Beyond the Baseline: Establishing the Value in Mobile Phone Based Poverty Estimates

Chris Smith-Clarke
University College London
Gower Street
London, UK
c.smith-clarke@ucl.ac.uk

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

ABSTRACT

Within the remit of `Data for Development' there have been a number of promising recent works that investigate the use of mobile phone Call Detail Records (CDRs) to estimate the spatial distribution of poverty or socio-economic status. The methods being developed have the potential to offer immense value to organisations and agencies who currently struggle to identify the poorest parts of a country, due to the lack of reliable and up to date survey data in certain parts of the world. However, the results of this research have thus far only been presented in isolation rather than in comparison to any alternative approach or benchmark. Consequently, the true practical value of these methods remains unknown.

Here, we seek to allay this shortcoming, by proposing two baseline poverty estimators grounded on concrete usage scenarios: one that exploits correlation with population density only, to be used when no poverty data exists at all; and one that also exploits spatial autocorrelation, to be used when poverty data has been collected for a few regions within a country. We then compare the predictive performance of these baseline models with models that also include features derived from CDRs, so to establish their real added value. We present extensive analysis of the performance of all these models on data acquired for two developing countries { Senegal and Ivory Coast. Our results reveal that CDR-based models do provide more accurate estimates in most cases; however, the improvement is modest and more significant when estimating (extreme) poverty intensity rates rather than mean wealth.

1. INTRODUCTION

In many parts of the world, the ability to acquire detailed and up to date knowledge of the distribution of wealth and poverty in a country remains an ambition rather than a reality. Traditionally, this has required manual collection of household survey data, the costs of which put this method beyond the means of some poorer nations. Towards mitigating this problem, recent research has highlighted the potential for producing estimates of the spatial distribution of poverty or socio-economic status from models incorporating features of mobile phone call activity.

Two broad types of approaches can be identified: the first assumes (often implicitly) that no ground truth data pertaining to wealth or socio-economic status distribution is available for any part of the country, as would be the case for countries in which no recent survey sampling has been undertaken. Research in this category has then produced general models from the study data that aimed to produce predictions for the entire country [23, 7, 21, 17]. The second assumes that a sample of the ground truth data is available instead, as could be the case if a survey had been undertaken in the past, or if a census is being conducted now but for a few selected regions only (e.g., to cut costs, with the plan to then interpolate the results to unsurveyed locations). Research in this latter category exploits such ground truth data to train models and make predictions for the remaining unknown locations only [24, 10].

Both types of approaches appear to demonstrate the value to be gained from mining Call Detail Records (CDRs) in terms of predictive power of poverty estimates; however, they all suffer from the same major limitation. Namely, that they have yet to establish a reasonable baseline against which a fair comparison can be made. The implicit assumption is that the best available baseline model would be a random guess, and therefore an improvement over this represents a positive result. Yet the reality is that socio-economic data is strongly correlated with population density; furthermore, it often contains a strong degree of spatial autocorrelation. Consequently, one would expect a baseline model that takes one or both of these factors into account to perform significantly better than a simple random one.

In order to measure the real added value of mining CDR features to estimate poverty in developing countries, we produce two fair baseline models, one exploiting correlations with population density only (to be used when no ground truth data is available), and one that also leverages spatial auto-correlations (to be used when partial ground truth data exists instead). By means of extensive comparative analysis of both baseline models and state-of-the-art CRD-based approaches, we then establish under what circumstances the latter do add value to poverty prediction models, and to what extent.

The remainder of the paper is structured as follows: we first provide a brief overview of related works that use CDRs to build poverty prediction models, and highlight their common limitation in terms of lack of realistic baseline com-

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
WWW2016 Montréal, Canada
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883076 .

parison. We then provide an overview of the Demographic and Health Surveys (DHS) data, which we use as ground truth for poverty in this work, and provide evidence of its relationship with population density, as well as its spatial dependency. Informed by this exploratory analysis, we then build and test baseline models, with the aim of predicting both average wealth and (extreme) poverty rates. Using CDR data from Senegal and Côte d'Ivoire, released as part of the D4D challenge series [6, 8], we then build models incorporating features mined from CDRs and compare their predictive performance to these baselines. We finally conclude the paper with a summary of the main findings, current limitations, and directions for future work.

2. RELATED WORK

In the last few years, there have been several works exploring the potential for CDRs to be used to predict poverty or socio-economic status, spurred in large part by the release of aggregated CDR datasets by Orange as part of the D4D (Data for Development) Challenges [6, 8], which are the datasets used in the current study.

In our own previous work [23, 22], we found that a number of features derived from such aggregated CDR data correlate strongly with poverty indices in two developing countries. However, due to the sparsity of poverty data at our disposal, we were limited to a straightforward correlation analysis, and the correlation coefficients were subject to wide confidence intervals. Other works have also found that characteristics of the call network reflect levels of socio-economic development [17, 21], but suffer from similar limitations. Bruckschen et al. [7] go further by building a more complex model of poverty at a finer level of spatial granularity, as well as extending their approach to numerous other demographic indicators. They successfully create a model that fits the observed data well, but stop short of validating their results on unseen data, which is required to establish the generalisability of their approach.

Other related works include that of Soto et al. [24] and Frias-Martinez et al. [12, 11], in which machine learning methods are used to predict the socio-economic status of a city's neighbourhoods using CDR data. Individual subscribers' mobile phone data is used as input, as opposed to the aggregated data used in the previously mentioned works, which allows for a richer set of features to be created, potentially leading to better predictive performance, but also raises privacy concerns. An apparently high level of accuracy is achieved in these works, and the method presented could potentially provide significant savings on the costs of conducting socio-economic surveys. However, in common with all research discussed here, no baseline is presented against which to compare the results, and therefore it remains unclear as to what the true benefit would be in implementing any of the proposed methods in the real world, and what the alternative would be.

In order to inform the design of realistic baseline models, we next present the results of an exploratory analysis of the Demographic and Health Surveys (DHS) data, which is used by governments and not-for-profit organisations worldwide as representative data of populations in developing countries. In this work, we will use DHS data as ground truth data for poverty estimates.

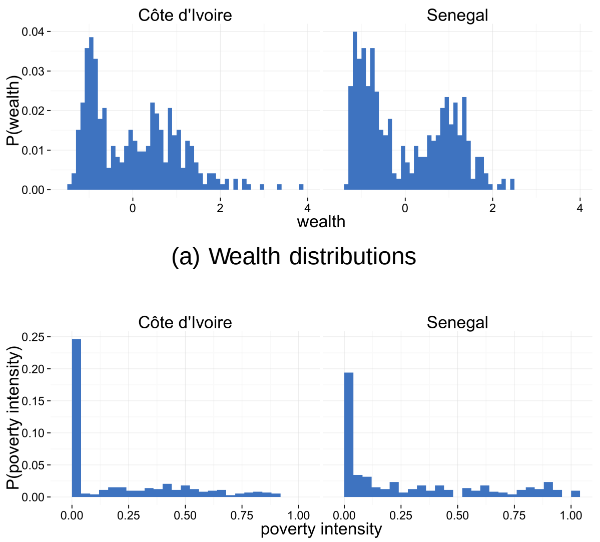
3. DHS WEALTH INDEX

Demographic and Health Surveys (DHS) are conducted in several developing countries, usually in collaboration with the national statistical agency and other organisations. The surveys take place with a sample of households that is designed to be representative at the largest subnational administrative region, of which there are 14 in Senegal and 11 in Côte d'Ivoire. The household sampling process consists of several stages. First, the country is stratified by an urban or rural designation within each subnational region; then, within each stratum, enumeration areas (EAs) are selected with a probability proportional to their size. EAs normally consist of neighbourhoods in urban areas and villages, or groups of villages in rural areas. Finally, households are randomly selected with uniform probability within each EA selected in the previous stage. The group of selected households within each EA are known as clusters. The GPS coordinates of the centroid of each cluster is provided with the DHS in order to enable spatial analysis of the survey data. However, in order to hide the identity of selected households, coordinates are randomly displaced with a circle of radius 2 km for urban clusters and 5 km for rural clusters, with 1% of rural clusters being displaced up to 10 km. Table 1 presents the number of clusters sampled for both Senegal and Côte d'Ivoire, together with their overall population and surface area.

Table 1: DHS and country summary statistics

	Senegal	Côte d'Ivoire
Population	20 million	15 million
Area	197 km^2	322 km^2
Clusters	385	341

The DHS includes questions regarding the ownership of certain assets, such as mobile phones, computers, vehicles and refrigerators, as well as questions related to living conditions, such as access to electricity, sanitation and material used for flooring. These factors are combined using Principal Components Analysis into an index representing the socio-economic level of each household (note this index is included in the DHS and not created by the current authors). When estimating poverty, we operate at the cluster level rather than that of individual households, we thus aggregate the wealth index in two different ways. In the first case, we take the median wealth index of households in the cluster to represent the average wealth at that location. In the second case, we take the percentage of households in the cluster that are among the poorest quintile of households nationally; the latter thus represents the intensity of poverty at a location, or the household poverty rate. The two aggregate measures differ in that the latter is invariant to the distribution of wealth among households within the cluster; in other words, extreme poverty within the cluster is not masked by the existence of wealth within a cluster too, as may be the case with average wealth (although use of the median will mitigate against the influence of extreme wealth it will still fail to reflect the existence of poverty in an otherwise wealthy area). Figures 1a and 1b respectively illustrate the distribution of wealth and poverty rates derived from DHS data, for the two countries under exam. We next perform an exploratory analysis of such data, with the



(a) Wealth distributions

(b) Poverty rate distributions

Figure 1: Distributions of average wealth and poverty rate.

aim to inform the design of well-grounded baseline predictor models of poverty.

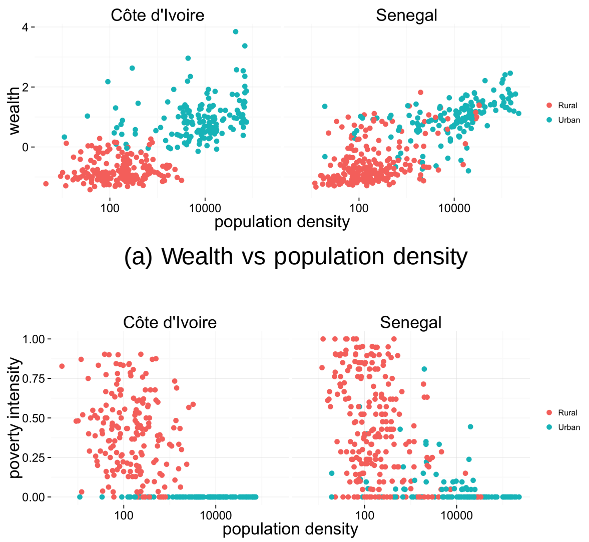
3.1 Wealth and population density

A link between population density and prosperity is is often posited, with many mechanisms proposed to explain this relationship, including efficiency of service provision and increased access to diverse sources of information and opportunity [20, 13, 5]. For Senegal and Côte d'Ivoire, this relationship can clearly be seen in Figure 2a, where wealth (as computed before) is plotted against population density, here defined as the population within 1 km circle centred on the cluster point. Similarly, Figure 2b plots poverty rate vs population density. As shown, denser areas tend to also be wealthier and have lower concentration of poverty. We can see a marked division between urban and rural locations, with urban locations tending to be wealthier and with a lower concentration of poverty. Indeed, in Côte d'Ivoire no urban cluster contains a household among the poorest 20% (Figure 2b).

Considering these simple observations, we stipulate that a realistic baseline prediction method ought to take population density into account. In particular, we propose to do so by computing the log of population density, instead of using a binary urban/rural indicator variable, both because the urban/rural designation may not always be readily available, and because it is more appropriate to use a continuous predictor to estimate a continuous outcome.

3.2 The spatial distribution of wealth

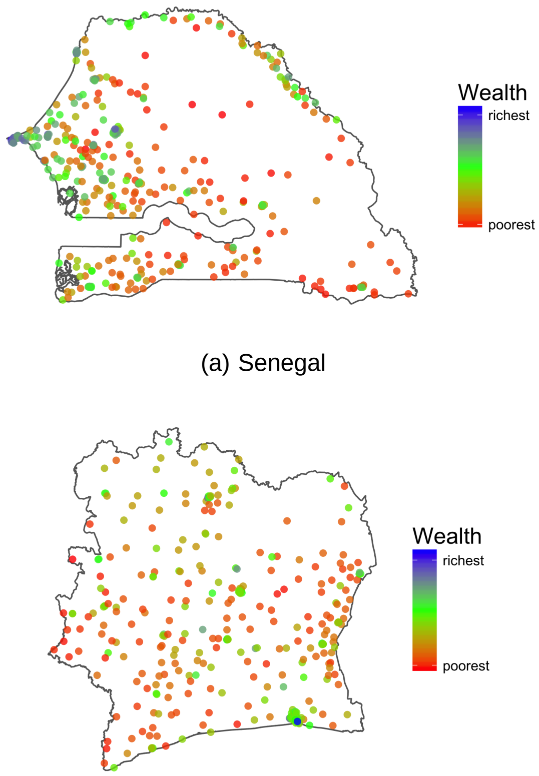
Next we look at the spatial distribution of average wealth and poverty rate in Senegal and Côte d'Ivoire. Figure 3 shows the average wealth at DHS cluster locations (we omit figures depicting poverty rate as they are very similar). A degree of spatial clustering of wealth is evident, with wealthier clusters tending to appear in close proximity, although a significant number of exceptions are apparent. These figures alone also fail to depict the level of clustering at smaller scales.



(a) Wealth vs population density

(b) Poverty rate vs population density

Figure 2: Average wealth and poverty rate in relation to population density.



(a) Senegal

(b) Côte d'Ivoire

Figure 3: Average wealth at DHS cluster locations

We therefore quantify the level of spatial clustering further with a correlogram, which measures the similarity of the variable of interest (i.e., wealth) at various distances. The similarity measure used is Moran's I [19]:

$$I = \frac{N}{\sum_{i,j} w_{ij}} \frac{\sum_{i,j} w_{ij} z_i z_j}{\sum_i z_i^2}$$

where N is the number of points, $z_i = (y_i - y)$ is the deviation from the mean in the quantity of interest (median wealth or poverty intensity in our case). The spatial weights w_{ij} are derived from the distance between pairs of points. To produce the correlogram, points pairs are divided into bins according the distance between them, at 2 km increments. Moran's I is then calculated separately for the members of each bin. Positive values of Moran's I indicate the presence of positive spatial autocorrelation and Figure 4 depicts the decrease in the strength of spatial autocorrelation of median wealth as the distance between points increases.

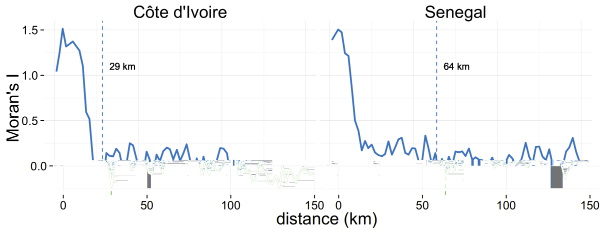
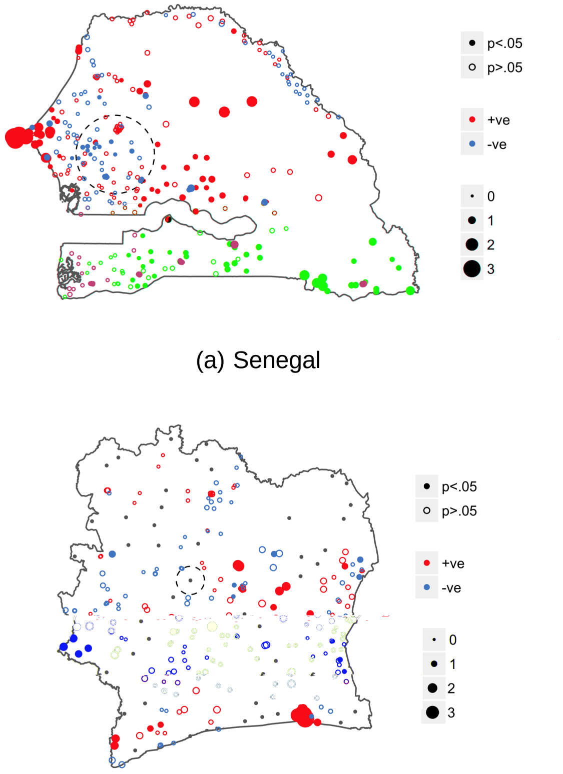


Figure 4: Correlograms showing the strength of spatial autocorrelation in wealth according to distance intervals. Distances are binned at 2 km increments with bins containing a minimum of 55 point pairs. Vertical lines indicate the distance at which spatial autocorrelation reduces to a level expected from randomly placed points.

It is clear from this simple analysis that estimates at unsampled points derived from nearby sampled points would be significantly more accurate than random guessing. Subsequently, a baseline against which to evaluate predictions from CDR data ought to take proximity to sampled points into account, if these were available. However, it is also clear that, in the case of Senegal and Côte d'Ivoire, many locations are not within range of sampled points for such an approach to be reliable on its own. Furthermore, estimating unsampled locations solely as a function of nearby sample points is likely to miss locations which are outliers relative to their neighbours, and these are arguably among the most important to identify. To establish the extent that this is likely to occur, we measure spatial autocorrelation within the neighbourhood of each point using local Moran's I, or Local Indicators of Spatial Autocorrelation (LISA) [2]:

$$I_i = \frac{z_i}{m} \sum_j w_{ij} z_j$$

where $m = \sum_i z_i^2 / N$. Figure 5 maps the local Moran's I of average wealth in the neighbourhood of each sample point at the spatial scale corresponding to the approximate distance at which the level of spatial clustering reduces to that expected from randomly placed points (i.e., 29 km in Côte d'Ivoire and 64 km in Senegal, indicated by the vertical lines in the correlograms of Figure 4). As can be seen, many points have too few neighbours to allow a statistically significant estimate to be computed at the 5% level ($p < .05$), suggesting that poverty estimates derived from proximity alone would perform poorly in these areas. Furthermore, the figures also reveal several sample points negatively correlated with their neighbourhood, which again indicates that relying solely on spatial dependency for estimating unsampled points would be inappropriate in this context.



(a) Senegal

(b) Côte d'Ivoire

Figure 5: Local Moran's I of wealth. The dashed circles show an example neighbourhood radius for reference, 64 km for Senegal and 29 km for Côte d'Ivoire. Points with too few neighbours to compute correlation are coloured grey.

4. BASELINE MODELS

We now leverage the above observations to construct two simple yet well-grounded baseline models to estimate poverty: one exploits the existing correlation between poverty and population density (as evidenced in Section 3.1); the other expands it, by adding the spatial auto-correlation of poverty (as evidenced in Section 3.2). The former is most suitable when no ground truth poverty data about a country exists; the latter is applicable when partial ground truth data exist instead for a subset of clusters. In order to later assess the predictive power of these new baseline models, relative to a traditional random baseline, we first produce a random baseline ourselves.

Random Baselines

We built two random baseline models that consist of values drawn from distributions that approximate those exhibited in Figure 1. Let us consider approximating wealth distributions first (Figure 1a). This requires two steps: we first take 5000 random draws from two distinct normal distributions, and concatenate the results to form a vector of length 10000. From this vector, we then take n uniformly random samples, where n is the number of observations in the dataset. The parameters of the normal distributions in the first step were chosen such that the probability density functions of the random vector in the second step resemble the density functions of the observed data. For Senegal, these are $\mathcal{N}(-0.9, 0.3)$ and $\mathcal{N}(1, 0.5)$; for Côte d'Ivoire, these are $\mathcal{N}(-1, 0.4)$ and $\mathcal{N}(0.65, 0.7)$.

Let us now consider approximating poverty rate distributions instead (Figure 1b). As was done for wealth, we proceed in two steps: first, we sample from a binomial distribution with probability of success equal to the proportion of clusters that have poverty rate greater than zero, which is 0.634 for Senegal and 0496 for Côte d'Ivoire. Next, all non-zero samples are replaced by sampling from a uniform distribution between 0 and 1.

To measure the predictive performance of these random baseline models, we have computed the mean absolute error (MAE), and also Spearman's rank coefficient (ρ) since we are interested in predicting the relative ordering of locations too. Results are are shown in Table 2. We will later compare the MAE and Spearman's ρ of our new baselines, as well as the models exploiting CDR-mined features, relative to these reference values.

Table 2: Random baseline metrics

		Wealth	Poverty Rate
Senegal	MAE	1.154	0.360
	Spearman's ρ	0.248	0.249
Côte d'Ivoire	MAE	1.143	0.308
	Spearman's ρ	0.249	0.236

Population-Density Baseline

The first baseline simply consists of a regression model, where the independent variable is the log of population density for cluster area i, and the response variable y_i will be either median wealth or poverty rate of area i.

Spatial-Lag Baseline

For the second baseline, we add a spatially-lagged dependent variable:

$$z_i = \sum_j w_{ij} y_j,$$

where y is the response variable, w is a weight inversely proportional to squared distance between points i and j, and $\sum_j w_{ij} = 1$. By using squared distance to calculate the weights, the effect of distant neighbours on the lagged variable will be negligible for those points with relatively close neighbours, whilst still allowing us to compute a lag for those points with no nearby neighbours.

5. CDR MODELS

We now turn our attention to CDR data and the features that we can extract from it to build predictive models of poverty. CDR data for Côte d'Ivoire and Senegal was released as part of the 1st and 2nd Orange Data for Development Challenge,[1] respectively. The data is summarised in Table 3. The data from Senegal covers a much longer period than that from Côte d'Ivoire (52 weeks verses 12 weeks); however, due to service providers' larger market share in Côte d'Ivoire, the average daily volume is much larger, at 1.4 calls per person compared to 02 calls per person in Senegal. To extract features from this raw data, we first need to build a graph, in which the base transceiver stations (BTS) are vertices, and the edges between pairs of vertices are weighted by call volume, or total number of calls, between those towers. In previous work [23], little difference was found between features of the call volume graph and call duration graph; for simplicity, we therefore only consider the call volume graph here.

Table 3: Descriptive statistics of CDR data.

	Senegal	Côte d'Ivoire
Country Population	20 m	15 m
Time span (weeks)	52	12
Number of BTS towers	1614	1217
Mean Daily Volume	4.0 m	10.8 m
Mean BTS Distance	236 km	228 km

Based on the state-of-the-art works reviewed in Section 2, we can summarise the list of possible CDR-mined features as follow (more details can be found in the referenced literature).

Total Call Volume and Total Call Volume Per Person . The level of call activity is likely to reflect wealth of an area in a straightforward way, i.e., the more affluent people are, the more likely they will own a phone and make more calls. Conversely, it also possible that mobile phone adoption could spur economic development by reducing the cost of accessing information and by improving the efficiency of supply chain management [1].

Introversion . This metric computes the ratio of internal call volume (source and target are one and the same) to external call volume (source and target are different) of a cell tower. The intuition here is that more introverted areas will have access to fewer resources and thus less opportunities for economic development.

Network Advantage. The 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. We include the entropy of a tower's edge weights [9] as a measure of the diversity of its contact locations. We also include pagerank and eigenvector centrality as measures of the tower's integration and importance in the communication network. These last two have previously been found to correlate with a poverty index in Côte d'Ivoire [17].

Gravity Residuals. Previous work [23] found that the difference between observed and expected flows between locations reflects the level of wealth in an area. To estimate flows, a gravity model was used, which expects the volume of interaction between two areas to be proportional to the mass (i.e., population) of those areas, and inversely proportional to the distance between them. The model had already been successfully used to describe macro scale interactions

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

(e.g., between cities, and across states), using both road and airline networks [4, 14] and its use has extended to other domains too, such as the spreading of infectious diseases [3, 25], cargo ship movements [15], and to model intercity phone calls [16]. When applied to estimate poverty, the underlying intuition is that, by examining the residuals between observed and expected flows, one can capture the restricting effect of poverty on an area's interactions with others.

With respect to [23], we make two improvements: previously, only the mean negative residual of a tower's edges, plus the residual of the sum of a towers edges, were included as features. Here, we extend the approach by computing also the other graph features on the gravity interaction network (i.e., pagerank, eigenvector centrality, entropy and standard deviation), and taking the residuals as additional features. The motivation is the same, that is, we surmise that differences between observed and expected values of, for example, entropy, could indicate that that location is experiencing a less diverse set of interactions with other areas, which in turn could indicate lower levels of wealth. Furthermore, in [23], a simple gravity model with a single scaling parameter was used: that is, $F_{ij} = \beta \frac{P_i P_j}{d_{ij}^2}$, where P_i is the population of area i, $d_{i,j}$ is the Euclidean distance between tower locations i and j, and β is the scaling parameter fitted to the data. However, more nuanced models exist to estimate flows, including: a 4-parameter version of the gravity model; a 9-parameter distance-varying version, in which the parameters are allowed to change at some distance threshold (this is designed to avoid the common tendency of gravity models to perform poorly at shorter distances); and the radiation model [18]. In this work we experimented with all three variants. Intuitively, we might expect residuals derived from the radiation model to perform best as predictors of wealth or poverty rate since, unlike all gravity models, the radiation model is not fitted to observed flows. By fitting the model to observed flows, we might mask the very signal we hope to uncover, that is, the error. However, we found that the 4-parameter gravity model:

$$F_{ij} = \beta_1 \frac{P_i^{\beta_2} P_j^{\beta_3}}{d_{ij}^{\beta_4}},$$

fitted by taking the logarithms of each element and performing Ordinary Least Squares regression on the resulting formula (i.e., $log(F_{ij}) = log(\beta_1) + \beta_2 log(P_i) + \beta_3 log(P_j) + \beta_4 log(d_{ij})$), not only had the best fit to the observed flows, but also produced residuals which performed best as predictor variables in the regression models. We therefore used the residuals from this model as predictors.

All the above listed features are aggregated to cluster points by taking an average of BTS towers within 30 km, weighted by squared Euclidean distance from the cluster point to the tower. The maximum range of a BTS tower is determined by many things, such as its design, configuration, the local terrain and climate. The real maximum distance of each BTS tower is unknown to the authors; therefore, we chose 30 km as a reasonable maximum distance so as to ensure that all cluster points would be assigned a value. Using squared distance as weights also means that, in denser environments where the maximum distance is likely to be much shorter than 30 km, many towers will be much closer to the cluster point, meaning that the effect of more distant towers on the computed mean will be negligible.

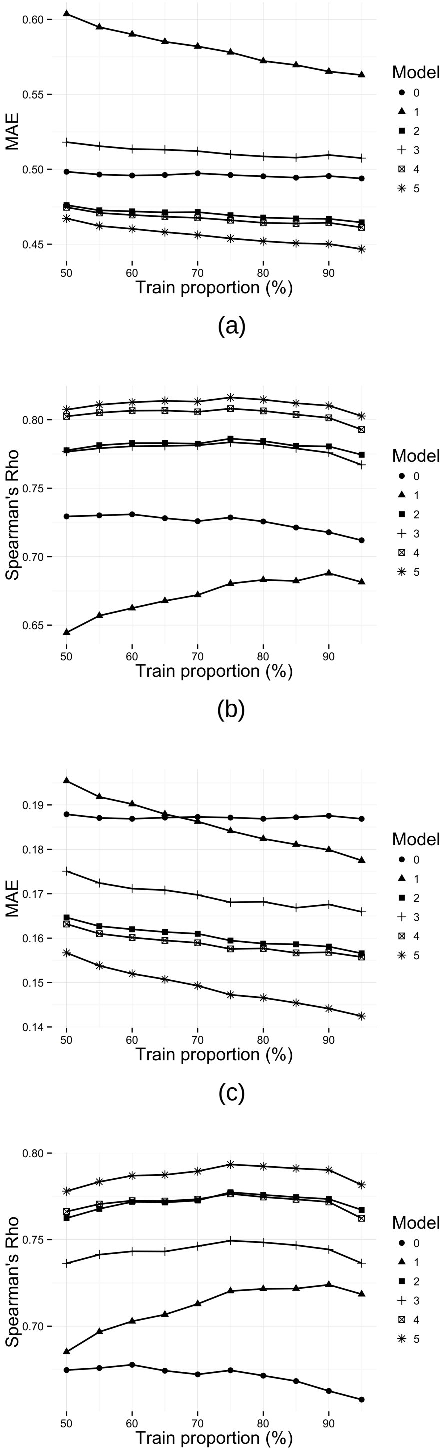
For completeness, we also compared results using raw distance as weights and also by simply taking the value of the closest tower, which for point data is equivalent to using Voronoi cells, as has been used in much previous work. We found that using a squared distance weighted mean gave best predictive performance. A likely explanation for this is that calls are not always routed through the nearest BTS tower. As well as distance, load balancing, directionality of tower cells, and whether and how fast a caller is travelling, can all affect which tower a call is connected to. This would render a closest tower, or Voronoi cell approach, less accurate. Although a distance weighted mean approach does not solve the problem of not knowing which tower a call is routed through, by effectively smoothing out the feature surface it may offer a more accurate representation. We leave a more rigorous investigation into these considerations for future work.

6. RESULTS

First we performed an initial comparison between the random baselines and the newly proposed ones. To predict $y =$ average wealth, we used Ordinary Least Squares regression; to predict $y =$ poverty rate, given its highly skewed distribution, we opted for a hurdle model instead. Hurdle models are applied to count data that contain a larger number of zero counts than would be expected if the data could be modelled by a simple discrete distribution, such as a Poisson distribution. The hurdle model consists of two stages: first, a binomial model is used to estimate whether a data point is zero or positive non-zero; then, Possion regression truncated to 1 is used to predict the value of those points estimated to be positive non-zero in the first stage. In our case, we count the number of poor households in a cluster, and we include the total number of households as an offset variable (coefficient is set to 1), which effectively means that we are estimating the household poverty rate in each cluster.

To obtain a robust measure of predictive performance, we ran 1000 iterations of random train/test splits with varying training proportion. When using the spatial-lag baseline, the lagged variable was computed using only members of the training set. Figures 6(a) to 6(d) report the mean absolute error (MAE) and Spearman's rank correlation coefficient computed with the test data, for both average wealth and poverty rate, for Senegal. Figures 7(a) to 7(d) do the same for Côte d'Ivoire.

Concentrating first on models 0 (population density baseline), 1 (spatial lag baseline) and 2 (population density with lag baseline), the spatial and population density baselines significantly outperform the random baselines in all cases, even with relatively few training points. For example, when estimating the poverty intensity in Senegal with 50% training data, which amounts to 192 training points, the random baseline achieved a MAE of 1.154, while we now reduce MAE to 0.188 for the population density baseline, 0.195 for the spatial lag baseline, and 0.164 for the baseline using both population density and spatial lag. As expected, the spatial lag baseline performs less well with fewer training examples, whereas the population density baseline performance metrics remain fairly stable as training size increases. A similar story can be told for the poverty rate baselines, and for baselines in Côte d'Ivoire.



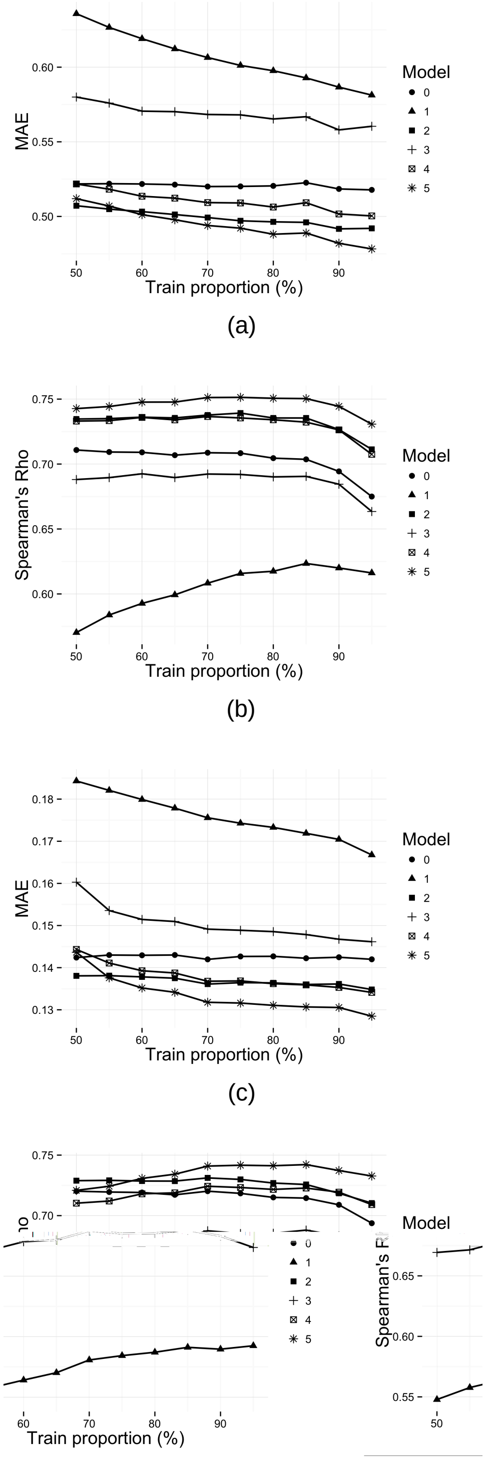
(a)

(b)

(c)

(d)

Figure 6: Regression test scores for average wealth (a, b) and poverty rate (c, d) in Senegal. Predictor variables in each model are, 0: Population density, 1: lag, 2: population density + lag, 3: CDR features, 4: CDR features + population density, 5. CDR features + population density + lag.



(a)

(b)

(c)

(d)

Figure 7: Regression test scores for average wealth (a, b) and poverty rate (c, d) in Côte d'Ivoire. Predictor variables in each model are, 0: Population density, 1: lag, 2: population density + lag, 3: CDR features, 4: CDR features + population density, 5. CDR features + population density + lag.

Next we compare the results of the CDR based models. Figures 6(a) to 6(d) report the mean absolute error (MAE) and Spearman's rank correlation coefficient computed on varying splits of train/test data, for both average wealth and poverty rate, for Senegal. Figures 7(a) to 7(d) do the same for Côte d'Ivoire. As well as our new baseline models (0 - population density, 1 - spatial lag, and 2 - population density with spatial lag), the figures report results for the following models: CDR features only (model 3), CDR features with population density (model 4), and CDR features with both population density and spatial lag (model 5).

In interpreting results, it is useful to think in terms of concrete situations in which different kinds of data are available, as described in the introduction. The two situations differ in the availability of ground truth poverty or socio-economic status data with which we can create a spatially lagged variable.

In case there is no such data available, we can compare baseline model 0 (population density only) with model 3 (CDR features only) and 4 (population density and CDR features). We can see that in Senegal model 3 out performs model 0, while in Côte d'Ivoire it fails to do so. In both countries, the model that combines population density with CDR features (model 4) performs the best.

In case there is some ground truth data from which to compute the spatially lagged variable, we can then compare models 1 (lag only), 2 (population density and lag), 3 (CDR features only), and 5 (CDR features plus population density plus lag). Baseline model 1 performs worst out of all models, which is somewhat surprising given the level of spatial autocorrelation present in the data. However, as expected, it closes the gap as more training data becomes available. When population density is included (model 2), accuracy improves and is indeed higher than a model with CDR data only (model 3). As before, the model combining CDR data with baseline features (model 5) is the one giving the highest performance gains.

The headline result here may be that the models including CDR features as predictors outperform those that do not, and indeed model 5, which contains all the CDR features as well as population density and the spatially lagged variable, outperforms our baselines when a sufficient amount of training data is provided. However, one has to be critical of the actual gains, as these appear to be rather small when analysed more closely.

Let us consider the case of Senegal first. The MAE of model 4 (CDR features and population density) is only 5% lower than model 0 (population density alone), when estimating average wealth in Senegal with 50% training data, and it only reaches 7% improvement at 95% training data. ρ in this case shows an improvement of 10-11%. If we look at poverty rates rather than average wealth, improvements are more significant: the MAE of model 4 (population density and CDR features) is 13% lower than model 0 (population density), with the gain climbing to 17% with 95% training data; ρ improves 14-16%. However, in this case the improvement brought in by spatial lag is less neat: comparing models 2 (spatial lag and population density) and 5 (all features), we find the addition of CDR features to only offer a reduction in MAE of 2-4% and an increase in ρ of around 4% for average wealth predictions, and a reduction of 5-10% in MAE and increase of just 2% in ρ for poverty rate predictions.

Let us now turn our attention to Côte d'Ivoire. When predicting average wealth, model 4 (CDR features and population density) provides an improvement of up to 4% in MAE and 3-5% in ρ over our baselines. However, when adding lag, model 5 (all features) performs marginally worse than baseline model 2 (population density and lag), with 50% training data according to MAE scores; only at 60% training data does model 5 begin to offer an improvement, though this reaches just 3% at 95% training data. When estimating poverty rate, model 4 has a higher MAE and lower ρ than our baseline with only 50% training data, and it becomes better as more data is available, decreasing MAE by up to 5% and increasing ρ by up to 2%. Similar figures can be given when comparing models that include spatial lag.

To conclude, the addition of CDR features to models that already account for population density (and spatial lag if available) does provide an improvement in predictive performance; more so in Senegal than in Côte d'Ivoire, and more so for poverty rate than average wealth. However, improvements appear to be modest across the board, and the results therefore much less striking than previous research suggests when comparing against random baseline models.

7. DISCUSSION

In this work, we have investigated the relationship between fine grained poverty data and population density estimates, as well as the spatial distribution of said poverty data. We have used the findings to inform the construction of baseline predictive models against which to compare more complex CDR-based models. Previous research had demonstrated a strong correlation between CDR-based features and poverty or wealth indices, and shown the predictive performance of CDR-based models to be good relative to random baseline models. However, by comparing these CDR-based models with our new baselines, we have gone beyond that, to establish the extent to which the inclusion of CDR features does offer additional value, if any. By means of a comparative performance analysis using data from two developing countries (namely, Senegal and Côte d'Ivoire), we have found that CDR features do in fact offer improved predictive performance, particularly when predicting poverty rate as opposed to the average wealth of an area. Although the improvement is modest in most cases, it is consistent nonetheless. Thus we can remain optimistic about the value that such an approach can offer to governments and organisations that lack the means to perform a more comprehensive survey of a country's socio-economic status. We have also found that results vary across the two studied countries, with the added value of CDR features being small in Côte d'Ivoire compared to Senegal. Continuous testing and refinement ought to be an integral part of any implementation of the methods outlined here and elsewhere.

This work is subject to some limitations, owing largely to characteristics of the available data. For our analyses, we have utilised DHS surveys and explored the effect of varying levels of training data. However, it should be noted that the survey clusters themselves represent only a fraction of the census enumeration areas within each country. For example, in Senegal, the survey clusters represent only 4% of a total of 9733 enumeration areas, and on average only 20% of households are surveyed within each selected enumeration area. Consequently, some caution is appropriate when extrapolating the results presented here to the entire country.

However, the sampling methodology employed by the DHS surveyors is such that there is nothing inherently different between the selected enumeration areas and unselected ones, therefore, we would not expect modelling outcomes to differ greatly were we to have access to fully representative data. An important direction for future work would be to perform a comprehensive analysis of a country in which poverty or socio-economic status data is available for every enumeration area. In addition, we have not explored the effect of the random displacement that is applied to DHS cluster and BTS tower locations. This will introduce a degree of uncertainty in both the predictor and response variables, but we expect this to be largely compensated for by the spatial smoothing that takes place when aggregating predictor variables. Nevertheless, a proper investigation of sensitivity of our modelling approach to such displacement would be prudent.

By presenting results from two different countries, we expanded the scope at which this approach can be evaluated; however, some differences exist between the datasets which may limit the degree to which results can be compared. Firstly, the time periods from which the CDR data is extracted differ in length, with 20 weeks from Côte d'Ivoire compared to a full year from Senegal. The shorter period covered in Côte d'Ivoire could contribute to the smaller gain from CDR features we see in this country, perhaps due to seasonal variation in calling patterns. In future work, we will examine the temporal stability of CDR features over time and test the sensitivity of predictions to differing time spans. Secondly, there is large difference between the population coverage in each country, with 3 million subscribers in Côte d'Ivoire, representing approximately 13% of the population, and 300,000 subscribers in Senegal, representing just 2.3% of its population. The fact that greater gains are achieved with CDR features in Senegal suggests that this small sample size is nevertheless representative enough to be utilised. However, the significance of the size of the subscriber base, as well as other factors such as cultural and environmental differences between countries, will only come to light as further such studies are produced.

8. ACKNOWLEDGMENTS

This work was supported in part by a Google Europe Doctoral Fellowship.

9. REFERENCES

[1] J. C. Aker and I. M. Mbiti. Mobile Phones and Economic Development in Africa. Journal of Economic Perspectives, 24(3):207{232, 2010.

[2] L. Anselin. Local indicators of spatial associationâATlisa. Geographical Analysis, 27(2):93{115, 1995.

[3] D. Balcan, V. Colizza, B. Goncalves, H. Hu, J. J. Ramasco, and A. Vespignani. 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):21484{9, Dec. 2009.

[4] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. Proceedings of the National Academy of Sciences of the United States of America, 101(11):3747{52, Mar. 2004.

[5] L. M. A. Bettencourt, J. Lobo, D. Helbing, C. KÃijhnert, and G. B. West. Growth, innovation, scaling, and the pace of life in cities. Proceedings of the National Academy of Sciences, 104(17):7301{7306, 2007.

[6] V. D. Blondel, M. Esch, C. Chan, F. Clerot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki. Data for Development: the D4D Challenge on Mobile Phone Data. page 10, Sept. 2012.

[7] F. Bruckschen, T. Schmid, and T. Zbiranski. Cookbook for a socio-demographic basket. In D4D Challenge Senegal Sessions Scientific Papers, Netmob '15, 2015.

[8] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel. D4d-senegal: The second mobile phone data for development challenge. 2014.

[9] N. Eagle, M. Macy, and R. Claxton. Network diversity and economic development. Science (New York, N.Y.), 328(5981):1029{31, May 2010.

[10] V. Frias-martinez, V. Soto, J. Virseda, and E. Frias-martinez. Computing Cost-Effective Census Maps From Cell Phone Traces. In Pervasive Urban Applications (PURBA), Newcastle, 2012.

[11] V. Frias-Martinez and J. Virseda. On the relationship between socio-economic factors and cell phone usage. In Fifth International Conference on Information and Communication Technologies and Development (ICTD '12), New York, New York, USA, Mar. 2012.

[12] V. Frias-Martinez, J. Virseda-Jerez, and E. Frias-Martinez. On the relation between socio-economic status and physical mobility. Information Technology for Development, 18(2):91{106, Apr. 2012.

[13] K. M. M. Gary S. Becker, Edward L. Glaeser. Population and economic growth. The American Economic Review, 89(2):145{149, 1999.

[14] W. Jung and F. Wang. Gravity model in the Korean highway. Europhysics Letters, 81, 2008.

[15] P. Kaluza, A. Kölzsch, M. T. Gastner, and B. Blasius. The complex network of global cargo ship movements. Journal of the Royal Society, Interface / the Royal Society, 7(48):1093{103, July 2010.

[16] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel. Urban gravity: a model for inter-city telecommunication flows. Journal of Statistical Mechanics: Theory and Experiment, 2009(07):L07003, May 2009.

[17] H. Mao, X. Shuai, Y.-Y. Ahn, and J. Bollen. Mobile communications reveal the regional economy in côte dâAZivoire. In D4D Challenge Book of Abstracts, Netmob, 2013.

[18] P. Masucci, J. Serras, A. Johansson, and M. Batty. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. Physics Review E, 8(2), August 2013.

[19] P. A. P. Moran. Notes on continuous stochastic phenomena. Biometrika, 37(1/2):17{23, 1950.

[20] W. Pan, G. Ghoshal, C. Krumme, M. Cebrian, and A. Pentland. Urban characteristics attributable to density-driven tie formation. Nature communications, 4:1961, 2013.

[21] N. Pokhriyal and W. Dong. Virtual networks and poverty analysis in senegal. In D4D Challenge Senegal Sessions Scientific Papers, Netmob '15, 2015.

[22] C. Smith, A. Mashhadi, and L. Capra. Ubiquitous sensing for mapping poverty in developing countries. In D4D Challenge Book of Abstracts, Netmob, 2013.

[23] C. Smith-Clarke, A. Mashhadi, and L. Capra. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In Proceedings of the 32nd annual ACM conference on Human factors in computing systems. ACM, 2014.

[24] V. Soto, V. Frias-Martinez, J. Virseda, and E. Frias-Martinez. Prediction of socioeconomic levels using cell phone records.User Modeling, Adaption and Personalization, pages 377{388, 2011.

[25] C. Viboud, O. N. Bj rnstad, D. L. Smith, L. Simonsen, M. A. Miller, and B. T. Grenfell. Synchrony, waves, and spatial hierarchies in the spread of influenza. Science (New York, N.Y.) , 312(5772):447{51, Apr. 2006.