Poverty on the Cheap: Estimating Poverty Maps Using Aggregated Mobile Communication Networks

Christopher Smith-Clarke
ICRI: Cities
London, UK
chris.smith@ucl.ac.uk

Afra Mashhadi
Bell Labs, Alcatel-Lucent
Dublin, Rep. of Ireland
afra.mashhadi@alcatel-lucent.com

Licia Capra
University College London
London, UK
l.capra@ucl.ac.uk

ABSTRACT

Governments and other organisations often rely on data collected by household surveys and censuses to identity areas in most need of regeneration and development projects. However, due to the high cost associated with the data collection process, many developing countries conduct such surveys very infrequently and include only a rather small sample of the population, thus failing to accurately capture the current socio-economic status of the country's population. In this paper, we address this problem by means of a methodology that relies on an alternative source of data from which to derive up to date poverty indicators, at a very fine level of spatio-temporal granularity. Taking two developing countries as examples, we show how to analyse the aggregated call detail records of mobile phone subscribers and extract features that are strongly correlated with poverty indexes currently derived from census data.

Author Keywords

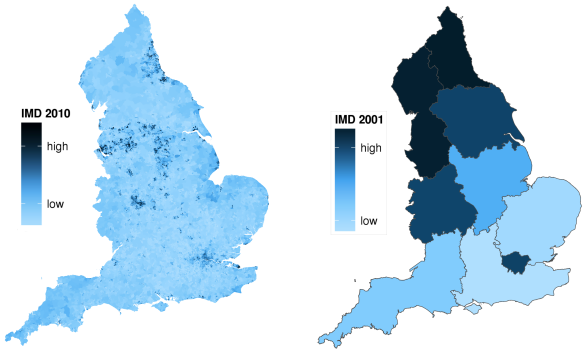
ICT4D; Data4D; Call detail records; Socio-economics

ACM Classification Keywords

H.1.2 User/Machine Systems: Human information processing; K.4.3 Computer and Society: Organizational Impacts

INTRODUCTION

Household surveys and censuses, periodically conducted by National Statistical Institutes and the like, collect information describing the social and economic well being of a nation, as well as the relative prosperity of its different regions. Such data is then used by agencies and governments to identify those areas in most need of intervention, for example, in the form of policies and programs that aim to improve the plight of their citizens. Interventions can take many forms, from national or regional policy, to local regeneration projects. To provide the most value socio-economic data needs to be up to date and it ought to be possible to disaggregate the data at each of these levels of granularity, and in between. However, the high cost associated with conducting surveys makes the data collection practices of many developing countries such that sample sizes are rarely large enough to provide statistically significant estimates for small geographical units, such as municipalities and villages.



(a) 32844 LSOAs, 2010 (b) 9 regions, 2001

Figure 1. IMD for England at different levels of spatial and temporal accuracy - darker shades represent more deprived areas.

To appreciate the problem that governments of developing countries face, let us first consider the situation of a developed country. In the UK, an Index of Multiple Deprivation (IMD) is computed using a combination of survey data and automatically collated government statistics pertaining to factors such as state benefit claims and crime. The government uses this index to allocate resources appropriately, and consequently mitigate against the detrimental effects of poverty and inequality. IMD is computed every three or four years, for each Lower layer Super Output Area (LSOA) in the country, with each such area having a minimum population of 1,000 and maximum of 3,000. Figure 1(a) visualises the IMD index computed in 2010 for each of the 32,844 LSOAs in England. As shown, deprivation information is captured at a very fine level of spatial granularity. Imagine, however, that survey data and statistics were sampled only at a much coarser granularity, such as England's 9 statistical regions, comprising of around 5 million people on average. Imagine further that the latest available raw data, and thus the IMD derived from it, had been collected more than 10 years ago. For illustrative purposes Figure 1(b) shows the IMD of 2001 aggregated to these 9 regions. For central government, when it comes to al-

CHI'14, April 26–May 1, 2014, Toronto, ON, Canada.

http://dx.doi.org/10.1145/2556288.2557358

locating resources to the most needy areas, such information provides little basis for any kind of policy or program targeting beyond this level. In many parts of the world this is the best that can be achieved, and in others the situation is even worse, with no estimates of living standards available at all.

In this paper we present work towards alleviating this problem. We show how, through analysis of patterns inherent in mobile phone users' collective behaviour, governments and NGOs can automatically compute proxy poverty indicators in a low cost and timely manner from unobtrusively collected call network data. Our methodology takes as input call detail records (CDR) aggregated to the cell tower level, which are collected by mobile phone operators primarily for billing purposes. Starting from this raw data, we describe features that can be extracted and used as proxy indicators of poverty levels. Taking two developing areas as examples, we quantitatively demonstrate the strong correlation these new features have with respect to poverty indicators derived from costly census data. Our results benefit the intended users (e.g., policymakers and NGOs) by providing interpretable results to act upon, in contrast to the black-box machine learning approach of previous work. We enable disaggregation at multiple levels of granularity thereby potentially influencing policy implemented at different levels, from neighbourhood to region. We offer an increased level of protection of mobile phone users' privacy by aggregating CDRs prior to any analysis, thus removing a barrier to wider adoption of this approach. Finally, we make significant progress towards generalisation by presenting results from two different developing regions. All that is required is for telecommunication providers to share anonymised, aggregated call detail records on a controlled basis. There are early signs of this already happening (e.g., D4D Challenge[1]), with researchers also starting to develop frameworks that would encourage even more providers to embrace this initiative [22]. We conclude the paper with a discussion of the practical implications of this methodology, its known limitations, and the next steps required to improve it further.

RELATED WORK

Research has been active over the last decade, within the HCI community and elsewhere, to understand the relation between, on the one hand, factors of human well-being and socio-economics, and on the other, technologically mediated social networks, such as online services like Facebook and Twitter, and social relations represented in telecommunications networks. For example, Burke *et al.* [6] examined the relationship between activity on Facebook and social capital and loneliness, and found that engagement with Facebook correlates positively with overall well-being. Quercia *et al.* found that deprivation in London, UK, varies geographically with topics of tweets discussed in different areas [24], and also that sentiment of tweets [23] correlates with deprivation. Eagle *et al.* [8] measured communication diversity from fixed line phone call records in England, and found that higher diversity (i.e., the more evenly dispersed a person's communication between people and places) correlates with socio-economic deprivation, aggregated to telephone exchange areas. Blumemstock [5] looked more closely at the relation between users' demographics, collected through personal interviews, and their mobile phone usage from a sample of company employees in Rwanda; they found that gender and social status had a direct correlation with the volume of their call activity.

The results of the afore mentioned works clearly have the potential to provide predictions of factors of well-being and socio-economic status. Indeed, Kramer *et al.* [18] found that sentiment of Facebook user content correlates with a person's happiness, and used this finding to develop a measurement of 'Facebook Gross National Happiness'. However, a followup study presented contradictory results, suggesting instead that expressed sentiment may play a role in regulating a Facebook user's mood rather than directly reflect it. Gutierrez *et al.* [14] hypothesised that mobile top-up behaviour reflects the wealth of the phone user, with poorer people likely to top-up their phone credit in small amounts fairly frequently, whereas wealthier people likely to top-up infrequently in larger amounts. They built a model based on this hypothesis, applied it to individual call records from Côte d'Ivoire (not the same dataset as studied in this paper), and derived a proxy wealth indicator from it. To date, the top-up model has not been validated against any established wealth indicator, therefore only speculative conclusions can be drawn. Furthermore, knowledge of individuals' financial data is required by the model, raising serious privacy concerns.

Moving away from exploitation of human interaction, other attempts to develop proxies for socio-economic factors include the use of satellite imagery to remotely identify the visual signs of economic development. The total area lit by Night Time Light (NTL) measured from satellite imagery was shown to correlate with a country's Gross Domestic Product [10]; this was later shown to hold for other countries too [7, 30, 9]. However, the geographic scale at which such methodology can be applied is rather coarse. Furthermore, output of a recent 'Datadive'[2] that looked at the relationship between NTL and small area poverty levels in Bangladesh suggests that the penetration of electrical lighting is approaching saturation, consequently removing the signal previously present.[3]

Work that directly explores the potetial for CDRs to provide estimates of socio-economic factors includes that of Soto *et al.* [29], who defined a comprehensive list of 279 features that could be extracted from CDRs of a South-American city, and measured the extent to which each of them correlates with known socio-economic levels (SEL) in that city. The most significant features were then used in a variety of machine learning techniques, and shown to achieve up to 80% accuracy when classifying areas according to three classes of SEL. The method developed in [29] was further investigated by Frias-Martinez *et al.* [13, 12], who then implemented it in a GUI-based system, designed to reduce the number of census

[1] http://www.d4d.orange.com/home

[2] http://blogs.worldbank.org/opendata/scenes-dc-big-data-dive-final-report

[3] https://hackpad.com/Predicting-Small-Scale-Poverty-Measures-from-Night-Illumination-f6RoPTY6IWB

areas that needed to be manually surveyed, by using surveyed data as training labels, and using the model to estimate the remainder [11]. This time, the highest accuracy quoted is 76% for a 3-class problem and 63% for a 6-class problem. This approach has the potential to provide savings in the cost of surveying, but also suffers from a number of drawbacks: as with [14], many of the features used in these works require detailed knowledge of individuals' call behaviour, thus raising privacy concerns and, consequently, limiting the possibility of obtaining such data from telecommunication providers in the first place. Even if anonymised individual data were made available, a methodology that includes so many features, embedded in a complex machine learning, 'black-box', predictor, would be difficult to interpret by nonexperts and would thus require a significant level of trust from policymakers and others who might act upon the results. Arguably, for such predictions to play a role in decision making processes, it is vital that governments can understand how the estimates are reached. A final limitation is that this approach still requires a significant financial investment, with up to two thirds of census areas needing to be surveyed to provide training examples.

With the aim to develop a practical methodology to derive poverty indicators for developing countries, at a fine level of spatio-temporal granularity and at very low cost, we kept these two requirements in mind: First, the source data our methodology relies upon must not infringe users' privacy; at such, rather than using CDRs at an individual user level, we only look at CDRs aggregated by the cell towers through which the calls are routed. Second, the features we extract from raw data ought to be relatively easy to interpret, which may in turn increase the confidence with which the resulting estimates can be acted upon.

DATASET DESCRIPTION

In order to develop a practical methodology that governments of developing countries can use to accurately infer poverty, we require two kinds of datasets: i) a dataset that is representative of the country's population, that is automatically collected so to contain always up to date information, and that is available at a fine level of geographical granularity; we will use this dataset to automatically extract features that signals poverty. ii) A ground truth dataset of geographically disaggregated poverty (or wealth indicators), to be used for validation purposes. For the former, we consider mobile call detail records obtained from local network operators. Primarily for billing purposes, mobile telecoms providers record details of each call and text message made over their network, including the time, duration, caller and callee IDs, and the base station towers (or 'cell towers') through which the call or text was routed. These call detail records thus provide a rich source of data, and given the high penetration rate of mobile technology in developing countries, they also offer a relatively unbiased picture in terms of demographics. For the latter, we use country specific socio-economic datasets, publicly available and presently derived from surveys and censuses.

Call Detail Records

We obtained two datasets of anonymised, aggregated voice calls: the first contains calls between 5 million Orange customers from Côte d'Ivoire over the period beginning December 1st 2011 until April 28th 2012 (referred to as Set 1 in the D4D Challenge [4]); the second contains call records of around $928,000$ customers of a network operator from another developing region, which we will refer to as Region B, spanning a period of 6 weeks in early 2012 (specific details about the location and time are omitted in order to preserve the anonymity of the network operator). The datasets contain the total volume (number of calls) and duration of calls between each pair of cell towers over the entire period. Note also an important difference in how data is aggregated in our work and in [11, 12, 28] is that the aggregation performed in previous work is an average of properties of individuals. That is, individual data is used to infer home locations and to measure properties of individuals' behaviour. In contrast, our data consists of total traffic between cell towers so no individual data is ever accessed. Features we derive are of the aggregated network and not of individuals which are then aggregated. Mobile phone penetration is high enough in many developing countries to make such datasets sufficiently representative of the population.[4] Indeed, both network operators hold a dominant position in their respective markets, with Orange having 48% market share in Ivory Coast[5] and the anonymous operator being the leading provider of mobile services in Region B. Table 1 provides summary statistics of each dataset.

	Côte d'Ivoire	Region B
Total Number of Calls	471 million	40 million
Total Call Duration (mins)	960 million	170 million
Time Period	20 weeks	6 weeks
Population	20 million	58 million
Total Area (hectares)	33 million	19 million

Table 1. Summary properties of the two CDR datasets.

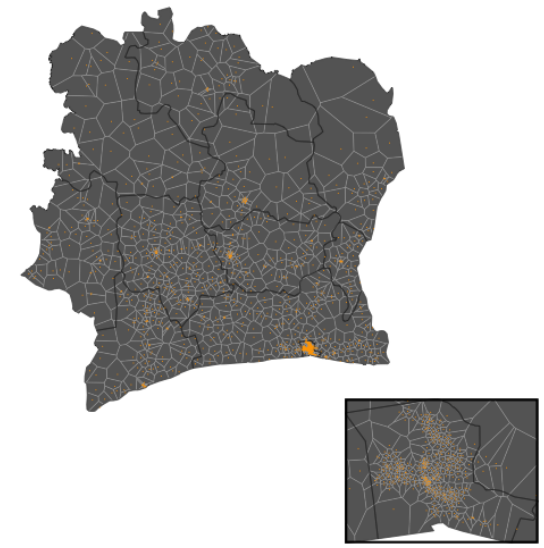
To provide some geographical context, Figures 2(a) and 2(b) show the locations of cell towers in relation to the regional boundaries of Côte d'Ivoire and Region B respectively. The insets of these figures show the dense concentration of cell towers in and around the largest cities of each region. The economic capital of Côte d'Ivoire, Abidjan, has a significantly higher population density than the rest of the country and is where most economic activity and trading takes place. A similar observation can be made of Region B shown inset in 2(b).

Socio-ecomonic data

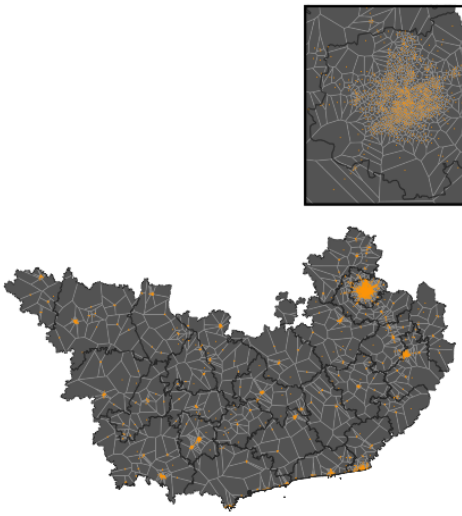
In order to validate the ability of our methodology to accurately estimate economic poverty, as currently defined by governments and international organisations, we use centrally managed datasets as ground-truth of poverty levels. In particular, for Côte d'Ivoire, we use poverty rate estimates provided by the International Monetary Fund, dating from 2008 [15].

[4] http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013.pdf

[5] http://www.orange.com/en/group/global-footprint/countries/Group-s-activities-in-Ivory-Coast



(a) Côte d'Ivoire

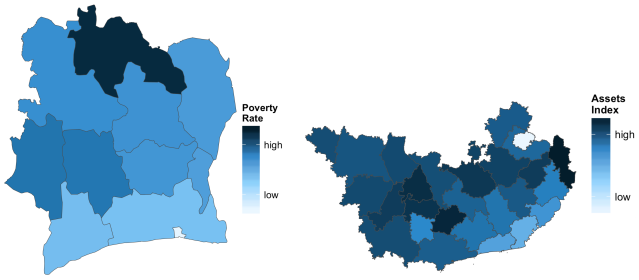


(b) Region B

Figure 2. Geography of cell towers, their Voronoi polygons (light lines) and regional boundaries (dark lines). Insets show a magnification of the economic centre of each region.

This is the most recent and most fine grained, freely available data pertaining to Côte d'Ivore that we have been able to obtain. Being at the level of 11 subnational regions, this data is also an example of the limitations imposed by low sample rates when collecting data to formulate poverty estimates. A more detailed census was carried out in 1998, but due to the recent civil strife afflicting the country the follow-up census has been twice postponed and is now planned for late 2013.

For Region B, we constructed an assets-based index, derived from 2011 census data. The census data contains 14 variables pertaining to household ownership of assets such as laptops, mobile phones and various kinds of vehicles. A single asset-based index was then derived, using Principal Component



(a) Côte d'Ivoire (b) Region B

Figure 3. Map of Côte d'Ivoire and Region B showing relative poverty rate and assets index respectively. Darker shades indicate poorer areas.

Analysis (PCA), following a methodology similar to [21]. We elected to construct such custom index since no specific poverty data was available at an appropriate level of spatial granularity. For convenience, we may allude to the 'poverty level' of Region B in the remainder of the paper, although we are always referring to the assets-based index. Figures 3(a) and 3(b) show the geographical distribution of poverty in Côte d'Ivoire and the assets-based index in Region B, respectively, at the finest level of spatial granularity currently possible.

METHODOLOGY

In this section we describe the features that we propose to automatically compute from CDRs. We provide both a formal definition of each feature, as well as the motivating interpretation of the way each feature may reflect an area's incidence of poverty.

Spatial aggregation of call data

Before discussing what features we have extracted from CDRs, we need to define a spatial unit of granularity at which we will operate. The raw data we start from contains location information (as latitude/longitude coordinates) of each cell tower. The finest level of spatial granularity that our methodology could operate upon is that of the Voronoi areas associated to the cell tower locations. By definition, the Voronoi area of a seed point (i.e., a cell tower) on a plane contains every location that is closer to that point than it is to any other seed point (i.e., any other cell tower). The Voronoi areas of the two regions under exam can be seen in Figures 2(a) and 2(b). As shown, such level of granularity is extremely fine grained; indeed, no socio-economic data is presently available at this resolution. To be able to later validate our methodology against available socio-economic data, we aggregate call data (and thus the features we will extract from it) at a coarser level of granularity as follows.

The features we will extract from call data are of two types: pertaining to a single tower i (e.g., the number of incoming/outgoing calls from i), or pertaining pairs of towers i and j (e.g., the flow of calls between them). We generally refer to features of the former type as f_i, and to features of the latter type as $f_{i,j}$. When operating at a coarser level of granularity than Voronoi cells vor_i, we associate features to such coarser areas u in a way that is proportional to the *population* within each area of the intersection $vor_i \cap u$. To this end, we first

estimate the population that falls within each Voronoi cell by intersecting them with a population density grid with a resolution of approximately 100m.[6] Then, for each feature f_i, the proportion of it associated to area u is:

$$P(f_i, u) = f_i \frac{M(\text{vor}_i \cap u)}{M(\text{vor}_i)} \quad (1)$$

where $M(.)$ is the estimated population of the area passed as parameter. For a feature $f_{i,j}$, such that vor_i interesects u and vor_j intersects v, we set:

$$P(f_{i,j}, (u, v)) = \frac{1}{2} f_{i,j} \frac{M(\text{vor}_i \cap u)}{M(\text{vor}_i)} \frac{M(\text{vor}_j \cap v)}{M(\text{vor}_j)} \quad (2)$$

In practice, when administrative areas are very large in comparison to the Voronoi areas, such as are the 11 sub-national regions of Côte d'Ivoire, the number of Voronoi cells intersecting regional boundaries is negligible with respect to the number of those fully contained in them, thus the above preprocessing step is not strictly necessary. However, in the case of Region B, and in general as spatial granularity increases, the above methodology may afford significantly higher accuracy.

Feature Extraction

We next formally describe the features we automatically extract from call data, as well the motivation behind them. For illustrative purposes, we abstract the CDR dataset as a graph, where the cell towers are nodes (or vertices), and the edges (or links) between pairs of nodes i and j are weighted with values $w_{i,j}$ that represent interchangeably the volume or duration of calls between them. wi,j = weight of two cell towers

Activity

We expect to find that the level of mobile communication activity in an area will reflect its social and economic activity, and thus its level of prosperity. Aker and Mbiti [1] outline a number of mechanisms by which mobile phone adoption could spur economic development, including by reducing the cost of searching for, and accessing information, and by improving the efficiency of supply chain management. They also find that mobile phone use is strongly linked to socioeconomic status, with early adopters being primarily young, educated, urban males. To capture this relationship, the first feature we compute is simply the strength, or weighted, undirected degree, of nodes aggregated to the administrative area:

$$\text{activity}(u) = \sum_i P(s_i, u) \quad (3)$$

where $s_i = \sum_j w_{i,j}$ is the strength of node i. However, as mobile technology becomes more and more ubiquitous the link between activity and wealth is eroding, with mobile phone use rapidly increasing among poorer people [1].

[6] Available from http://www.afripop.org/ and http://www.asiapop.org/.

wij = volume of calls between CT i and j. i.e. sum of i to j and j to i.
Activity = sum of the proportion of the strength of a CT i associated to region u.
Strength of the node = sum of volumes between i and other cell phone towers.

This trend motivates the search for more robust signals of economic well being which will survive market saturation. The simple relationship between total activity and poverty may quickly dissipate just as appears to be the case with the link between night time lights and poverty. For this reason we present three further features that go beyond simply measuring usage and capture ways in which areas relate to each other.

Gravity Residual

We next hypothesise that the difference between observed and expected flows between areas reflects the level of social and economic activity in those areas, and thus will be related to poverty. To estimate flows we use a gravity model. First introduced by Zipf in 1946 [33], gravity models rest on the hypothesis that the size of flow between two areas is proportional to the mass (i.e., population) of those areas, but decays as the distance between them increases. Despite some criticisms ([26, 32]), the model has been successfully used to describe macro scale interactions (e.g., between cities, and across states), using both road and airline networks [3, 16] and its use has extended to other domains, such as the spreading of infectious diseases [2, 31], cargo ship movements [17], and to model intercity phone calls [19]. By examining the residuals between observed and expected flows we aim to capture the restricting effect of poverty on an area's interactions with others.

We use the following equation to find the expected flows between areas: For the voronoi, it will be the proportion of the population for the region

$$F^{\text{est}}_{u,v} = g \frac{M(u)M(v)}{d(u,v)^2} \quad (4)$$

population

where $M(u)$ is the ~~mass~~ of area u and $d(u, v)$ is the as-the-crow-flies distance between centroids of areas u and v. For Region B, we use the number of subscribers in each area as the mass, but for Côte d'Ivoire this information is not available, therefore we use the actual population of each area. The factor g scales the estimates to bring them in line with observed weights and is fitted to each dataset separately. In general, g depends only on the period of observation. The residual between each pair of areas is then the result of subtracting the estimated flow from the observed flow, $F_{u,v} = \sum_i P(w_i, u)$, so overestimates will result in negative residuals. Finally, for each area, we calculate the mean of all negative residuals connecting the area, so that areas involved in more overestimates will have lower values. Formally, the gravity residual is:

$$\text{gResidual}(u) = \frac{1}{n_u^{\text{-ve}}} \sum_v \text{negRes}(u, v) \quad (5)$$

where

$$\text{negRes}(u, v) = \begin{cases} F_{u,v} - F^{\text{est}}_{u,v} & \text{if } F_{u,v} < F^{\text{est}}_{u,v} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Recent work modelled the flow of passengers in London's rail system in a similar fashion, and found that the gravity residuals were related to deprivation of neighbourhoods[27].

Network Advantage

Our next set of features aims to capture the opportunity for economic development afforded by an advantageous position in the network with respect to the flow of information. In studying a social network represented by a fixed-line telephone call dataset, Eagle *et al.* [8] showed that the average normalised entropy (referred to as *diversity*) of the social ties of people living in a neighbourhood correlates strongly with the level of socio-economic deprivation (a concept closely related to poverty) in that neighbourhood. In this work we are constrained by the aggregation of the call records to cell tower and are unable to look directly at the underlying individual social network. Instead, we hypothesise that the structure of a cell tower's links will also reflect the poverty level in its location. We thus extract two measures of a cell tower's network advantage: the degree of the cell tower after discarding links with a weight below a certain threshold, and the normalised entropy. For the former, if links of any weight are considered, the cell tower network is extremely dense and the degree distribution is largely uniform; we therefore drop links with weight below the median, denoted k_i^{med} in order to magnify actual heterogeneity in degree distribution. For the latter, normalised entropy is computed using the following formula from [8]:

$$S_i = \frac{-\sum_j q_{i,j} \log q_{i,j}}{\log(k_i)} \tag{7}$$

where $q_{i,j} = w_{i,j}/w_i$ is the fraction of cell tower i's total weight on its link with j, and k_i is the degree of i. Then our area level features are:

$$\text{entropy}(u) = \frac{1}{n_u} \sum_i P(S_i, u) \tag{8}$$

and

$$\text{medDegree}(u) = \frac{1}{n_u} \sum_i P(k_i^{\text{med}}, u) \tag{9}$$

Introversion

Finally, we hypothesise that an area's level of *introversion* may be a signal of its poverty level. In other words, if an area has relatively low volume of traffic to other areas compared to the traffic within it, the less likely it will be able to benefit from new sources of opportunity arising further afield. This is similar in spirit to the theory of open economies, albeit on a different scale, which expects nations that close their borders to international trade to fare less well than those that are more open [25]. It is also related to the idea of network advantage, except that we now explicitly take into account geography and consider only a binary property of flow, that is, whether it is internal or external to the area. A caveat to the above hypothesis is that we may expect denser areas to naturally exhibit higher introversion given that there will be a higher likelihood of communications taking place within the vicinity. However, since density of cell towers tends to follow population density, the size of Voronoi cells in dense areas is smaller, thus mitigating somewhat against the higher likelihood of internal communications. We first calculate the introversion of cell towers with the following equation:

$$\text{H}_i = \frac{w_{i,i}}{\sum_{i \neq j} w_{i,j}} \tag{10}$$

We then compute the average introversion of all cell towers within each region and define the feature:

$$\text{introversion}(u) = \frac{1}{n_u} \sum_i P(H_i, u) \tag{11}$$

Intuitively, introversion values below 1 indicate more introverted areas (i.e., internal flow is higher than external flow), and conversely, values above 1 indicate more extroverted areas.

To see that these last three features are not simply alternative ways to measure overall usage, consider that activity levels could be flat across the entire region and yet the other features could vary greatly. For example, an area might be making/receiving many fewer calls than expected and thus have a very large gravity residual, and its neighbours might receive/make more calls and have a smaller residual. We next establish the potential for each of these features to be used as proxies for poverty rate.

RESULTS

Feature Validation

For the two regions under exam (Côte d'Ivoire and Region B), we compute Pearson correlation coefficients between each of the previously extracted features and their poverty rates. Table 2 shows the overall results. Note that we are measuring properties of an aggregated network, for which we would not necessarily expect to find the same correlations as in an individual call network. For example, human networks tend to have very low edge density (proportion of all possible edges present in the network) and the degree of nodes will be a tiny fraction of the total number of nodes. Subsequently, the average degree of individuals within an area will also be small, and indeed, a correlation between degree and poverty will be retained when averaging. In contrast, the aggregated cell tower network is extremely dense and the degree of each node tends to be much closer to the total number of nodes. This is because as we aggregate we reduce the number of nodes whilst accumulating all the edges. We cannot therefore assume that a correlation between individual degree and poverty will also be present between the degree of cell towers and average poverty (or poverty rate) of the area. Similar arguments apply to other network properties.

Activity

To begin with, we see strong negative correlations between the total volume and total duration of calls within an area and its poverty level, both in Côte d'Ivoire and Region B. This confirms that aggregated communication activity provides a simple proxy for poverty level; however, as mentioned in the previous section, this relationship may depend in part on the maturity of the mobile telecoms market. Therefore we are particularly interested in the results of the remaining features since these are potentially more robust in the face of market saturation.

Hypothesis	Country	Feature	Pearson's r	95% Conf. Int.	p-value
Activity	Region B	activity: volume	$-.776$	.561, .893	$< 1e{-5}$
		activity: duration	$-.775$	.560, .892	$< 1e{-5}$
	Côte d'Ivoire	activity: volume	$-.834$	$-.956, -.469$	.001
		activity: duration	$-.830$	$-.955, -.458$	.002
Gravity Residual	Region B	gResidual: volume	.686	.848, .407	$< .001$
		gResidual: duration	.701	.856, .430	$< 1e{-4}$
	Côte d'Ivoire	gResidual: volume	.831	.460, .955	.002
		gResidual: duration	.830	.458, .955	.002
Network Advantage	Region B	entropy: volume	$-.746$	$-.877, -.511$	$< 1e{-5}$
		entropy: duration	$-.726$	$-.867, -.478$	$< 1e{-4}$
		medDegree: volume	$-.440$	$-.702, -.072$	.021
		medDegree: duration	$-.430$	$-.696, -.059$	.025
	Côte d'Ivoire	entropy: volume	$-.774$	$-.938, -.326$	.005
		entropy: duration	$-.750$	$-.931, -.273$	.008
		medDegree: volume	$-.801$	$-.946, -.388$	.003
		medDegree: duration	$-.797$	$-.945, -.379$	.003
Introversion	Region B	introversion: volume	$-.784$	$-.897, -.575$	$< 1e{-5}$
		introversion: duration	$-.782$	$-.896, -.573$	$< 1e{-5}$
	Côte d'Ivoire	introversion: volume	.710	.190, .918	.015
		introversion: duration	.644	.072, .897	.032

Table 2. Correlations between features derived from mobile phone data and poverty level at the presently available (coarse) spatial granularity.

Gravity Residual

We found strong correlation between the mean negative residual of the gravity model and poverty level. This suggest that when communication flows in or out of an area are lower than expected, higher poverty levels are expected to be found. Figure 4 illustrates the relationship between gravity residual and geography for Côte d'Ivoire. The majority of negative residuals can be seen to connect to northern regions, which are also known to be poorer. Although we cannot posit a causal relationship, the results suggest a clear link between poverty or wealth of an area and its level of interaction with other areas. Note, however, the anomalously large negative residual between Abidjan and the South, the two wealthiest regions in the country; this initially surprising result may explained away by the well-known poor fit of the gravity model at short distances (e.g., [20]). To gain confidence in this assessment, we plot in Figure 5 the absolute error of the gravity model as a function of distance: the model is less accurate at shorter ranges, which in turn means that the mean negative residual is likely to be less effective as a signal of poverty in these cases. To overcome this limitation it may be pertinent to replace the gravity model with alternative interaction models, such as the radiation model, which takes into account population density [26], or the more recent heat-conduction model, which has been shown to be more accurate at the city scale [32]. With a more reliable model of flows we would hope to more accurately capture the effect of socioeconomic well being in the residuals.

Network Advantage

The two features extracted in relation to network advantage exhibit significant negative correlations with poverty, both in Côte d'Ivoire and Region B. We find that the more evidence of network advantage, as given by the entropy and degree measures, the lower the poverty levels. Although we cannot posit a causal relationship, these results are in line with previous work which found a link between network advantage and socioeconomic deprivation at the individual level [8], suggesting that similar forces are at play in bestowing greater opportunity to those areas with increases access to sources of information.

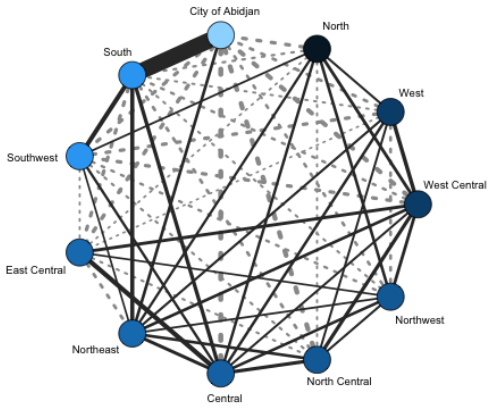


Figure 4. Network visualisation of the gravity model residuals in Côte d'Ivoire. Nodes represent regions with shade corresponding to poverty rate decreasing clockwise from node North. Line thickness corresponds to magnitude of gravity model residual, with negative residuals as solid lines and positive as dashed lines.

Introversion

The last feature we extracted is introversion, of both volume and duration. Here we find contrasting results for the two countries under examination: for Côte d'Ivoire, the introversion of an area correlates strongly with poverty level (albeit with higher p-values than for other features), in support of the hypothesis we put forward in the previous section. However, in Region B we see the reverse relationship, with higher levels of introversion being associated with lower values on deprivation. There may be several reasons for this, including cultural differences such as a tendency in Region B for relations, both

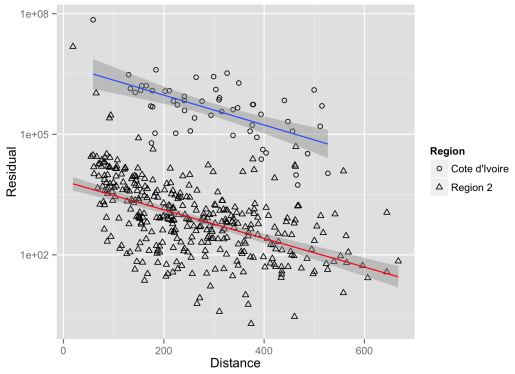


Figure 5. Absolute error of gravity model as a function of distance. The regression lines show the general trend for residuals to be larger at shorter distances.

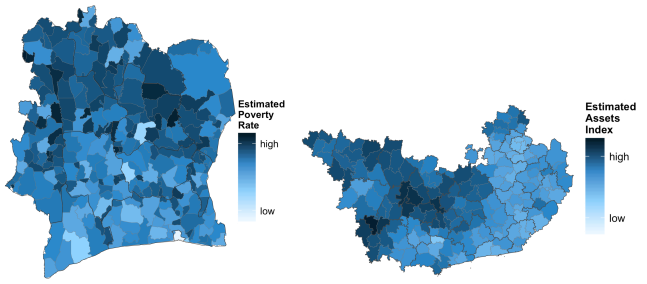
social and economic, to remain within social strata, which could in turn affect the spatial distribution of such relations. Alternatively, we could also attribute the contrasting results to differences in the representativeness of the datasets, with the dataset of Region B containing a smaller proportion of the total number of mobile phone users compared to Côte d'Ivoire. Speculations aside, these results highlight the importance of placing the results in a local context, as opposed to relying on a 'black box' predictor.

The above correlation analysis suggests that the features we have extracted from aggregated call data are indeed meaningful and could be used as relatively easy-to-interpret proxies of poverty rates in the developing regions under focus. We also claim that despite the fact that correlations are lower for the latter three features, their inclusion is valuable since they can in combination increase confidence in the ability of our predictions to accurately track poverty. Whereas in isolation we could be less confident since each suffers from potential confounds. Thus these results represent significant progress towards developing robust metrics.

Next we discuss in more detail the practical implications that follow from the above methodology and the limitations of these results.

DISCUSSION

We have outlined and tested a methodology for estimating poverty levels that has the potential to impact the practices of policymakers and NGOs working to improve the living standards of people in countries that lack the resources to manually collect socio-economic data on a frequent basis and at sample rates that would allow fine spatial disaggregation. Tools built upon these results would be relatively low cost to implement and could provide interpretable results (in contrast to a black-box machine learning approach) to act upon in a timely manner. Furthermore, we enable disaggregation at multiple levels of spatial granularity thus potentially influencing policy implemented at different levels, from neighbourhood to region. The results also benefit mobile phone users by protecting their privacy from the outset, thus removing a barrier to wider adoption of this approach.



(a) Côte d'Ivoire (b) Region B

Figure 6. Map of Côte d'Ivoire and Region B showing relative poverty rate and assets index respectively. Darker shades indicate poorer areas.

In discussion with the United Nations Population Fund (UNFPA) to determine how to put the methodology to actual use, an important need identified is the availabilty of maps at different levels of spatial granularity, so to provide information as required for different purposes. For example, national governments determining the allocation of a development budget to regional governments would require coarser grained information at the level of the administrative division in question. At the other end of the scale, regeneration or aid projects implemented at the local level for the benefit of small communities would require much finer resolution poverty maps to ensure the most needy areas are targetted. The methodology we have presented provides for both situations, with the ability to aggregate data at multiple levels of granularity, unlike sparsely sampled survey data that must be aggregated to a certain minimum (and often impractically coarse) level in order to achieve statistical significance.

To demonstrate this, we estimate poverty at the level of 255 sub-prefectures in Côte d'Ivoire, and of 176 areas at the next administrative level down in Region B, by deriving a simple linear model from the features above, using ordinary least squares regression. To visualise the granularity of information that such a model would give governments and agencies, we provide an estimated poverty map for Côte d'Ivoire in Figure 6(a), and for Region B in Figure 6(b). Notice the dramatic change in the spatial pattern of poverty information, compared to the regional maps previously shown in Figures 3(a) and 3(b) respectively. The coarser grained map of Côte d'Ivoire depicts poverty increasing as we radiate out from the city of Abidjan. Instead, our finer grained estimates complicate the picture, suggesting that the south-east of the country may contain areas of high poverty near Abidjan and conversely the north-west may contain areas of low poverty. Similarly in Region B we see areas estimated to be low on the assets index adjacent to areas estimated to be high.

LIMITATIONS AND FUTURE WORK

Difficulty in obtaining CDR data currently prevents us from establishing the global applicability of our work, but by replicating results in two developing economies that differ greatly in contextual factors, such as culture, migration patterns and social and family relationships, we are able to claim that our results are not simply chance correlations. This represents a significant advance towards general application compared to

related work. It might be suggested that validity is threatened by variation in adoption rates, but rather, this is a factor which will partly determine the values of the properties we derive (others being individual usage, infrastructure, etc.). Consequently, use of our metrics would not disadvantage groups with low adoption rates, but in fact they would show up as black spots in our models (most intuitively when measuring activity, i.e., low adoption will mean low activity, but also in the other features) and would thus be identified much sooner than with traditional methods.

The lack of up to date and spatially accurate socio-economic ground truth data also represents a significant hurdle toward a rigorous evaluation of the results. In order to be confident that the features we extract can be used to accurately track poverty in a timely and spatially accurate manner, we initially require knowledge of real poverty rates that also fulfil these constraints. Instead we have a lag of 4 years in Côte d'Ivoire and 2 years in Region B between the socio-economic data we use as ground truth and the mobile phone data from which we derive our proxies. Although this temporal lag will undoubtedly affect the accuracy of predictive models based on our proxies, such as the simple linear model we present above, we argue that the legitimacy of the methodology we have developed is not compromised. Rather, we would expect its accuracy and utility to increase were this lag removed. In future work we will take steps towards overcoming these limitations by acquiring ground truth data that is both more recent and has a more precise level of geo-location. This will allow us to fully investigate the relationship between geographical hierarchies and validate estimates at finer granularity.

Further practical usages of our methodology identified include the potential for identifying trends, thereby providing early warning of conditions worsening in specific areas, and the ability to evaluate the effect of policy and projects in a reasonable time frame (i.e., as soon as changes occur and before policy is due for renewal). Indeed, tools built upon the methods we have described would be a useful augmentation to socio-economic data collection processes in any country. The cost of producing estimates from passively and automatically collected communication data is negligible compared to that of manual surveying, thus a main barrier to obtaining up to date poverty estimates has been removed. Côte d'Ivoire is a perfect example of a country in which timely and accurate information regarding poverty is severely lacking. In cases such as this, the ability to obtain estimates of poverty levels on a continuous basis would represent a vast improvement. UNFPA has stressed the value that *any* indicative estimates would provide in certain situations where none are currently available; even if they carried with them a significant level of uncertainty such estimates would still represent a large improvement in many cases. Indeed, novel methods to provide low cost poverty indicators would represent significant value to many governments and NGOSs working to improve people's lives. Limited resources could be allocated in much more efficient manner thereby helping to alleviate some of the detrimental effects of poverty and inequality.

ACKNOWLEDGMENTS

We would like to thank Sabrina Juran of UNFPA, Olivia De Backer, Miguel Luengo-Oroz and René Clausen Nielsen of UN Global Pulse for their helpful feedback on previous versions of this document. This research was funded by a Google Europe Doctoral Fellowship in Data Mining and Intel Collaborative Research Institute: Cities.

REFERENCES

1. Aker, J. C., and Mbiti, I. M. Mobile Phones and Economic Development in Africa. *Journal of Economic Perspectives 24*, 3 (2010), 207–232.
2. Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., and Vespignani, A. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America 106*, 51 (Dec. 2009), 21484–9.
3. Barrat, A., Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America 101*, 11 (Mar. 2004), 3747–52.
4. Blondel, V. D., Esch, M., Chan, C., Clerot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z., and Ziemlicki, C. Data for Development: the D4D Challenge on Mobile Phone Data. 10.
5. Blumenstock, J., and Eagle, N. Mobile divides: gender, socioeconomic status, and mobile phone use in rwanda. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, ACM (2010), 6.
6. Burke, M., Marlow, C., and Lento, T. Social network activity and social well-being. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, ACM (New York, NY, USA, 2010), 1909–1912.
7. Doll, C. H., Muller, J.-P., and Elvidge, C. D. Night-time imagery as a tool for global mapping of socioeconomic parameters and greenhouse gas emissions. *AMBIO: a Journal of the Human Environment 29*, 3 (2000), 157–162.
8. Eagle, N., Macy, M., and Claxton, R. Network diversity and economic development. *Science (New York, N.Y.) 328*, 5981 (May 2010), 1029–31.
9. Ebener, S., Murray, C., Tandon, A., and Elvidge, C. C. From wealth to health: modelling the distribution of income per capita at the sub-national level using night-time light imagery. *International Journal of Health Geographics 4*, 1 (2005), 5.
10. Elvidge, C. D., Baugh, K. E., Kihn, E. A., Kroehl, H. W., and Davis, E. R. Mapping city lights with nighttime data from the dmsp operational linescan system. *Photogrammetric Engineering and Remote Sensing 63*, 6 (1997), 727–734.

11. Frias-martinez, V., Soto, V., Virseda, J., and Frias-martinez, E. Computing Cost-Effective Census Maps From Cell Phone Traces. In *Pervasive Urban Applications (PURBA)* (Newcastle, 2012).
12. Frias-Martinez, V., and Virseda, J. On the relationship between socio-economic factors and cell phone usage. In *Fifth International Conference on Information and Communication Technologies and Development (ICTD '12)*, ACM Press (New York, New York, USA, Mar. 2012).
13. Frias-Martinez, V., Virseda-Jerez, J., and Frias-Martinez, E. On the relation between socio-economic status and physical mobility. *Information Technology for Development 18*, 2 (Apr. 2012), 91–106.
14. Gutierrez, T., Krings, G., and Blondel, V. D. Indicators of wealth, economic diversity and segregation in cøte divoire using mobile phone datasets. In *Netmob 2013 Book of Abstracts* (2013).
15. International Monetary Fund. Coîe divoire: Poverty reduction strategy paper. Tech. rep., 2009.
16. Jung, W., and Wang, F. Gravity model in the Korean highway. *Europhysics Letters 81* (2008).
17. Kaluza, P., Kölzsch, A., Gastner, M. T., and Blasius, B. The complex network of global cargo ship movements. *Journal of the Royal Society, Interface / the Royal Society 7*, 48 (July 2010), 1093–103.
18. Kramer, A. D. I. An Unobtrusive Behavioral Model of Gross National Happiness. In *Proceedings of the 28th ACM CHI*, ACM Press (New York, New York, USA, Apr. 2010), 287–290.
19. Krings, G., Calabrese, F., Ratti, C., and Blondel, V. D. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment 2009*, 07 (May 2009), L07003.
20. Masucci, A. P., Serras, J., Johansson, A., and Batty, M. Gravity vs radiation model : on the importance of scale and heterogeneity in commuting flows.
21. Noor, A. M., Alegana, V. a., Gething, P. W., Tatem, A. J., and Snow, R. W. Using remotely sensed night-time light as a proxy for poverty in Africa. *Population health metrics 6* (Jan. 2008), 5.
22. Parate, A., and Miklau, G. A framework for safely publishing communication traces. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (2009), 1469–1472.
23. Quercia, D., Ellis, J., Capra, L., and Crowcroft, J. Tracking Gross Community Happiness from Tweets. In *Proceedings of ACM CSCW 2012* (2012).
24. Quercia, D., Seaghdha, D. O., and Crowcroft, J. Talk of the City : Our Tweets , Our Community Happiness. In *Proc.of AAAI ICWSM* (2012).
25. Sachs, J. D., and Warner, A. M. Source of Slow Growth in African Economies. *Journal of African Economics 6*, 3 (1997), 335–376.
26. Simini, F., González, M. C., Maritan, A., and Barabási, A.-L. A universal model for mobility and migration patterns. *Nature 484*, 7392 (Apr. 2012), 96–100.
27. Smith, C., Quercia, D., and Capra, L. Finger on the pulse. In *Proceedings of the 2013 conference on Computer supported cooperative work - CSCW '13*, ACM Press (Feb. 2013), 683.
28. Soto, V., and Frías-Martínez, E. Automated land use identification using cell-phone records. In *Proceedings of the 3rd ACM international workshop on MobiArch*, ACM (2011), 17–22.
29. Soto, V., Frias-Martinez, V., Virseda, J., and Frias-Martinez, E. Prediction of socioeconomic levels using cell phone records. *User Modeling, Adaption and Personalization* (2011), 377–388.
30. Sutton, P., Roberts, D., Elvidge, C., and Baugh, K. Census from heaven: an estimate of the global human population using night-time satellite imagery. *International Journal of Remote Sensing 22*, 16 (2001), 3061–3076.
31. Viboud, C., Bjø rnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., and Grenfell, B. T. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science (New York, N.Y.) 312*, 5772 (Apr. 2006), 447–51.
32. Yan, X.-y., Zhao, C., Fan, Y., Di, Z., and Wang, W.-x. Universal Predictability of Mobility Patterns in Cities. 1–19.
33. Zipf, G. The P 1 P 2/D hypothesis: On the intercity movement of persons. *American Sociological Review 11*, 6 (1946), 677–686.