Facebook Ads as a Demographic Tool to Measure the Urban-Rural Divide

Daniele Rama
University of Turin & ISI Foundation
Turin, Italy
daniele.rama@unito.it

Yelena Mejova
ISI Foundation
Turin, Italy
yelenamejova@acm.org

Michele Tizzoni
ISI Foundation
Turin, Italy
michele.tizzoni@isi.it

Kyriaki Kalimeri
ISI Foundation
Turin, Italy
kkalimeri@acm.org

Ingmar Weber
Qatar Computing Research Institute
Doha, Qatar
ingmarweber@acm.org

ABSTRACT

In the global move toward urbanization, making sure the people remaining in rural areas are not left behind in terms of development and policy considerations is a priority for governments worldwide. However, it is increasingly challenging to track important statistics concerning this sparse, geographically dispersed population, resulting in a lack of reliable, up-to-date data. In this study, we examine the usefulness of the Facebook Advertising platform, which offers a digital "census" of over two billions of its users, in measuring potential rural-urban inequalities. We focus on Italy, a country where about 30% of the population lives in rural areas. First, we show that the population statistics that Facebook produces suffer from instability across time and incomplete coverage of sparsely populated municipalities. To overcome such limitation, we propose an alternative methodology for estimating Facebook Ads audiences that nearly triples the coverage of the rural municipalities from 19% to 55% and makes feasible fine-grained sub-population analysis. Using official national census data, we evaluate our approach and confirm known significant urban-rural divides in terms of educational attainment and income. Extending the analysis to Facebook-specific user "interests" and behaviors, we provide further insights on the divide, for instance, finding that rural areas show a higher interest in gambling. Notably, we find that the most predictive features of income in rural areas differ from those for urban centres, suggesting researchers need to consider a broader range of attributes when examining rural wellbeing. The findings of this study illustrate the necessity of improving existing tools and methodologies to include under-represented populations in digital demographic studies – the failure to do so could result in misleading observations, conclusions, and most importantly, policies.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**; • **Information systems** → **Online advertising**; • **Applied computing** → **Sociology**.

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380118

KEYWORDS

digital demography, online advertising, social networks, urban-rural divide

ACM Reference Format:
Daniele Rama, Yelena Mejova, Michele Tizzoni, Kyriaki Kalimeri, and Ingmar Weber. 2020. Facebook Ads as a Demographic Tool to Measure the Urban-Rural Divide. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3366423.3380118

1 INTRODUCTION

In a rapidly urbanizing world, living in a rural community may present disadvantages from potentially residing far from a healthy food source (a "food desert") [27], to lower wages [54], to poorer health outcomes [28]. Disadvantages continue when considering the study and measurement of these populations to motivate appropriate policies. Demographers have long acknowledged the instability of measures concerning rural populations due to sparsity [32], often methodologically mitigated by substituting statistics from larger areas with similar population characteristics, making trends observed in specific rural areas, in fact, synthetic [33].

A possible solution may lie in a wealth of new digital data sources that has prompted a rise in recent research under the umbrella of "Digital Demography" [2]. The digitization of censuses [50], digital traces from online social networks [25], crowd-sourced data from participatory platforms [38], and internet-enabled devices [41] present new exciting opportunities for demographers by providing several advantages with respect to traditional sources [23]. Digital traces are generally accessible in high volume, often carry geographic information, and can be collected in real-time, allowing for the study of populations and their behaviors with unprecedented temporal and spatial granularity.

One such resource becoming popular in demography studies is Facebook's Advertising platform [2], which provides advertisers with an estimate of a potential advertisement's reach, given the location, demographic, or behavioral constraints of the target audience[1]. Thus providing a digital "census" of its estimated 2.23 billion monthly active user base, this resource has prompted studies in tracking health conditions [31, 39], migration [49, 57], crime [20], and gender inequality [19, 21, 29]. These studies illustrate the

[1] https://developers.facebook.com/docs/marketing-api/buying-api/targeting

benefits of using massive social media platforms for the examination of, especially, difficult-to-reach populations such as women in India [29], or difficult-to-count ones such as migrants in Spain [49]. Furthermore, the ability to track user "interests" beyond standard demographics, such as dietary habits, entertainment preferences, and technology use can extend the observations well beyond the standard demographic indicators [31]. Thus, the application of this data to rural populations promises both a higher granularity and a richer palette of potential variables. Nevertheless, given the nature of the service, its internal construction is a "black box", and a slew of biases, including Facebook user base self-selection, the algorithmic bias in extracting user attributes, and advertisement revenue incentives, must be taken into account by researchers [13].

In this study, we examine Facebook Advertising as a resource for measuring the rural-urban divide in Italy, a country in which 30% of its population is living in rural areas (76% of total landmass)[2]. In particular, we address three major research questions:

RQ1: How reliable are Facebook Advertising audience estimates across time and population density, especially considering rural or sparsely populated areas?
RQ2: How well do these estimates correspond to the official demographic figures?
RQ3: How can we enrich the current measurement of the urban-rural divide using Facebook Advertising?

We contribute to the current state of the art in four ways. First, we provide a systematic stability and spatial coverage analysis of the Facebook Advertising audience estimates for a wide set of user interests and behaviors, evidencing how both introduce particular disadvantages for rural municipalities. Second, we propose an alternative use of the platform to overcome limitations of sparsely populated areas posed by the platform itself due to its rate limits. Our methodology drastically improves coverage, especially of the rural municipalities, nearly tripling their coverage and making possible sub-dividing these populations for further analysis. Third, evaluating our method on official national census data, we quantify urban-rural inequalities, regarding both standard demographic attributes such as socio-economic indicators, as well as novel behavioural estimates available through the Facebook platform. We confirm a significant urban-rural divide, with Facebook users in urban areas having higher educational attainment and using higher-end cellphones (a proxy for income), while those in rural areas showing higher interest in gambling and Catholic church, doing more commuting, and using 3G or 4G networks instead of WiFi (pointing to a difference in internet access). Finally, we model per capita income in rural, suburban, and urban areas using the Facebook indicators, and show that the variables most predictive of income in rural areas (commuting, use of 4G network) are different from those in urban areas (marital status, educational attainment, interest in fitness and wellness, etc.).

Insights provided in this study are applicable to any re-purposing of digital resources for demographic research, pointing to the necessity of a cautious examination of any platform's coverage and stability before using information it provides. However, we also show that careful use of the platform's flexible querying process extends the usability of the data and allows for rich modeling of rural-urban inequality, augmenting analysis with Facebook users' many behavioral attributes.

2 RELATED WORK

Assessing global trends in the urban-rural divide is an essential topic of research in economics and demography, as almost everywhere in the world living standards of urban areas remain superior to those in rural areas [43]. Such divide can be observed across several different socioeconomic dimensions, ranging from per capita income to child mortality rates, persisting even as countries develop into industrialized economies, as demonstrated by the cases of China and India [44, 47]. Historical trends in development, which tend to benefit those in already privileged positions, have resulted in, for example, technological disparities in terms of internet access and technological literacy – trends which ongoing rural development policies attempt to address [17, 36]. Even in OECD countries, persistent disparities between large metropolitan centres and rural areas are being recorded every year, as migration patterns intersect with other population attributes [35]. According to a report of the World Bank, the urban-rural disparities increase quickly in early development until countries reach upper-middle-income levels [40]. Then, as countries grow, the gap becomes smaller, but convergence is usually slower.

From a theoretical point of view, recent studies have investigated the structural advantage of cities in a wide range of output indicators, from patent production to personal income, through the robust framework of scaling [8, 16]. In particular, superlinear scaling of cities' growth has been explained as a consequence of increased social interactions with population density [7] and by the process of selective migration of highly productive individuals into larger cities [26, 54]. Understanding the mechanisms underlying the urban-rural divide remains an essential issue for policymaking, especially to make progress towards the "Leaving no one behind" pledge of the United Nations 2030 Agenda[3].

Leveraging on the immense amount of digital data produced daily, the field of Digital Demography emerged, addressing vital research questions of demographic research via innovative data sources. These new sources of data are demonstrated to be particularly powerful in monitoring a series of demographic phenomena such as birthrates [5], mortality [4], unemployment [10], daily commuting [6], international and internal migration [55], but also modelling more complex socio-demographic issues such as psychological well-being and attitudes towards health [24, 30]. Digital data are particularly useful in cases where official data are sparse, incomplete, or even impossible to obtain. For instance, Adler et al. [1] assessed the issue of suicide underreporting via query data, focusing on the Indian context, where social stigma and the only recent decriminalization of suicides, hampered the official agencies' data collection. Interestingly, digital sources can be used to investigate social inequalities. In particular, social media data are proven to be useful in studying gender differences in access to technology [21, 29] and parenting biases favouring male children mentions on social media [48].

[2] https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS?locations=IT

[3] https://www.un.org/sustainabledevelopment/sustainable-development-goals/

Among all online social networks, Facebook is the most popular one. In June 2018[4], Facebook reported 1.47 billion daily active users (DAUs) and 2.23 billion monthly active users (MAUs), with an increase of 11% year-over-year. On average, three out of ten people used Facebook in 2018, and this estimate, considering the current trend, is destined to grow.

Researchers have taken note of Facebook's massive user base, and, for instance, used it in the health domain for the recruitment of people affected by not-so-common conditions [14], or a particular health behaviour [12]. With the passing of the European General Data Protection Regulation (GDPR) more restrictions have been put on the processing and exploitation of personal data of individual users [11], such as political orientation, religious beliefs, ethnic origin, etc., due to the apparent privacy risks that may be derived from a malicious use of such type of information.

Yet, collecting individual user data is not necessary for the demographic study of populations. Instead, demographers recently began to use Facebook Advertising platform to gather statistics on select populations by querying for the number of users who an advertisement could reach. In this fashion, it is possible to communicate a count of users matching specific socio-demographic characteristics at various geographic scales without revealing personal information of individual users.

Seminal works using this approach focused on the health domain [31, 39]. For example, Araújo et al. [3] extracted data in 47 countries to track health conditions associated with lifestyle diseases. They showed that, within each country, Facebook data could provide insights into different trends of health awareness across demographic groups. Zagheni et al. [57] proposed the use of Facebook Advertising data to monitor stocks of migrants inside the US with promising results, paving the way for similar studies in the European context [38, 49], but also more in depth studies on the assimilation of migrants in society [18].

Only recently, Facebook Advertising data were considered to examine social inequalities, and in particular gender inequalities in Internet access both at national [19, 21] and sub-national levels [29]. In this direction, a study by Gil-Clavel and Zagheni [22] extended the analysis of the gender gap in Facebook adoption by adding the dimension of age.

All these works provided evidence that Facebook Ads data can indeed be used as a source of information for the study of digital disparities. However, little effort was made to understand the stability and representativity of such data across the several attributes available through the platform. As several studies pointed out, big observational data are not always representative of larger populations in the way that randomized surveys are [37, 51, 53, 56]. Coverage can also be an issue, as access to the internet is more restricted in low and middle-income countries, which can lead even to risks of re-identification [11, 15, 45].

This study contributes to the Digital Demography literature providing a thorough examination of the stability and coverage of Facebook Advertising data. We particularly focus on behavioral signals related to social inequalities, especially those that may be more difficult to track officially, such as interests and hobbies.

[4] https://investor.fb.com/investor-news/press-release-details/2018/Facebook-Reports-Second-Quarter-2018-Results/default.aspx

3 DATA COLLECTION & METHODS

3.1 Census data

Italy is divided into 20 regions, which are subdivided into provinces, and then further into municipalities, or *comuni*, which are the smallest administrative units. In order to differentiate between the urbanization within Italy, we consider the scale of municipalities, of which there were 7,978 as of February 20, 2019. Note that this number fluctuates, with municipalities merging, breaking up, and being redefined over time.

The Italian national institute of statistics (Istat[5]) adopts the definition of urbanization from Eurostat, the statistical office of the European Union, and separates municipalities into three categories: cities, towns and suburbs, and rural areas[6]. The assignment of these categories is based on population density, as measured using 1 km^2 grid. In Italy, there are 270 urban, 2,303 suburban, and 5,405 rural municipalities, having an average population of 74.9K, 11.2K, and 2.7K, respectively (See Figure 1).

In order to work with the municipalities via Facebook Marketing API, we first request their IDs by specifying municipality name, region, and state and, in case more than one match is returned, use string matching to choose that with closest name. Out of the 7,978 Italian municipalities, we are able to match 6,891, excluding 24 (8.9%) urban, 282 (12.2%) suburban, 781 (14.4%) rural municipalities. Note that matching of urban ones is easier, showing for a bias to densely populated areas even at this stage of data collection.

For all Italian municipalities, we download the demographic and socio-economic indicators from Istat. Unlike aggregate indicator values for larger geographic regions, only few are available at the fine-grained level of municipalities. Thus, we are able to collect data on the overall population (overall and split by gender), education (high school attainment and college attainment, from last census in 2011), income (net income per capita, 2018), and migration (Italian residents who are not Italian citizens, 2018).

3.2 Facebook Marketing API

Facebook Advertising audience estimates are available via Facebook Marketing API, which we access using a Python package[7]. For all queries, we request a count of "People who live there" (technically, setting `location_type` parameter to `home`), as we are interested in the population living in the municipality of interest, not those working there or passing through as tourists. The API also allows the querying of users using other services owned by Facebook, including Instagram, Messenger, and "Audience network". However, we choose to constrain the query to Facebook users, for simplicity of interpretation of results.

Furthermore, several advertising campaign types are available, focusing on either "brand awareness" or "reach". As we are interested in the most complete count of the users on the platform, we choose the "reach" option, which targets the "maximum number of people"[8]. Finally, in the reply to our query, we save the Monthly Active Users (MAU) (a Daily Active Users count is also available, but we do not use it, as it is less stable over time). Once we compose

[5] https://www.istat.it/en/
[6] https://ec.europa.eu/eurostat/web/degree-of-urbanisation/background
[7] https://github.com/facebook/facebook-python-business-sdk
[8] https://www.facebook.com/business/help/197976123664242

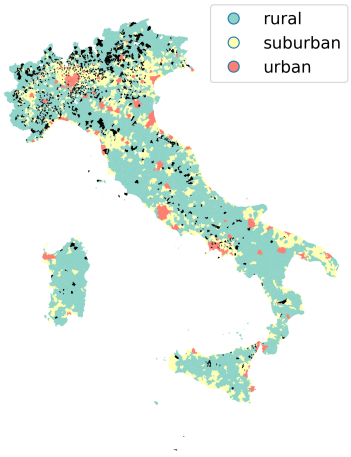

Figure 1: Italian municipalities colored according to degree of urbanization. They are colored in black if not available on Facebook marketing platform.

the queries combining the options above with various combinations of targeting options (described below), we query the Facebook Marketing API via Python, with a delay of 8 seconds empirically determined to avoid passing the rate limits.

3.3 Targeting

We build on previous literature to choose attributes for comparison of urban and rural municipalities, as well as Facebook-specific ones dealing with user interests (as inferred from user profile and activity) and technological aspects of user interactions, such as what kind of phone and connection they use to access it. We list the attributes below:

- Gender (male, female)
- Marital status (single, married)
- Education (high school grad, college grad)
- Cell network (3G, 4G, Wi-Fi)
- Cellphone operating system (Android, iOS)
- Newness of cellphone ("Technology early adopter")
- Travel (living abroad, away from hometown, frequent travel, frequent international travel)
- Interests pertaining to culture (Catholic church, gambling)
- Interests pertaining to health (cooking, fast food, restaurants, fitness and wellness)

Some of these are inferred by Facebook from the self-disclosed information and from the user interactions on the platform (such as marital status and education). Others are determined automatically from the metadata associated with the connection, such as which cell network or phone is being used, which may be more reliable. The resulting query consists of 6,891 municipalities, each queried once to estimate the total population, plus 22 times for populations with the above attributes. Note that by restricting our focus to these sub-populations, the problem of coverage becomes even more dire. In the next section we discuss our approach to solving it. An online interactive map showing the Facebook population estimates for each of the above attributes can be found at http://www.datainterfaces.org/projects/facebookMap/.

3.4 Exclusion Query

If the combination of targeting options for the query is too specific, the resulting MAU estimate may be limited to a lower threshold of 1,000 users. Given the FB variables that we chose, the standard querying process excludes 95% rural, 67% suburban and 38% urban municipalities from the dataset.

To overcome this limitation, we propose an "exclusion" query, wherein in order to get an attribute-constrained population estimate of a small municipality `S`, it is first queried with another, larger, "reference" municipality `R` resulting in a combined query `S+R`. The difference between the combined query and the known reference municipality population (`(S+R)-R`) provides us an estimate for the small municipality `S` with a possible range of down to 100. This lower range is possible due to the finer resolution of results, which is not in 1,000s, but in the 100s.

Below we summarize the steps taken to query the Facebook API for municipality estimates matching a set of attributes:

(1) Query Facebook API for all municipalities using the standard query.
(2) Choose 5 reference municipalities with MAUs in the range between 2,000 and 10,000
(3) For each of the municipalities that previously hit the 1,000 threshold, we run combined queries 5 times, each one with a reference municipality
(4) Compute the difference of combined query (only if it did not also hit 1,000 threshold) and the reference municipality alone, and take the average across all the valid (non-negative, non-zero) responses: resulting in the "exclusion query" estimate.

Thus, for each collection, some queries may result in valid responses using the standard query, while others will need an exclusion query, and even these may not result in a valid estimate. However, with this approach we aim to improve the coverage of sparsely populated municipalities as well as the resolution of the estimates.

In order to assess the accuracy of the exclusion query estimates with respect to the standard ones, we select 20 municipalities for each Facebook variable and degree of urbanization for which standard estimates are known and we compared them with the same estimates extracted using the exclusion query approach. For such municipalities, the mean Pearson's correlation coefficient across all Facebook variables between the two types of estimates is 0.99, showing that exclusion query closely tracks the results of the standard one.

4 COVERAGE / STABILITY TRADE-OFF

After we select the municipalities of interest and the population attributes we want to acquire from Facebook Advertising, we perform a data quality study. As a black box, we ask, how much does the advertisement audience estimates vary across time and populations? Upon initial experimentation, we find the estimates change within a week of original query. Thus, we perform 5 data collections every

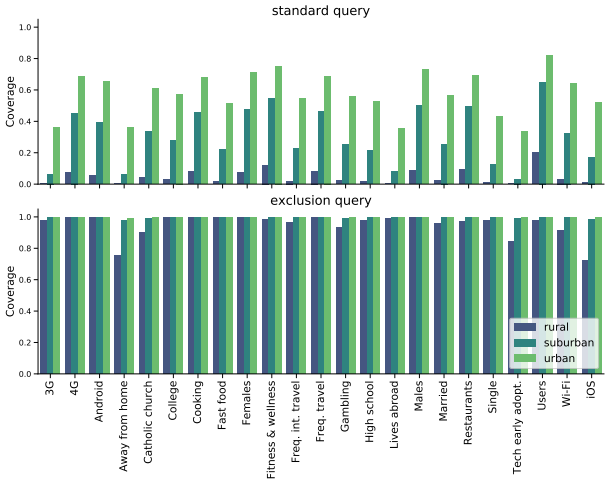

Figure 2: Coverage of municipalities reached using "standard" and "exclusion" querying with respect to the Facebook variables selected.

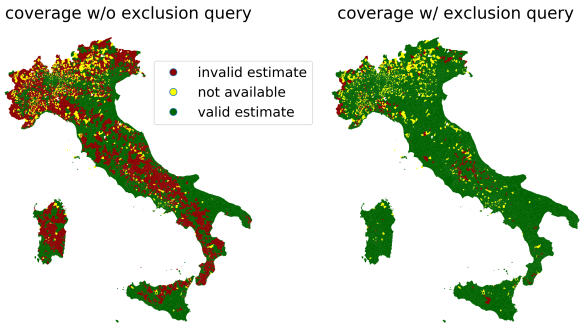

Figure 3: Coverage of municipalities within Italy, querying population without variable constraints.

two weeks in the time span between April 7th, 2019 and June 2nd 2019, and examine the variability of the results.

4.1 Spatial coverage

First, we measure the spatial coverage for each Facebook attribute as the proportion of municipalities for which Facebook provides a valid estimate. We start by considering the standard query estimates, which are limited to a minimum threshold of 1,000 users. In this case, we consider a municipality having a valid estimate if, considering a Facebook variable, it receives at least one response above the threshold over 5 runs of the same query. Figure 2 (top) shows the percentage of municipalities covered by the standard query by degree of urbanization for each Facebook variable considered. The variable *Users* refers to the total number of Facebook users living in a municipality (without any constraints applied). Focusing on this variable, we observe that 81%, 64%, and 19% of urban, suburban, and rural municipalities respectively are covered, showing that rural municipalities are indeed hard to reach. For

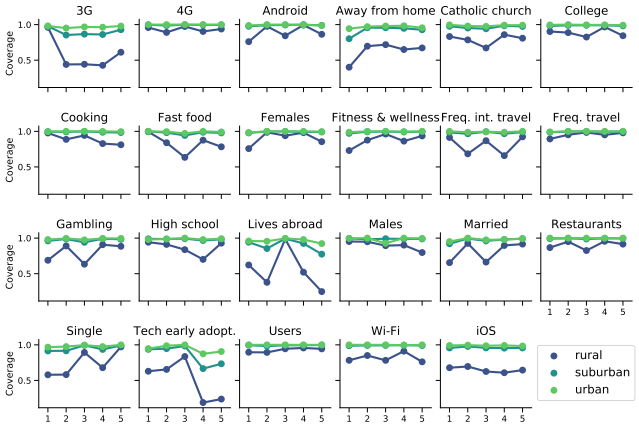

Figure 4: Percentage of municipalities having valid response for each attribute over 5 runs.

example, in Figure 3 the map on the left shows the coverage of the standard query when querying municipality population without any attribute constraints. If we restrict the analysis to subgroups of Facebook users matching certain constraints, this disparity grows even more. While the coverage of urban municipalities is in the range between 40% and 80% for almost any variable, in the case of rural municipalities it is always below 15%.

In the case of the exclusion query, the estimates now have a possible range down to 100 users. Therefore, to calculate the spatial coverage, we consider a municipality having a valid estimate if, given a Facebook variable, it receives at least one response of 100 or more, over 5 runs of the same query. Figure 2 (bottom) illustrates the substantial improvement in the coverage for every Facebook variable selected, and 3 (right) shows the geographical coverage in the case of the generic query, suggesting that this approach can effectively be used to reach even sparsely populated municipalities.

Note that, in calculating the coverage, we considered all estimates greater or equal to 100 users as valid. Nevertheless, we may choose different cut-offs, e.g. 200, 300, 400, etc., with the corresponding change in the coverage. To understand which threshold is most suitable, we perform a stability analysis of the results in the following section.

4.2 Stability of query results

To assess the stability of the estimates, we examine the values of the same variables collected five times, each two weeks apart. First, considering a threshold of 100 users, we calculate the coverage for each collection as the percentage of all municipalities which pass the threshold. The result is shown in Figure 4. While the spatial coverage is somewhat stable and close to 100% for suburban and urban municipalities, it shows large fluctuations in the case of rural municipalities. For instance, variable *Lives abroad* ranges from almost full coverage on third day to almost no coverage for rural municipalities on the fifth.

However, a threshold of 100 is the most optimistic, and we may want to consider stricter ones to improve the quality of the estimate. We check the impact of the threshold on coverage by selecting

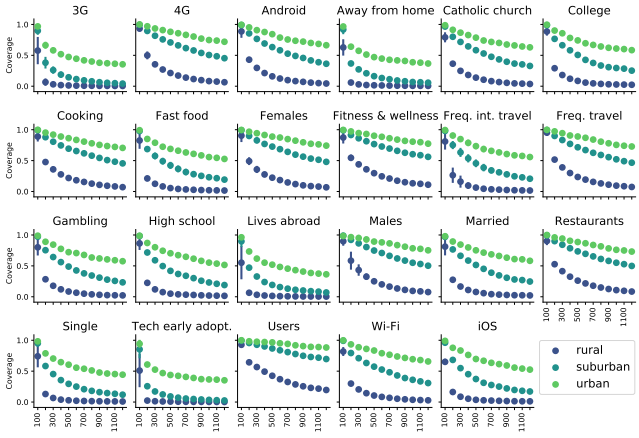

Figure 5: Coverage of municipalities for each attribute over different definitions of threshold for valid response.

twelve thresholds from 100 to 1,200 users with a resolution of 100 users. Figure 5 shows the coverage decreases smoothly for urban and suburban municipalities as the threshold increases. For rural (and sometimes more populated) municipalities, we observe a drop between 100 and 200 users, while after 200 users the coverage decreases smoothly, indicating the threshold of 100 may be artificially high.

Finally, we check the variability of the estimates at different thresholds. To do this, for each Facebook variable, we compute the proportion of sub-population P with specific characteristic c defined as $P_m[c] = FB_m[c]/FB_m$ where FB_m is the total number of users who live in municipality m and $FB_m[c]$ is the total number of users who live in municipality m and also match the characteristic c. Given the variability of estimates across days, $FB_m[c]$ is calculated as the median across all the valid estimates over five runs of the same query. Now, we calculate the variance of the distribution of this index for the three degrees of urbanization, which is shown in Figure 6 (colors of the 22 attributes omitted for clarity). We observe that the variance rapidly decreases in the range from 100 to 200 users, while after the latter it remains stable for most of the Facebook variables. Combined with our earlier observation of coverage drastically falling from 100 to 200 users, we choose the threshold of 200 for the following experiments in order to achieve the best coverage while ensuring the stability of the estimate. With this threshold, we are able to cover 55% rural, 84% suburban, and 90% urban municipalities, nearly tripling the coverage of the rural ones. The improvement is even more drastic for specialized queries: the average coverage of 4%, 29%, and 57% for rural, suburban, and urban municipalities now becomes 31%, 67%, and 81% with the use of exclusion query.

5 DATA QUALITY ASSESSMENT

5.1 Relating to Government Statistics

As described in Section 3.1, some demographic and socio-economic indicators are available at the municipality level. To understand how well Facebook Advertising audience estimates track the figures

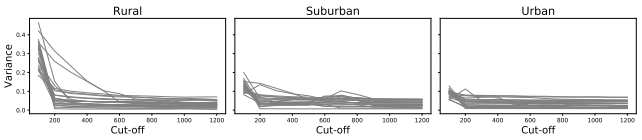

Figure 6: Variance of Facebook variables over different thresholds for valid response.

gathered by the Italian Government (Istat), we correlate the Facebook variables most closely related to each of the Istat demographic variable in this section.

First set of plots in Figure 7 shows the relationship between municipality size, as estimated using Facebook (x axis), and the population given by Istat (y axis), shown separately for rural, suburban, and urban municipalities. Note the different scale, as populations in the three plots differ. Despite high Pearson correlations of 0.93, 0.89, 0.99 for rural, suburban, and urban, respectively, we find that Facebook under-counts people in smaller municipalities, with almost all data points above the diagonal. The estimates become more accurate for the highly populated municipalities, which appear on the diagonal. The under-counting, in fact, affects the rural municipalities at a higher rate, with Facebook estimates being 71% off on average, compared to 55% for urban municipalities. Note that the figure distinguishes between data points acquired using standard query (in dark marks) and extended (in light marks), illustrating the drastically improved coverage of rural municipalities from 19% to 55% (with the same correlation).

Considering the gender, the raw numbers are again highly correlated to the Istat population (graphs omitted for brevity, but they look much like population ones). Instead, we examine the gender ratios within each municipality. Figure 7b shows the comparison between the gender ratio estimated by Facebook (female/male), to that estimated by Istat, with dashed lines showing parity (50% females and 50% males). Despite actual Istat statistics showing a trend toward municipalities with more women then men, we observe that Facebook tends to overcount males, and especially so in rural communities.

Encouraged by the substantial correlation of population statistics, we examine a more challenging case of monitoring particular characteristics, measured as a proportion of population. For instance, Figure 7c shows the proportion of Facebook users who have attained a college degree, compared to the Istat estimate of the same statistic. The Pearson correlation between the rural, suburban, and urban areas are 0.36, 0.46, 0.61, respectively. Checking the outliers in the upper right of rural plot, we find Urbino and Camerino, municipalities containing universities which have a high proportion of educated residents which may be more captured by Facebook and less by formal residency requirements of Istat. Figure 7d shows the proportion of Facebook users marked as "Living abroad", compared to the Istat estimates of "foreigners" residing in each municipality. We find substantial correlations of 0.72, 0.72, and 0.82 for rural, suburban, and urban, respectively. Similarly, checking outliers in the three graphs, we find that in some cases Facebook drastically over-counts the number of those "living abroad", and upon manual inspection a month later we find the numbers to come down,

indicating a high variability either due to actual measurement of human mobility, or internal platform changes. Finally, Figure 7e shows two proxies of wealth – proportion of Facebook users who use iOS or Android devices – compared to Istat estimates of income per capita. We find a clear signal that the use of iOS is positively related with income (correlations of 0.57, 0.62, 0.78) and use of Android is negatively related (-0.73, -0.77, -0.86 for rural, suburban, and urban).

Note that, unlike the gender, educational attainment, and migration, the last two attributes of the phone's operating system are detected directly and unobtrusively, instead of inferred from self-reported data. Thus, it may not be suffering from as much measurement error as the others, and thus provide a clearer signal.

5.2 Measuring Inequality via Facebook

Next, we use Facebook signals to measure potential inequalities between the urban and rural communities in Italy – those that can be also tracked via Istat, and those that may be more difficult to track officially, such as interests, hobbies, travel habits, etc.

For this experiment, we exclude suburban municipalities for clarity, and leave for future work a more continuous analysis. For each Facebook attribute, we consider a geographically-cohesive comparison wherein in each province (there are 107 provinces in Italy, but only 71 have both rural and urban areas to compare) we subtract the median attribute value of its rural from median of its urban municipalities. We then plot the distribution of these differences in Figure 8, in which the differences at the significance level of $p < 0.05$ (chosen with the small number of data points in mind) are in dark grey, and the means of distributions are indicated by the blue triangle. Note that the significance level is affected both by the magnitude of the difference, as well as the number of provinces that have enough coverage to capture both urban and rural areas of each province, thus ranging in coverage from 27 for technology early adopters to the maximum of 68, on average 61 provinces per attribute. The coverage is still markedly better than if the data was used without the exclusion queries, with an average of 36, almost half as many, provinces having enough data for analysis.

The most drastic inequalities, we find, are those associated with education (there are fewer college graduates in rural areas), and income (with fewer iOS and more Android usage in rural areas). These findings confirm the official Istat numbers, which show significant differences in education and income. In a directly comparable case, Istat shows a mean difference of 7.2%, while the difference in College graduation Facebook attribute is 5.0% (both significant at $p < 0.001$). Interestingly, the Facebook data does not show a significant difference between people "Living abroad", whereas Istat numbers show a slight difference at 4.0%, which may be either due to Facebook usage bias or the possible ability of the platform capturing different populations, as we discuss in Discussion section.

When we examine the Facebook attributes which cannot be directly confirmed by the Istat municipality-level data, we find that in urban areas, the Facebook users tend to be more interested in fitness & wellness, be frequent international travelers, and be single. In the rural areas, they tend to be interested in cooking and restaurants, to commute (be frequent travelers), to be married, be Catholic, and use 4G or 3G networks instead of Wifi (a sign

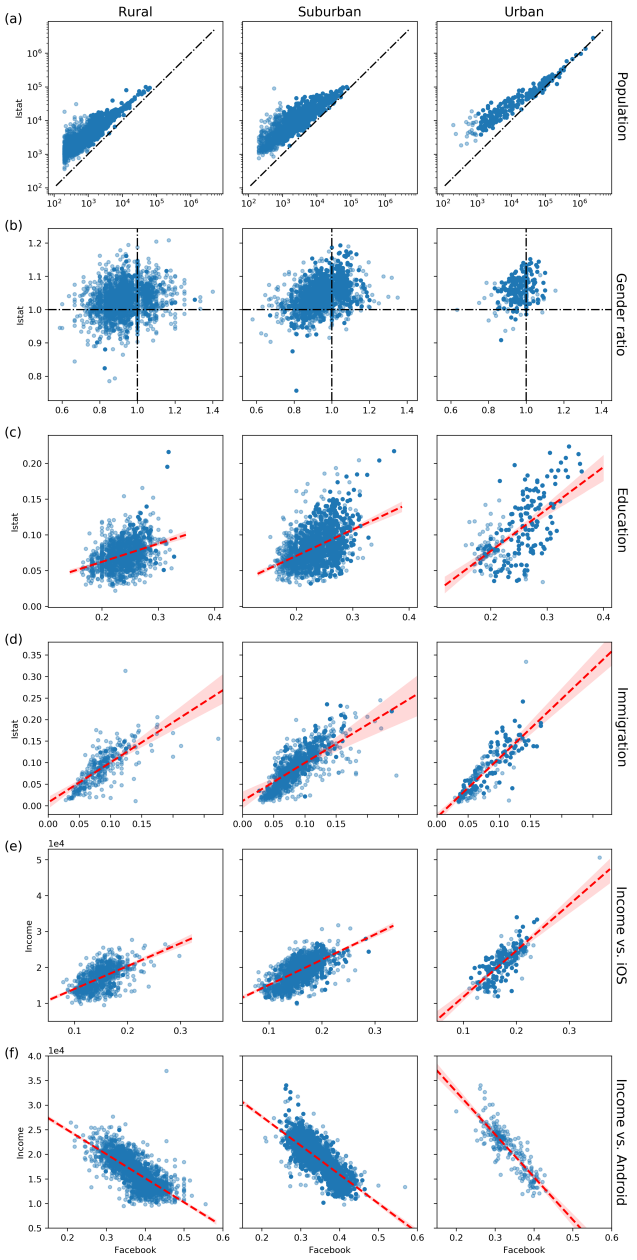

Figure 7: Facebook sub-population estimates (x axis) versus relevant Istat statistics (y axis). Dark marks are estimates of standard query, light – exclusion query. Dashed black lines show diagonals (or parity in case of gender) and red lines show regression line with 95% confidence.

of necessity for the latest efforts by Italy to extend its broadband network[9]. Thus, Facebook allows us to peer into inequalities which are not captured by the governmental agencies, for instance those of relationship status, interests, and daily technology use – all of

[9] https://ec.europa.eu/digital-single-market/en/country-information-italy

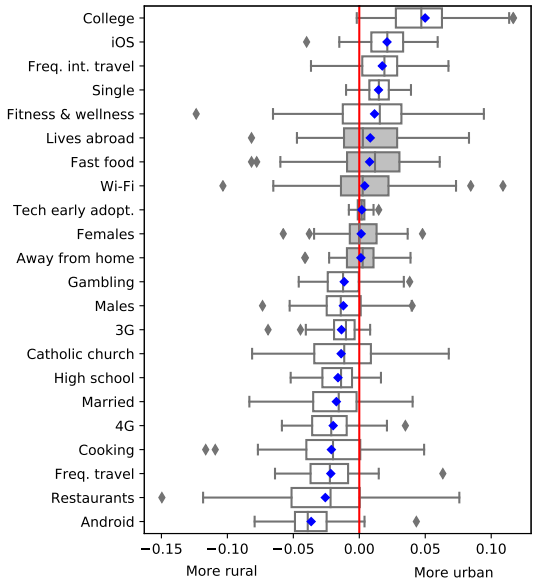

Figure 8: Distribution of difference in medians between urban and rural areas in Italian provinces for each attribute. Statistically significant boxplots are in white ($p < 0.05$), insignificant in grey. Blue points show means.

which could be used to examine the well-being of the population holistically, in addition to the standard demographics.

5.3 Modeling Inequality

In the previous section we find several Facebook indicators showing inequalities between rural and urban communities, many of which are related to the socio-economic state of its residents. We ask whether the combination of these signals may be useful in modeling the financial well-being, as measured by income per capita. Not only would such a model provide an alternative, up-to-date estimate of financial well-being, it would also provide an explanatory power to gauge the possible factors associated with income inequality.

We begin by building three baseline models using the municipality-level variables made available by Istat, one for each kind of municipality (rural, suburban, and urban). Table 1 shows the coefficients and their significance levels for the three models, as well as the number of municipalities used in the dataset (n), coverage of all possible municipalities (f), and the Adjusted R^2 (which corrects for the number of attributes in the model). The best performance is attained for the Urban municipalities at R^2=0.660, despite the smaller dataset size. The relatively poor performance of the models in rural and suburban areas (R^2=0.231 and R^2=0.281, respectively) may signify that more information is needed to differentiate between high and low income areas. Note that the performance of these small models is limited by the data available on Istat website, and may be drastically better if other variables are added. The aim of this exercise is to convey the difference in difficulty of the task between the kinds of population densities.

Table 1: Linear regression model, predicting income (in Euros) using standardized Istat variables. For each model, number of municipalities (n), coverage of all municipalities (f), and Adjusted R^2 are shown. Confidence levels: $p < 0.001$ ***, $p < 0.01$ **, $p < 0.05$ *.

	Istat variables		
	Rural_{istat}	Suburban_{istat}	Urban_{istat}
	n=5,405 f=1.00 R^2=0.231	n=2,303 f=1.00 R^2=0.281	n=270 f=1.00 R^2=0.660
(Intercept)	16,830 ***	19,600 ***	21,250 ***
males	211 ***	555 ***	1,442 ***
high school	1,236 ***	978 ***	2,537 ***
college	202 ***	914 ***	2,226 ***
migrants	773 ***	1,009 ***	984 ***

We follow a similar setup for modeling Istat income variable using the Facebook attributes. The first three models of Table 2 show the performance of linear regressions modeling the income (as measured by Istat) in rural, suburban, and urban municipalities, using all available attributes. The models achieve a more uniform performance with R^2=0.772, R^2=0.798, and R^2=0.856, respectively, with similar intercepts as the Istat models. The coefficients for the attributes which are similar to those in Istat model now reverse their sign in some cases, and lose their significance. For example, in the rural areas the model favors the information about frequent travelers, marital status, and the access to cellphone networks. In the urban areas, instead, information about college attainment, use of iOS, and interest in fitness & wellness are more significant. Unfortunately, the increase in performance comes at a cost of coverage: only 2% of rural and 21% of suburban municipalities contain complete values for all features. The number of complete records is also not sufficient to perform missing value imputation, as our additional experiments reveal.

In order to improve coverage, we examine the trade-off between including features and the number of municipalities which can be used in the model. Figure 9 shows the Adjusted R^2 performance (red line) as the number of features in the model increases (ordered by magnitude of the coefficients in the complete model), and the coverage in dashed blue line. A baseline of Istat model performance is shown as the horizontal. We observe that for rural and suburban municipalities, there comes a point when the coverage falls precipitously, with rural areas falling within the first three features. Gauging this trade-off, we select a cutoff where a reasonable performance can be achieved without discarding most of the data (shown by vertical lines). In this study we do not propose a particular metric for selecting such a cutoff, and defer the selection of such a metric to the experts in the particular issue and population being studied (where either coverage or precision may be more important).

The best trade-off models are shown in the right-most three columns of Table 2. The slight loss in performance is accompanied by substantial gains in coverage: from 2% to 46% for rural, from 21% to 67% in suburban, and from 53% to 78% in urban municipalities. In

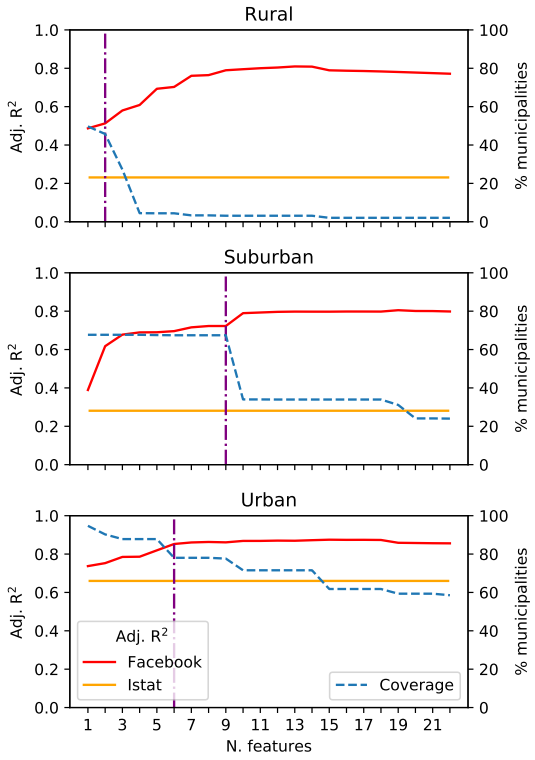

Figure 9: Coverage in % of municipalities (right y axis) vs. performance in Adj. R^2 (left y axis), as features are added to the models with largest coefficients first. Baseline performance of Istat indicated by horizontal line, and the model with best trade-off indicated by vertical line.

the case of urban model, the drop in performance is negligible (from R^2=0.856 to R^2=0.853). Also, while the rural model contains only two attributes, it shows substantial fitness to the data at R^2=0.513. Note the difference in the attributes selected for each model, showing different characteristics may be important for different levels of urbanization.

Finally, if these models were computed on the data without using exclusion queries, there would not be enough coverage of rural municipalities to build one at all (n=11), and extremely poor coverage for suburban (n=70) and urban (n=86) municipalities.

We would caution the reader to seek a fully automated machine-learning style of optimization in this task, as the trade-off between coverage, performance, and complexity of the model (number of features) must be determined by experts in the case-by-case basis. Instead, we hope these experiments encourage the reader to consider sources of data alternative to the statistics gathered by governments.

6 DISCUSSION & CONCLUSIONS

In this work we explore an increasingly popular data source in the field of Digital Demography, the Facebook Advertising platform, which, in addition to cost-effective population estimates, provides rich behavioral data at an unprecedented scale and granularity. We find that, much like the standard demographic research, it takes more effort to obtain reliable statistics for the rural populations. Nonetheless, the rich behavioral and technical insights Facebook is able to collect on its users have a potential to extend the study of wellbeing of populations. For instance, health-related behaviors such as having interests in cooking at home or exercise may help in contextualizing the ongoing obesity and diabetes epidemics, much of which has been recently attributed to the rural communities [9]. Further, we find the variables connected to the use of technology, and especially of mobile devices, a strong proxy to financial wellbeing of the population. In the case of internet access, the fact that the rural residents are more likely to connect to Facebook via mobile data network instead of WiFi (land-based internet connection) may point to a persisting digital infrastructure divide between the rural and urban areas [34]. In aggregate, an expanded view of population's behaviors, interests, and demographics would be useful in creating compound measures of wellbeing (such as in [10]).

However, our analyses uncover serious instability and coverage issues in the signal Facebook Advertising provides, especially when it pertains to the rural communities. The volatile behavior of this data source should be a cautionary tale to any demographers using digital platforms as "black boxes" that have opaque implementation and unpredictable update schedules. The method we propose to improve the quality of data for smaller populations drastically increases the coverage of, for instance, the income model, raising the number of rural municipalities with complete records from 2% to 46%, allowing us to take advantage of the finer-grained attributes Facebook provides. Though we caution the reader not to focus on the particular numbers achieved here, as they depend also on the model construction and other methodological choices, all of which should be adjusted when working on other domains and variables of interest. Nonetheless, a choice must be made between the coverage and signal stability, the margins of which may be best determined by the knowledge of experts in the under-represented demographic group of interest. When choosing the parameters, robustness checks must ensure a "researcher degree of freedom" [52] does not lead to misleading observations and conclusions that could impact real-world policies.

Besides the coverage and stability issues, we also uncover several biases in the Facebook Advertising data. Even when we employ the "exclusion query" methodology, Facebook audience estimates consistently under-count populations in rural areas and over-count males (with the gender imbalance being more pronounced in the rural municipalities). Many sources of bias may be at play: (1) self-selection bias in the user base of Facebook, thought to be younger and more tech savvy (but which may be shifting toward older users[11]), (2) measurement bias in the sensitivity of Facebook's user attribute extraction pipeline (for example, the gender statistic may be swayed by numerous fake accounts[12]), and (3) financial incentives to inflate the number of users who may see an advertisement (being an important revenue stream, Facebook's advertising revenue exceeded

[10] http://lab24.ilsole24ore.com/qdv2018/
[11] https://www.techspot.com/news/79082-facebook-rapidly-losing-millennials-us-user-base-down.html
[12] https://www.buzzfeednews.com/article/craigsilverman/facebook-fake-accounts-afd

Table 2: Linear regression models, predicting income (in Euros) using standardized Facebook variables. For each model, number of municipalities (n), coverage of all municipalities (f), and Adjusted R^2 are shown. Confidence levels: $p < 0.001$ ***, $p < 0.01$ **, $p < 0.05$ *.

	All Facebook attributes included						Coverage vs. performance selection					
	Rural_{all}		Suburban_{all}		Urban_{all}		Rural_{cut}		Suburban_{cut}		Urban_{cut}	
	n=95 f=0.02 R^2=0.772		n=485 f=0.21 R^2=0.798		n=144 f=0.53 R^2=0.856		n=2,115 f=0.46 R^2=0.513		n=1,363 f=0.67 R^2=0.723		n=192 f=0.78 R^2=0.853	
(Intercept)	16,300	***	18,760	***	20,590	***	16,380		19,200	***	20,940	
males	-288		-56		216				-612	***		
married	613	**	-143	***	646							
single	-20		-159		-886	*					-1,424	***
high school	-133		150		-356							
college	265		455	***	1,486	***			120		2,057	***
lives abroad	307		-11		-455							
away from hometown	-286		52		250							
frequent international travelers	15		-281	**	-32							
frequent travelers	-1,212	***	-804	***	-229		-2,325	***	-841	***		
android	27		-360		-1,539	*			-748	***	-2,011	***
ios	384		1,140	***	1,181	**			982	***	1,362	***
technology early adopters	159		37		-168							
3g	-468	*	-283	**	14							
4g	-693	*	-607	***	180		-148	*	-476	***		
wi-fi	274		97		-394							
retaurants	159		-71		-1,094	*					-2,122	***
fast food	-194		-76		97							
cooking	-261		-596	**	-448				-445	***		
catholic church	-114		-404	***	-664				-436	***		
fitness and wellness	378		335	**	1,130	**			100		1,284	***
gambling	-241		25		-689							

$55 billion in 2018[13]), among others already explored in literature on online data representativeness [37, 42, 46, 51]. Although it is not likely that Facebook will release details of its user attribute inference code, demographers may be able to adjust for the larger sample biases of Facebook user base, as recommended in [13].

Nevertheless, with the appropriate handling of certain biases of Facebook Advertising data, its benefits in terms of coverage and attribute diversity may contribute to the ongoing efforts in the theoretical understanding of the attribute variability within urbanization spectrum via the Urban Scaling Theory [7]. Our own preliminary experiments showed scaling trends of Facebook attributes in the range of those in the existing literature [8]. Expanding the application of this theoretical framework to cultural and wellbeing aspects of populations, as measured through Facebook, would be an exciting future research direction.

Finally, despite the aggregate nature of this data, Facebook Advertising (and many similar platforms) pose several ethical and privacy issues. The methodology proposed in this study allows for the tracking of smaller demographic groups at a higher resolution, introducing risks especially for the more vulnerable populations and minorities, and hence should be applied with caution. It may be the case that Facebook needs to limit the exposure of user "interests" that may result in government censorship or prosecution, or targeting by other groups. Conversely, the platform may undercount people with impairments or disabilities who are not able to use the website and who may be under-represented in its user base. Thus, the ethics rules already established for sociological and demographic studies (such as those published by the American Sociological Association[14]) must be applied to those using new data sources, such as to protect the subjects of the study.

ACKNOWLEDGMENTS

We gratefully acknowledge the support from the Lagrange Project of the ISI Foundation funded by the CRT Foundation.

REFERENCES

[1] Natalia Adler, Ciro Cattuto, Kyriaki Kalimeri, Daniela Paolotti, Michele Tizzoni, Stefaan Verhulst, Elad Yom-Tov, and Andrew Young. 2019. How search engine data enhance the understanding of determinants of suicide in India and inform

[13] https://newsfeed.org/facebooks-revenue-exceeded-55-billion-in-2018/

[14] https://www.asanet.org/code-ethics

prevention: observational study. *Journal of medical internet research* 21, 1 (2019), e10179.
[2] Diego Alburez-Gutierrez, Emilio Zagheni, Samin Aref, Soƿa Gil-Clavel, Andre Grow, and Daniela V Negraia. 2019. Demography in the Digital Era: New Data Sources for Population Research. https://doi.org/10.31235/osf.io/24jp7
[3] Matheus Araujo, Yelena Mejova, Ingmar Weber, and Fabricio Benevenuto. 2017. Using Facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. In *Proceedings of the 2017 ACM on Web Science Conference*. ACM, 253–257.
[4] Anna Baranowska-Rataj, Kieron Barclay, and Martin Kolk. 2017. The effect of number of siblings on adult mortality: Evidence from Swedish registers for cohorts born between 1938 and 1972. *Population Studies* 71, 1 (2017), 43–63.
[5] Kieron J Barclay and Martin Kolk. 2017. The long-term cognitive and socioeconomic consequences of birth intervals: A within-family sibling comparison using Swedish register data. *Demography* 54, 2 (2017), 459–484.
[6] Mariano G Beiró, André Panisson, Michele Tizzoni, and Ciro Cattuto. 2016. Predicting human mobility through the assimilation of social media traces into mobility models. *EPJ Data Science* 5, 1 (2016), 30.
[7] Luís MA Bettencourt. 2013. The origins of scaling in cities. *science* 340, 6139 (2013), 1438–1441.
[8] Luís MA Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B West. 2007. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences* 104, 17 (2007), 7301–7306.
[9] Honor Bixby, James Bentham, Bin Zhou, Mariachiara Di Cesare, Christopher J Paciorek, NCD Risk Factor Collaboration, et al. 2019. Rising rural body-mass index is the main driver of the global obesity epidemic. *Nature* 569 (2019), 260–264.
[10] Andrea Bonanomi, Alessandro Rosina, Ciro Cattuto, and Kyriaki Kalimeri. 2017. Understanding Youth Unemployment in Italy via Social Media Data. In *28th IUSSP international population conference*.
[11] José González Cabañas, Ángel Cuevas, and Rubén Cuevas. 2018. Facebook use of sensitive data for advertising in Europe. *arXiv preprint arXiv:1802.05030* (2018).
[12] Lisa Carter-Harris, Rebecca Bartlett Ellis, Adam Warrick, and Susan Rawl. 2016. Beyond traditional newspaper advertisement: leveraging Facebook-targeted advertisement to recruit long-term smokers for research. *Journal of medical Internet research* 18, 6 (2016), e117.
[13] Nina Cesare, Hedwig Lee, Tyler McCormick, Emma Spiro, and Emilio Zagheni. 2018. Promises and pitfalls of using digital traces for demographic research. *Demography* 55, 5 (2018), 1979–1999.
[14] Benjamin Sage Crosier, Rachel Marie Brian, and Dror Ben-Zeev. 2016. Using Facebook to reach people who experience auditory hallucinations. *Journal of medical Internet research* 18, 6 (2016), e160.
[15] Yves-Alexandre De Montjoye, Laura Radaelli, Vivek Kumar Singh, et al. 2015. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* 347, 6221 (2015), 536–539.
[16] Jules Depersin and Marc Barthelemy. 2018. From global scaling to the dynamics of individual cities. *Proceedings of the National Academy of Sciences* 115, 10 (2018), 2317–2322.
[17] Paul DiMaggio, Eszter Hargittai, et al. 2001. From the 'digital divide'to 'digital inequality': Studying Internet use as penetration increases. *Princeton: Center for Arts and Cultural Policy Studies, Woodrow Wilson School, Princeton University* 4, 1 (2001), 4–2.
[18] Antoine Dubois, Emilio Zagheni, Kiran Garimella, and Ingmar Weber. 2018. Studying migrant assimilation through facebook interests. In *International Conference on Social Informatics*. Springer, 51–60.
[19] Masoomali Fatehkia, Ridhi Kashyap, and Ingmar Weber. 2018. Using Facebook ad data to track the global digital gender gap. *World Development* 107 (2018), 189–209.
[20] Masoomali Fatehkia, Dan O'Brien, and Ingmar Weber. 2019. Correlated impulses: Using Facebook interests to improve predictions of crime rates in urban areas. *PLOS ONE* 14, 2 (2019), 1–16. https://doi.org/10.1371/journal.pone.0211350
[21] David Garcia, Yonas Mitike Kassa, Angel Cuevas, Manuel Cebrian, Esteban Moro, Iyad Rahwan, and Ruben Cuevas. 2018. Analyzing gender inequality through large-scale Facebook advertising data. *Proceedings of the National Academy of Sciences* 115, 27 (2018), 6958–6963.
[22] Sofia Gil-Clavel and Emilio Zagheni. 2019. Demographic Differentials in Facebook Usage Around the World. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13(01). AAAI, 647–650.
[23] Kyriaki Kalimeri, Mariano G Beiró, Matteo Delfino, Robert Raleigh, and Ciro Cattuto. 2019. Predicting demographics, moral foundations, and human values from digital behaviours. *Computers in Human Behavior* 92 (2019), 428–445.
[24] Kyriaki Kalimeri, Mariano G Beiró, Alessandra Urbinati, Andrea Bonanomi, Alessandro Rosina, and Ciro Cattuto. 2019. Human Values and Attitudes towards Vaccination in Social Media. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, 248–254.
[25] Joshua D Kent and Henry T Capello Jr. 2013. Spatial patterns and demographic indicators of effective social media content during theHorsethief Canyon fire of 2012. *Cartography and Geographic Information Science* 40, 2 (2013), 78–89.
[26] Marc Keuschnigg, Selcan Mutgan, and Peter Hedström. 2019. Urban scaling and the regional divide. *Science advances* 5, 1 (2019), eaav0042.
[27] Steph Larsen. 2011. Welcome to the food deserts of rural America. https://grist.org/article/2011-01-21-welcome-to-the-food-deserts-of-rural-america/.
[28] Arch G Mainous and Francis P Kohrs. 1995. A comparison of health status between rural and urban adults. *Journal of Community Health* 20, 5 (1995), 423–431.
[29] Yelena Mejova, Harsh Rajiv Gandhi, Tejas Jivanbhai Rafaliya, Mayank Rameshbhai Sitapara, Ridhi Kashyap, and Ingmar Weber. 2018. Measuring Subnational Digital Gender Inequality in India through Gender Gaps in Facebook Use. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. ACM, 43.
[30] Yelena Mejova and Kyriaki Kalimeri. 2019. Effect of Values and Technology Use on Exercise: Implications for Personalized Behavior Change Interventions. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, 36–45.
[31] Yelena Mejova, Ingmar Weber, and Luis Fernandez-Luque. 2018. Online health monitoring using Facebook advertisement audience estimates in the United States: Evaluation study. *JMIR public health and surveillance* 4, 1 (2018), e30.
[32] S Murdock and DA Swanson. 2008. Applied demography in the twenty-first century.
[33] Steve H. Murdock, Michael Cline, and Mary Zey. 2012. *Challenges in the Analysis of Rural Populations in the United States*. Springer Netherlands, Dordrecht, 7–15. https://doi.org/10.1007/978-94-007-1842-5_2
[34] Somen Nandi, Saigopal Thota, Avishek Nag, Sw Divyasukhananda, Partha Goswami, Ashwin Aravindakshan, Raymond Rodriguez, and Biswanath Mukherjee. 2016. Computing for rural empowerment: enabled by last-mile telecommunications. *IEEE Communications Magazine* 54, 6 (2016), 102–109.
[35] OECD. 2018. *OECD Regions and Cities at a Glance 2018*. 168 pages. https://doi.org/10.1787/reg_cit_glance-2018-en
[36] OECD. 2018. *Rural 3.0 A framework for rural develppment*. Technical Report. OECD. https://www.oecd.org/cfe/regional-policy/Rural-3.0-Policy-Note.pdf
[37] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.
[38] Steffen Pötzschke and Michael Braun. 2017. Migrant sampling using facebook advertisements: A case study of polish migrants in four European countries. *Social Science Computer Review* 35, 5 (2017), 633–653.
[39] Francesco Rampazzo, Emilio Zagheni, Ingmar Weber, Maria Rita Testa, and Francesco Billari. 2018. Mater certa est, pater numquam: What can Facebook Advertising Data Tell Us about Male Fertility Rates?. In *Twelfth International AAAI Conference on Web and Social Media*.
[40] Jonathan Rigg, Anthony Bebbington, Katherine V Gough, Deborah F Bryceson, Jytte Agergaard, Niels Fold, and Cecilia Tacoli. 2009. The World Development Report 2009'reshapes economic geography': geographical reflections. *Transactions of the Institute of British Geographers* 34, 2 (2009), 128–136.
[41] Kevin M Roessger, Arie Greenleaf, and Chad Hoggan. 2017. Using data collection apps and single-case designs to research transformative learning in adults. *Journal of Adult and Continuing Education* 23, 2 (2017), 206–225.
[42] Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346, 6213 (2014), 1063–1064.
[43] David E Sahn and David C Stifel. 2003. Urban–rural inequality in living standards in Africa. *Journal of African Economies* 12, 4 (2003), 564–597.
[44] Nandita Saikia, Abhishek Singh, Domantas Jasilionis, and Prof Faujdar Ram. 2013. Explaining the rural-urban gap in infant mortality in India. *Demographic Research* 29 (2013), 473–506.
[45] Matthew J Salganik. 2019. *Bit by bit: social research in the digital age*. Princeton University Press.
[46] Indira Sen, Fabian Floeck, Katrin Weller, Bernd Weiss, and Claudia Wagner. 2019. A Total Error Framework for Digital Traces of Humans. *arXiv preprint arXiv:1907.08228* (2019).
[47] Terry Sicular, Yue Ximing, Björn Gustafsson, and Li Shi. 2007. The urban–rural income gap and inequality in China. *Review of Income and Wealth* 53, 1 (2007), 93–126.
[48] Elizaveta Sivak and Ivan Smirnov. 2019. Parents mention sons more often than daughters on social media. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 2039–2041. https://doi.org/10.1073/pnas.1804996116 arXiv:https://www.pnas.org/content/116/6/2039.full.pdf
[49] S Spyratos, M Vespe, F Natale, I Weber, E Zagheni, and M Rango. 2018. Migration Data using Social Media. *JRC Science Hub* (2018).
[50] G Thorvaldsen. [n.d.]. Handbook of international historical microdata for population research.
[51] Zeynep Tufekci. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Eighth International AAAI Conference on Weblogs and Social Media*.
[52] Jelte M Wicherts, Coosje LS Veldkamp, Hilde EM Augusteijn, Marjan Bakker, Robbie Van Aert, and Marcel ALM Van Assen. 2016. Degrees of freedom in

planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology* 7 (2016), 1832.

[53] Dilek Yildiz, Joanna Munson, Agnese Vitali, Ramine Tinati, Jennifer Holland, et al. 2017. Using Twitter data for demographic research. *Demographic Research* 37 (2017), 1477–1514.

[54] Alwyn Young. 2013. Inequality, the urban-rural gap, and migration. *The Quarterly Journal of Economics* 128, 4 (2013), 1727–1785.

[55] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, et al. 2014. Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 439–444.

[56] Emilio Zagheni and Ingmar Weber. 2015. Demographic research with non-representative internet data. *International Journal of Manpower* 36, 1 (2015), 13–25.

[57] Emilio Zagheni, Ingmar Weber, Krishna Gummadi, et al. 2017. Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review* 43, 4 (2017), 721–734.