MAPPING FINE-SCALE SPATIAL DISTRIBUTION OF POPULATION USING REMOTE SENSING DATA AND POIS WITH DEEP LEARNING

Tran Thanh Dan (1,2), Manzul Kumar Hazarika (1), Hiroyuki Miyazaki (2)

[1] Asian Institute of Technology, P.O. Box 4, Klong Luang, Pathum Thani 12120, Thailand.
[2] Center for Spatial Information Science, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba, 277-8568, Japan
Email: ttdan@ait.asia

KEYWORDS: Population mapping, randomForest algorithm, Open multi-source geospatial data

ABSTRACT: Spatial distribution of population map at a finer scale is useful for planning and policy development. Therefore, this research is one of those attempts to improving high resolution on human population distributions, by presenting a new approach to map the population using open multi-source geospatial and ancillary data. This research is conducted through two main steps: (1) to disaggregate census data and predict population density at commune level, gridded population dataset, (2) To map population distribution at building level using population-building gravity model. The data were processed through five main steps: (i) data collection and pre-processing including: population and building footprints extraction from census data and cadastral map and/or satellite data, respectively; and ancillary data collection, including: topographic, infrastructure, river network, road network, satellite data, and night-time light imagery; (ii) covariates preparation for fitting and predicting randomForest model; (iii) model adjustment and estimation population at building level; (iv) geospatial population distribution mapping at 30m spatial resolution; and (v) map population distribution at building level. Validation of results was made by comparing the estimation with the census population, which showed a good correlation with R^2 larger than 0.9. This study was successfully applying a new population mapping approach by using open multi-source geospatial data. An advantage with the approach is that we can aggregated population can be re-distributed to a fine scale. The produce fine-scale population map can offer a more thorough understanding of inner-city population, which can thus help makers optimize the allocation of resources.

1. INTRODUCTION

Fine-scale population distribution data, especially at the building level, play an important role in many fields, for example migrant population monitoring, resource allocation optimization and the analysis of city structures (Wu and Murray 2005, Lu et al. 2006, Bhaduri et al. 2007, Gaughan et al. 2013, Langford 2013, Bakillah et al. 2014, Deville et al. 2014). Moreover, population map itself shows the location and pattern of the settlement of the population, and socioeconomic characteristics. Indeed, representation of population in spatial units different from the census data

may be essential for a better performance of various spatial applications (Ural et al., 2011). Some of these applications include criminal investigation, public health, natural hazards risk, environmental risk and accessibility analysis, facilities and retail planning, land use planning, resource allocation, emergency planning, and spatial interaction modeling (Chen, 2002; Langford, 2006; Mennis, 2009).

The strong correlation between remote sensing observations and large-scaled population distributions has been revealed with the rapid development of remote sensing and geographical information system (GIS) technology (Zha et al. 2003, Lu et al. 2006). Currently, the most popular approaches of extracting fine-scale population distributions still use remote sensing products, such as impervious surfaces and night-time light data (Azar et al. 2010, Ural et al. 2011, Gaughan et al. 2013, Stevens et al. 2015, Yao et al. 2016). Azar et al. (2010) built a linear model between impervious surfaces and the population distribution and then obtained a refined population distribution map by extracting impervious surfaces in Haiti from Landsat images. Gaughan et al. (2013) proposed a spatial weighting logistical regression model that was based on Landsat derived settlement maps and land cover data to map the population distribution in Southeast Asia at a spatial resolution of 100 m. Stevens et al. (2015) then used a random forest algorithm (RFA) to establish a nonparametric predictive model that could downscale census data and map fine-scale population distributions in Kenya, Vietnam and Cambodia. A few studies focused on mapping population distributions at the scale of buildings by using residential building footprints and census data to build empirical weighting models (Lwin and Murayama 2009, Ural et al. 2011).

Population mapping at building level to grid cells with high spatial resolution is needed for socioeconomic management. Bakillah (2014) used volunteered geographic information (VGI) to map the population distribution at the building level in Hamburg and produced a satisfactory mapping result. However, Bakillah's method only relied on POIs and fine land use/land cover data (LULC) and did not consider the spatial heterogeneity of the population distribution when computing the population inside buildings. However, none of the above studies could allocate population distributions at the building level by using multisource geospatial big data because of a lack of an effective model. We suggest that a model that can effectively fuse information from multisource geospatial data, including official survey data and big data, can better reveal the actual population distribution at a fine scale.

2. METHODOLOGY

2.1 Study area

Hue City, the capital of Thua Thien Hue province in central Vietnam, is known as an international tourism destination with significant historical assets, such as the Citadel and the Imperial City. The city is located at longitude of 107° 31' 29.29"– 107° 37' 49.25" east and latitude of 16° 30' 27.8"– 16° 23' 54.32" north (**Figure 1.**).

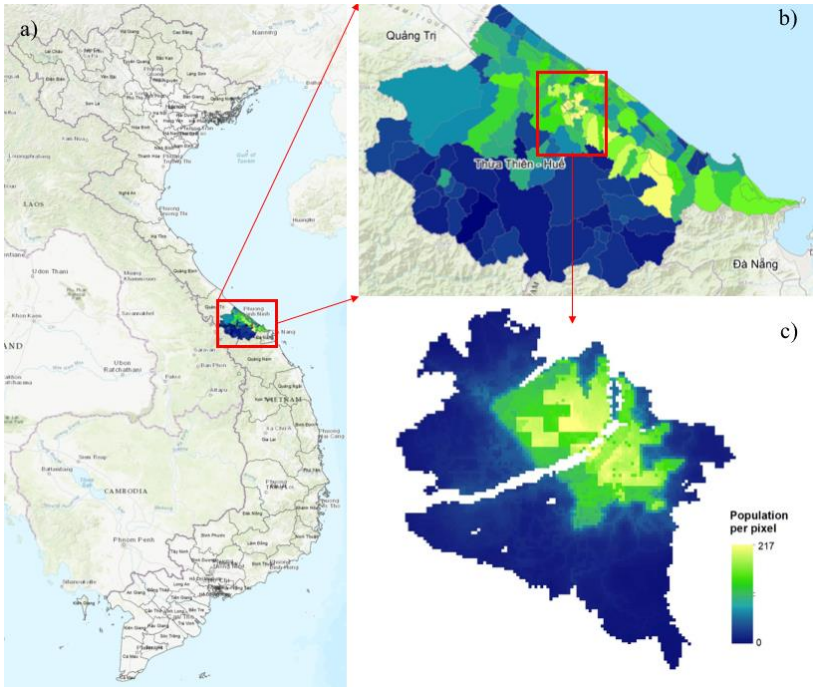


Figure 1. The study area, a) Vietnam administrative map, at provincial level; b) Thua Thien Hue population map which extracted from 2009 census data; and c) spatial distribution of population in the study area (WorldPop, http://www.worldpop.org.uk).

The population in 2009 was 335,575 people, accounting for 31.01% of the provincial population. Population density in the city was 4,778.9 people/km^2, 21.8 times higher than the average level of the whole province. Phuoc Vinh ward had the highest density at 20,705 people/km^2 and Huong Long ward had the lowest density at 1,411 people/km^2. The urban population increased by about 100,000 people in the 10 years period from 2001 to 2011, while the rural population decreased from about 60,000 people in 2001 to about 32,000 people in 2008 (M-BRACE, 2014).

2.2 Method

To predict and map population at building level, this research was adopted by two main steps, including population distribution mapping, gird-based 30x30m, and population distribution at building level, using iterative gravity model.

Population in 2009 was extracted from census data and linked with administrative map at ward level (**Figure 3.**). And other datasets which are often highly presented correlated with spatial distribution of population such as: land use land cover (LULC), night time light (high intensity of light means high of population density), geospatial data such as: road networks, rivers/stream systems, infrastructure (water supply, electricity system, …), topography are collected.

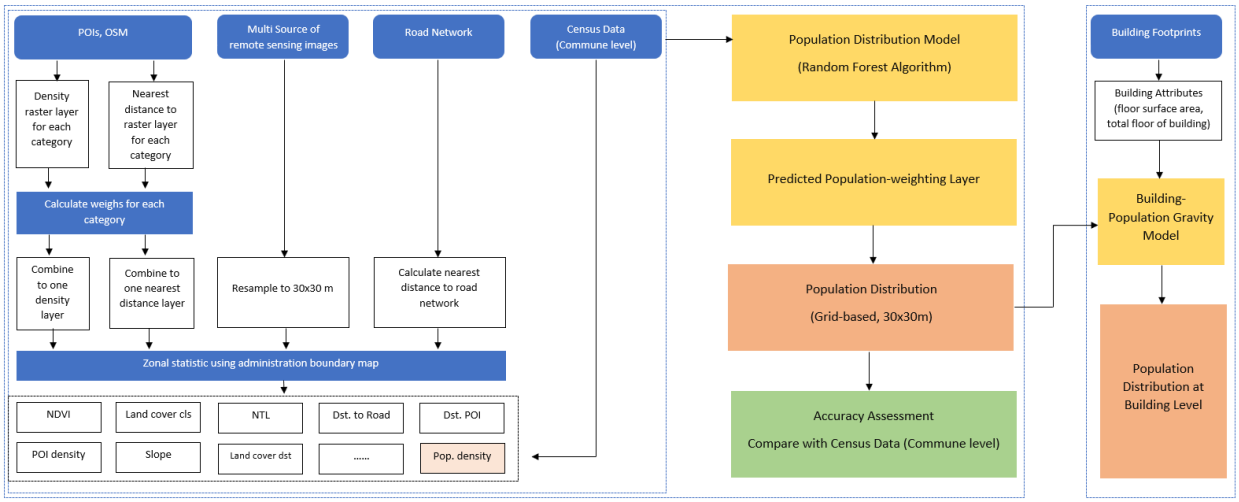
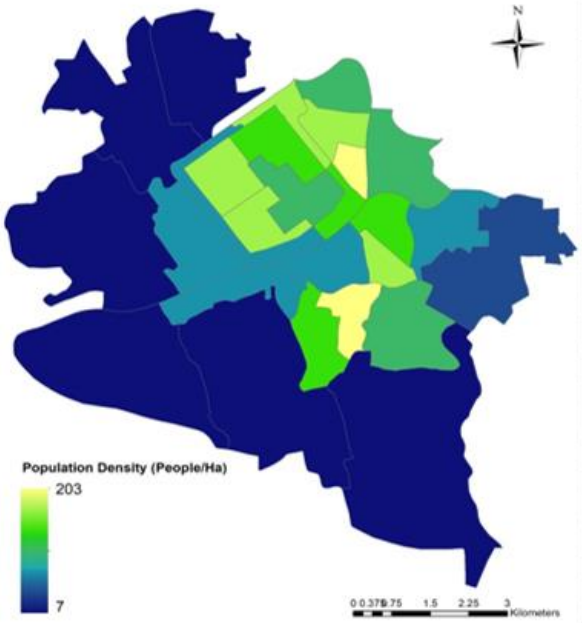


Figure 2. Flowchart of the population density mapping approach.



Com_name	Pop_2009	Male_2009	Female_2009	Pop_Urban	Pop_Rural
Phường Phú Thuận	7366	3690	3676	7366	0
Phường Phú Bình	9664	4787	4877	9664	0
Phường Tây Lộc	19473	9479	9994	19473	0
Phường Thuận Lộc	15250	7388	7862	15250	0
Phường Phú Hiệp	13774	6781	6993	13774	0
Phường Phú Hậu	9758	4722	5036	9758	0
Phường Thuận Hòa	14821	7182	7639	14821	0
Phường Thuận Thành	13421	6525	6896	13421	0
Phường Phú Hòa	6080	2937	3143	6080	0
Phường Phú Cát	8859	4248	4611	8859	0
Phường Kim Long	14656	7087	7569	14656	0
Phường Vĩ Dạ	19423	9121	10302	19423	0
Phường Phường Đúc	11307	5479	5828	11307	0
Phường Vĩnh Ninh	7553	3679	3874	7553	0
Phường Phú Hội	12038	5372	6666	12038	0
Phường Phú Nhuận	8821	4007	4814	8821	0
Phường Xuân Phú	12641	6031	6610	12641	0
Phường Trường An	15745	7627	8118	15745	0
Phường Phước Vĩnh	20959	9645	11314	20959	0
Phường An Cựu	22268	10311	11957	22268	0
Phường An Hoà	9648	4755	4893	9648	0
Phường Hương Sơ	6908	3473	3435	6908	0
Xã Thủy Biều	9433	4654	4779	0	9433
Xã Hương Long	10059	5082	4977	0	10059
Xã Thủy Xuân	13100	6527	6573	0	13100
Phường An Đông	15653	7579	8074	15653	0
Phường An Tây	6897	3405	3492	6897	0

Figure 3. Spatial population distribution in the city in 2009.

Based on statistical data from the census in 2009, the total area of 27 wards is around 80 km^2. The total population in the city in 2009 was 335,575 people. In our study area, Phuoc Vinh ward has highest population density (approximately 190 persons/ha), followed by Phu Cat (168 persons/ha), Phu Binh (157 persons/ha). Those wards have highest density because of its location, in or nearby the central of the city. An Tay has the smallest population density in the study area (approximately 7 persons/ha).

For data processing, the administrative boundary which linked with census population count at ward level is converted to raster with 30m spatial resolution. Then, the class data of land use land cover (LULC) are converted to binary mask with same resolution as census population raster.

After that, these binary masks a distance to class is calculated for each dataset. Data sets representing raster data, for example: night-time light imagery, DEM,… are projected, resampled and aggregated to match the gird census. Besides, distance to feature line data (roads, waterways,…) is also calculated. We are also created density as well as calculated distance of POIs for creating input variables to RF model.

The input data was used to create RF model to predict log of population. RF model is an ensemble, nonparametric modeling approach that grows a "forest" of individual classification or regression trees and improves upon bagging by using the best of a random selection of predictors at each node in each tree (Breiman et al., 2001). In many cases the predictive performance for RF is on par with boosted regression trees but have advantage of having fewer tuning parameters (Sikonja, 2004). In methodology this is especially important part of the fitting process.

Model estimation, fitting and prediction were all completed using R and the randomForest package. First, we fit a series of models using the tuneRF function with all available covariates. The tuneRF function uses a step function to tune the mtry parameter. This parameter determines the number of covariates to randomly select and choose from the best covariate for each node during the tree growing process. Prediction accuracies can be sensitive to the mtry parameter and tuneRF uses the minimization of OOB prediction error as an objective function to select an appropriate value for mtry (Stevens et al., 2015). In this case, the RF model showed the best result when mtry = 10, and ntree = 200.

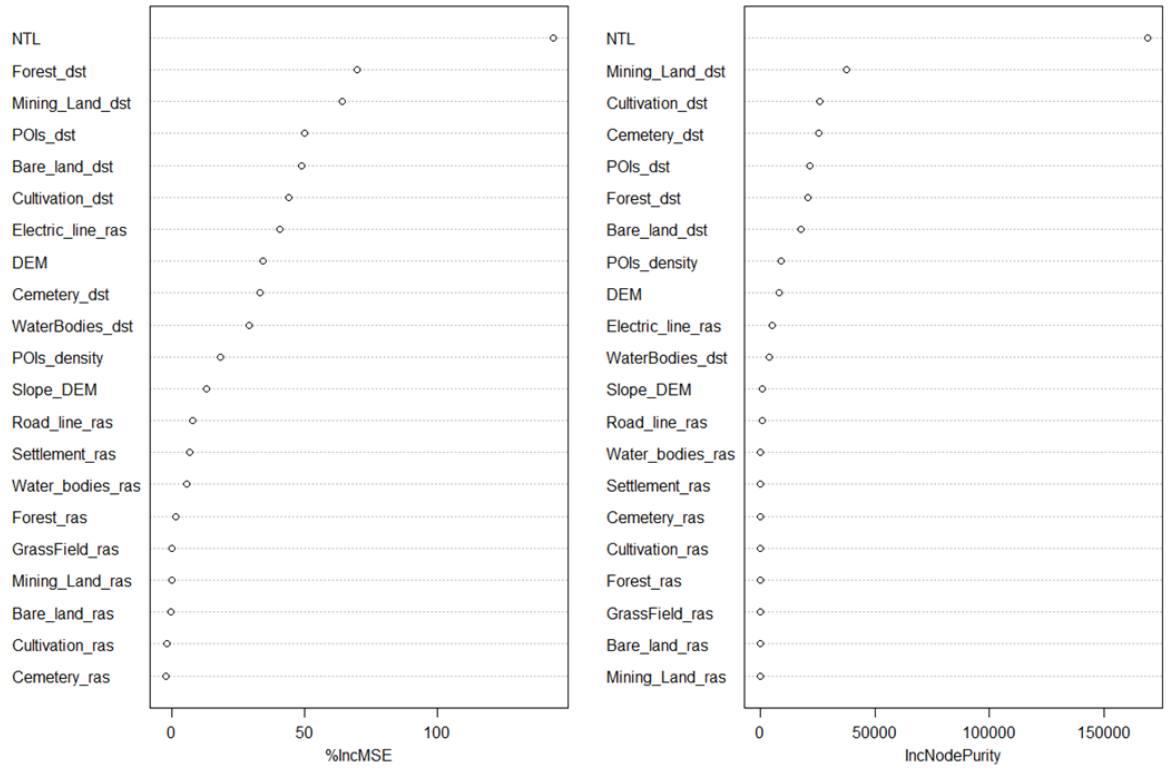


Figure 4. Top covariates important to fit the RF model. Those covariates are highly correlated with population distribution such distance to: night-time light, POIs, cultivation,....

The next step is a covariate selection process for the resulting of RF model. We performed this step to reduce the number of covariates in final RF model. For any covariate that has a variable less importance, we remove it from the list of covariates and rerun tuneRF with the reduced set of data. This is iterated until only positive importance scores remain for every covariate included in the modeling process. After fitting the RF model, it is applied for predicting a RF model using prediction data. Finally, we get density of population. The final step was to allocate the population at the building level. The iterative gravity model was applied, using seed-growing algorithm.

$$\begin{cases} Gravity(i,j,k,t+1) = \dfrac{F_{ij} * A_{ij} - \left[pop_{ij}(t) + pop_{ik}\right] * \bar{A} - C_i}{\left(\sqrt{(x_{ij} - x_{ik})^2 + (y_{ij} - y_{ik})^2}\right)^{\beta}} j,k \in i \\ j' = argmax_j[Gravity(i,j,k,t+1)] \end{cases}$$

where: c_i is an adjustment factor; $c_i = \frac{\sum F_{ij} * A_{ij} - pop_{total} * \bar{A}}{n_i}$

F_{ij} is total floor of building

A_{ij} is floor surface area

$pop_{ij}(t)$ is residential population in building $loc_{ij}(x_{ij},\ y_{ij})$ after t interactions

pop_{ik} is residential population in grid, predicted result $loc_{ik}(x_{ik},\ y_{ik})$

pop_{total} is total population

β set as 1.5 (Liu et al. 2015)

$\bar{A}$ is the per capita housing area of the census unit ($\bar{A}$ = 16.9m2, according to government statistical data in 2009)

n_i is total number of buildings in the ith census unit

3. RESULT AND DISCUSSION

As we built the RF algorithm-based population-fitting model, we selected 21 categories of spatial variables as inputs, including LULC, OSM and basic GIS data. During this process, 36000 training samples were selected. We implemented the RF algorithm-based fitting model with 200 decision trees, and the percentages of the training data set and out-of-bag data set were set to 0.7 and 0.3, respectively, for cross-validation. Next, the fitting model of the spatial variables and census population was generated, the number of people in each building estimated using predicting a RF model. The correlation has been compared between census data and estimation result. The result shows good linear relationship with R^2 larger than 0.9 (**Figure 5**).

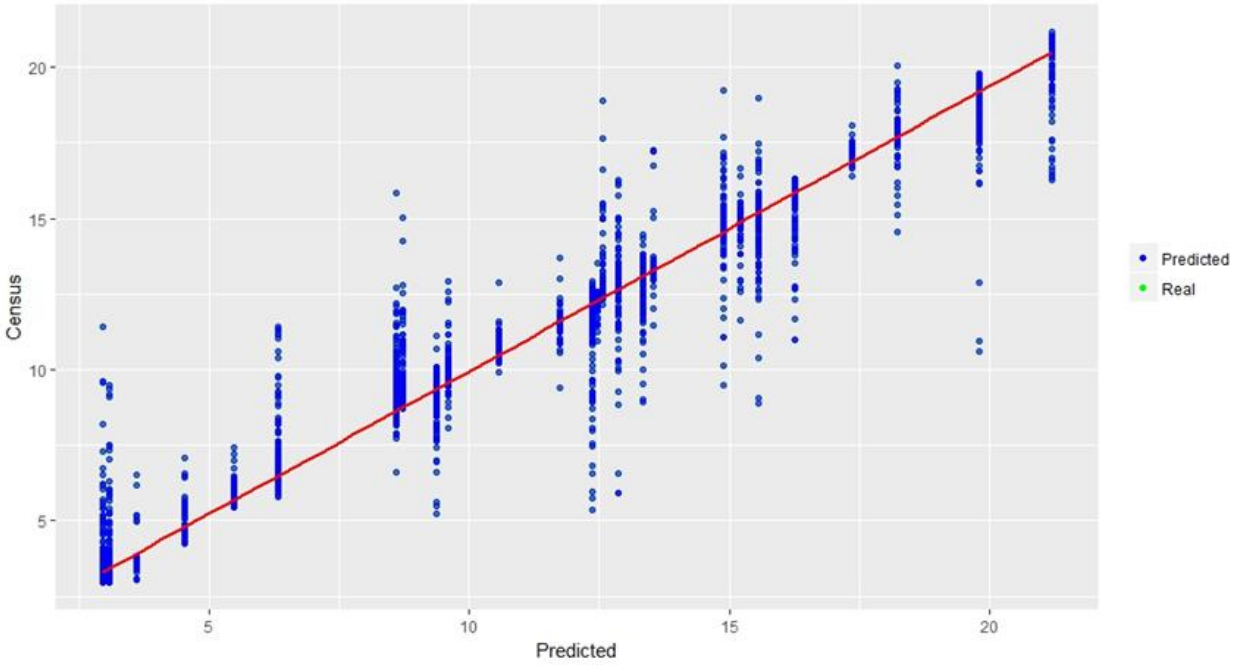


Figure 5. Linear relationship between population data and estimation data in 2009.

Figure 6 shows the mapping result of the simulated population distribution when downscaled from statistical data at the ward level at a spatial resolution of 30 m. The blue color means less populated and the yellow color shows high populated. The most population is concentrated surrounding central ward such as: Phu Cat, Phuoc Vinh, Truong An, Phu Hoi, etc.

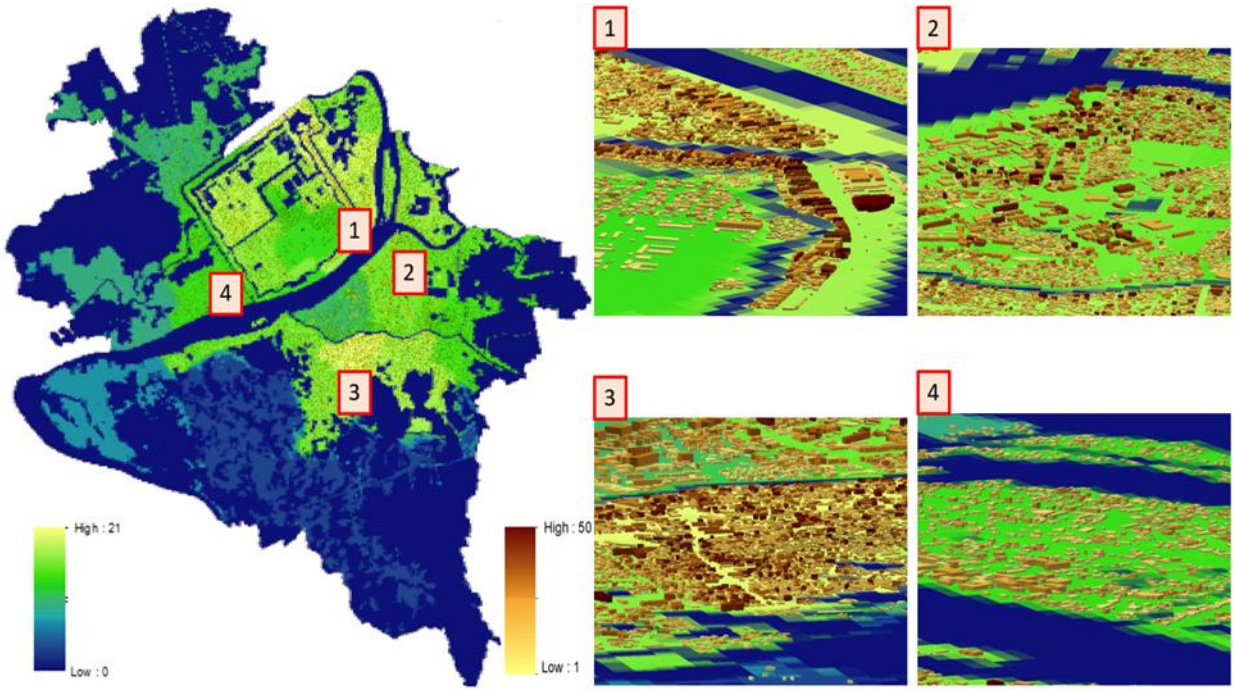


Figure 6. Spatial distribution of population in the Hue city. The blue color presents that area less population density and the yellow color shows high concentrated of population. And the 3D of

building in the right figure presented number of people in the building. The yellow presents less number of people and brown color presents high number of people in the building.

Figure 6, for which the building-population gravity model specified by equation above was used, shows the estimated housing area per capita of all the residential buildings at the city. By repeatedly adjusting Ci in Equation above, we produced a reasonable population distribution at the building level.

The crucial impact on the quality of population redistribution (especially in a fine scale) have ancillary data sources. However, many findings concerning data quality are commonly available in the literature. In addition, data that are derived by native government are generally more reliable than those from open source data (i. e., OSM, POI, etc.). The data are generally freely available and some available for a payment. This research used official data, that are accurate and credible but for some areas are not up-to-date (i. e., building footprints in An Dong and Xuan Phu wards, or no attribute information in An Dong and Phu Hau wards). A limitation of this research is time consuming and high cost of obtaining information about the buildings, in particular building structural type, volume and height (building survey). One other limitation is that we did not focus on number of level(s) and building use (public, commercial, industrial, and residential) when estimated the number of population. The overall accuracy of estimation result reduced due to those limitation. In further studies, we will try to overcome disadvantages here. This certainly contributes significantly to increase the accuracy of estimation the population totals in buildings.

4. CONCLUSION

The RF model performs substantially better than several other commonly used. An assessment of which of the ancillary data covariates are important for accurately estimating population at the building level is produced by the RF algorithm. During the variable selection phase of the algorithm, the values of variable importance may fluctuate as the number of covariates is reduced. However, the relative ranking is quite stable among the top covariates. This indicates that ancillary datasets are extremely valuable.

The appropriateness and quality of the ancillary data used influence the accuracy and level of detail of population distribution techniques and algorithms. One of the advantages with the approach put forward in this paper, is that the data set (building footprints) used has the capacity to provide information about the characteristics of the population distribution in a fine scale. Thus, the elaborated population surface could be used in risk exposure and risk evacuation or mitigation plans.

References from Journals:

Alegana V. A., Atkinson P. M., Pezzulo C., Sorichetta A., Sorichetta D., Bird T., Erbach-Schoenberg E., Tatem A. J. (2015) Fine resolution mapping of population age-structures for health and development applications. J. R. Soc. Interface, 12.

Bakillah, M., et al., 2014. Fine-resolution population mapping using OpenStreetMap points-ofinterest. International Journal Of Geographical Information Science, 28 (9), 1940–1963. doi:10.1080/13658816.2014.909045.

Bielecka, E. (2005). A Dasymetric Population Density Map of Poland.

Bhaduri, B., et al., 2007. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. GeoJournal, 69 (1–2), 103–117. doi:10.1007/s10708-007-9105-9.

Chen K. (2002). An approach to linking remotely sensed data and areal census data. International Journal of Remote Sensing, 23 (2002), pp. 37-48

Deville P., Linard C., Martin S., Gilbert M., Stevens F. R. (2014). Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences of the United States of America, National Academy of Sciences, 111 (45), pp.15888-15893

Eicher, C. L., and C. A. Brewer. (2001). Dasymetric mapping and areal interpolation: implementation and evaluation. Cartography and Geographic Information Science 28 (2): 125-38.

Gallego J., S. Peedell. (2001). Using CORINE Land cover to map population density. Towards agri-environmental indicators. EEA Topic report 6/2001:94-105.

Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P., Tatem, A. J. (2013). High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015. PLoS ONE, 8(2). https://doi.org/10.1371/journal.pone.0055882.

Langford M. (2006). Obtaining population estimates in non-census reporting zones: an evaluation of the 3-class dasymetric method. Computers, Environments and Urban Systems, 30, pp. 161-180

Langford, M., 2013. An evaluation of small area population estimation techniques using open access ancillary data. Geographical Analysis, 45 (3), 324–344. doi:10.1111/gean.2013.45.issue-3.

Liaw A, Wiener M, others. Classification and regression by randomForest. R news 2002;2(3):18–22.

Lu, D., Weng, Q., and Li, G., 2006. Residential population estimation using a remote sensing derived impervious surface approach. International Journal Of Remote Sensing, 27 (16), 3553–3570. doi:10.1080/01431160600617202.

Lwin, K., Murayama, Y., 2009. A GIS estimation of building population for micro- spatial analysis. Transactions in GIS 13, 401–404.

Maantay, J.A., Maroko, A.R., Herrmann, C., 2007. Mapping population distribution in the urban environment: the Cadastral-based Expert Dasymetric System (CEDS). Cartography and Geographic Information Science 34, 77–102.

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. The Professional Geographer 55(1):31-42.

Mennis J., Hultgren T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. Cartography and Geographic Information Science, 33 (2006), pp. 179-194.

Robnik-Sikonja, M. Improving random forests. Mach. Learn 2004, 3201, 359–370.

Sorichetta, A., Honrby G. M., Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). High resolution grid-ded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. Scientific data, 2. DOI 10.1038/sdata.2015.45.

Stevens F. R., Gaughan A. E., Linard C., Tatem A. J. (2015). Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. PLoS ONE 10(2): e0107042. https://doi.org/10.1371/journal.pone.0107042.

Sutton P. (2003). Estimation of human population parameters using night-time satellite imagery. Remotely Sensed Cities, London: Taylor and Francis, 301-334.

Tatem, A. J., Noor, A. M., von Hagen, C., Di Gregorio, A., & Hay, S. I. (2007). High resolution population maps for low income nations: Combining land cover and census in East Africa. PLoS ONE, 2(12). https://doi.org/10.1371/journal.pone.0001298.

Ural, S., Hussain, E., & Shan, J. (2011). International Journal of Applied Earth Observation and Geoinformation Building population mapping with aerial imagery and GIS data. International Journal of Applied Earth Observations and Geoinformation, 13(6), 841–852. https://doi.org/10.1016/j.jag.2011.06.004.

Wu, C. and Murray, A.T., 2005. A cokriging method for estimating population density in urban areas. Computers, Environment and Urban Systems, 29 (5), 558–579. doi:10.1016/j. compenvurbsys.2005.01.006

References from Books:

Breiman, L. (2001). Random forests. Mach. Learn. 45, 5–32.

Breiman L. (2002). Manual on setting up, using, and understanding random forests v3.1.

References from Other Literature:

M-BRACE Project Management Board in Thua Thien Province and Institute for Social and Environmental Transition-International Vietnam (2014). “Climate Action Plan for Hue City: Responding to Climate Change from 2014-2020”. Publication funded by the United States Agency for International Development (USAID).