available from Orange, but restrictions apply to the availability of these data, which were used under license for the current study, and so are
not publicly available. Data are however available from the authors upon reasonable request and with permission of Orange.

References

[1] Bettencourt LMA, Lobo J, Helbing D, Kühnert C, West GB. Growth, innovation, scaling, and the pace of life in cities. Proceedings of the National Academy of Sciences. 2007;104(17):7301 – 7306. Available from: https://www.pnas.org/content/104/17/7301.

[2] Chetty R, Hendren N, Katz LF. The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment. The American economic review. 2016;106(4):855 – 902.

[3] Soto V, Frias-Martinez V, Virseda J, Frias-Martinez E. Prediction of socioeconomic levels using cell phone records. In: International conference on user modeling, adaptation, and personalization. Springer; 2011. p. 377–388.

[4] Gao J, Zhang YC, Zhou T. Computational socioeconomics. Physics Reports. 2019;817:1–104. Computational Socioeconomics. Available from: https://www.sciencedirect.com/science/article/pii/S0370157319301954.

[5] Dong L, Ratti C, Zheng S. Predicting neighborhoods’ socioeconomic attributes using restaurant data. Proceedings of the National Academy of Sciences. 2019 07;116(31):15447–15452.

[6] Blondel VD, Decuyper A, Krings G. A survey of results on mobile phone datasets analysis. EPJ Data Science. 2015 08;4(1):10.
[7] Blumenstock JE. Fighting poverty with data. Science. 2016;353(6301):753–754. Available from: https://science.sciencemag.org/content/353/6301/753.

[8] Pappalardo L, Vanhoof M, Gabrielli L, Smoreda Z, Pedreschi D, Giannotti F. An analytical framework to nowcast well-being using mobile phone data. International Journal of Data Science and Analytics. 2016 Jun;2(1-2):75–92. Available from: https://link.springer.com/article/10.1007/s41060-016-0013-2.

[9] Steele JE, Sundsøy PR, Pezzulo C, Alegana VA, Bird TJ, Blumenstock J, et al. Mapping poverty using mobile phone and satellite data. Journal of Royal Society Interface. 2017;14(127):20160690. Available from: http://rsif.royalsocietypublishing.org/content/14/127/20160690?utm_source=social&utm_medium=hootsuite&utm_campaign=standard.

[10] Llorente A, Garcia-Herranz M, Cebrian M, Moro E. Social Media Fingertips of Unemployment. PLoS ONE. 2015;10(5):e0128692.

[11] Hashemian B, Massaro E, Bojic I, Arias JM, Sobolevsky S, Ratti C. Socioeconomic characterization of regions through the lens of individual financial transactions. PLoS ONE. 2017;12(11):e0187031. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0187031.

[12] Jean N, Burke M, Xie M, Davis WM, Lobell DB, Ermon S. Combining satellite imagery and machine learning to predict poverty. Science. 2016;353(6301):790 – 794. Available from: http://science.sciencemag.org/content/353/6301/790.
[13] Abitbol JL, Karsai M. Interpretable socioeconomic status inference from aerial imagery through urban patterns. Nature Machine Intelligence. 2020 Oct;2(11):684–692. Available from: https://doi.org/10.1038/s42256-020-00243-5.

[14] Gebru T, Krause J, Wang Y, Chen D, Deng J, Aiden EL, et al. Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. Proceedings of the National Academy of Sciences. 2017;114(50):13108–13113.

[15] Engstrom R, Hersh J, Newhouse D. Poverty from Space: Using High-Resolution Satellite Imagery for Estimating Economic Well-Being. The World Bank; 2017. Available from: https://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-8284.

[16] Xu Y, Belyi A, Bojic I, Ratti C. Human mobility and socioeconomic status - Analysis of Singapore and Boston. Computers, Environment and Urban Systems. 2018;72:51 – 67.

[17] Hargittai E, Hsieh YP. Digital Inequality. In: Dutton WH, editor. The Oxford Handbook of Internet Studies. Oxford University Press; 2013. Available from: http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780199589074.001.0001/oxfordhb-9780199589074-e-7.

[18] Tsetsi E, Rains SA. Smartphone Internet access and use: Extending the digital divide and usage gap. Mobile Media & Communication. 2017;5(3):239–255.

[19] Aral S. The Hype Machine: How Social Media Disrupts Our Elections, Our Economy, and Our Health—And How We Must Adapt. Crown/Archetype; 2020. Available from: https://books.google.es/books?id=N1v6DwAAQBAJ.
[20] Rahmati A, Tossell C, Shepard C, Kortum P, Zhong L. Exploring iPhone usage. In: the 14th international conference. New York, New York, USA: ACM Press; 2012. p. 11. Available from: http://dl.acm.org/citation.cfm?doid=2371574.2371577.

[21] Vilella S, Paolotti D, Ruffo G, Ferres L. News and the city: understanding online press consumption patterns through mobile data. EPJ Data Science. 2020;9(1):10.

[22] OECD. Are there differences in how advantaged and disadvantaged students use the Internet? PISA in focus. 2016:64.

[23] Walsh JL, Fielder RL, Carey KB, Carey MP. Female College Students’ Media Use and Academic Outcomes. Emerging Adulthood. 2013;1(3):219–232.

[24] Balassa B. Trade Liberalisation and “Revealed” Comparative Advantage. The Manchester School. 1965;33(2):99–123. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9957.1965.tb00050.x.

[25] Barthélemy M, Gondran B, Guichard E. Large scale cross-correlations in Internet traffic. Phys Rev E. 2002 Nov;66:056110. Available from: https://link.aps.org/doi/10.1103/PhysRevE.66.056110.

[26] Jamalova M, Constantinovits M. The Comparative Study of the Relationship Between Smartphone Choice and Socio-Economic Indicators. International Journal of Marketing Studies. 2019;11(3):11.

[27] Bowd K. In: Social media and news media: Building new publics or fragmenting audiences? University of Adelaide Press; 2016. p. 129–144. Available from: http://www.jstor.org/stable/10.20851/j.ctt1t304qd.13.
[28] How social media is reshaping news. Pew Research Center, Washington, D.C.; 2014. Available from: https://www.pewresearch.org/fact-tank/2014/09/24/how-social-media-is-reshaping-news/.

[29] News Use Across Social Media Platforms in 2020. Pew Research Center, Washington, D.C.; 2021. Available from: https://www.journalism.org/2021/01/12/news-use-across-social-media-platforms-in-2020/.

[30] Measuring News Consumption in a Digital Era. Pew Research Center, Washington, D.C.; 2020. Available from: https://www.journalism.org/2020/12/08/measuring-news-consumption-in-a-digital-era/.

[31] Mitchell A, Jurkowitz M, Olinphant J, Shearer E. Americans who mainly get their news on social media are less engaged, less knowledgeable; 2020. Accessed: 2021-01-20. https://www.journalism.org/2020/07/30/americans-who-mainly-get-their-news-on-social-media-are-less-engaged-less-knowledgeable/.

[32] Bourdieu P, Nice R. Distinction: A Social Critique of the Judgement of Taste. Polity Short Introductions. Harvard University Press; 1984. Available from: https://books.google.es/books?id=nVaS6gS9Jz4C.

[33] Micheli M. What is new in the digital divide? Understanding internet use by teenagers from different social backgrounds. In: Communication and information technologies annual. Emerald Group Publishing Limited; 2015. .

[34] Hoch D. App Usage Peaks At 8 P.M.; 2015. Available from: https://www.business2community.com/mobile-apps/app-usage-peaks-8-p-m-01153062.
[35] Murnane EL, Abdullah S, Matthews M, Kay M, Kientz JA, Choudhury T, et al. Mobile manifestations of alertness: Connecting biological rhythms with patterns of smartphone app use. In: Proceedings of the 18th international conference on human-computer interaction with mobile devices and services; 2016. p. 465–477.

[36] Kumar A, Sengupta J, Liu Yf. 3GPP LTE: The future of mobile broadband. Wireless Personal Communications. 2012;62(3):671–686.

[37] Alani MM. Guide to OSI and TCP/IP Models. Springer International Publishing; 2014. Available from: https://doi.org/10.1007/978-3-319-05152-9.

[38] Belgorodski N, Greiner M, Tolksdorf K, Schueller K. rriskDistributions: Fitting Distributions to Given Data or Known Quantiles; 2017. R package version 2.1.2. Available from: https://CRAN.R-project.org/package=rriskDistributions.

[39] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2021. Available from: https://www.R-project.org/.

[40] Imdadullah M, Aslam M, Altaf S. mctest: An R Package for Detection of Collinearity among Regressors. The R Journal. 2016;8(2):499–509. Available from: https://journal.r-project.org/archive/2016/RJ-2016-062/index.html.

[41] Cribari-Neto F, Zeileis A. Beta Regression in R. Journal of Statistical Software. 2010;34(2):1–24.

[42] Bivand RS, Pebesma E, Gomez-Rubio V. Applied spatial data analysis with R, Second edition. Springer, NY; 2013. Available from: http://www.asdar-book.org/.
[43] Bivand R, Wong DWS. Comparing implementations of global and local indicators of spatial association. TEST. 2018;27(3):716–748. Available from: https://doi.org/10.1007/s11749-018-0599-x.

Acknowledgments

This work has been supported by the research project CANCAN (Content and Context based Adaptation in Mobile Networks), grant number ANR-18-CE25-0011, funded by the French National Research Agency (ANR). The work of M.F. was partially supported by the Atracción de Talento Investigador grant number 2019-T1/TIC-16037 NetSense, funded by Comunidad de Madrid. E.M. and I.U. acknowledge partial support by Ministerio de Economía, Industria y Competitividad, Gobierno de España, grant number FIS2016-78904-C3-3-P and PID2019-106811GB-C32.

Supplementary materials

Materials and Methods
Supplementary Text
Figs. S1 to S4
Tables S1 to S6
References (36-43)
Fig. 1: Areal interpolation infographic. The mobile traffic data set comprises mobile service usage statistics for 25,000 geo-located Base Stations (BS, bottom layer). The coverage areas of BS are approximated by Voronoi polygons where mobile traffic is assumed to be uniformly distributed (middle layer). The mobile traffic is weighted and interpolated into France administrative areas (IRIS zones, top layer). The top plot depicts the average daily time series of downlink traffic per inhabitant at the richest 5% (dashed lines) and the poorest 5% IRIS zones in Paris for two representative mobile services: Facebook (red) and News (blue). As can be seen, time series of raw byte counts in the same area are highly correlated and reveal little information. However, the relative traffic generated by the two services in different areas exposes unique patterns that can be exploited for SES prediction.
Fig. 2: Correlation matrix of the consumption of each pair of mobile services, across all considered IRIS zone after RCA scaling. Variables are clustered using First Principal Component order. Non-significant coefficients are crossed out.
Fig. 3: Regression results for the three SES responses considered, i.e., median income, higher education ratio and local inequality. a. Performance metrics for each response and four different sets of regressors: (1) population structure variables, (2) traffic features, (3) both population and traffic features (All), and (4) population and traffic features combined with spatial filtering (All+SF). The left pane shows the adjusted pseudo-$R^2$ obtained from the linear relationship between the observed versus predicted values. The right pane shows the standardized mean absolute error (MAE), computed as the MAE divided by the mean response. b. Map of observed (left) versus predicted (right) responses using the best model (All+SF) in the Paris metropolitan area.
Fig. 4: Relative effect sizes for the three SES responses considered: median income, higher education ratio and local inequality. Model estimates with 95% CIs are shown for traffic features (top pane), population variables (middle pane) and the spatial term (bottom pane), in logarithmic scale. Non-significant estimates are grayed out. The model for the higher education ratio presents the stronger (positive and negative) effects overall. The median income response shows the higher spatial correlation.