Technical Report

Monitoring multidimensional poverty using mobile phone metadata: A feasibility study from Al Gezira Sudan

May 2018

Draft Version: Do not circulate without the consent of the authors

PREPARED FOR

UNITED NATIONS DEVELOPMENT PROGRAMME SUDAN

PREPARED BY

FABIAN BRUCKSCHEN

TILL KOEBE

Acknowledgements

The authors of this report would like to thank Anisha Thapa and Azza Nabil (UNDP) and Mohammed Abdalla Rahamtalla and his team (Telecommunications and Post Regulatory Authority) for making this project possible. The authors would like to extend their thank you to Dr. Karamallah and his team (Central Bureau of Statistics). Furthermore, special thanks goes to Mukhtar Isameldin, Hafiz Alhaj Mustafa and Abdullah Saieed from Zain for their technical support.

Executive Summary

1. Monitoring needs are growing rapidly, both for policymaking as well as for international development interventions. The standard set of tools to gather information - surveys and censuses - face a strong trade-off between timeliness and granularity on one hand and costs on the other hand.

2. This study is a proof-of-concept that sets out to explore mobile phone metadata as a scale-invariant method to be added to the standard set of tools. Mobile phone metadata, specifically call detail records, reflect human behaviour and are available for a large part of the population.

3. In the study, a household survey - conducted in Q1 2018 in the state of Al Gezira, Sudan - is augmented with covariates extracted from anonymized call detail records from a major Sudanese mobile network operator and used to calculate proxy indicators for the multidimensional poverty index.

4. The multidimensional poverty index and covariates from call detail records show high correlation (> 0.7). The derived proxy indicators favourably compare with the survey benchmark in terms of bias, root mean square error and adjusted R^2 (>0.7).

5. The use of call detail records allow for a higher geographical resolution of the results. Antenna-level estimates of multidimensional poverty show - as expected - higher variability compared to locality-level values. Indirect validation via aggregation show good antenna-level performance; direct validation, however, is not possible in this study setup.

6. For a comprehensive assessment of the feasibility and scalability of the approach it is recommended to verify the proxy estimates using additional temporal and cross-sectional validation data.

Contents

Introduction

Lifting people out of poverty has been and still is a major goal of both national policy-making and international development. In many parts of the world significant progress has been made, however, one of the lessons looking back at the millenium development goals is that anti-poverty programmes need to become more effective. This call directly challenges monitoring efforts in their current form. Censuses and surveys heavily draw on resources, reliable administrative data systems require a robust institutional infrastructure and long time intervals between data capture and the release of the results have created a lack of good monitoring data.

The rapid change of the last two decades due to digitization has now created an opportunity to expand the tools of monitoring: passively collected data on large parts of the population. Mobile phones have spread across the globe and have become ubiquitous in many people's life. Calls are made, money is transferred, moments are captured - always with the mobile phone within reach. Consequently, the data created hereby can form digital shadows of the users, providing information on a user's social network and their mobility profiles. When handled responsibly, data created from these interactions can be harnessed for societal progress by providing better information for evidence-based policy-making.

The study at hand follows this line of thought by augmenting survey data with anonymized mobile phone metadata to create proxy indicators for multidimensional poverty in Al Gezira, Sudan. It is intended to provide a first informed look whether mobile phone metadata can be used (more widely) to inform poverty interventions. Therefore, the study is structured as follows: First, since the study draws on multiple data sources, they are described in detail. Second, the methodology of the approach is laid out. For details on the methodology, the study refers to two accompanying technical guidance documents. In a third step, the results of the study are presented. An assessment of the feasibility and a recommendation for the scalability of the approach is given in section four, the conclusion.

Data

The feasibility study draws on four data sources:

- Call detail records
- Antenna locations
- Household survey information
- Geographic information on administrative boundaries

While the first two data sources are in many cases held by mobile network operators, the latter are often governed by the national statistical office.

Call Detail Records

Call detail records are data generated within mobile networks mainly for billing purposes. They keep track of calling and texting activities of the network users. Consequently, normal users generate multiple records each day, leading to millions of new records in the mobile network on a daily basis. In the case of this study, the CDRs are stored as daily .csv-files in a highly distributed file system (HDFS) on the premises of the mobile network operator. Table 1 shows an example how call detail records usually look like.

Table 1: Example structure of CDRs

call_record_type	**caller_msisdn**	**call_date**	**basic_service**	**cell_id**	**call_partner_identity_type**	**call_partner_identity**	**tac_code**	**call_duration**
2	8723541620	2018-01-01 0:00:00	1	608-01-09004-02971	1	6202316723	---	1

The *call_record_type* defines the status of an event, the *call_date* provides the time and date the event took place and the *basic_service* defines the nature of an event, e.g. a call or a message. The *cell_id* refers to a Base Transceiver Station (i.e. antenna) and the *call_partner_identity_type* shows if it is a national or international event. The *tac_code* describes a user's mobile handset and the *caller_msisdn* as well as the *call_partner_identity* identify the initiator and receiver of an event.

The data available for this study are CDRs covering approx. four billion records for the Al Gezira state from one major Sudanese mobile network operator in the time period of 26 days (January 10th, 2018 to February 4th, 2018).

Antenna locations

The antenna locations are the essential link between call detail records and administrative boundaries. One antenna location usually hosts multiple cells in order to ensure coverage. The geographic location is provided as GPS coordinates. Table 2 presents an example how antenna location information are stored:

Table 2: Example structure of antenna locations

cell_id	antenna_id	longitude	latitude
608-01-09004-02971	1	-33.57443	11.75922

The data available for this study included the geographic locations of approx. 400 antennas in Al Gezira state.

Household survey

Household survey information are used to calibrate and evaluate prediction models built with big data. They are a common means to gather in-depth information on a variety of topics from the population. The household survey used in this study is a two-stage sample survey for Al Gezira state representative on the locality-level. It covers the indicators required to calculate the multidimensional poverty index as defined in the Oxford Poverty and Human Development Initiative (OPHI) methodology. The sample survey holds information from 811 households representing 4053 household members. For more details on this, please refer to the final survey report.

Table 3: MPI, Poverty Incidence (H) and Poverty Intensity (A), by locality

Locality	MPI	H	A
Eastern Algazira	.03	.07	.43
Alkamleen	.20	.36	.55
Alhasahisa	.08	.20	.42
Um Algura	.07	.15	.47

Almanagil	.28	.50	.56
Greater Wadmadani	.06	.16	.38
Southern Algazira	.03	.07	.36
Algurashi	.21	.47	.44
Gezira State	**0.12**	**0.24**	**0.48**

The data available for this study includes the poverty incidence, the poverty intensity and the multidimensional poverty index for the eight localities of Al Gezira state (Table 3).

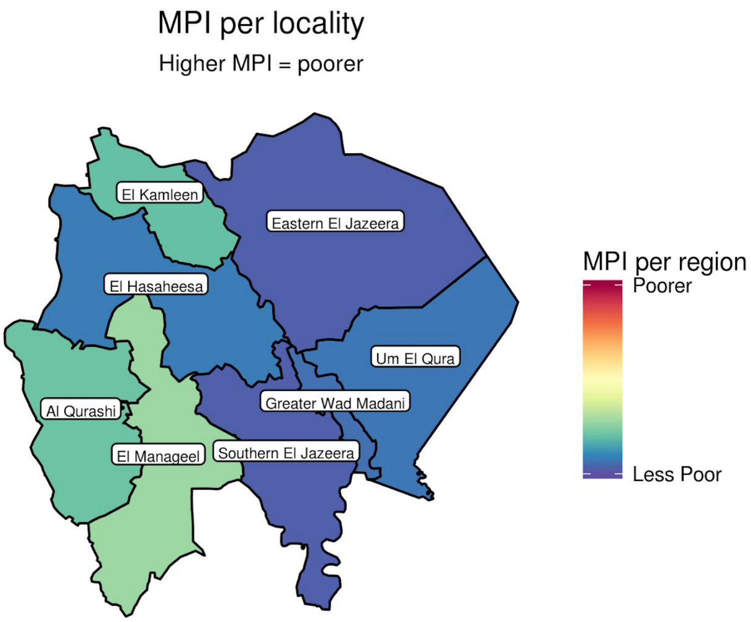
Geographic information on administrative boundaries

Geographic information on the administrative boundaries provide geospatial reference for both the antenna locations and the household survey. Sudan has following administrative divisions:

- NUTS0: National
- NUTS1: States
- NUTS2: Localities
- NUTS3: Administrative Units

In order to combine survey information with call detail records, it is necessary to define a common geospatial reference level. This study uses the level of localities (see Figure 1), since the household survey provides representative information at this level.

Figure 1: Multidimensional Poverty Index, by locality



The data available for this study includes geographic information for Al Gezira and its localities (NUTS1 - NUTS2).

Methodology

The methodology used to derive proxy indicators for multidimensional poverty from call detail records builds on the concept of small area estimation: estimates for small areas can be improved by taking auxiliary data into account. The quality of auxiliary data can be judged by two main characteristics:

- Is the data available for a large part of the population?
- Does it carry informational value concerning the target variable?

Since call detail records clearly answer the first question with a ‘yes’ in most parts of the world (assuming the unit of observation are individuals), it is the goal of this study to provide an answer to the second question in the context of multidimensional poverty in Al Gezira, Sudan.

The methodology can be divided into two parts: First, the preprocessing focuses on creating antenna-level covariates from raw call detail records that can be used as auxiliary data. Second, the analysis combines these aggregated CDR covariates with the survey data in order to estimate, evaluate and visualize proxy indicators for multidimensional poverty.

Preprocessing

For details on this, please see ‘Technical Guide Part I: Preprocessing.html’.

The goal of the preprocessing part is to create a dataset with covariates (also called features) derived from raw call detail records aggregated to the antenna-level. Therefore, user-level features are created (Level 1) and then aggregated onto the antenna-level thereby creating additional antenna-level features (Level 2&3). While currently stored per day, the call detail records need to be made available per user for Level 1 in a first step.

Level 1: The user

- **user_metrics**: basic metrics aggregated per user, e.g.: *og_calls*, *ic_calls*, *og_sms*, *etc.*
- **user_home_antenna**: monthly estimate of antenna with most events during the night (between 7pm and 7am) per user

Level 2 & 3: The antenna

- **antenna_metrics_week**: metrics aggregated per home antenna of individual users, week and part of the week *(antenna_id*, *week_part*, *week_number*, *og_calls*, *ic_calls*, *og_sms*, *ic_sms*, *og_vol*, *ic_vol)*
- **antenna_metrics_hourly**: metrics aggregated per home antenna of individual users and hour *(antenna_id*, *hour*, *og_calls*, *ic_calls*, *og_sms*, *ic_sms*, *og_vol*, *ic_vol)*
- **antenna_interactions**: all-time interactions between antennas based on the users' behavior to which a certain antenna is the homebase (antenna_id1, antenna_id2, sms_count, calls_count, vol_sum)

The preprocessing is performed using PySpark remotely in a virtual environment on the premises of the mobile network operator to ensure no personally identifiable information leaves the operator's systems.

Analysis

For details on this, please see 'Technical Guide Part II: Analysis.html'.

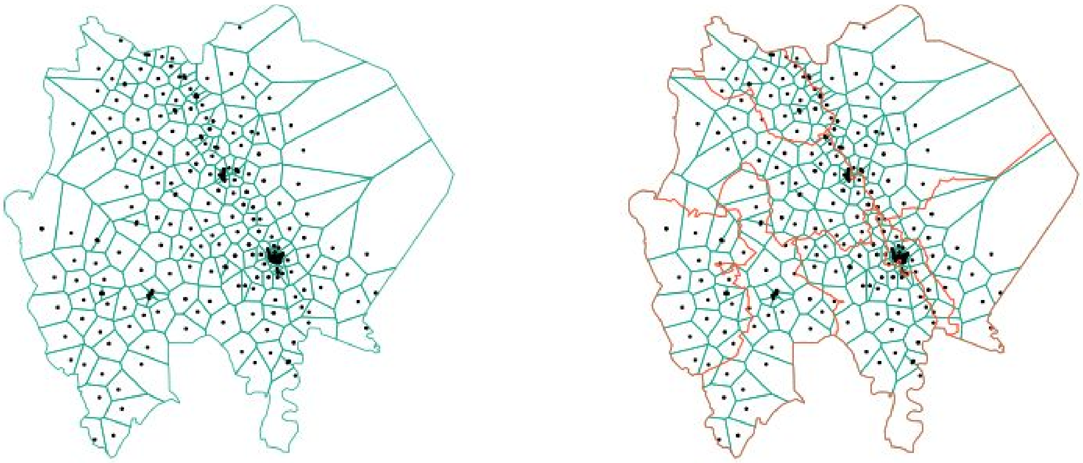
The analysis takes the output of the preprocessing and combines it with survey data on various geographical levels (Data synthesis). In a next step, feature selection is performed to avoid over-specification when constructing the model (Modelling). In a last step, proxy indicators are derived for available geographical levels, evaluated against locality-level survey data using bias and rmse and finally visualized. The analysis is conducted using the statistical programming language *R*.

Data synthesis

Data synthesis involves finding and defining common identifiers across data sources. While for surveys and administrative areas this can be straightforward, aligning call detail records and administrative areas poses a challenge: In an naïve approach, antenna-level CDR covariates would be allocated to the administrative area in which the point coordinate of the antenna location falls. However, in cases where the antenna is located close to one or multiple borders, its coverage area may largely lie outside its designated administrative area.

One approach to improve on this is to approximate the coverage area using voronoi tessellation based on the GPS coordinates of the antenna locations (see Figure 2, LHS) and overlay them with the administrative areas (see Figure 2, RHS).

Figure 2: Voronoi tessellation of Al Gezira using antenna locations



Voronoi tessellation divides a space into tiles based on the distance between points - here the antenna locations. These tiles are taken as rough approximations of the ‘true’ coverage area of the antennas. This allows to calculate weighted averages of the CDR covariates for the administrative areas.
So far, antennas have been treated equally although they might show different levels of activity. Additionally, while household characteristics are spatially associated with the location of their home, the CDR covariates of a user are allocated across the visited antennas. One way to align the data capture between antennas and surveys is to:

1) Define a home antenna (‘home location’) for each user
2) Assign the CDR covariates of a user entirely to their home location
3) Weight the CDR covariates of an antenna by the number of active users that have the respective antennas as their home location (‘active users’)

Consequently, in order to derive CDR covariates for the administrative level a weighted average is applied, using the approximations of the coverage area and of the number of active users to whom this antenna is their home location.

Modelling

By moving from individuals to small areas as unit of observation the number of observations is reduced. With a view on the number of extracted CDR covariates, dimensionality reduction likely becomes necessary to avoid over-specification. There are multiple ways to do this. This study, however, uses a two-step approach consisting of 1) preprocessing and 2) variable selection.

The preprocessing is done by filtering out covariates with:

- (almost) no variation
- strong collinearity
- strong correlation with other covariates

The resulting data is a set of covariates which are distinct from each other showing a minimum degree of variation. Nevertheless, it may still include variables with no or little predictive power concerning multidimensional poverty. Therefore, stepwise variable selection based on the Akaike Information Criterion (AIC) is applied.

Prediction, evaluation & visualization

This study - as a proof-of-concept - sets out to answer the question whether call detail records carry informational value concerning multidimensional poverty in Al Gezira, Sudan. The informational value is determined by the envisioned purpose, e.g.: Can multidimensional poverty be described with call detail records? Does this relationship hold over time? Does it hold over space? Does it hold across spatial dimensions? The scope of this study is limited to its first aspect.

A simple linear model is chosen to quantify the relationship between CDR covariates and MPI values as more complex models do not seem adequate with respect to the available survey data. The small sample size (n = 8), the lack of MPI disaggregation and sparse additional survey information restricts the range of applicable modelling options. Furthermore, this also limits the use of evaluation mechanisms such as cross-validation and comparing the results to disaggregated survey data. Consequently, the informational value is evaluated using three standard metrics:

- Bias
- Root mean square error
- Adjusted R^2

At this point, it is important to note that the results of this study have to be interpreted as a guide for further exploration, not as official MPI figures.

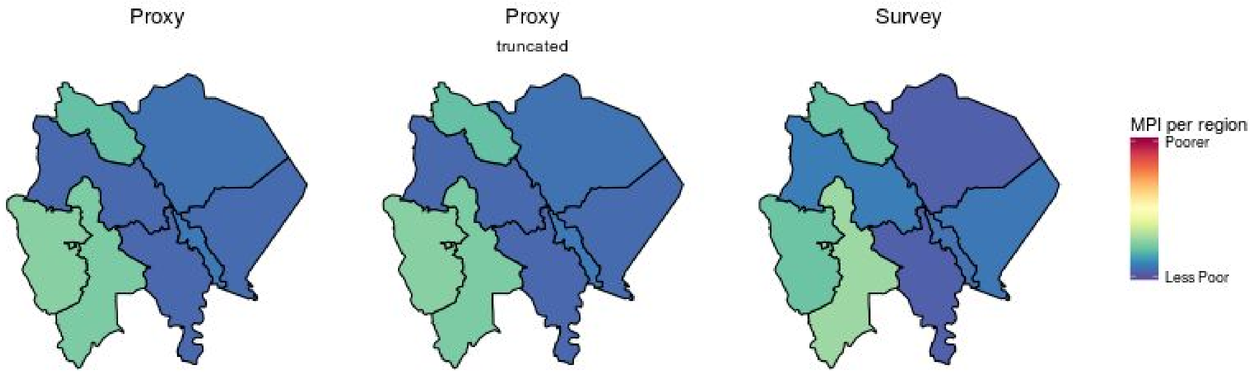
Proxy indicators are created by applying the estimated regression parameters fitted on the locality-level on CDR covariates from different geographical levels, i.e. localities and voronoi cells of the antennas. In case some estimates go out of bounds, also truncated estimates [0,1] are provided. While locality-level estimates can directly be compared to survey values, lower-level predictions can only be validated indirectly by aggregating them up to the localities.

Finally, MPI proxy indicators and performance metrics are visualized.

Results

The results of the study are promising. Proxy indicators for multidimensional poverty compare well against survey data (see Figure 3).

Figure 3: MPI proxies and MPI survey values on the locality-level

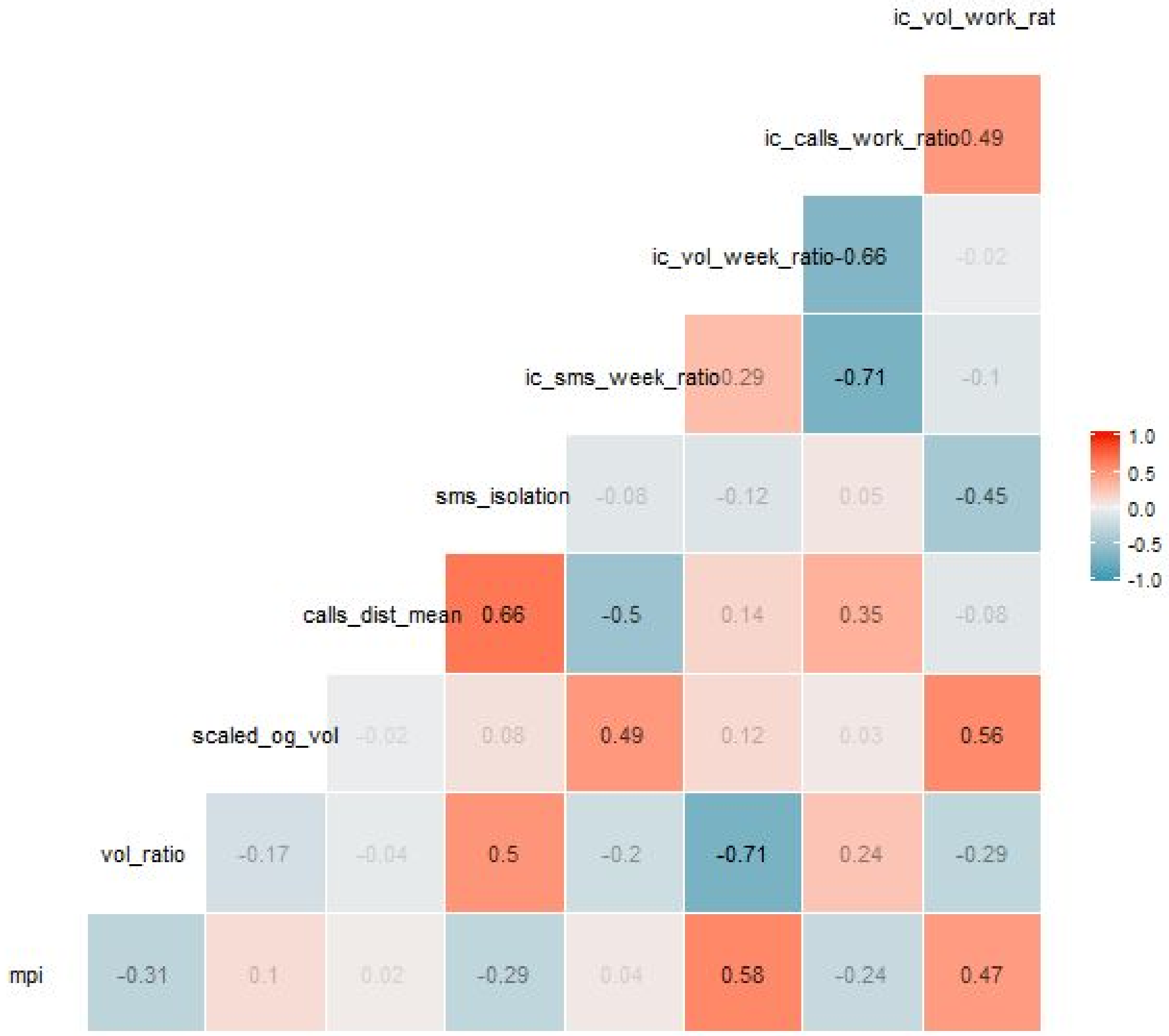


After doing a stepwise variable selection using AIC, four CDR covariates are left for further modelling. Running a repeated 8-fold cross-validation of a simple linear model using the four CDR covariates, the R² reaches high levels of fit:

- $adj.\ R^2 \approx 0.75$
- $R^2 \approx 0.9$

Single CDR covariates also demonstrate strong correlations with the locality-level MPI data from the survey (see Figure 4).

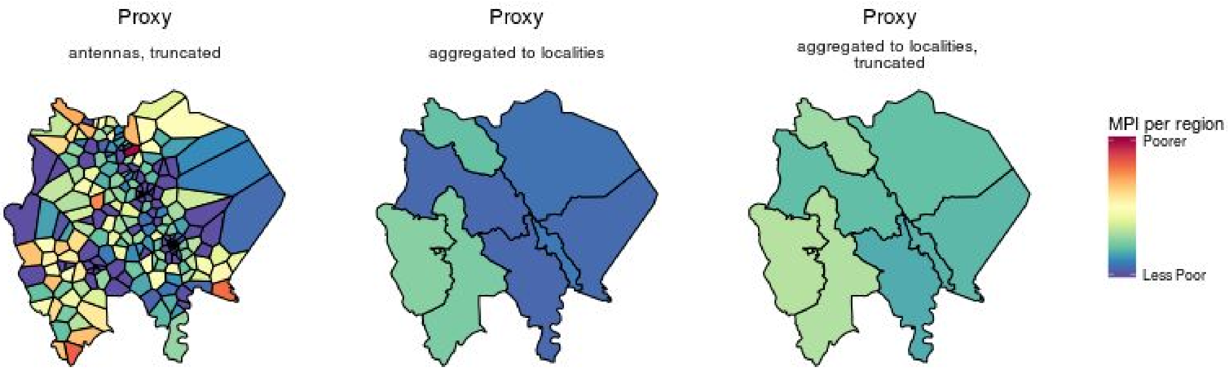
Figure 4: Correlation plot between the survey MPI and preprocessed CDR covariates



As reliable map data for the level of administrative units is not available, the study goes directly down on the level of antennas. Direct validation data on this level is generally only available from a census. Consequently, validation in this study is done indirectly using antenna-level MPI proxy values aggregated to the locality-level and evaluated against MPI data from the survey.

The MPI proxy indicator on this level shows a higher variability than the locality-level MPI. This is expected, since aggregation usually cancels out variation. While untruncated data produces a significant share of antennas with infeasible MPI values (<0), their locality-level aggregates present a good fit (see Figure 5). In contrast, truncation applied to the antenna-level data cancels out variation in the locality-level aggregates leading to a worse fit.

Figure 5: MPI proxies on the antenna-level (LHS), aggregated to localities



The better fit without truncation is also reflected in the point estimates as presented in Table 4.

Table 4: Survey and proxy indicators for multidimensional poverty

Locality	Survey MPI	Proxy indicators			
		MPI…	truncated	aggregated	agg. + trunc.
Al Qurashi	0.21	0.25	0.25	0.25	0.32
Eastern El Jazeera	0.03	0.06	0.06	0.06	0.2
El Hasaheesa	0.08	0.05	0.05	0.05	0.19
El Kamleen	0.2	0.2	0.2	0.2	0.28
El Manageel	0.28	0.23	0.23	0.23	0.31
Greater Wad Madani	0.06	0.07	0.07	0.07	0.17
Southern El Jazeera	0.03	0.05	0.05	0.05	0.16
Um El Qura	0.07	0.05	0.05	0.05	0.18

Conclusion

This study has provided a first informed view on whether mobile phone metadata can improve poverty monitoring in Al Gezira, Sudan. The results look promising: proxy indicators for the multidimensional poverty index constructed on anonymized call detail records have shown results for the eight localities of Al Gezira very similar to the survey estimates. Additionally, CDR covariates have shown high correlation with the MPI and overall model performance has clearly been higher compared to randomly generated CDRs (Proof of Concept). The geographical granularity of call detail records allowed a spatially disaggregated view on multidimensional poverty down to the level of antennas. Evaluated using their locality-level aggregates, the results appeared plausible for the untruncated estimates.

Building on these first insights, the ability of this study to assess the feasibility and to evaluate the scaling potential of the approach in a comprehensive fashion can be improved with additional validation data along the temporal and the cross-sectional dimension and with contextual data from the survey. Specifically, this would enable the authors to:

- apply more sophisticated estimation models taking the error variance into account
- sufficiently test the predictive power for out of sample areas
- assess the bias-variance trade-off
- directly evaluate spatially disaggregated predictions

Furthermore, besides technical questions, institutional challenges around mobile phone metadata access and privacy legislation have to be answered to establish mobile phone metadata as a standard source for evidence-based policy-making in the future.