ESTIMATING FINE-GRAINED POPULATION GROWTH RATES FROM COARSE CENSUS DATA

John E. Vargas-Muñoz[1], *Nando Metzger*[2], *Rodrigo C. Daudt*[2], *Konrad Schindler*[2], *Devis Tuia*[1]

[1]École Polytechnique Fédérale de Lausanne, Switzerland
[2] ETH Zurich, Switzerland

ABSTRACT

Fine-grained population estimation is important for several domains, such as urban planning, public health, and humanitarian action. Due to limited resources, population maps with sufficient spatial resolution and temporal frequency are not available for many developing countries. Population estimates are often based on statistics available at the national or provincial level, which moreover are updated infrequently. The United Nations produce estimates of population growth rates, which are widely used to project population numbers to a target year, starting from the most recent census data. However, they use a simplified model of population growth with uniform growth rates for all urban, respectively rural areas within a country. This neglects the complex dynamics of population growth (e.g., growth rates in big cities are usually larger than in smaller urban areas) and leads to significant errors in population projections. In this work, we propose a methodology to estimate fine-grained population growth rates and present experimental results for Mozambique.

***Index Terms*—** population estimation, growth rate, fine-grained, Markov random fields

1. INTRODUCTION

Precise population information is critical for urban planning, disaster relief, and humanitarian actions. Population growth is usually correlated with an increase in infrastructure, built-up area, and socio-economic activities [1, 2]. Human population growth is a complex process that differs in urban and rural places, with a tendency toward higher growth rates for the former. But even within urban agglomerations, different population growth rates are common, depending on their sizes [3]; additionally, the borders between areas shift, as smaller settlements merge into urban centers. Another common process is the in-situ urbanization of rural areas [4]: as a consequence, it is expected that in the next three decades, the main population growth will occur in developing countries, where precise growth rates are often not known [2].

Although population censuses are performed regularly in developed countries, the same cannot be said for many developing countries. An additional problem is that the census data collected in developing countries is often spatially coarse (e.g., only a single population count per region or province). Recently, some works have explored how fine-grained population maps can be derived from coarse census data with the help of machine learning models [5, 6, 7]. They use multimodal, fine-grained geographical data as inputs to spatially distribute the coarse population counts. Examples of inputs that are predictive of population density are, among others, building counts (extracted from remote sensing images), nightlight images, and the distance to infrastructures such as roads [6]. In particular, buildings were – unsurprisingly – found to be highly predictive for population counts.

While the above methods are highly effective in disaggregating coarse census data, they remain dependent on the availability of such census surveys, thus creating the need for estimating population growth to be extrapolated beyond census dates. The United Nations regularly publish population growth rates of urban and rural areas in several countries, which are used to extrapolate the most recent census data to a desired target year.

A common limitation of population estimation methods is that they are trained with UN projections as the regression target. Those projections assume a uniform population growth rate in urban and rural areas, which is an oversimplification of the reality: almost everywhere, population dynamics are more complex, and even two different urban areas of comparable size, within the same country, may experience very different growth rates. This variability can be seen in Mozambique: Figure 1 shows the population growth rates in different provinces of Mozambique, computed from the last censuses of 2007 and 2017.

In this paper, we propose a method to predict population growth rates from maps of built-up areas and historical census data. We use a computational scheme based on the Markov Random Field model to predict the growth rates, using built-up data from two different points in time, namely (1) the last census year and (2) the target year. We defined a number of built-up agglomerations and we enforce consistent growth rates for agglomerations with similar characteristics in terms of area, growth, and nightlights intensity. Our method is evaluated for the country of Mozambique, a developing country for which census information is available at

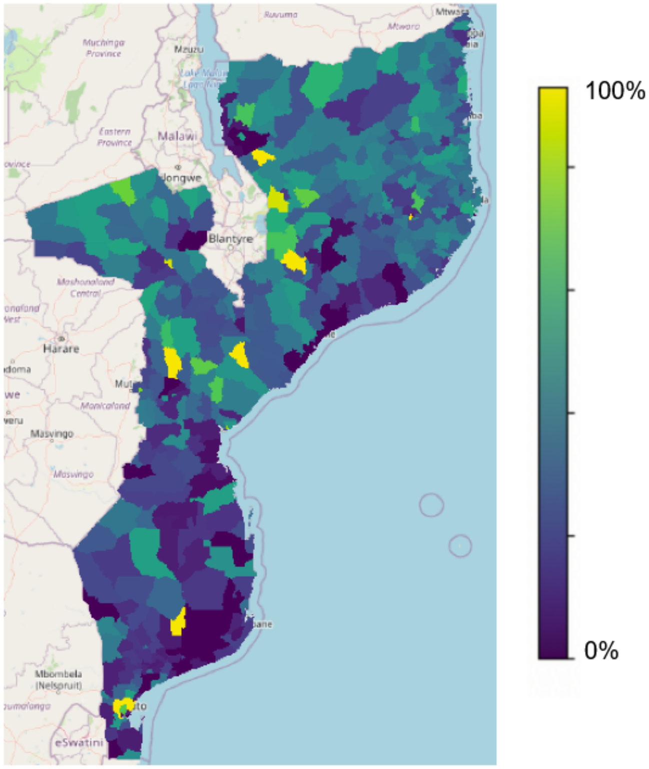

Fig. 1. Growth rate in provinces of Mozambique between the census of 2007 and 2017

two different points in time. The experiments show that our predicted growth rates depict the true population growth more accurately than those used for the UN projections and lead to better regional population forecasts.

2. METHODOLOGY

The objective of our proposed method is to compute a population growth rate specific to a series of locally defined agglomerations. In the following, we detail how these agglomerations were defined, and the computational approach to estimate their growth rates using a Markov Random Field model based on agglomeration similarity and regional consistency.

2.1. Defining urban agglomerations

To obtain a fine-grained estimation of population growth rates, we define a series of urban agglomerations based on built-up areas. These area can be located anywhere in the country, both in urban or rural areas, and are defined by clustering pixels related to built-up areas. To do so, we clustered the spatial coordinates of the pixels containing built-up areas from a rasterized urban settlements layer. We used DBSCAN to perform the clustering, leading to N_A different agglomerations. The building data was obtained from the open repository World Settlement Footprints [8, 9]. The remaining (non-building) pixels were assigned to the spatially closest agglomeration.

2.2. Estimating growth rates

Once the N_A agglomerations have been defined, we want to learn a population growth rate, α_i, for each agglomeration $i \in [1, \ldots, N_A]$. The aim of our model is then to estimate local α_i's, one per agglomeration. We start from the assumption that the growth rate of the population is highly correlated with the increase in built-up area between the two time steps $t-1$ and t. Indeed, creating built-up area layers is a good proxy task that can be approached successfully with remote sensing [10]. We use this intuition to create a Markov Random Field (MRF)-based method [11] that will estimate growth rates starting from:

- the built-up area map at the time of the last census b_k^{t-1};
- the built-up area map for the target year b_k^t;
- the known population counts of the last census p_k^{t-1};
- an estimate of similarity between the different agglomerations, obtained by clustering the agglomerations in a set of Q prototypes, obtained by K-Means. To cluster every agglomeration i, we used as variables the built-up total area, the built-up area growth, and the mean nightlight intensity. This led to the cluster assignment q_i.

We assume the following relation between the populations at the two time steps:

$$p_i^t = p_i^{t-1} \times \left(\frac{b_i^t}{b_i^{t-1}}\right)^{\alpha_i}, \tag{1}$$

where the α_i parameter rules the growth speed for agglomeration i:

$$\alpha_i = \frac{log\left(\frac{p_i^t}{p_i^{t-1}}\right)}{log\left(\frac{b_i^t}{b_i^{t-1}}\right)}. \tag{2}$$

Using the four input variables above and relations in Eqs. (1) and (2), the MRF model can be employed to minimize the following energy function U, aiming at finding the most accurate population estimation for each agglomeration i at time t, $\hat{\mathbf{p}}^{\mathbf{t}}$:

$$\begin{aligned}\hat{\mathbf{p}}^{\mathbf{t}} &= \arg\min_{\mathbf{p}^{\mathbf{t}}} \sum_{i=1}^{N_A} U(p_i^t | p_i^{t-1}, b_i^{t-1}, b_i^t, q_i) \qquad (3)\\ &= \arg\min_{\mathbf{p}^{\mathbf{t}}} \underbrace{\sum_{i=1}^{N_A}\sum_{k=1}^{N_A} \delta_{ik}^S |\alpha_i - \alpha_k|}_{\text{Similarity-based growth}} + \underbrace{\lambda \sum_{j=1}^{N_P} |c_j^t - \sum_{i=1}^{N_A} \delta_{ij}^G p_i^t|}_{\text{Geographical consistency}},\end{aligned}$$

where c_j^t is the population projection of the UN for administrative region j. In the equation, the following two energy terms balance two effects:

- *Similarity-based growth*: this term favors consistency between agglomerations with similar characteristics. The term

$$\delta_{ik}^S = \begin{cases} 1, & \text{if } q_i = q_k \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

  ensures that agglomerations assigned to the same cluster prototype q receive similar growth rates.

- *Geographical consistency*: this term favors that the total predicted population in each administrative region (province) $j \in [1, \dots, N_P]$ corresponds to the regional statistics. We do so by comparing c_j^t (the projected population given by the UN for province j) to p_i^t (the sum of the populations estimated for all the agglomerations within the province) for each agglomeration i in province j. The term

$$\delta_{ij}^G = \begin{cases} 1, & \text{if } i \in P_j \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

  ensures that only agglomerations geographically located in province P_j are accounted for.

3. EXPERIMENTS

3.1. Dataset and experimental setup

To validate our proposed method we use official census data of $t-1 = 2007$ and $t = 2017$ from Mozambique. For validation purposes, we assume that the census data of 2017 is not known during the training of the MRF. As stated above, the built-up area maps were extracted from the World Settlement Footprints [8, 9]. We set the value of $\lambda = 1$ for the experiments. We use as administrative regions the $N_P = 440$ districts of Mozambique.

3.2. Results

We summarize our results in Table 1. After training the MRF, we obtained the estimated populations for 2017 and compared it the UN projections for 2017, using the metrics R2, MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentual Error). Additionally, we also report of using our method constraining the projections by the total population of the country (which basically corresponds to turning off the second term of the energy function, the *Geographical consistency*).

It can be observed that the MRF-based methods achieves smaller error compared to the UN projections, and also that the MRF constrained by district or the whole country obtain similar results. Figure 2 shows a map of percentual errors of the population estimated for several districts of the capital city of Maputo. We can observe that in some districts the error of the UN projections is very high, and the MRF can considerably reduce the error in those regions.

Table 1. Performance of methods for population estimation. 'MRF - country' correspondos to our proposed method without the geographical consistency potential. All methods are evaluated at the province level using the census data of $t = 2017$ as ground reference.

Method	R2 ↑	MAE ↓	MAPE ↓
UN extrapolation	0.901	$9'806$	15.45
MRF - district (**ours**)	**0.939**	$\mathbf{8'646}$	**14.29**
MRF - country	0.932	$8'941$	14.55

These results could potentially be improved by using a built-up areas layer extracted from images temporally closer to the census date. However, one needs to keep in mind the considerable effort to create that layer both in terms of data involved and computing power required [8]. Alternatively, one could also use projections in existing built-up area data, such as those contained in the Global Human Settlement Layers [12].

4. CONCLUSION

Recently proposed methods for fine-grained population estimation in developing countries often depend on the presence of census surveys, which are updated irregularly at best. In between census dates, one needs to use coarse growth statistics to estimate populations, but the official growth statistics (e.g. from FAO) are often inaccurate. In this paper, we have presented a method to estimate fine-grained population growth rates that can then be used to update population estimates for years without census data. Our method depends mainly on historic census data and built-up area maps that can be obtained regularly from free satellite imagery. We propose to estimate the growth rates in fine-grained units, which we call agglomerations, obtained by performing spatial clustering. The MRF energy function that we used tries to obtain similar growth rates for agglomerations that are similar, while constraining the results to respect the original population projections for a particular administrative region. In the results we showed that our proposed method can attain results with lower estimation errors than the UN population projections.

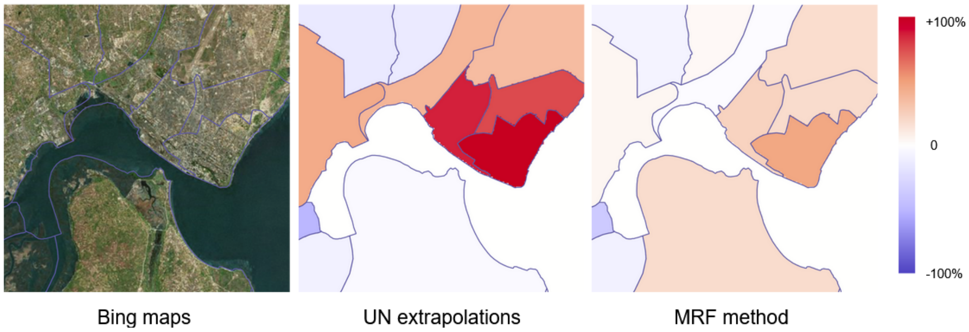

Fig. 2. Map of population estimation errors for the Metropolitan area of the capital city of Maputo, Mozambique.

5. REFERENCES

[1] Marcello Schiavina, Michele Melchiorri, Christina Corbane, Sergio Freire, and F Batista e Silva, "Built-up areas are expanding faster than population growth: regional patterns and trajectories in europe," *Journal of Land Use Science*, vol. 17, no. 1, pp. 591–608, 2022.

[2] Richa Mahtta, Michail Fragkias, Burak Güneralp, Anjali Mahendra, Meredith Reba, Elizabeth A Wentz, and Karen C Seto, "Urban land expansion: The role of population and economic growth for 300+ cities," *Npj Urban Sustainability*, vol. 2, no. 1, pp. 5, 2022.

[3] Philipp Heinrigs, "Africapolis: understanding the dynamics of urbanization in africa," *Field Actions Science Reports. The journal of field actions*, , no. Special Issue 22, pp. 18–23, 2020.

[4] Amadu Jacky Kaba, "Explaining africa's rapid population growth, 1950 to 2020: Trends, factors, implications, and recommendations," *Sociology Mind*, vol. 10, no. 4, pp. 226–268, 2020.

[5] Forrest R Stevens, Andrea E Gaughan, Catherine Linard, and Andrew J Tatem, "Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data," *PloS one*, vol. 10, no. 2, pp. e0107042, 2015.

[6] Nando Metzger, John E Vargas-Muñoz, Rodrigo C Daudt, Benjamin Kellenberger, Thao Ton-That Whelan, Ferda Ofli, Muhammad Imran, Konrad Schindler, and Devis Tuia, "Fine-grained population mapping from coarse census counts and open geodata," *Scientific Reports*, vol. 12, no. 1, pp. 20085, 2022.

[7] Nando Metzger, Rodrigo Caye Daudt, Devis Tuia, and Konrad Schindler, "High-resolution population maps derived from sentinel-1 and sentinel-2," *arXiv preprint arXiv:2311.14006*, 2023.

[8] Mattia Marconcini, Annekatrin Metz-Marconcini, Soner Üreyen, Daniela Palacios-Lopez, Wiebke Hanke, Felix Bachofer, Julian Zeidler, Thomas Esch, Noel Gorelick, Ashwin Kakarla, et al., "Outlining where humans live, the world settlement footprint 2015," *Scientific Data*, vol. 7, no. 1, pp. 242, 2020.

[9] Mattia Marconcini, Annekatrin Metz-Marconcini, Thomas Esch, and Noel Gorelick, "Understanding current trends in global urbanisation-the world settlement footprint suite," *GI_Forum*, vol. 9, no. 1, pp. 33–38, 2021.

[10] Sebastian Hafner, Yifang Ban, and Andrea Nascetti, "Unsupervised domain adaptation for global urban extraction using Sentinel-1 SAR and Sentinel-2 MSI data," *Remote Sensing of Environment*, vol. 280, pp. 113192, 2022.

[11] Julian Besag, "On the statistical analysis of dirty pictures," *J. Royal Stat. Soc.*, vol. 68, pp. 259–302, 1986.

[12] Martino Pesaresi, Daniele Ehrlich, Aneta J Florczyk, Sergio Freire, Andreea Julea, Thomas Kemper, and Vassilis Syrris, "The global human settlement layer from landsat imagery," in *2016 IEEE international geoscience and remote sensing symposium (IGARSS)*. IEEE, 2016, pp. 7276–7279.