# A Tool for Estimating and Visualizing Poverty Maps

**Myeong Lee[1], Rachael Dottle[2], Carlos Espino[2], Imam Subkhan[3], Ariel Rokem[3], Afra Mashhadi[3]**
[1]iSchool, University of Maryland, College Park, USA
[2]Columbia University, New York, USA
[3]University of Washington, Seattle, USA
myeong@umd.edu, rcd2127@barnard.edu, carlos.espino@columbia.edu, imams@uw.edu, arokem@gmail.com, afra.mashhadi@gmail.com

## ABSTRACT
"Poverty maps" are designed to simultaneously display the spatial distribution of welfare and different dimensions of poverty determinants. The plotting of such information on maps heavily relies on data that is collected through infrequent national household surveys and censuses. However, due to the high cost associated with this type of data collection process, poverty maps are often inaccurate in capturing the current deprivation status. In this paper, we address this challenge by means of a methodology that relies on alternative data sources from which to derive up-to-date poverty indicators, at a very fine level of spatial granularity. We validate our methodology for the city of Milano and demonstrate how it could be used to implement a poverty mapping tool for policy makers.

## INTRODUCTION
Small area estimation *poverty maps* are a recent innovation that provide detailed estimates of poverty levels in highly disaggregated geographical units [2]. The visualization of deprivation information in this form has been shown to be extremely effective in empowering policy makers and local municipalities to identify those areas in most need of interventions and revitalization programs. Poverty maps not only improve readability compared to traditional tabular data format by simultaneously preserving the spatial distribution of welfare, but also are powerful tools for capturing relations between deprivation and geographical factors such as city infrastructure and offerings.

In order for poverty maps to be impactful in determining and designing interventions, the poverty data ought to be up-to-date and presented in disaggregate level of granularity. However, due to the high costs associated with the household survey and censuses, National Statistical Institutes and the like often possess social and economic well being information that is out-of-date (i.e., collected infrequently) and only inclusive of a rather small sample of the population.

Within the remit of 'Data for Development' there have been a number of promising recent works, whereby researchers have relied on alternative sources of data to estimate deprivation. For example, models exploiting Call Detail Records (CDRs) from mobile phones have shown to be good indicators of the spatial distribution of socio-economic status in developing countries [7, 4]. Other sources of data including readily available open datasets such as those of Volunteer Geographical Information (VGI) have been shown to successfully predict the poverty level based on the offering advantages of the cities [8]. While both these alternative sources have been shown to successfully estimate deprivation level, their results have only been presented in isolation rather than in comparison. Furthermore, each has various shortcomings. The CDR are often hard to obtain due to their commercial and privacy sensitive nature. The Open Data sources on the other hand are readily available but could suffer from biases in coverage [5].

In this paper, we propose a methodology that leverages both open source and proprietorial datasets to compute and offer a more complete spatial distribution of deprivation at the city scale. In so doing, we extract features corresponding to the functional offerings and connectedness of the urban areas from OpenStreetMap (OSM), and network and activity related features from the CDR for the city of Milano. We quantitatively draw a comparison between the poverty estimation offered by each source with respect to poverty indicators derived from costly census data (ground truth). Our results, based on a Random Forest Classifier, indicate that the combinations of features that were extracted from CDR and OpenStreetMap data predicted poverty level better than the baseline models and complement each other. Based on the proposed methodology we built a visualization tool that displays the estimated poverty level for each area of the city as well as the determinants that contribute to the predicted poverty level.

## FEATURE EXTRACTION AND BASELINES
For the purpose of this study we used CDR dataset made available by Telecom Italia [1] and freely available OSM data. We extracted features from these datasets following the measures that are often considered as proxy of the poverty level. Insights for the dynamics of a city from CDR data are usually based on individual-level mobility, people's communication network, and the amount of activity. Due to the anonymization of the CDR data, however, it is not possible to extract mobility information. We thus used network advantages and activity signatures as major predictors from the dataset. Network advantages include (1) call volume (the extend to which people are active in each region); (2) introversion (the amount of calls made inside of a region divide by the outgoing call, which is related to access to resources and social capital outside of a neighborhood); (3) PageRank and eigenvector centrality (relative popularity based on communication network structure); and (4) entropy (the diversity of a region that may imply the potential capability to reach diverse resources). Furthermore, we generated different CDR activity signatures by separating weekdays and weekends due to different human behavioral patterns. As a result, a vector with 288 elements (144 for weekdays and 144 for weekends in 10-minute interval) was

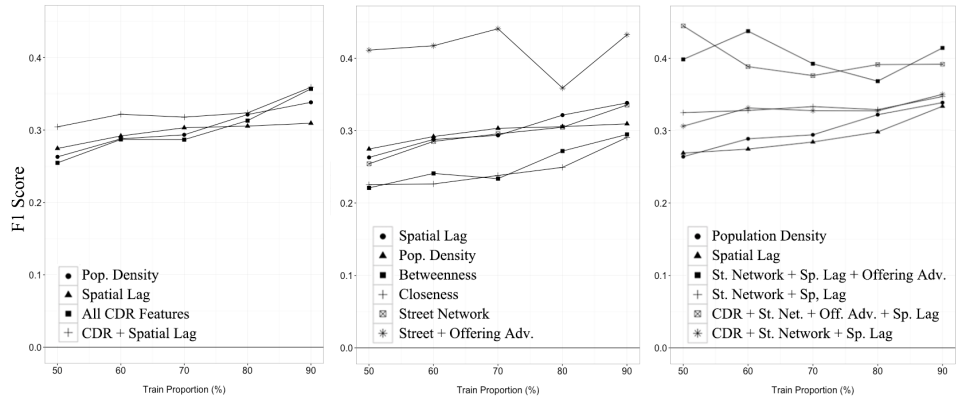


**Figure 1. Average F1-score for models based on CDR features (left), OSM features (middle) and combined features (right) for varying train/test set.**

constructed for each region. By conducting K-means clustering for these vectors, we classified every census tract into three categories, i.e., business, residential, and mixed regions. The number of clusters, $K$, was determined by using Davies-Bouldin cluster index [3] ($K = 3$).

While CDR provides information about hidden communication structure and dynamic human activities, OSM allows to extract features about physical resources (i.e., offering advantage) and access structure (i.e., street network) in a city. For offering advantage, we extracted amenity nodes for the city of Milano and grouped them by their categories. We then calculated the *offering advantage* metric for each sub-category tag based on the distribution of the city. The offering advantage metric was utilized to understand what PoIs are present in a given neighborhood, distinct from other neighborhoods [8]. This metric weighs each category by its presence, so that categories that are not very popular are more significant in the analysis of a neighborhood compared to a category that occurs frequently. These features enable a greater understanding of the distribution of resources across the city, and the extent to which areas of the city have less access to such amenities as hospital, police station, and bicycle parking, as opposed to other PoIs, like bars and fast food restaurants. Additionally, street layers were extracted for the city of Milano alongside OSM amenity data. These street layers were converted into graph networks, representing the intersections as nodes and street segments as edges. Two centrality measures, closeness and betweenness, were calculated to understand each urban area's global and local centrality, which is an indicator of that space's social, economic and spatial prosperity and accessibility. These features further contribute to our model as indicative of the spatial distribution of resources and access in the city.

In order to compare the performance of our models against benchmarks, we borrow from the methodology proposed by [6] and implement the two baselines: popularity density and spatial-lag based on past poverty level. Population density is a well-known indicator of poverty as it has been shown to negatively correlate with socio-economic level, especially in developing countries. That is, the populated areas in a city are more likely to exhibit a lower welfare. This measure is calculated from the population of a SEZ region, a smallest census block in Milano, divided by its area. For the second baseline and as an alternative to the above, we target scenarios where no current census information exists and we are limited to the older information (e.g., previous census). For this baseline we exploit the poverty information provided by ISTAT corresponding to 2001 Italian census. While not suitable for prediction due to the changes of census block boundaries as well as urban landscapes, the past poverty data could help us understand the spatial autocorrelation in a given city as poverty often contains a strong degree of spatial autocorrelation. We thus create a second baseline model based on the spatial-lag of this independent variable.

## PREDICTION MODEL

In order to test the predictability, we first examined the correlation between poverty level of Milano and each extracted feature by using linear regression. The Spearman's Rho correlation value and Mean Absolute Error (MAE) suggest that each feature by itself is very weakly correlated with the poverty level. Furthermore the baseline models also suffer from low accuracy ($rho = 0.35$ for population density, $rho = 0.15$ for spatial lag regarding 2001 poverty level). This observation indicates that prediction of poverty in such a fine grain level of granularity in a developed city is fundamentally a difficult problem even when we possess knowledge about how the poverty is spatially distributed (spatial-lag baseline) and strong determinant of the poverty (population density). We thus treat this problem as a classification problem where rather than predicting exact numerical poverty level, we aim to accurately classify which areas fall within different poverty levels. For this purpose we categorise the poverty distribution into 7 bins based on standard deviation, $\sigma$. We then use Random Forest classification with cross-fold validation. To ensure the robustness of our models, we run each model 20 iterations of random train/test splits with varying training proportion.

We used $F1$ score to measure the performance of the models. By combining the concepts of precision and recall, it provides an intuitive measure for prediction power. Figure 1 reports the $F1$ scores for models based on different features against varying proportions of train sets. As it can be observed, the performance of models are consistently better than the baselines but with varied differences depending on features and combinations. Based on these figures we can observe that

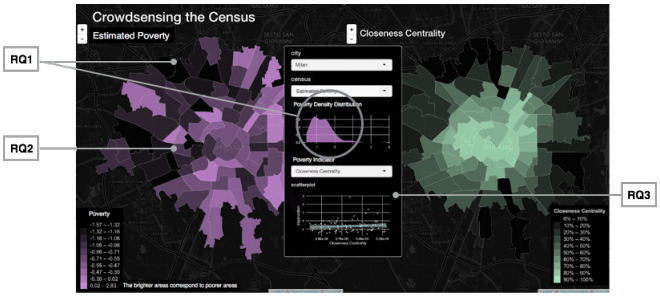


**Figure 2. A screenshot of the Poverty Mapping Tool illustrating disaggregated distribution of poverty estimated by our model (left) and the determinants contributing to the estimated poverty (right).**

while CDR and OSM features individually provide an improvement over the baseline prediction, they best perform when combined together and boosted with the knowledge of poverty autocorrelation (i.e., spatial lag).

## POVERTY MAPPING VISUALIZATION TOOL

In order to put our methodology into actual use, we collected a set of requirements through discussions with the Global Pulse, UN research lab. These requirements led us to the design and development of a poverty mapping tool that leverages our prediction model. Figure 2 (RQ1) presents a screenshot of this tool which enables the user to select a city and view the estimated poverty level presented by a simple chloropleth color ramp. Furthermore, it visualizes the density distribution of poverty across the city as a whole by a bar chart. This representation captures the requirement of simultaneously preserving the spatial distribution of welfare, and provide this information in a highly disaggregated geographical unit. Therefore it enables easy interpretation and comparison across different areas and in relation with the geographical distribution of the areas (e.g., suburban areas vs central). The tool also allows the user to view the poverty level at various spatial granularity catering for diverse information needs of different users. For example, a district commissioner would be interested in sub-district poverty and require a more abstract view of the poverty map. This is illustrated in Figure 2 (RQ2) where poverty is displayed in the sub-municipal area level, in contrast to the finest resolution, SEZ level.

Furthermore the tool also caters for the need for transparency. Indeed many policy makers in the past have viewed the poverty estimation methodology as a black box [2], and thus have often lacked trust in the poverty mapping algorithms behind the scene. Based on this observation we have created a transparent design where the determinants and indicators that contribute to the estimated poverty values are clearly communicated to the user. As illustrated in Figure 2 our tool allows the user to select a poverty determinant from the drop down menu (e.g., community services) and explore the spatial distribution of the selected determinant across the city (right map).

Finally, the last design requirement that our tool caters for is the presentation of the urban infrastructure and elements, allowing the stakeholders to interpret the potential impact of their policies in relation to the existing elements. This information could help with the placement of amenities that

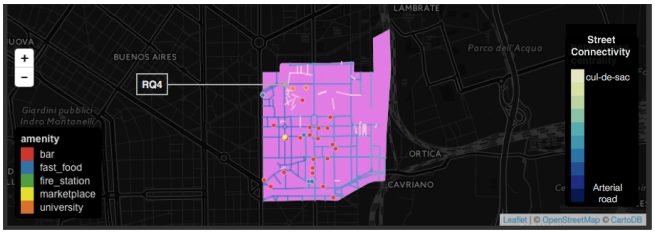


**Figure 3. A screenshot of our poverty mapping tool visualizing the urban elements and infrastructure of a given area.**

are vital but missing in a neighborhood. For example, as part of gentrification interventions one could decide where to create third places that would encourage a sense of community in isolated neighborhoods. Our tool allows the user to select an area and display the spatial distribution of existing PoIs and street connectivity for that selection. Figure 3 illustrates this feature for one of the poorer areas of Milano, where the map indicates the sheer presence of bars and fast food places, resembling the previous findings reported in [8].

## CONCLUSION

We proposed a methodology for estimating poverty levels that relies on alternative data sources than census and thus is relatively low in cost. We have demonstrated how these poverty estimates could be presented as poverty maps and offer spatially fine grained and up-to-date information that is easy to interpret. Our results indicate that our model is able to predict poverty more accurately than the known baselines.

## REFERENCES

1. Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. 2015. A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Sci Data* 2 (27 Oct. 2015), 150055.
2. Tara Bedi, Aline Coudouel, and Kenneth Simler. 2007. *More than a pretty picture: using poverty maps to design better policies and interventions*. World Bank Publications.
3. David L Davies and Donald W Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence* 2 (1979), 224–227.
4. Vanessa Frias-Martinez and Jesus Virseda. 2012. On the relationship between socio-economic factors and cell phone usage. In *Proceedings of the fifth international conference on information and communication technologies and development*. ACM, 76–84.
5. Giovanni Quattrone, Afra Mashhadi, and Licia Capra. 2014. Mind the map: the impact of culture and economic affluence on crowd-mapping behaviours. In *Proceedings of the 17th CSCW*. ACM, 934–944.
6. Chris Smith-Clarke and Licia Capra. 2016. Beyond the Baseline: Establishing the Value in Mobile Phone Based Poverty Estimates. In *Proceedings of the 25th International Conference on World Wide Web*. 425–434.
7. Christopher Smith-Clarke, Afra Mashhadi, and Licia Capra. 2014. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proceedings of CHI*. ACM, 511–520.
8. Alessandro Venerandi, Giovanni Quattrone, Licia Capra, Daniele Quercia, and Diego Saez-Trumper. 2015. Measuring Urban Deprivation from User Generated Content. In *Proceedings of the 18th ACM CSCW*. ACM, 254–264.