March 2023

Unlocking the Power of Mobile Phone Data: A Novel Approach for Sociodemographic Estimation in Low-Income Countries

Abstract

Policy-making and social welfare improvement require an accurate comprehensive picture of a country's health through national statistics. However, traditional data-gathering methods require significant time and infrastructure which low-income countries lack the ability to do. Thus it forces low-income countries to rely on statistical estimations based on outdated national statistics that lead to biases and inaccurate conclusions about the health of their country. In an effort to improve accuracy, current researchers have analyzed the viability of mobile phone data in these statistical estimations. Mobile phone data offers to ability to obtain up-to-date data about a user's behavior and activity. Current research incorporates mobile phone data in well-established statistical modeling that showed better performance while also revealing significant biases in mobile phone data. To combat these shortcomings, researchers have developed specialized statistical estimations designed for mobile phone data with the inclusion of satellite imagery that led to higher accuracy. However, specialized methods still require significant knowledge and updating of a country's current economic literature. Thus dynamic statistical modeling like machine learning techniques arose to unclear results in improvement. Suggesting possible research is needed in improving machine learning techniques and the inclusion of more nontraditional datasets—like satellite imagery.

Introduction

A year removed from the heart of the COVID-19 pandemic, the World Health Organization announced, for 2020-2021, the estimated full global death toll to be 14.9 million. 9.5 million off from the initially reported death toll by government health institutions. A stark reminder of the deficiencies of death registrations globally. In fact, the scale of this issue points to a bigger problem: the incomplete traditional systems to gather sociodemographic attributes—like death toll.

This problem is even more evident in low-income countries. These countries lack the significant infrastructure costs and investment required in traditional person-to-person data-gathering methods such as censuses, civil registration, and household surveys. Furthermore, bias is inevitable. Rural areas are commonly left out and significant time lag can occur—data can take years to publish (Bwambale et al., 2020). All of this culminates into an incomplete comprehensive outlook of a country that can lead to misdirection in policymaking. For instance, a misappropriated geographic distribution of income can be used to make decisions on where resources are allocated. Thus, researchers have developed model-based estimation methods to fill the gaps in traditional data.

During the last decade, in an effort to remedy out-of-sample populations, like rural areas, researchers successfully developed small-area techniques for estimating sociodemographic indicators (Schmid et al., 2017). Small area estimation is a modeling technique used to estimate attributes in subpopulations with little to no samples. However, this technique still requires the availability of variables dependent on up-to-date traditional data. Making it susceptible to time

lags which is common among low-income countries. Therefore, research has now turned to non-traditional data sources such as satellite imagery and mobile phone data.

Mobile phone ownership and usage have increased significantly in low-income countries. This trend has spurred interest in the use of mobile phone data to estimate sociodemographic attributes in low-income countries. Call detail records (CDR) or mobile phone data commonly can feature a user's and receiver's call location via the nearest cell tower ping, the call duration, and the time of the call. These variables make it possible to provide up-to-date insight into a user's behavior and activities. Thus making it a possible solution to the deficiencies in traditional data common in low-income countries.

Therefore, it is imperative to evaluate the potential benefits and limitations of incorporating mobile phone data in estimating sociodemographic attributes in low-income countries. In this literature review, we aim to answer them under three contexts. We first evaluate its performance in statistical modeling traditionally used for sociodemographic indicators. Then evaluate it under more advanced and specialized statistical modeling specific to mobile phone data. Finally, examine mobile phone data through machine learning techniques in estimating indicators. Afterward, we will identify potential areas for future research and development in this field.

Traditional Statistical Modeling

The reproducibility of modeling estimation is crucial in understanding the viability of mobile phone data for estimating sociodemographic indicators. For statistical modeling to be viable in this context, it must be commonly used by most low-income countries. Thus it is important to first evaluate the performance of mobile phone data in statistical modeling techniques already established in sociodemographic estimation.

One of the most common and well-established modeling techniques is linear regression analysis. It aims at modeling a linear relationship between a dependent variable and its explanatory variables. Regression analysis offers the ability to predict variables, like sociodemographic indicators, based on various potential factors. The use of traditional data is commonly used in regression analysis to predict sociodemographic indicators in countries.

However, Milusheva (2019) evaluated the predictive power of mobile phone-based regression analysis in estimating the size and temporal changes in short-term movements in Senegal. Understanding short-term movement is critical for the government as it can lead to health consequences like the introduction of novel diseases. Thus there is a necessity for timely data to provide a scope in short-term movement—a shortcoming of census and household data. So, Milusheva turned to Senegal's largest mobile phone operator to provide call detail records data (CDR) of its subscribers in 2013. Milusheva defined the dependent variable through CDR data as residential movement: the change in a person's cell tower pings from one district to another in a given month. Then using ten percent of the sample in old census data and historical data on determinants of movement, like vegetation index or rainfall, to be its explanatory variables. Through the fitted regression analysis, it found the factors of the predicted value and actual short-term movement to be equal with a confidence level of 95%—a statistical threshold of significance. This demonstrates that linear regression even without readily available short-term movement data, is still possible to estimate the dynamics of movements through CDR data and out-of-date traditional data.

Limitations are still present. The analysis first required country-specific variables like vegetation index and rainfall that may not be a driver of short-term movement for non-

agriculturally dependent low-income countries. In fact, when removing country-specific variables, the predicted values were no longer significant enough to be equal to the actual value. Signifying that in order for the incorporation of mobile phone data to be viable in this case, researchers need significant research and analysis in understanding country-specific variables. A task itself that requires a notable amount of infrastructure and investment.

To remedy this limitation, Bwambale (2020)—in an effort to predict short-term movement in Dhaka Metropolitan Area in Bangladesh for 2013—compared discrete choice models based on noncountry-specific variables and CDR data. A discrete choice model is a traditional economic model that predicts decisions based on certain attributes. It attempted to predict two decisions: movement over time (temporal transferability) and movement over a region (spatial transferability). Along with CDR data, Bwambale used noncountry-specific variables like income and marital status from out-of-date censuses and household surveys between 2009-2012 as its attributes. They compared two discrete choice models. A model with just the noncountry-specific variables, and another that includes variables from CDR data. With the predicted values from the two discrete choice models, Bwambale calculated its root mean square error (RMSE)—a measure of the difference between the actual value and the predicted squared. It found that the RMSE of the CDR-included model was lower than that of the model without it. Suggesting a higher accuracy found in the CDR-included model. This presents discrete choice models demonstrating the ability to use out-of-date general traditional data and CDR to accurately predict sociodemographic indicators. Alleviating the pressure of providing adequate infrastructure in updating traditional data and research in country-specific variables.

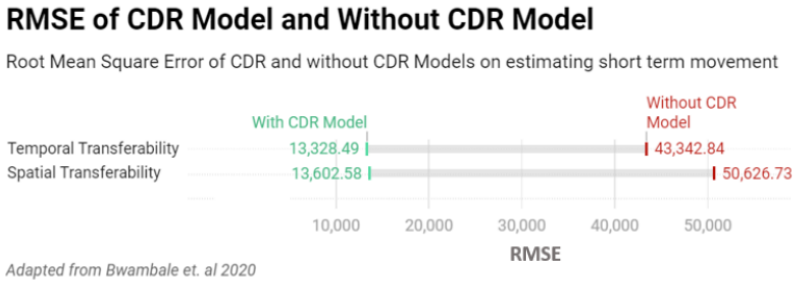

Figure 1: Root Mean Square Error of CDR Model and without CDR Model.
Created by DataWrapper.

However, the reliance on using both CDR and out-of-date traditional data still has its shortcomings in these modeling techniques. Beine (2020) when attempting to use CDR data and survey data to identify determinants of refugee mobility in Turkey notes potential biases in linking an individual's CDR data to their survey data. Linkage is needed when using multiple data sources in modeling. Researchers must be able to connect an individual's observations from one data source to the same individual's observation from another data source. For Beine, they noted that some refugees may share a non-refugee phone to conduct business—causing a linkage issue. This is largely due to refugees' phone contracts containing limitations on the number of calls they can make. Therefore, it signifies a potential underlying problem: phones shared amongst an impoverished commune can lead to potential biases in linking its CDR data to survey data.

Specialized Statistical Modeling

In order to reduce these biases, research turned to specialize statistical models designed for CDR data. In Beine's case, he modified an already-established technique of gravity modeling to remedy biases in CDR data. Gravity modeling uses two sets of variables to estimate spatial interaction such as movement between places. The standard literature on refugee mobility gravity modeling uses two sets of mobility determinants as its explanatory variables: GDP and factors relating to the attractiveness of a region—like political alliances. Similar to Milesheva's, Beine used CDR data to define movement, the dependent variable, from one area to another by cell tower ping. However, instead of being defined by individuals like previous models, they were aggregated by region to combat CDR biases. To do so, Beine employed the Nomenclature of Territorial Units for Statistics Administrative Level 2 (NUTS-2)—a geographical system used by the European Union (EU) that divides Turkey into 26 sub-regions. NUTS-2, designed for policy application, divides regions by similar socio-economic attributes. Beine argues the use of this spatial statistic aggregates data to remedy the bias of variability in linking issues in phone data as it summarizes into entire regions. Furthermore, he claimed the ability to now capture the movement of a whole family—as children and mothers typically never owned phones. The inclusion of spatial statistics such as NUTS-2 shows a possible proof-of-concept to further improve CDR data in sociodemographic estimation, but it still has flaws.

Figure 2: NUTS-2 Regions in Turkey. *Adapted from Beine et. Al (2020).*

Although NUTS-2 provided solutions to the limitations of traditional statistical modeling, the use of NUTS-2 is only designed for EU—largely higher-income countries. Since it divides and updates regions by socioeconomic attributes, NUTS-2 relies annually on traditional data. This is difficult for low-income as it already does not have the necessary infrastructure to derive sociodemographic statistics consistently. However, this imposed a question of if spatial statistics is the answer to combat limitations in mobile phone data.

Steele (2017) answered this question in Bangladesh using the nation's geographic information system data (RS) as its spatial statistics. RS data divides regions based on satellite imagery through remoting sensing—a process that detects and collects physical characteristics of an area via satellite radiation. RS data does not rely on traditional data sources and is relatively simple to gather. Steele with the RS data used a hierarchical Bayesian geostatistical modeling (BGM) to predict poverty metrics—like income—in Dhaka, Bangladesh. BGM is a statistical framework that combines spatial modeling—an approach to describing the interaction of spatial features—and Bayesian inference, which uses prior information like RS data as its basis. Steele compared a model with just CDR data and another with both RS and CDR data included as its factors. The result showed a high correlation—correlation coefficient threshold of above 0.7—between CDR-RS models' predicted income values to the actual statistics at urban and national levels. However, the CDR-only models performed significantly worse at those levels. Steele

suggested a possible cause is using only CDR along with traditional data does little to reduce the variability in measuring short-term data. Short-term data is susceptible to sudden changes in this case like the time between switching phones. Unlike CDR data, which retrieves data on a minute basis, RS data is taken over the course of a month—remedying the variance. Thus signifying a possible solution to the limitation of CDR data with the inclusion of satellite imagery like RS data.

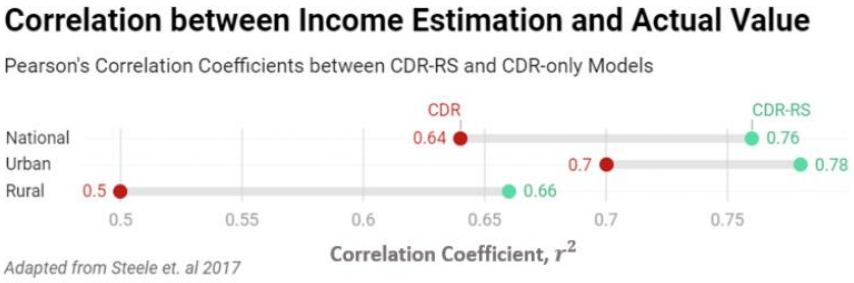

Figure 3: Correlation Coefficient between CDR-RS Models and CDR Models. *Created by DataWrapper.*

It is also important to note, for both CDR-RS and CDR-only models, the correlation was insignificant at the rural level as both failed to cross the threshold of high correlation. A large reason due to the lack of samples in rural areas. A problem shared by traditional data and mobile phone data.

The significant improvement in resolving out-of-sample regions came from researchers at the Free University of Berlin. Schmid (2017) proposes the inclusion of mobile phones and satellite imagery in a popularized small area estimation technique—Fay-Herriot estimation. The Fay-Herriot model estimates a certain subregion's variable—like literacy rate—through linear regression and random effects. Random effects are the assumption of randomness to occur in a given area. Adapting to mobile phone data, Schmid aggregated regions through the nation's geographic information system (GIS) satellite imagery. Using mobile phone data and old census data as its factors, the results found only a slight underestimation of the official statistics of literacy rates in Senegal for 2011. More significantly, the results showed estimations in communes that were out-of-sample to be similar to estimations derived from up-to-date 2011 census and household surveys via traditional small area estimation techniques. Remedying another bias of traditional data—out-of-sample regions.

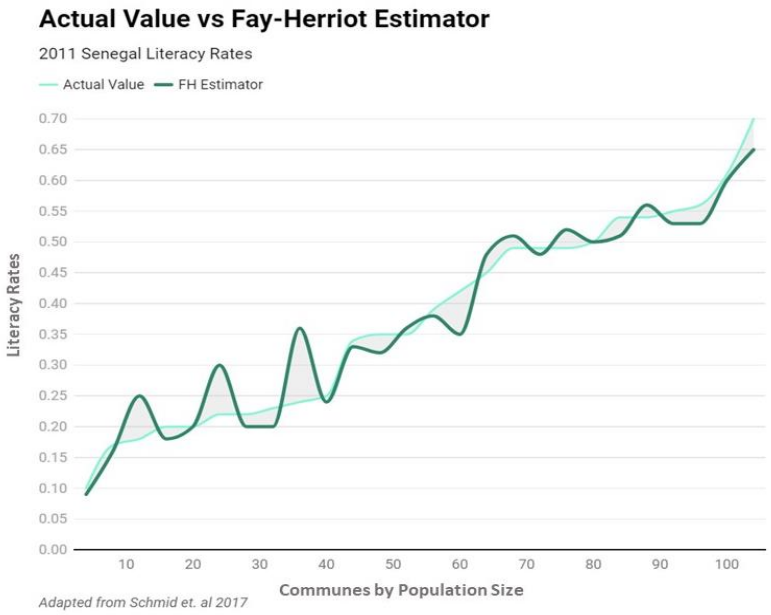

Figure 4: Literacy Rates between the Actual Value and the Estimated across Senegal Communes. *Created by DataWrapper.*

Machine Learning Techniques

However, one limitation still persists throughout these statistical models: it requires significant sensitivity analysis. Sensitivity analysis is a trial-and-error approach to understanding which variables among CDR and other data sources are useful for estimating sociodemographic indicators. For instance, it is impossible to know how movement captured in phone data can lead to a certain conclusion. A movement to a region in one month can mean something totally different after the region is inflicted with economic toil the next month. This requires a statistics institute already in place among low-income countries to consistently update its models as the landscape of their countries changes over time. A possible solution is developing an adaptive and dynamic statistical model that automatically adjusts to a given landscape—machine learning models.

Aiken (2020) explored implementing a machine learning model using CDR data and household survey data to accurately measure poverty in Afghanistan. Similar to all previous models, Aiken fitted old household survey data (2015-2018) and CDR variables as the explanatory variables. Then subjected it to a gradient boosting model—a popular machine learning technique. The gradient boosting model finds the best model at each iteration minimizing the prediction error. They do this by finding the gradient (the direction) of the significance of each variable and adjusting it towards that direction at each iteration. Thus, the model optimizes which variable is significant for a given indicator it wants to estimate. This allows the model to automatically adjust when new data is fed into the system. The results confirmed prior model findings that combining CDR and survey data performs better than just one of the data sources. The Area Under the Curve (AUC) for the singular data source models ranged between 0.68-0.72. This implies around 68-72% of the time the model can accurately segregate between two data points. Whereas the combined method yielded an AUC range between 0.75-0.76—a statistically accepted range of accuracy. However, the author noted a shortcoming to this model as the CDR-based models did not perform as well with populations that do not own phones. This is where satellite imagery enters the picture again.

Aiken and Blumenstock (2022) following previous trends added satellite imagery to a gradient-boosting model. This time, the focus was on evaluating poverty in Togo. Aiken and Blumenstock developed high-resolution satellite imagery that subjected Togo to designated grid cells. This aggregation, as previously stated, ideally hopes to dilute missing data like individuals without phones. The results led to a drop in performance with AUC scores ranging from 0.59 to 0.64. Thus signifying the reemerging issue with CDR data estimation—each model is heavily dependent on the region. Although satellite imagery proved successful in regions like Turkey and Bangladesh, it did not under a machine-learning model in Togo.

Conclusion

Consistently throughout the various statistical modeling, mobile phone data have shown higher accuracy when combined with out-of-date field surveys than without it. Attribution to better performance comes from the nature of the accessibility of CDR data. Traditional modeling techniques without CDR data rely solely on out-of-date traditional data, whereas the inclusion of CDR data offers these models the ability to adjust to a more up-to-date outlook on the country. Furthermore, research has shown statistical models that have been specialized for mobile phone data performed significantly better in identifying sociodemographic indicators than traditional models without the specialization.

However, there are significant exclusions that come with mobile phone data. Although mobile phone usage continues to increase across countries, for many low-income countries it is still inaccessible for the poorest part leading to inaccurate estimations. Mobile phone data is also measured on a short-term basis, thus it is susceptible to high variance. A possible solution comes in the form of including other forms of non-traditional data that are measured at longer intervals like satellite imagery. However, the improvements in performance are still unclear for all models. Thus, further research is necessary on combining CDR data with other forms of nontraditional data that can possibly dilute the limitations that are evident in mobile phones.

Finally, current statistical modeling with mobile requires significant sensitivity analysis. That in its own right requires the assumption of a proper statistical institute in each low-income country to perform timely analysis. Unfortunately, current solutions like machine learning is still relatively young and yields unclear results. Therefore further research is necessary for investigating comprehensive dynamic modeling—like machine learning models—to counteract this shortcoming.

References

Aiken, E. L., Bedoya, G., Blumenstock, J. E., & Coville, A. (2020). Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in Afghanistan. *Journal of Development Economics*, *161*, 103016. https://doi.org/10.1016/j.jdeveco.2022.103016

Aiken, E., Bellue, S., Karlan, D., Udry, C., & Blumenstock, J. E. (2022). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, *603*(7903), 864–870. https://doi.org/10.1038/s41586-022-04484-9

Beine, M., Bertinelli, L., Cömertpay, R., Litina, A., & Maystadt, J.-F. (2021). A gravity analysis of refugee mobility using mobile phone data. *Journal of Development Economics*, *150*, 102618. https://doi.org/10.1016/j.jdeveco.2020.102618

Bwambale, A., Choudhury, C. F., Hess, S., & Iqbal, M. S. (2020). Getting the best of Both worlds: A framework for combining disaggregate travel survey data and aggregate mobile phone data for trip generation modelling. *Transportation*, *48*(5), 2287–2314. https://doi.org/10.1007/s11116-020-10129-5

Milusheva, S. (2019). Predicting dynamic patterns of short-term movement. *The World Bank Economic Review*, *34*(Supplement_1), 26–34. https://doi.org/10.1093/wber/lhz036

Schmid, T., Bruckschen, F., Salvati, N., & Zbiranski, T. (2017). Constructing sociodemographic indicators for national statistical institutes by using Mobile Phone Data: Estimating literacy rates in Senegal. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(4), 1163–1190. https://doi.org/10.1111/rssa.12305

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.-A., Iqbal, A. M., Hadiuzzaman, K. N., Lu, X., Wetter, E., Tatem, A. J., & Bengtsson, L. (2017). Mapping poverty using mobile phone and Satellite Data. *Journal of The Royal Society Interface*, *14*(127), 20160690. https://doi.org/10.1098/rsif.2016.0690

World Health Organization. (2022, May 5). *14.9 million excess deaths associated with the COVID-19 pandemic in 2020 and 2021*. World Health Organization. Retrieved February 22, 2023, from https://www.who.int/news/item/05-05-2022-14.9-million-excess-deaths-were-associated-with-the-covid-19-pandemic-in-2020-and-2021

Rhetorical Analysis

The audience for my literature review is probably undergraduates and possibly graduates interested in economics and statistical methods. Therefore the audience is probably two-fold: economists focused on low-income countries and statisticians. Due to the difference in audience, it will be important to define certain economic and statistical terminologies, so that it is comprehensible for both audiences. This would also require more emphasis on the big picture of each statistical modeling process, as focusing on the jargon and process of the certain statistical model may not be of interest to the general audience compared to those in my department.

I plan on organizing my literature review by first establishing the need for this research in the introduction. Then explain the possible benefits and limits of mobile phone data under three modeling frameworks. We will look at it under the context of traditional statistical methods, then specialized methods, and finally machine learning methods. The reason for this organization is that each method builds on the shortcomings of the previous methods.

I think I received some very good feedback from Karen and Cooper. Karen suggested improving the transitions in my Traditional Statistical Modelling section. I adjusted the order of the paragraphs and also moved a body paragraph from the next section to this one, so the text flows more intuitively. Furthermore, I took Cooper's suggestion of breaking up my paragraph about linear regression as it was incredibly long. From their suggestions, I clearly defined the subjects in some of my sentences—a lot of sentences were ambiguous with just "it" as the subject.

I think the final version was decent but not as good as I hoped it would be. I think the cohesions sometimes lacked given it was jumping from one statistical model to another. But overall, the assignment was pretty fun. It was really nice actually exploring ways my statistical research can practically impact society. I think the AB and Matrix contributed a lot to the assignment. It definitely relieved a lot of the pressure of the contents of the lit review and allowed me to focus on the narrative/cohesion of telling a story. If had more time, I would have hoped to do more research into the machine-learning aspects of this topic.