Mapping multi-dimensional poverty by combining satellite and mobile phone data: Challenges and Opportunities

Neeti Pokhriyal

[1]Department of Computer Science
Dartmouth College, NH

neeti.pokhriyal@dartmouth.edu

Abstract. *This paper presents innovative scientific research in predicting and mapping multi-dimensional poverty using computational methods that combine non-traditional data sources, including satellite imagery and mobile phone data, with traditional data sources, like census and surveys. These methods can provide poverty estimates at a higher frequency and at finer spatial resolution for improved targeting of policy interventions. These poverty estimates can be produced for intercensal/intersurvey periods when no or scant survey data are available and can be used in regions of conflict or fragility or those recovering from natural disasters. Despite the technical feasibility of such alternate data and methods, the paper also discusses important challenges related to data privacy and the assessment of algorithmic biases in the model's estimates that need consideration. To accelerate global actions for poverty eradication, the paper ends with key recommendations on using novel data and methods that can assist in targeted policy interventions.*

1. Introduction

Timely, accurate, and spatially fine-grained poverty estimates at policy planning microregions are essential to determining policies for reducing poverty. Existing ways to estimate poverty rely on socioeconomic household surveys or censuses, which are costly and time-consuming, and hindered by pandemics and wars/conflicts. These estimates are generally available at coarse spatial resolution and are delayed for poorer economies by many years, making timely updates on poverty challenging.

There is growing research in realizing the potential of combining alternate data using computational methods to understand poverty and associated deprivations [Pokhriyal and Jacques 2017, Steele et al. 2017, Njuguna and McSharry 2017]. Poverty is a complex phenomenon, and understanding it from the *lens* of disparate datasets has been demonstrated to provide better accuracies in predicting the deprivations than a single-source dataset.

Satellite data[1] has been widely studied to predict socioeconomic deprivations as the data can capture various geographic factors, including night-time lights, vegetation cover, agro-meteorological conditions, access to natural resources and man-made structure, including the urban build-up, and accessibility and proximity to markets, schools,

[1] "Satellite data" is used as an encompassing term to denote all data and data products produced from earth observations satellites. Also referred to as remote sensing data and earth observation data.

etc. However, they lack information about population structure, especially the socioeconomic ties, cultural interactions, and micro-/macro-behavior essential to understanding poverty.

The widespread use of **digital technologies**, especially mobile phones, provides one such *lens* on societal interactions [Eagle et al. 2010]. A call data record (CDR) is generated each time a mobile call is placed. **CDRs**[2] to capture how, when, where, and with whom the individual communicates. These data, traditionally used by telecommunication companies for billing purposes, capture both micro/macro patterns of human interaction. The information, typically captured from mobile phone data for predicting poverty metrics includes aggregated statistics about mobility, social ties, data on phone usage, and data on the amount and frequency of recharges (top-ups).

Existing works have studied various definitions of poverty, including income-based poverty, asset-based wealth index, and OPHI's Multidimensional poverty index (MPI). Some of the case studies demonstrating that using CDR and satellite imagery provides better poverty performance compared to using either of the datasets are done for Senegal [Pokhriyal and Jacques 2017], Haiti [Pokhriyal et al. 2020], and Bangladesh [Steele et al. 2017].

Combining these disparate datasets with surveys and census to predict poverty is a **computationally challenging** task, and some of the salient challenges are highlighted below:

- Each dataset has **heterogeneous statistical properties** and exists at different spatial and semantic granularity. CDRs are available for each subscriber, satellite data exists as image pixels with varying spatial resolution, and census and survey data are available for either individuals or households.
- Satellite imagery is publicly available to researchers and typically has no privacy constraints, especially at the resolution analyzed in extant works for poverty prediction. **Accessibility** to CDR data is contentious owing to privacy and/or business/proprietary concerns of telecom providers.
- **Validation of the estimates** needs the existence of ground-truth data. Most existing works validate their poverty estimates for time points coinciding with surveys when training and validation data are readily available. There is scant work in studying the temporal evolution of these estimates beyond the survey/census years and for fragile regions, paradoxically when these are most needed.

The rest of the paper is organized as follows: In Section 2, a case study on Senegal [Pokhriyal and Jacques 2017] is described which is done in collaboration with the National Statistics Office of Senegal and Sonatel (the predominant telecom provider in Senegal). Section 2 focuses on the results of poverty mapping using mobile phone data and satellite data. Section 3 discusses some challenges to putting this research into practice, and Section 4 puts forth key recommendations so that these innovative methods can drive poverty eradication efforts.

[2] CDR or mobile phone data are used interchangeably in the literature, and, here, we use the term CDR.

Summary Statistics	CDRs	Environment Data	Census	Poverty Index
Timeline	Jan-Dec 2013	1960-2014	2013	2013
Total calls & text	11 Billion	N/A	N/A	N/A
Unique individuals	9.54 M	N/A	1.4 M	N/A
Spatial granularity	Antenna-level (1666)	vector data - 100 m -1 km	Household-level	Region-level (14)
Cost incurred in data collection & preparation	Low/no cost (data exhaust)	Low/no cost (data exhaust)	USD 29 Million	Very high cost, and human expertise
Frequency of update of data	Real-time	∼1 year	10 years	3-5 years

Table 1. Summary statistics and characteristics of the data used - CDRs, environment, census, MPI poverty index.

2. Case study for poverty prediction and mapping in Senegal using CDR and satellite imagery

In this case study, the “satellite imagery” data capture information related to **food security** (availability and access components), **economic activity**, and **access to services**, like proximity to schools, markets, etc. The CDR data captures the **basic phone usage statistics** of a user, and also the **regularity, diversity, and spatiotemporal variability** in the user’s mobile interactions. The details are given in Table 1.

The study is done in the year 2013. The CDRs span the entire year of 2013 for 9.54 million subscribers (the population of Senegal in 2013 was 14.13 million), and the satellite data is based on remote sensing data covering concurrent time periods. The poverty maps are produced at the spatially finest level of policy planning, called *communes* in Senegal, and validated at that level using the 2013 census data.

Processing the mobile phone and satellite imagery data Significant computational resources are needed to store and process (i.e., extract quantitative features) mobile phone data and satellite imagery.

For CDRs, it involves **localizing the subscribers** to their home antenna and then extracting the features of interest for these subscribers, and later aggregating the antenna-level features at communes. Figure 1 depicts the antenna locations used in the Senegal study. This work focuses on well-studied metrics capturing the individualistic, spatial, and temporal patterns of the subscriber [de Montjoye et al. 2016, Steele et al. 2021]. The individual aspects, namely the number of active days, the number of contacts, the average call duration, and the percent nocturnal quantify the typical use pattern of an individual. Spatial metrics, like the radius of gyration, and entropy of antennas, quantify the typical movement pattern of an individual. All features were calculated at monthly granularity capturing the temporal patterns of a subscriber.

Note that this approach does not depend on explicitly linking individuals from mobile phone data to census records, since information is anonymized for both datasets and there is no way to link the records across these two data sets. Rather, we localize individuals (and households) to their respective communes using census information and localize the mobile phone subscribers to communes via their home antenna information.

Three broad categories of features are extracted from satellite data: food security (mainly quantified by agrometeorological measurements (temperature, precipitation, slope, elevation, and soil type) that drive agricultural production), economic activity (intensity of urbanization, night-time lights), and access to services (proximity to school, water towers, and hospitals).

Computational Model The class of computational models used in this (and related) works belongs to the class of geo-statistical models, which have been known to work well with the characteristics of geolocated datasets (modeling spatial autocorrelation, handling missing data, etc.).The idea, used in this work, is to build individual non-linear regression models [Rasmussen and Williams 2006] on each data source and then combine the probabilistic outputs from these models. There are two distinct **advantages of this combination approach**. First, each data source **remains private** to its ecosystem, as only the output predictions are shared. This mitigates some of the critical privacy concerns when employing mobile phone data. Second, **more disparate data sources** (e.g. local data, or aerial imagery at different resolutions) can be easily added to the computational framework.

Results This study on Senegal uses Multidimensional Poverty Index (MPI) as a measure of poverty and provides estimates for the headcount of poverty (H) and the average intensity (A) among the poor. The study also provides estimates along the three dimensions of MPI — education (years of schooling, school enrollment), health (malnutrition, child mortality), and standard of living.

The **predicted map of MPI for Senegal** at the commune level is depicted in Figure 2 top. The map on the bottom depicts MPI at the commune level calculated using OPHI's methodology and employed census data. We refer to this map as the ground truth and compare the model's estimates against it.

Each individual deprivation indicator is taken as the regression target of our computational framework, and the averaged spatial cross-validated results along the three dimensions are reported in Table 2. As a comparative study of how the model performs using multi-source, and single-source data, we experimented with three datasets - Multi-source, CDR, and Environment to predict the headcount of poverty (H), the intensity of poverty (A), and poverty index (MPI) at commune-level (see Table 2). Highly accurate results for all three targets (H, A, and MPI) were reported. Rank correlations are preserved, with Spearman's correlation of 0.85 for both H and A. The results demonstrate

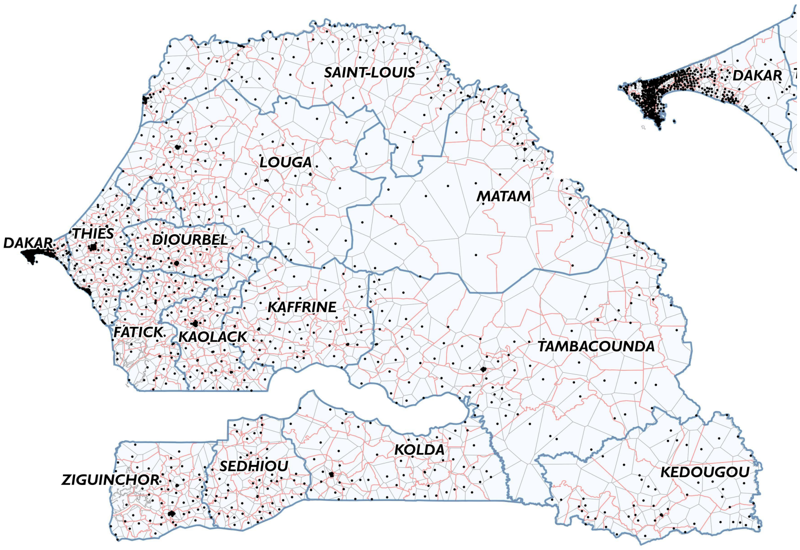

Figure 1. A composite map of Senegal, with black dots depicting the location of mobile towers (antennas). The Voronoi tessellation formed by these towers is shown in gray. The commune (which is the finest administrative unit in Senegal) boundaries are shown in red. There are 552 communes with 431 rural communes and 121 urban centers. The navy blue boundaries are those of regions, the coarsest administrative unit in Senegal. There are 14 regions, which are named on the map.

that combining multiple data sources (CDRs and satellite imagery) results in a consistent improvement of accuracy over using individual data sources.

As seen in Table 2, the model performs poorly for the indicators within the health dimension, i.e., child mortality and nutrition. This is attributed to the fact that our data are not representative of the children population, and, thus, the features extracted from CDR data do not capture this deprivation. Similar inference can be drawn for poorer correlations for nutrition.

The left panel of Figure 3 plots the relationship between MPI values predicted by our model and those estimated from the census. Since the points lie along the diagonal, a linear relationship, in general, is observed for MPI, with lower poverty values for urban areas (shown in red) and higher values for rural areas (shown in blue). Figure 3 (top right panel) depicts the scatter plot of our predictions for asset ownership, while the bottom right panel depicts the scatter plot for the education dimension of MPI.

3. Discussion

Beyond the technical feasibility of these research works, some of the **important challenges** that need consideration are described below, along with their mitigation steps

Data Governance issues related to responsible data collection, data management, and data sharing A important question that comes up is *How to incentivize sharing of mobile phone data that is owned and monetized by private companies, and does not contain subscriber's consent for its use for driving national statistics?* Mobile phone

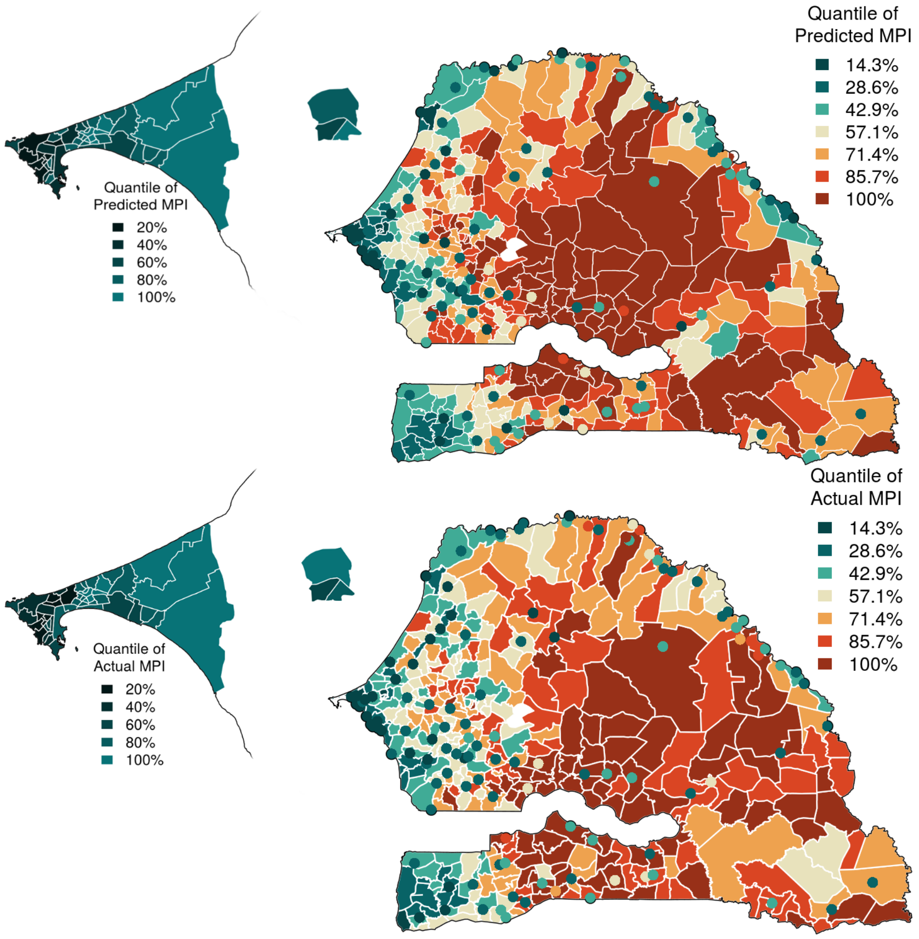

Figure 2. Quantiles of predicted (top) and actual (bottom) MPI at the commune level. The urban centers are depicted in small circles on the map. The communes in Dakar and Thiès regions are shown enlarged.

data use might be misconstrued as surveillance. It is to be emphasized, here, that current works do not use the content of the call and work only with anonymized CDR data. A lot of research is currently underway in striking a balance between data privacy and its usability to derive national statistics [Hotz et al. 2022], as well as in developing sustainable models of responsible data sharing [Stoyanovich et al. 2022], which will, likely, benefit in achieving the **twin goals** of using data that have privacy related sensitivities and their use for estimating national statistics.

Ways to facilitate inclusion of potentially missing sub-demographics like children or ultra-poor Some demographic sub-groups like children and ultra-poor, are left out by the analysis while only using only mobile phone data. To mitigate this challenge, **integrating** more data sets, such as (digital) data related to hospitals, schools, and those collected by the local governments in response to the poverty eradication and intervention programs (via e-governance initiatives) have the potential to provide a more holistic analysis on the dimensions of poverty.

Table 2. Spatially-cross-validated results of the predictions of MPI, Headcount of poverty (H), and Intensity of poverty (A), along with the individual indicators for poverty given by our model using disparate datasets. The results are compared when single source data is available. corr. – Pearson's r correlation, rank corr. – Spearman's rank correlation, and Error – Root Mean Square Error. For both types of correlations, all p-values were less than 10^{-20}. A standard deviation associated with the multiple runs for each measurement is reported within simple brackets.

	Multi-source Data		CDR		Satellite	
Poverty Indicators & Dimensions	rank corr.	Error	rank corr.	Error	rank corr.	Error
MPI	0.88 (0.06)	0.08 (0.01)	0.86 (0.07)	0.08 (0.01)	0.80 (0.10)	0.10 (0.02)
H	0.85 (0.08)	10.79 (3.96)	0.84 (0.08)	10.76 (2.60)	0.75 (0.11)	13.65 (4.86)
A	0.85 (0.07)	4.71 0.96)	0.82 (0.08)	4.98 (1.14)	0.79 (0.08)	5.36 (0.75)
Education	0.84 (0.05)	11.84 (1.88)	0.81 (0.07)	13.08 (1.68)	0.74 (0.07)	14.98 (3.03)
Health	0.50 (0.16)	12.76 (2.12)	0.52 (0.12)	12.91 (1.92)	0.35 (0.23)	13.91 (2.32)
Standard of Living	0.75 (0.13)	14.82 (3.92)	0.74 (0.11)	15.24 (3.45)	0.64 (0.20)	17.88 (4.50)

Mitigating potential biases in data A key technical concern associated with using mobile phone data for deriving population-wide metrics is the selection bias arising from the mobile phone ownership of a particular telecom provider; or bias in satellite imagery (mostly along the urban-rural divide). Better coverage of CDR data, data from more telecom providers in the country, and higher resolution satellite data should benefit the modeling task.

Careful **assessment** of biases existing in input data and data products is essential to ensure that the poverty predictions based on alternate data do not introduce systematic errors. Major research efforts are underway in this direction [Aiken et al. 2023, Pestre et al. 2020], which will, again, benefit in contextualizing the estimates of these models (one possible way is by providing the error bounds with the estimates).

4. Key recommendations on going forward

Given the multiple crises and the urgent need to explore innovative approaches for poverty monitoring and eradication, this paper puts forth the following recommendations.

1. There is a need to build **public-private partnerships** that involve the National Statistics Offices (NSOs), researchers, government agencies (that traditionally focus on poverty eradication efforts), as well as the private companies, like the telecom providers and those in possession of satellite imagery, to discuss **participatory mechanisms of responsible data sharing** among these entities that can support providing accurate estimates of poverty while preventing the misuse of data or models. It is imperative that in an effort that no one is left behind in addressing poverty and inequality diverse digital datasets that are collected

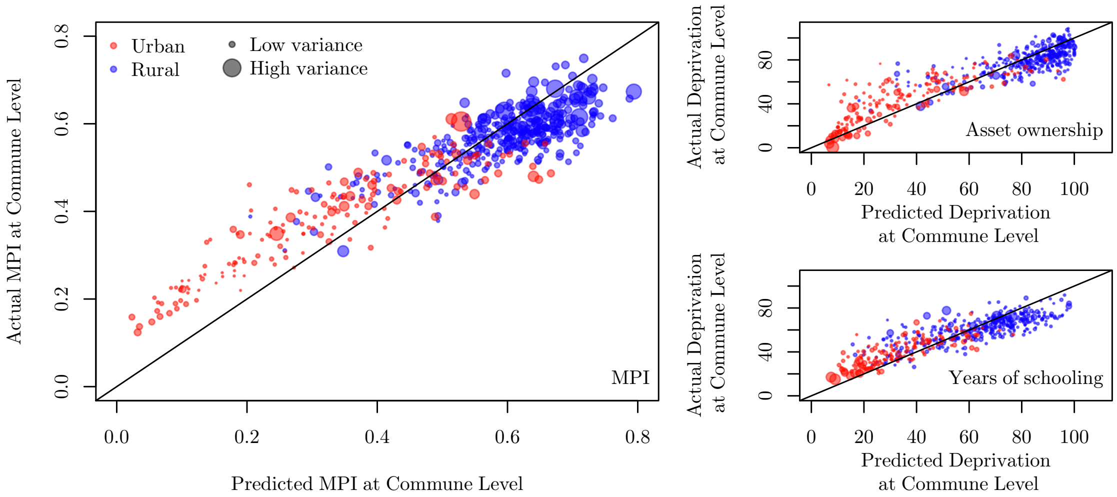

Figure 3. The left panel denotes the comparison of actual and predicted MPI values for all communes and urban areas of Senegal. The rural and urban areas are differentiated using blue and red colors respectively. The size of the circle denotes the variance of MPI prediction for that commune. The top right panel shows how the actual and predicted values compare for asset ownership, while the one on the bottom shows the comparison for years of schooling.

by the local governments in response to poverty eradication and intervention programs (via e-governance initiatives) be explored in conjunction with satellite and mobile phone data.

2. There is an urgent need to develop the **workforce and infrastructural, computing, and technical capacity of the National Statistical agencies** for African countries. Since the innovative approaches to poverty monitoring and eradication are data-intensive, empowering the NSOs with the technical know-how will have dual benefits. First, it will **strengthen** their capacity to produce poverty estimates from alternate data. Second, it will open new ways to **enhance the traditional survey and census-based data** collection and produce frequent and interim statistics that can drive improved decision-making, facilitating recovery efforts from the poly-crises.

3. Concerted efforts that **scale** the successful methodologies described here to **other countries and to more intercensal time periods** should be explored to accelerate coordination efforts in achieving the SDGs. This will require **multilateral partnerships** among the NSOs/government agencies of different countries and, also, importantly sustained collaborations with the researchers and participatory involvement with the local universities and communities.

Acknowledgement

This paper was developed for the UN Inter-agency Expert-group Meeting on “Implementation of the Third United Nations Decade for the Eradication of Poverty (2018-2027)” held in Addis Ababa on May 10-12 2023. Responsibility for the content rests entirely upon the author.

References

[Aiken et al. 2023] Aiken, E., Rolf, E., and Blumenstock, J. (2023). Fairness and representation in satellite-based poverty maps: Evidence of urban-rural disparities and their impacts on downstream policy. *The International Joint Conference on Artificial Intelligence, 2023*.

[de Montjoye et al. 2016] de Montjoye, Y.-A., Rocher, L., and Pentland, A. S. (2016). bandicoot: a python toolbox for mobile phone metadata. *Journal of Machine Learning Research*, 17(175):1–5.

[Eagle et al. 2010] Eagle, N., Macy, M., and Claxton, R. (2010). Network diversity and economic development. *Science*, 328(5981):1029–1031.

[Hotz et al. 2022] Hotz, V. J., Bollinger, C. R., Komarova, T., Manski, C. F., Moffitt, R. A., Nekipelov, D., Sojourner, A., and Spencer, B. D. (2022). Balancing data privacy and usability in the federal statistical system. *Proceedings of the National Academy of Sciences*, 119(31):e2104906119.

[Njuguna and McSharry 2017] Njuguna, C. and McSharry, P. (2017). Constructing spatiotemporal poverty indices from Big Data. *Journal of Business Research*, 70:318–327.

[Pestre et al. 2020] Pestre, G., Letouzé, E., and Zagheni, E. (2020). The abcde of big data: assessing biases in call-detail records for development estimates. *The World Bank Economic Review*, 34(Supplement_1):S89–S97.

[Pokhriyal and Dong 2015] Pokhriyal, N. and Dong, W. (2015). Virtual network and poverty analysis in Senegal. *D4D Challenge Senegal Scientific Papers, Netmob, MIT.*

[Pokhriyal and Jacques 2017] Pokhriyal, N. and Jacques, D. C. (2017). Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences*, (46).

[Pokhriyal et al. 2022] Pokhriyal, N., Letouze, E., and Vosoughi, S. (2022). Accurate nowcasts of energy access at policy-planning microregions to track sustainable development goals. *EPJ Data Science*.

[Pokhriyal et al. 2021] Pokhriyal, N., Valentino, B., and Vosoughi, S. (2021). An interpretable model for real-time tracking of economic indicators. *ACM Transactions on Data Science*.

[Pokhriyal and Vosoughi] Pokhriyal, N. and Vosoughi, S. Assessing countrywide socioeconomic deprivations using auxiliary data sets. *AI for Africa for Sustainable Economic Development Workshop, ACM International Conference on AI in Finance (ICAIF '20).*

[Pokhriyal et al. 2020] Pokhriyal, N., Zambrano, O., Linares, J., and Hernandez, H. (2020). Estimating and forecasting income poverty and inequality in haiti using satellite imagery and mobile phone data. *Working Paper, Inter-American Development Bank.*

[Rasmussen and Williams 2006] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

[Steele et al. 2021] Steele, J. E., Pezzulo, C., Albert, M., Brooks, C. J., zu Erbach-Schoenberg, E., O'Connor, S. B., Sundsøy, P. R., Engø-Monsen, K., Nilsen, K., Graupe, B., et al. (2021). Mobility and phone call behavior explain patterns in poverty at high-resolution across multiple settings. *Humanities and Social Sciences Communications*, 8(1):1–12.

[Steele et al. 2017] 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., et al. (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127):20160690.

[Stoyanovich et al. 2022] Stoyanovich, J., Abiteboul, S., Howe, B., Jagadish, H., and Schelter, S. (2022). Responsible data management. *Communications of the ACM*, 65(6):64–74.