Regional Population Estimation Using Satellite Imagery

Aikaterini Koutsouri[1], Ilektra Skepetari[2], Konstantinos Anastasakis[3], and Stefanos Lappas[4]

[1] National Technical University of Athens, Iroon Politechniou 9, Athens, katerina@fsu.gr
[2] National Technical University of Athens, Iroon Politechniou 9, Athens, electra@fsu.gr
[3] National and Kapodistrian University of Athens, Panepistimiou 30, Athens, kon.anastasakis@gmail.com
[4] National and Kapodistrian University of Athens, Panepistimiou 30, Athens, steflappas@gmail.com

Abstract. Population data provides statistical information that can in turn support decision making processes. It provides essential information regarding many practices such as rescue operations or humanitarian actions, which require the estimation of local population. While traditional approaches are possible, they tend to be time consuming and expensive. Copernicus Earth Observation data (Sentinel 2 satellite images) provides high resolution satellite imagery that could be a useful alternative to collecting census data since it is significantly cheaper than traditional methods. The purpose of the following study examines a population extraction method that is based on open source high resolution images retrieved from Sentinel 2, as proposed by the remote task[1] by the imageCLEF2017[2] campaign. This paper describes a methodology for exporting a population estimation using satellite imagery based on the use of classification techniques coupled with a statistical forecast on historical data.

Keywords: Satellite Image Analysis, Supervised Classification, Population Estimation Mining

1 Introduction

The work described in this paper is directed at the automated estimation of population using Supervised Classification through a GIS platform applied to satellite imagery. The motivation for the work is that the collection of population data, when conducted in the traditional manner can be time consuming and inefficient. This is especially the case in rural areas that lack of sophisticated communication and transport infrastructure, especially in developing countries. The solution proposed in this paper is based on the concept of using satellite imagery in order to estimate the population. The concept includes using a small sample of satellite images to build a classifier that can then be used to predict

populated areas. The main issue to be addressed is how to best use the classified result towards matching it to the number of people in the study area. In order to correspond to the task described above, the following methodology is proposing (i) an image supervised classification technique , (ii) a statistical forecasting procedure and (iii) a regional disaggregation technique which will be furtherer explained in the following sections.

2 Gathering and Processing Sentinel Data

With this chapter, the methodology of downloading and analyzing Sentinel2 data shall be described. The Semi-Automatic Classification Plug-in (SCP)[3] is a free open source plug-in for QGIS that is used for semi-automatic classification or supervised classification of remote sensing images as well as tools for image preprocessing, the classification post processing, and the raster calculation and is used for obtaining the imagery needed for the purposes of this study.

2.1 Gathering Raster Files

In order to gather the Satellite imagery, The Copernicus Open Access Hub can be used, which provides complete, free and open access to Sentinel-1, Sentinel-2 and Sentinel-3 user products. The SCP Sentinel-2 tab allows the searching and downloading of the desired imagery (Level-1C) using the Data Hub API. The search area is defined accordingly for each one of the areas of interest to include all of the respective subregions. For the purposes of this study, the date of acquisition was defined as of 2016 or later in order to make the classification process relevant to the desired results. The maximum cloud cover shall be set not to exceed 10 % to help with the process whereas bands 1, 9, and 10 shall not be included.

2.2 Supervised Classification

In this section, the steps followed when classifying the satellite images will be explained. For the purposes of this study, several compounds of Uganda and Lusaka are examined. The two regions are classified separately. The same approach is followed for both of the study regions. It is noted that multiple different satellite images might be necessary to cover the entire study area. For the Sentinel-2 images, the bands are converted to reflectance and clipped to an extend large enough to cover each area examined so as to reduce the computational time of the process. The band set which is the input image for SCP is defined and the regions of interest are created.

Iterating through the different color composites, allows completion of the classification. The defined buildings macro-class is the area of interest that will be extracted. This shall provide us with high resolution two colored pixelated images that hold information regarding the buildings of the areas. Such data shall be used for completing the methodology through the process that will be furtherer explained in the next sections.

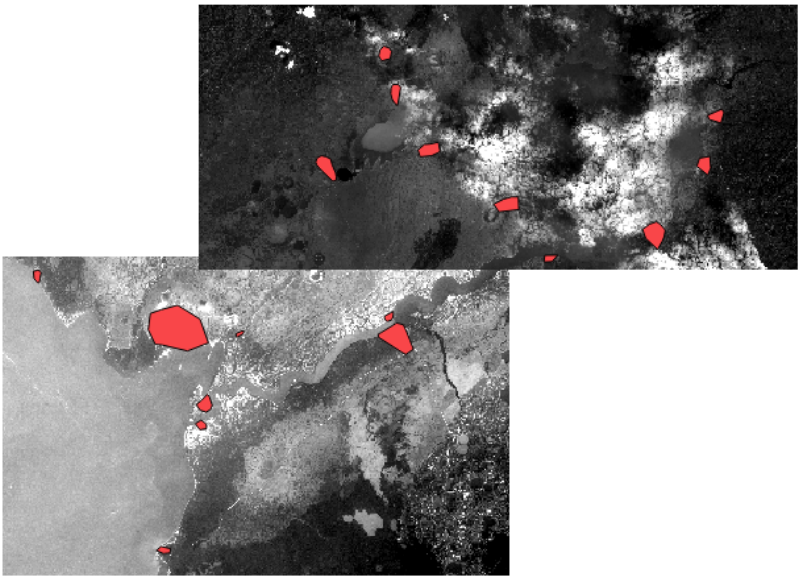

Fig. 1. Clipped Bands of the examined areas.

2.3 Supervised Classification Output

The classification results can be validated through comparison with the original raster files as well as old shape-file layers provided by various open data sources (e.g. OpenLayers Plug-in, Geofabrik).

Following the completion of the process described above, the two classified regions are separated using polygons that describe the sub-areas of interest. For the purposes of this study, the PNG images of the clipped images (17 for Uganda, 89 for Lusaka) are extracted, each one now containing a certain number of blank pixels corresponding to the living areas of each district. Concerning their usage in the estimation algorithm that will be described, it is important to note that all extracted classified files of each study area shall correctly correspond to the respective area's land coverage ratio.

Fig. 2. Classification output displaying the region of Katwe-Uganda.

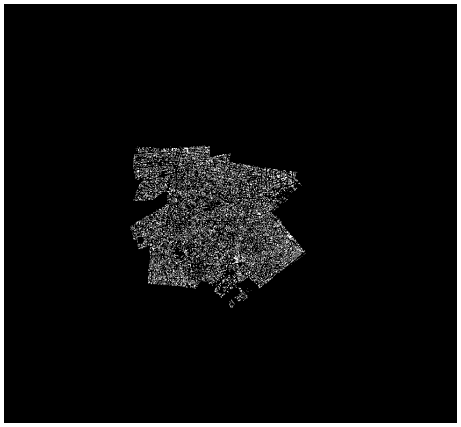

Fig. 3. Classification output displaying the region of Old Kanyama-Lusaka.

3 Statistical Forecast

While the results of the classification process can provide a decent overview of the population distribution, the buildings information shall be used in a way that can allow to extract a population estimation. The historical Census data of the countries can be used to predict the aggregated population result that can be later split into the study areas accordingly through the classified imagery. The statistical analysis of the Census data consists of two basic steps, (i) constructing the population growth time-series and (ii) a forecasting model in order to pertain to the out of sample observation; the current year population number.

3.1 Gathering and Processing Census Data

Various Census data for demographic purposes are used for both areas [12] obtained from The World Bank organization. Concerning the areas of interest of the areas of interest of the present study, the latest population data that was available for both regions of Uganda and Zambia was that of 2015.

3.2 Forecast

Following the process of collecting the population data time series, we can proceed to choosing the best fitting model to forecast the 2017 population of the two countries based on previously observed values. Regarding choosing the correct forecasting technique we shall examine several error metrics (ME, MPE, MAPE, MSE, sMAPE). The trend of our time-series shall also not be overlooked and, taking into consideration the growth ratio of the examined African regions, slightly optimistic methods are suitable for the purposes of this study. The time-series dataset that was obtained by the analysis procedure described

in the previous section can be used as input in the forecasting support system OMEN[10], a fully customizable web-based forecasting tool.

A cross-validation competition between the methods Naive, MAPA[9], ETS[8], Theta[5], ARIMA[7], SES, Holt and Damped [6] is performed through the platform. The rolling origin evaluation procedure [11] is used with the validation window matching the forecasting horizon and the performance of each method is evaluated based on the minimization of the squared errors. Thus, the 'best fitted' model is selected, which in this case was the autoregressive integrated moving average (ARIMA) model for both Uganda and Zambia; suitable for low frequency and short horizons.

Table 1. Uganda Insample - Forecast Error Metrics

	ME	MPE	MAPE	MSE	sMAPE
Theta	362,352.64	1.04 %	1.85 %	394,477,077,721	1.85 %
Naive	928,936.47	2.99 %	2.99 %	937,103,090,150	3.03 %
ETS	-4,143.2	-0.01 %	0.31 %	20,334,332,759	0.31 %
ARIMA	-29,783.84	-0.08 %	0.09 %	14,403,091,721	0.09 %
SES	701,607.45	2.01 %	3.96 %	1,816,289,360,487	3.93 %
Holt	-4,140.01	-0.01 %	0.31 %	20,334,001,686	0.31 %
Damped	1,361.34	0.01 %	0.3 %	20,354,628,986	0.3 %
MAPA	-4,683,085.6	-17.93 %	19.79 %	42,915,827,697,557	17.12 %

Table 2. Zambia Insample - Forecast Model Error Metrics

	ME	MPE	MAPE	MSE	sMAPE
Theta	132245.72	0.91 %	1.5 %	50,440,089,294	1.5 %
Naive	335,253.88	2.53 %	2.53 %	125,691,193,414	2.56%
ETS	-9,335.3	-0.06 %	0.3 %	6,423,231,745	0.3 %
ARIMA	-6,156.78	-0.02 %	0.21 %	5,608,027,543	0.21 %
SES	256,146.93	1.77 %	3.28 %	232,161,298,596	3.27 %
Holt	-9,334.91	-0.06 %	0.3 %	6,423,208,336	0.3 %
Damped	-5,292.38	-0.03 %	0.3 %	6,161,968,390	0.3 %
MAPA	-1,742,891.55	-15.24 %	16.82 %	5,825,850,536,317	14.89 %

The results extracted are 40,989,197 and 16,932,235 for the two regions respectively.

Table 3. Statistical Forecasts for the population of Zambia and Uganda

Year	Uganda	Zambia
2001	24,146,152	10,723,229
2002	24,945,231	11,000,608
2003	25,786,777	11,282,992
2004	26,666,251	11,575,821
2005	27,578,578	11,884,613
2006	28,521,669	12,212,550
2007	29,496,442	12,560,093
2008	30,503,193	12,926,628
2009	31,540,776	13,311,214
2010	32,608,271	13,712,644
2011	33,704,880	14,130,483
2012	34,830,481	14,565,054
2013	35,987,004	15,016,334
2014	37,178,179	15,483,715
2015	38,407,677	15,966,555
2016	39,675,318	16,418,477
2017	40,983,129	16,885,858

3.3 Regional Disaggregation and Adjustments

The processes described above provide with an arithmetic output regarding Uganda's and Zambia's total population. Such information is used to proceed to the following step which is a top-down division of the countries' population in order to come to a result about each specific area of interest in the given shapefiles. It is however important to note that the subareas of interest might not add up to the entire countries' extend and therefore a straightforward division of the entire population to the acreage of the living areas (as obtained from the classification process) might not be possible. Because of this difficulty, a different approach shall be examined. In order to split the total population into the different areas, a weight variable will be calculated for each pixel corresponding to living areas in the classified images. To achieve this, we gather Census data about housing and population in a small subregion of each one of the countries of interest and estimated the multiplier for each pixel.

Uganda For the purposes of this study, we shall first examine the areas of Uganda. Regarding the region of Katwe Lake, according to Census data obtained

from the Uganda Bureau of Statistics[13], the region had a total population of about 23,559 people in 2014. The quotient of the compound's population and Uganda's total population is estimated to be equal to ~ 0.00063368 in 2014. Assuming that the population density of the region has not changed significantly, it is estimated that the compound's population for the year of 2017 shall equal to $\sim 25,970$.

Table 4. Population Distribution for the region of Lake Katwe

Year	Uganda	Lake Katwe Compound
2014	37,178,179	23,559
2017	40,983,129	25,970

We proceed to examine the classification result extracted by the methodology that was described in the previous section.

Fig. 4. Classification output displaying the region of Katwe-Uganda.

We shall proceed to sum the pixels of the living areas (rgb(255,255,255)) in the region; The above process gives us a result of 2,242 total pixels that represent living areas in the compound of Lake Katwe. Thus, we calculate that the density multiplier of each pixel that represents living area in the compounds of Lusaka, shall equal to ~ 11.5834.

Lusaka The same process is followed for the regions of Lusaka, where the district of Kanyama is examined. According to Census data from 2010[14], the Kanyama compound had a total population of about 366,170 people in 2010 leading us to the conclusion that the quotient of the compound's population and Zambia's total population equaled to ~ 0.026703 in that year. Assuming that the population density of the region has not changed significantly, the compound's population for the year of 2017 is estimated to be equal to $\sim 388,002$.

Table 5. Population Distribution for the region of Old Kanyama

Year	Zambia	Kanyama Compound
2010	13,712,644	366,170
2017	16,885,858	450,903

We then proceed to examine the classification results for Lusaka; it is noted that the Kanyama compound consists of three polygon shapes: (i) Old Kanyama (ZMB_TW0029), (ii) New Kanyama (ZMB_TW0050) and (iii) Kanyama West (ZMB_TW0078). The aggregated result for those three subregions can be examined.

Fig. 5. Classification output displaying the region of Old Kanyama-Zambia.

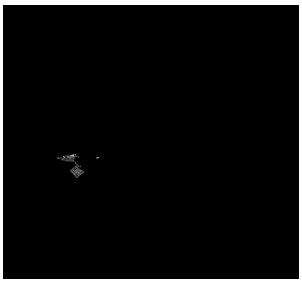

Fig. 6. Classification output displaying the region of New Kanyama-Zambia.

Fig. 7. Classification output displaying the region of Kanyama West-Zambia.

We shall proceed to sum the pixels of the living areas (rgb(255,255,255)) in the three districts using the same methodology. It is estimated that a result of total 64,199 pixels represents living areas in the compound of Kanyama. Therefore we calculate that the density multiplier of each pixel that represents living area in our Lusaka study area shall equal to ~ 7.02352.

3.4 Final Estimations

After calculating the weight variables for both regions as described above, the clipped pixelated images along with the multipliers can be used as input to calculate the population of each region. For each compound, each white pixel is multiplied by the estimated weight variable of the respective study area (11.5834 for Uganda - 7.02352 for Lusaka). The outputs provide with the final dataset containing the population estimations for all the subregions of each country.

Table 6. Estimation Output

Study Area	Estimated Population
Kisenyi	616
Kazinga	654
Kutunguru HCII	1148
Kutunguru HCIII	1592
⋮	⋮
Mtendere	102140
Helen Kaunda	6705
Kabulonga	5483
Ibex Hill	4858
⋮	⋮

It is important to note that the accuracy of the procedure described, depends largely on the classification output. A well classified image similar to the one obtained from the regions of Uganda during this study, can improve the results whereas a poorly classified region similar to the one obtained from Lusaka shall result to a dataset that differs greatly from its actual population.

Table 7. Error Metrics

Uganda	Lusaka
Delta: 10,160	Delta: 1,476,753
RMSE: 770.60	RMSE: 38,072.60
Pearson: 0.95	Pearson: 0.25

4 Notes and Comments.

The results obtained from the described methodology shall provide us with the final dataset containing information on the population data for each one of the study areas. However, the same procedure can be followed for estimating the dwellings of the regions of interest by appropriately parameterizing the census data input along with the forecasting model described in the previous sections.

References

1. Helbert Arenas, Bayzidul Islam, and Josiane Mothe. Overview of the ImageCLEF 2017 Population Estimation Task. In *CLEF 2017 Labs Working Notes*, CEUR Workshop Proceedings, Dublin, Ireland, September 11-14 2017. CEUR-WS.org <http://ceur-ws.org>.
2. Bogdan Ionescu, Henning Müller, Mauricio Villegas, Helbert Arenas, Giulia Boato, Duc-Tien Dang-Nguyen, Yashin Dicente Cid, Carsten Eickhoff, Alba Garcia Seco de Herrera, Cathal Gurrin, Bayzidul Islam, Vassili Kovalev, Vitali Liauchuk, Josiane Mothe, Luca Piras, Michael Riegler, and Immanuel Schwall. Overview of ImageCLEF 2017: Information extraction from images. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction 8th International Conference of the CLEF Association, CLEF 2017*, volume 10456 of *Lecture Notes in Computer Science*, Dublin, Ireland, September 11-14 2017. Springer.
3. Congedo, L., 2014: Semi-Automatic Classification Plugin User Manual
4. Rajendran, P., Mani, K., 2015: Quantifying the Dynamics of Landscape Patterns in Thiruvananthapuram Corporation Using Open Source GIS Tools
5. Assimakopoulos, V., Nikolopoulos, K., 2000: The theta model: a decomposition approach to forecasting. International Journal of Forecasting 16 (4), 521 - 530.
6. Gardner, E. S., 2006: Exponential smoothing: The state of the art part II. International Journal of Forecasting 22 (4), 637 - 666
7. Hyndman, R., Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for r. Journal of Statistical Software 26 (3), 1-22.
8. Hyndman, R. J., Koehler, A. B., Snyder, R. D., Grose, S., 2002: A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting 18 (3), 439 - 454
9. Kourentzes, N., Petropoulos, F., 2016: Multiple Aggregation Prediction Algorithm, Version: 2.0.1 https://CRAN.R-project.org/package=MAPA
10. Skepetari, I., Spiliotis, E., Raptis, A.,Assimakopoulos, V., 2016: OMEN: Promoting Forecasting Support Systems. 37th International Symposium on Forecasting
11. Tashman, L. J., 2000: Out-of-sample tests of forecasting accuracy: an analysis and review. InternationalJournal of Forecasting 16 (4), 437 - 450.
12. The World Bank Group, Global Development Data
13. Uganda Bureau of Statistics (UBOS), 2014: NPHC 2014 FINAL RESULTS REPORT
14. Republic of Zambia, Central Statistical Office: 2010 CENSUS OF POPULATION AND HOUSING