data

MDPI

Article

Improving Urban Population Distribution Models with Very-High Resolution Satellite Information

Taïs Grippa [1,*], Catherine Linard [2], Moritz Lennert [1], Stefanos Georganos [1], Nicholus Mboga [1], Sabine Vanhuysse [1], Assane Gadiaga [2] and Eléonore Wolff [1]

[1] Department of Geoscience, Environment & Society, Université Libre De Bruxelles (ULB), 1050 Bruxelles, Belgium; mlennert@ulb.ac.be (M.L.); sgeorgan@ulb.ac.be (S.G.); nmboga@ulb.ac.be (N.M.); svhuysse@ulb.ac.be (S.V.); ewolff@ulb.ac.be (E.W.)

[2] Department of Geography, University of Namur, 5000 Namur, Belgium; catherine.linard@unamur.be (C.L.); assanegadiaga@gmail.com (A.G.)

* Correspondence: tgrippa@ulb.ac.be; Tel.: +32-2-650-6803

Received: 5 December 2018 ; Accepted: 14 January 2019; Published: 16 January 2019

Abstract: Built-up layers derived from medium resolution (MR) satellite information have proven their contribution to dasymetric mapping, but suffer from important limitations when working at the intra-urban level, mainly due to their difficulty in capturing the whole range of variation in terms of built-up densities. In this regard, very-high resolution (VHR) remote sensing is known for its ability to better capture small variations in built-up densities and to derive detailed urban land use, which plead in favor of its use when mapping urban populations. In this paper, we compare the added value of various combinations of VHR data sets, compared to a MR one. A top-down dasymetric mapping strategy is applied to reallocate population counts from administrative units into a regular 100 × 100 m grid, according to different weighting layers. These weighting layers are created from MR and/or VHR input data, using simple built-up proportion or reallocation "weights", obtained from a set of multiple ancillary data used to train a Random Forest regression model. The results reveal that (1) a built-up mask derived from VHR can improve the accuracy of the reallocation by roughly 13%, compared to MR; (2) using VHR land-use information alone results in lower accuracy than using a MR built-up mask; and (3) there is a clear complementarity between VHR land cover and land use.

Keywords: population modelling; dasymetric mapping; top-down approach; very-high resolution data; remote sensing; random forest; African city

1. Introduction

The less developed regions of the world have reached a symbolic milestone: Half of the population is now living in urban areas [1]. Even though this ratio is much lower in the least developed countries, most of which are located in sub-Saharan Africa (SSA), urbanization rates are increasing rapidly (where about 33% of the population is urban and are expected to face the highest growth rates during the next decades). It is expected that 40% of the population will live in urban areas by 2030 and 50% by 2050 [1]. As a consequence of these rapid transformations, SSA cities are exposed to increasing urban poverty and intra-urban inequalities [2], while a large part of the urban population is extremely vulnerable to health and disaster risks. In this context, detailed population data is essential in improving evidence-based decision-making by relevant authorities and organizations [3–5], as well as for any application relying on a human population denominator, such as estimating the population at risk, assessing vulnerability, and deriving health or development goals indicators [6–8]. However, this knowledge is often very limited in SSA and population data are regularly outdated and criticized

regarding their reliability [6,7]. While collected at a household or individual level, census data are generally aggregated and released in administrative units for privacy reasons [5,9], and do not match the requirements for different fields of research [4]. With regards to population data aggregated in administrative units, we can further mention some issues related to the fact that (1) the real spatial patterns of the population distribution are blurred by an impression of homogeneity within entities [10], (2) the aggregated values and subsequent analysis are very dependent on the choice of the administrative limits, which is also known as the modifiable areal unit problem (MAUP) [11], and (3) administrative units create subjective spatial discontinuities that sometimes change from one census to another [9,10].

When the spatial extension of a phenomenon does not correspond to any existing administrative limits, official population data are often unexploitable. In such a situation, a gridded population product—a raster layer where the pixel value refers to the (estimated) number of inhabitants—can provide a more useful estimate of population counts [12], by summing up all the pixels falling into the area under investigation. Creation of these population grid products is usually achieved using dasymetric mapping [9,13]. This modeling technique relies on the assumption that the knowledge of the territory—places more densely populated than others—can be used to spatially disaggregate the official census data provided at the administrative level to a finer scale [5]. Ancillary geoinformation data, such as land cover (LC) and land use (LU) maps, can provide valuable information for estimating the potential of different locations within the administrative units to be inhabited. Even though they are different by nature—LC is related to the physical characteristics of earth surface elements (e.g., vegetation, water, built-up, ...), while LU refers to the functions and activities that humans decided to carry out in certain locations (e.g., agricultural land, residential area, industrial area, ...)—they can provide complementary information valuable for population modelling purposes; for example, by combining building density (from LC) with the distinction between residential and commercial areas (from LU). For example, the built-up density and the land use information of a location can be combined and used as proxies for population density.

The major challenge in dasymetric mapping resides in the determination, from a set of ancillary data, of the relative distribution of the population within the administrative units. This information can be seen as spatial reallocation "weights", which are used in dasymetric mapping to disaggregate (redistribute) the population count known for the administrative units into a finer subunit level. When a simple built-settlement layer is available, a common strategy is to homogeneously allocate the population counts of the administrative unit within areas identified as built-up (binary dasymetric method). When the ancillary data are thematically more detailed than just a binary built-up layer—e.g., with a distinction between urban core, periurban, and rural areas—the weights can be adjusted to better correspond with the expected relative distribution of the population. For a long time, these weights were subjectively determined based on expert knowledge, or according to existing information [12], such as land-use information or household characteristics, combined with the use of quantitative methods, such as correlation analysis and multivariate regression [13]. Recent research has shifted this paradigm by taking advantage of the power and the efficiency of machine learning algorithms to model the distribution of population densities, without any prior knowledge. In the case of the WorldPop project, the popular Random Forest (RF) algorithm [14] is used to predict the weights for reallocation of population in 100 $\times$ 100 m grid layers [15]. In this work, the RF algorithm is used in a similar fashion.

Irrespective of the approach (expert-based or using machine-learning), built-settlement layers are consistently among the most important predictors for population models [16]. These layers are typically extracted from satellite imagery, and have been commonly used to estimate population densities at large spatial scales. However, both the quality [5] and the spatial resolution [4] of ancillary information have a strong influence on the accuracy of the predictions. In an urban context, the potential of finer resolution products for population redistribution is largely unexplored. We hypothesize that, by

utilizing high and very-high resolution information (i.e., land cover and land use), the accuracy of the dasymetric reallocation might be significantly improved.

In this paper, we compare the contribution of three data sets with different spatial and thematic resolutions (built-up mask, land cover, and land use) for disaggregating population counts into 1 hectare grid cells. The availability of extremely detailed census data for the city of Dakar (Senegal) enables the assessment of the added value of very-high resolution (0.5 m) data, compared to medium resolution (10 m) data, in the context of a top-down dasymetric approach. Different levels of information are extracted from these data sets to create different weighting layers and perform dasymetric mapping. While very-high resolution data are expected to increase the accuracy of the dasymetric mapping procedure, their acquisition and processing costs might hinder their applicability for large-scale population mapping in Africa. It is, therefore, important to evaluate the loss in accuracy when using freely-available medium resolution data.

2. Materials and Methods

2.1. General Workflow

A visual representation of the different administrative levels and geographical scales used in this study as well, as the major steps of the workflow, is provided in Figure 1. Level 1 represents the finest level available with the reference population count (census data). It is reserved for validation purposes, and is kept completely independent from the dasymetric mapping procedure itself. It is important to understand that the total volume of the population of each administrative level is maintained during the disaggregation, meaning that, if the predictions at grid level are re-aggregated back to the original units, the initial population count is preserved (pycnophylactic property, [17]). As a consequence, the population counts and administrative units in level 1 cannot be used directly for dasymetric mapping, since they are the finest official level at which a validation is possible. In order to keep level 1 units available for validation purposes, the first step aims at creating coarser administrative units, hereafter referred to as 'level 0', by aggregating level 1 units (more details about this aggregation procedure are provided in Section 2.4). The second step consists of the dasymetric reallocation of population counts from level 0 units to a regular grid layer of 100 × 100 m. The different tests performed as well as the procedure for the creation of weighting layers used for dasymetric reallocation are further described in Section 2.5. The purpose of the third step is validation. The grid level predictions are summed for each of the level 1 units, in order to compare them against the reference count kept available at this level. More details about the validation procedure is provided in Section 2.6.

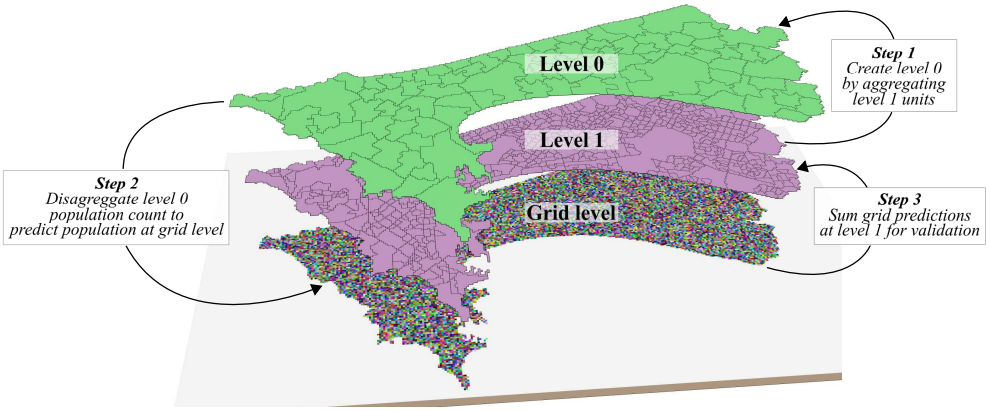

Figure 1. Representation of the major steps of the workflow presented and the different scales used in the analysis. A population count at level 0 is reallocated at grid level using weighting layers. For accuracy assessment, the predictions at grid level are aggregated to reach level 1 boundaries, and are compared with the official reference data available at this level.

2.2. Remote Sensing Derived Data

We used three ancillary data sets, to design different approaches for creating weighting layers to be used for the reallocation of population counts at a finer-scale using dasymetric mapping. All were derived from earth observation (EO) data, but are different regarding their spatial and thematic resolution. These data sets were previously produced in the context of the MAUPP (http://maupp.ulb.ac.be) and REACT (http://react.ulb.be/) projects, and are publicly available (see Appendix A).

The first data set consists of a binary built-up mask (see Figure 2A), recently published in [18]. It was derived from Landsat 8 and Sentinel-1 imagery from 2015, processed using a recently developed automated pixel-based fusion framework [19,20]. The main advantage of this product is that it is accessible and reproducible at no cost, since the EO data it relies on are free of cost. This built-up mask has a spatial resolution of 10 m and, with regards to its accuracy, an F1-score of 0.92. As its resolution could be described as medium, this product is referred to hereafter as "MR-BU".

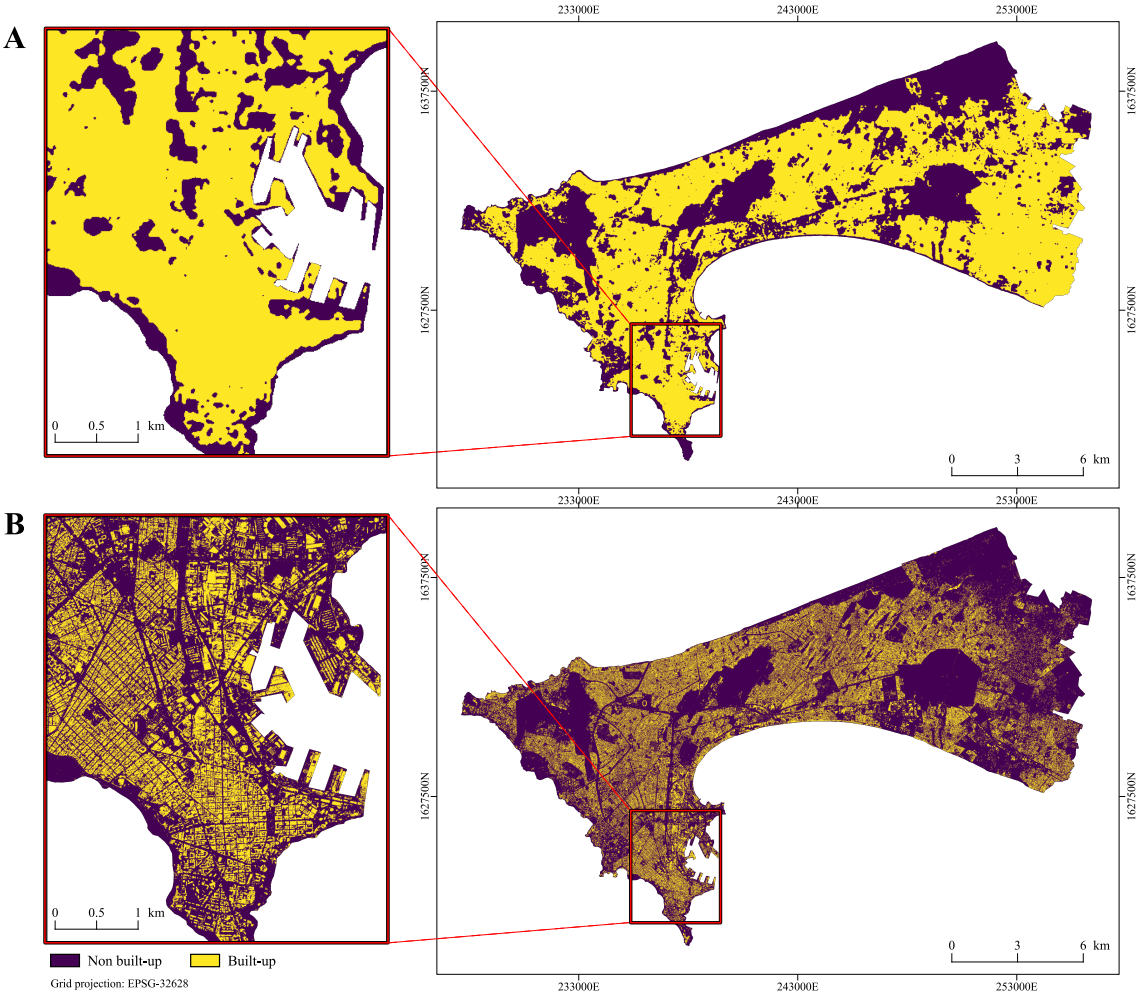

Figure 2. Built-up masks used: (**A**) Medium resolution (10 m), derived from freely available satellite images (MR-BU). (**B**) Very-high resolution (0.5 m), derived from commercial satellite images (VHR-BU).

The second data set consists of a very-high resolution land-cover map, derived from the Pleiades pan-sharpened tri-stereo images of 2015, with a spatial resolution of 0.5 m (see Figure 3). This map was previously produced thanks to a semi-automated open-source framework for object-based image analysis and supervised classification [21–23]. The overall accuracy (OA) of this product achieved 89.5%, and the building class reached 94% and 97% for user and producer accuracy, respectively (more details about the validation of this product are provided in [24]). A post-processing step was used, to further reclassify the built-up class into three classes of buildings by applying a threshold on the

height information provided by a photogrammetricaly-derived normalized digital surface model (nDSM). From the land-cover product, we extracted different combinations of land-cover classes in order to produce three very-high resolution layers for the analysis: (1) The land-cover map itself, including all the classes, as illustrated in Figure 3 (referred to as "VHR-LC"), (2) a built-up mask (referred to as "VHR-BU") (see Figure 2B), and (3) a layer containing three building classes, categorized by height (referred to as "VHR-3BU").

The last data set used in this study consists of a map providing the dominant land use at the street block level (see Figure 4). This map, reaching an OA of 79%, was produced in a recent study [25], in which street blocks were automatically created using OpenStreetMap data, and were further classified based on spatial metrics (also called 'landscape metrics') allowing characterization according to their composition and organization, in terms of land cover. This data set is an important complement to the land-cover or built-up masks, as it provides a distinction between residential and non-residential areas (e.g., commercial areas). Moreover, the map used here contains several residential classes that should help better estimate intra-urban population distribution. This product is referred to hereafter as "VHR-LU".

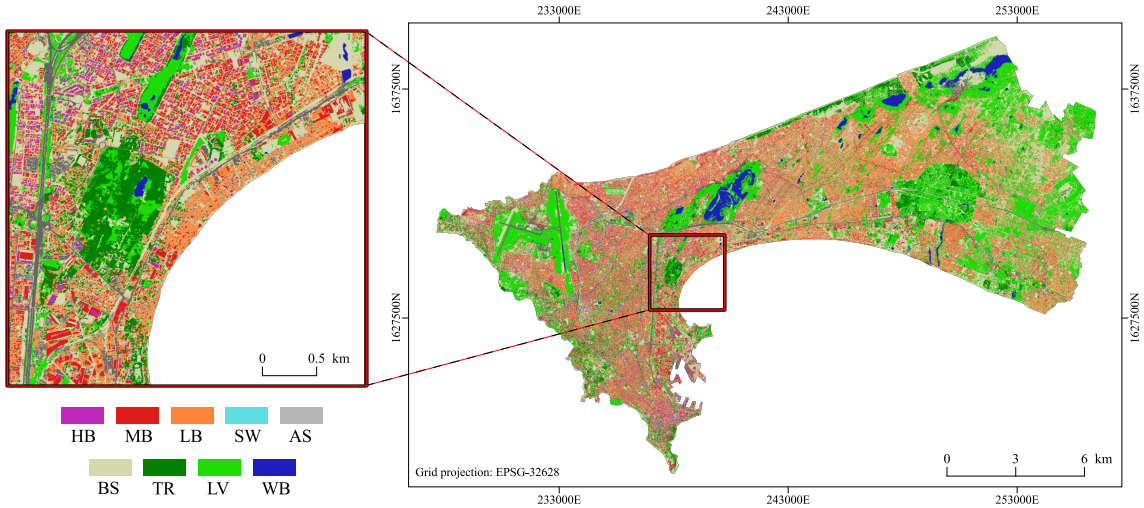

Figure 3. The very-high resolution land-cover map derived from commercial satellite images. Legend classes: HB: High buildings (>10 m); MB: Medium buildings (5–10 m); LB: Low buildings (<5 m); SW: Swimming pools; AS: Artificial ground surfaces; BS: Bare soils; TR: Trees; LV: Low vegetation; WB: Inland waters.

2.3. Population Data

Population data used in this research comes from the most recent national census, performed in 2013 [26], and provided by the National Agency for Statistics and Demography of Senegal (ANSD). A limited temporal shift of 2 years exists between census data and the imagery used. Here, we are working under the assumption that the urban expansion and/or densification that occurred during this two-year period is marginal and should not impact the main findings. Population counts are available at the 'neighborhood' level (admin-5 in the Senegalese scheme), consisting of 1347 administrative units for the whole extent of the Dakar agglomeration. The spatial extent of the population data was reduced, in order to only keep administrative units fully covered by the different map products to be used for the dasymetric mapping. This resulted in dropping 154 units (1193 remaining).

2.4. Creation of Validation and Training Levels

For this research, we had the opportunity to access population data at a very detailed scale. Figure 5A illustrates how fine this administrative level is, in terms of spatial resolution. It is important to point out that accessing such a fine-scaled population data, linked with the delineation

of the corresponding administrative units (i.e., polygon geometries provided as Shapefile), is rather exceptional for SSA cities, as the connection between population data and the finest administrative units is usually limited [6]. In this research, we take advantage of a citywide coverage with spatially detailed data, to implement a top-down dasymetric approach and perform fine-scale validation in an intra-urban environment.

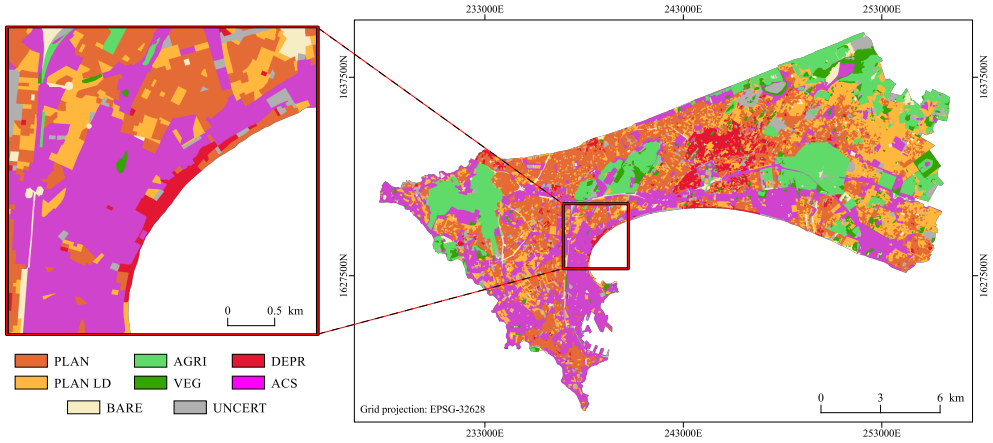

Figure 4. The very-high resolution land-use map derived from the VHR land-cover map. Legend classes: AGRI: Agricultural vegetation; VEG: Natural vegetation; BARE: Bare soils; ACS: Non-residential built-up (administrative, commercial, services, etc.); PLAN: Planned residential built-up; DEPR: Deprived residential built-up.

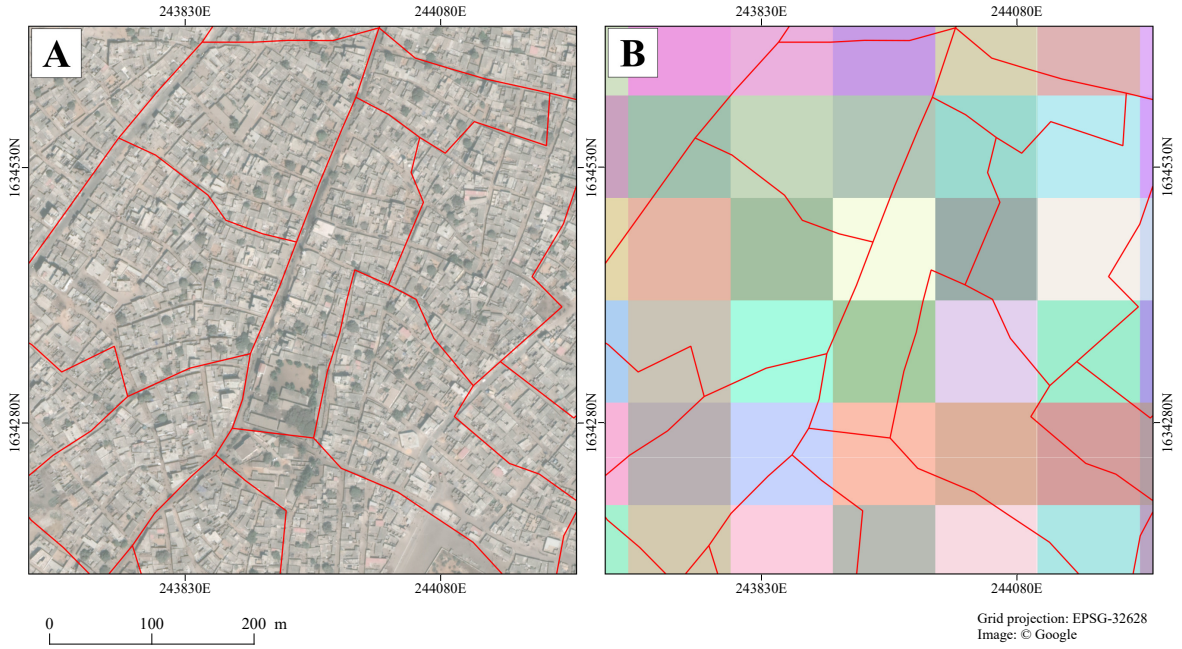

Figure 5. The administrative units (neighborhoods of level admin-5) are very small. Administrative unit limits (in red) are superimposed on (**A**) Google Map imagery, and (**B**) a 100 × 100 m grid.

In the core urban area, some administrative units are so small that they include only portions of a few 100 × 100 m grid cells, as illustrated in Figure 5B. This could create some issues when summing (aggregating) the predicted values from grid level to unit level. To mitigate these potential issues and ensure that units at level 1 cover a sufficient number of grid cells, a simple procedure was performed in order to automatically merge administrative units smaller than 8 hectares with the neighbor with which they share the longest border. This minimum size was chosen as a compromise between keeping enough administrative units at the finest level and having sufficiently big units to avoid potential

issues during the validation procedure. The resulting layer, referred to as "level 1", consisted of 677 units and was used in our analysis as the validation level. Figure 6A gives an overview of the population density at this level.

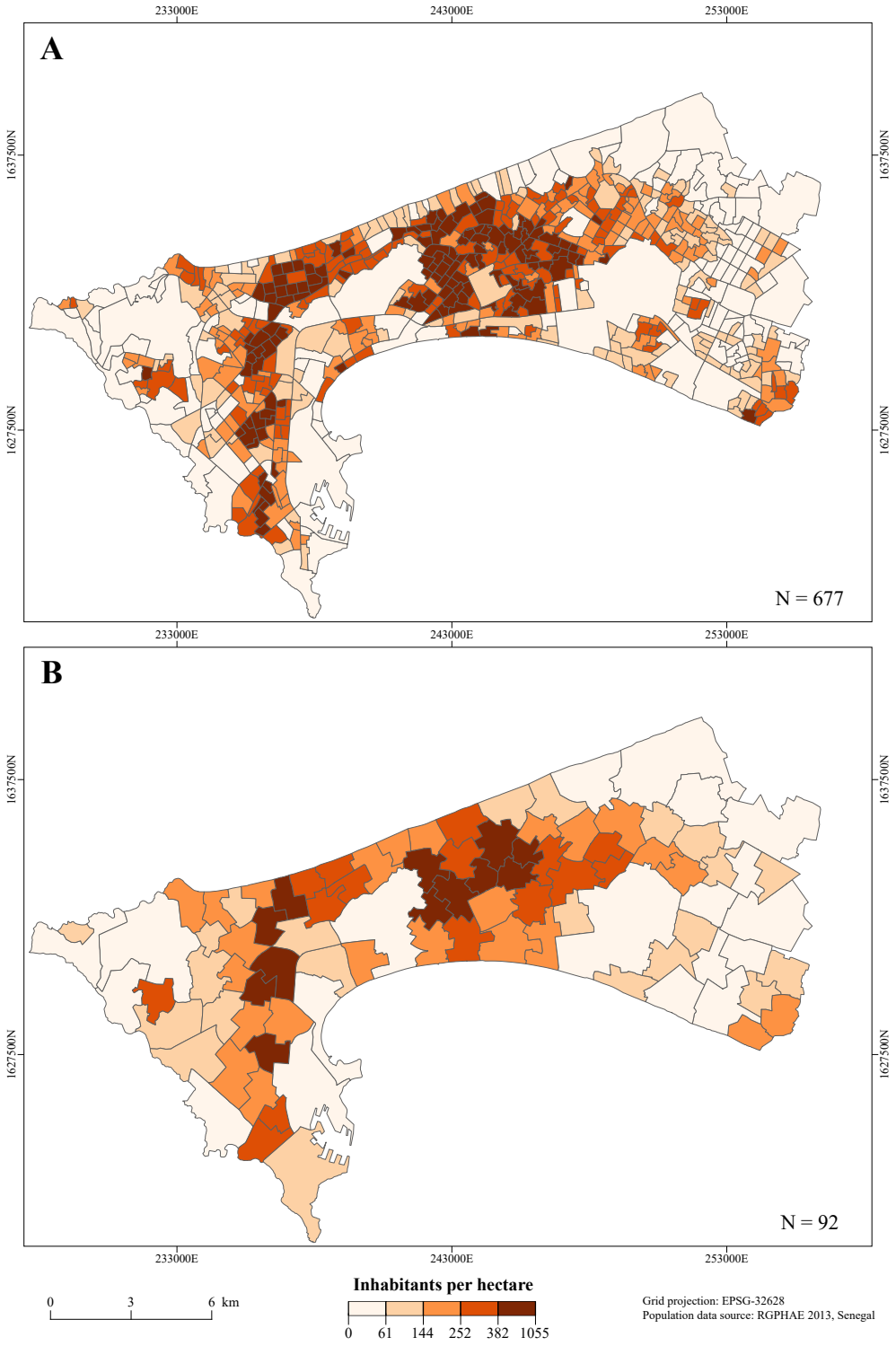

Figure 6. Reference population densities. (**A**) Level 1: Finest level reference population data, used here for validation purposes. (**B**) Level 0: A coarser level, created by aggregating contiguous level 1 units together. This level is used as a basis for dasymetric mapping and for training the RF model.

The official admin-4 level in the Senegalese administrative scheme could have been used to create a coarser base level for the dasymetric mapping procedure . However, only 40 admin-4 units were completely included in our area of interest, which would have dramatically affected the ability to train the RF model properly. Instead, we aggregated the level 1 administrative units to create a coarser level, to be used as basis for the disaggregation procedure (represented as step 1 in Figure 1). It was performed using K-means unsupervised clustering on the X/Y coordinates of the polygons' centroid. The desired number of clusters is specified by the user. Further refinement was performed to guarantee that the level 0 units consisted of at least 4 contiguous units from level 1. This procedure allowed for reduction of the 677 units available for validation (level 1) to 92 units, composing the level 0 on which to perform dasymetric mapping procedures. The reason of this drastic reduction is double: Firstly, the access to such spatially detailed data is quite rare for SSA cities. Therefore, having a limited number of units at level 0 gives a more realistic point of view regarding the data that are usually available. Secondly, the aim of the validation procedure is to assess the accuracy of the population reallocation from level 0 to level 1. With respect to the relevance of the validation scheme, it is important to have a sufficient ratio between the number of administrative units at the training and validation level. The average size of level 0 units and level 1 units are 167.42 hectares and 16.49 hectares, respectively (see Table 1).

Table 1. Descriptive statistics of both administrative levels used in the analysis.

Level	Area (ha.)	Population Density (inhab./ha.)		
	Mean	Minimum	Median	Maximum
Level 1 (677 units)	16.49	0.81	184.59	1047.23
Level 0 (92 units)	167.42	5.89	164.04	541.99

2.5. Analysis Design and Weighting Layer Creation

The aim of this study is to assess the contribution of very-high resolution land-cover and land-use products, in comparison to a medium resolution (10 m) built-up mask. Table 2 gives a snapshot of the different test layouts. When single binary built-up/non-built-up information is used as ancillary data, the built-up proportion is computed and directly used as weights for dasymetric mapping (tests A and B). On the other hand, when a larger amount of ancillary information is used, weights are derived through the use of a RF model, such as in tests C to J. Because they are derived from the same single source and consequently represent redundant information, a combination of VHR-BU, VHR-3BU, and/or VHR-LC layers are not used together in the same test.

Table 2. Layout of the different tests performed. The "X" marks inform about which layer(s) is used in each test.

Test	Weights Creation	MR-BU	VHR-BU	VHR-3BU	VHR-LC	VHR-LU
A	Simple proportion	X				
B	Simple proportion		X			
C	RF-derived weights			X		
D	RF-derived weights				X	
E	RF-derived weights					X
F	RF-derived weights	X				X
G	RF-derived weights		X			X
H	RF-derived weights			X		X
I	RF-derived weights				X	X
J	RF-derived weights	X			X	X

Expert knowledge could be sufficient for identifying evident trends in a data set, or for pointing out specific proxies for a well-known phenomenon. However, when numerous non-linear input data

are used to derive a weighting layer, as in dasymetric mapping strategies, it can become very difficult to rely on expert knowledge. In this context, advantage can be taken from machine learning methods which have proved their efficiency in finding relevant relationships between data for predicting a response variable (e.g., the population density), as is in the WorldPop project that utilises the Random Forest (RF) regression algorithm [15].

RF is a non-parametric supervised machine learning algorithm, which is efficient in handling noisy and highly correlated input data, in addition to its relative resistance to overfitting. It belongs to the category of "ensemble learning" strategies, and consists of an aggregation of several individual and independent trees (CART), each of them trained on a random bootstrapped sample of the training data. RF has a low number of (hyper-)parameters to be set when looking for model optimization. Usually, the number of trees to grow and the number of randomly selected features at each node within a tree are the most common. In the procedure we developed, these parameters are automatically fine-tuned, using a grid search procedure that considers all possible combinations from a range of potential parameter values, to train different models and assess their performance through a k-fold cross-validation. It should be noted that the RF algorithm can be used either for classification or, as in the case in this paper, for regression tasks. Interested readers can refer to the original publication of the RF algorithm [14] for a deeper understanding of its principles.

From the different ancillary data sets, we compute the proportion of each available class available in the different layers. This refers to the built-up proportion for the two binary mask products (MR-BU and VHR-BU), the proportions of three built-up classes categorized by elevation for the VHR-3BU layer, the proportion of each of the land cover classes for the VHR-LC layer, and the proportion of each land-use class provided by the VHR-LU layer. These results constitute the set of covariates used to train the RF model, according to the layout of the different tests designed (see Table 2). The proportions are computed at two levels: For each polygon at the administrative unit level, and for each pixel of the 100×100 m grid layer. After training on administrative unit level, the fitted model is used to predict the population density for each pixel of the grid layer. This means that this modeling strategy relies on the assumption that the relations that exist between the covariates and the response variable at the administrative unit level (training level) are the same that exist at the grid level (prediction level), which is unlikely to be completely true and is very dependent on the importance of the scale factor between these levels (MAUP effect). The natural log of the population density is used as response variable of the RF model, as previous research suggested it improved the quality of the weight prediction [15], and a back-transformation is applied on the predicted values to retrieve population densities. The values predicted by RF can only be in the range of the response variable it was trained on. It is, thus, incapable of predicting zero values (there are no zero densities in our population data, but even if there were, the log transformation requires the removal of all units with zero population count). Prior to using the predicted weights for dasymetric mapping, we make the choice to force all the grid pixels with a 0% built proportion (in case of a test using LC information) or with 0% potentially inhabited areas (in case of a test using only LU), to have a zero weight value for the grid. This strategy was already used in previous studies [15,27].

2.6. Validation Scheme

The prime rationale for performing dasymetric mapping is to estimate population distribution at a finer scale than the one at which official reference data are released. Usually, when the finest reference data is not sufficiently detailed, validation procedures to assess the accuracy of the spatial reallocation are not performed. Here, we take advantage of the sufficient details of the population data (see description in Section 2.3) to systematically assess the contribution of input data with different spatial and thematic resolution in a dasymetric reallocation procedure.

As already mentioned in Section 2.1, the validation design consists in aggregating the grid-level predictions to get the total estimates (sum) for each of the level 1 units, and comparing them against the reference population count available at this level. It is important to note that this validation procedure

only assesses the efficiency of the different weighting layers for reallocating the population count from level 0 to level 1. The validation of the grid level predictions cannot be achieved here, since official population counts do not exists at this level.

Two different metrics are used to evaluate the performance of the dasymetric models by confronting the population count estimates at level 1 against the reference counts: (1) The normalized version of the commonly used metric root-mean-square error (%RMSE), which uses the mean reference population of administrative units for normalization [15]; and (2) the relative total absolute error (RTAE), which is the ratio between the sum of all absolute errors and the total reference population [28]. These metrics are computed as follows:

$$\%RMSE = \frac{\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(pred_i - ref_i\right)^2}}{\frac{1}{n}\Sigma_{i=1}^{n} ref_i} * 100, \text{ and} \quad (1)$$

$$RTAE = \frac{\Sigma_{i=1}^{n} \mid pred_i - ref_i \mid}{\Sigma_{i=1}^{n} ref_i} * 100, \quad (2)$$

where ref_i is the reference population count of the administrative unit i, and $pred_i$ is the sum (aggregation) of all the predictions at grid level that fall within the administrative unit i.

2.7. Software Environment and Computer Code Availability

All the analyses were performed in Python, using common libraries for the manipulation of geospatial data (GeoPandas, Fiona, Shapely) and machine learning (Scikit-learn [29]). In addition, the "GRASS Python scripting library" enabled us to take advantage of the efficiency of the open-source software GRASS GIS [30] for raster processing and manipulation. All computer codes produced for the analysis are distributed in a "Jupyter notebook" format [31] and are available from a dedicated repository (see Appendix A). Where possible, the code was designed to support parallel processing on multiple cores, to save computing time.

3. Results

The analysis of the results, hereafter, is based on the validation performed by comparing the reference data against the aggregated estimates at administrative level 1. Several conclusions can be drawn, when analyzing the results (see Table 3). First, when relying only on built-up/non-built-up masks, the impact of the very-high resolution data is notable, since it allows the RTAE to decrease by 13% (from 36.7 to 31.7). As highlighted in Figure 7, VHR-BU also leads to an important reduction of extreme relative errors of prediction. It is consistent with recent research which shows that VHR settlement layers systematically have better feature importance than lower resolution layers, in a RF-based dasymetric approach [27].

Second, when taking advantage of the detailed spatial and thematic information provided by the VHR-LC layer, the accuracy of the dasymetric reallocation is significantly improved, with the RTAE reduced to 0.308; corresponding to a drop by 16%, relative to the results obtained using only the built-up mask at medium resolution (MR-BU). Third, when considered as a single source of ancillary data, the VHR-LU layer performs poorly, compared to VHR-LC alone, and is even worse than when only using the binary information provided by the VHR-BU layer. Regarding the data used, it is probable that the lower spatial resolution (characterization of land use at the street block level) and lower classification accuracy of the VHR-LU, compared to VHR-LC and VHR-BU, have a strong influence on this result. Fourth, combined use of VHR-LC and VHR-LU provides better result than when using either of them alone, which confirms that these data are complementary. Figure 8 depicts the feature importance provided by the RF model for the best-performing test (J). It supports the conclusion that VHR-LC and VHR-LU are complementary, since there is a clear alternation of land-cover and land-use variables in the sixth most important features. In addition, it is interesting to

see that both the building classes from the LC layer ("Low buildings (> 5 m)" and "Medium buildings (5 to 10 m)") appear in the most important variable, as well as the distinction between planned residential areas and deprived residential areas from the LU layer.

Table 3. Accuracy assessment for the different tests performed. RF internal OOB score refers to the Out-Of-Bag score, computed during the internal cross-validation of the Random Forest. External validation refers to the validation scheme described in the methods section. As tests A and B did not use RF for the creation of the weighting layer, an OOB score is not available ("NA").

Test	Input Data	RF Internal OOB Score		External Validation	
		Level 0	Level 1	%RMSE	RTAE
A	MR-BU	NA	NA	61.00	36.7
B	VHR-BU	NA	NA	54.54	31.7
C	VHR-3BU	0.767	0.715	52.22	33.9
D	VHR-LC	0.759	0.759	49.31	30.8
E	VHR-LU	0.789	0.757	54.37	33.5
F	MR-BU, VHR-LU	0.808	0.766	47.59	29.7
G	VHR-BU, VHR-LU	0.842	0.768	46.21	28.2
H	VHR-3BU, VHR-LU	**0.850**	0.802	45.22	28.8
I	VHR-LC, VHR-LU	0.833	**0.815**	45.24	28.4
J	MR-BU, VHR-LC, VHR-LU	0.836	0.813	**44.40**	**27.9**

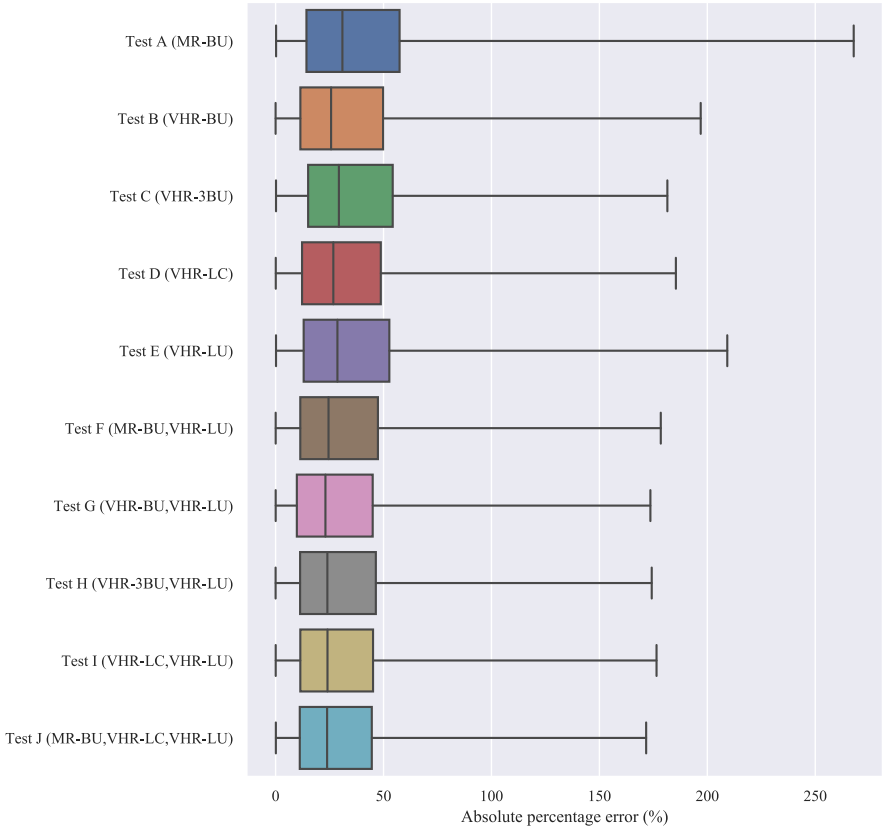

Figure 7. Boxplots representing the distribution of the absolute errors of prediction for level 1 units. The values refer to the percentage of absolute prediction error. The vertical line in the box corresponds to the median value. The left and right limits of the boxes refer to the first and third quartile, respectively. The right limit of the whiskers corresponds to the last observation whose value is below the 95th percentile. Observations with higher values are considered as outliers and are not represented here.

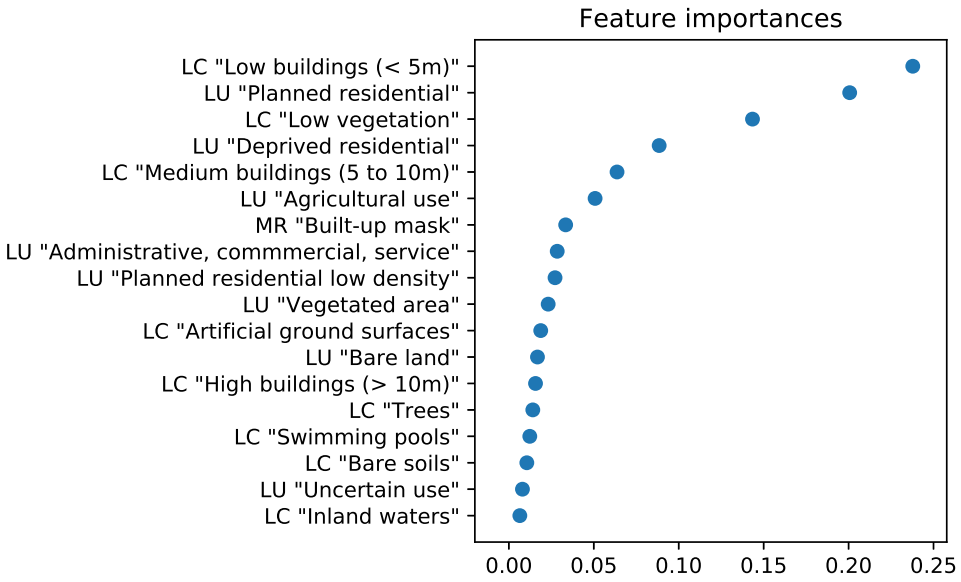

Figure 8. Feature importances from the Random Forest model for test J.

Finally, in our analysis, the best-performing dasymetric reallocation was obtained when using all available data; that is, test J, whose predictions at the grid level are illustrated in Figure 9. Surprisingly, the accuracy is improved when using MR-BU in addition to VHR-LC and VHR-LU. As shwon in Figure 10, the majority of the large relative errors (in terms of percentage of the reference population) are located, not surprisingly, in less populated administrative units.

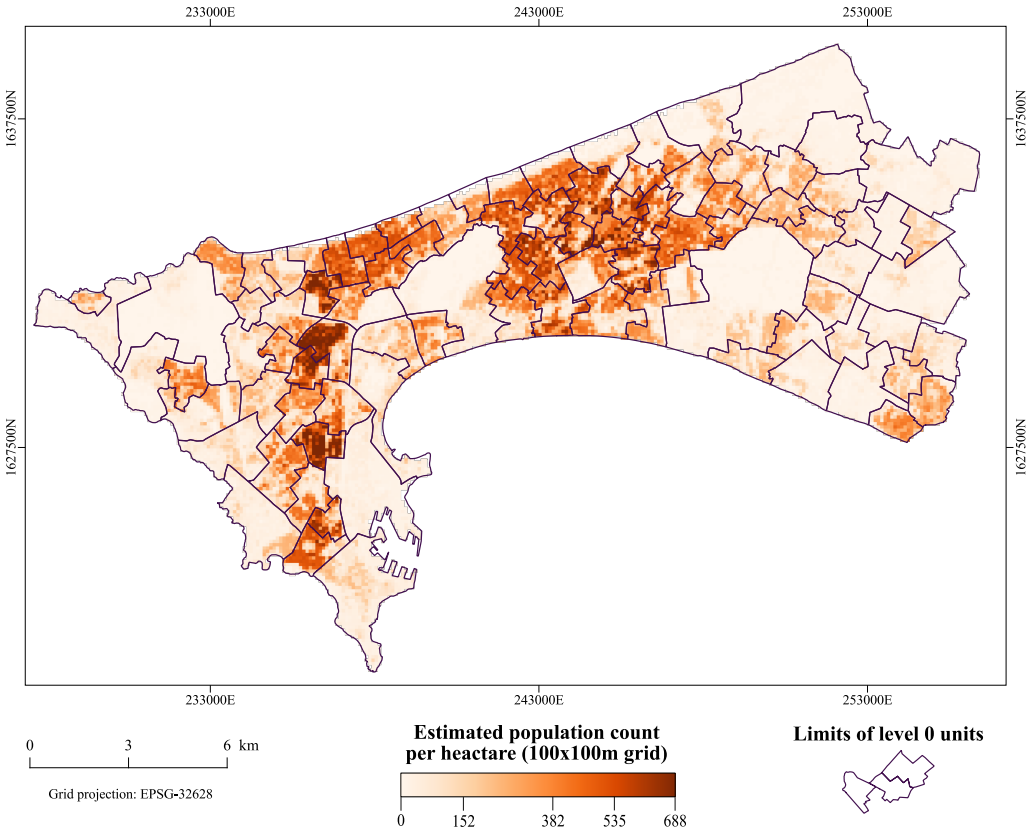

Figure 9. Test J—Prediction of population count at grid level (100 × 100 m), superimposed with the limits of units of level 0.

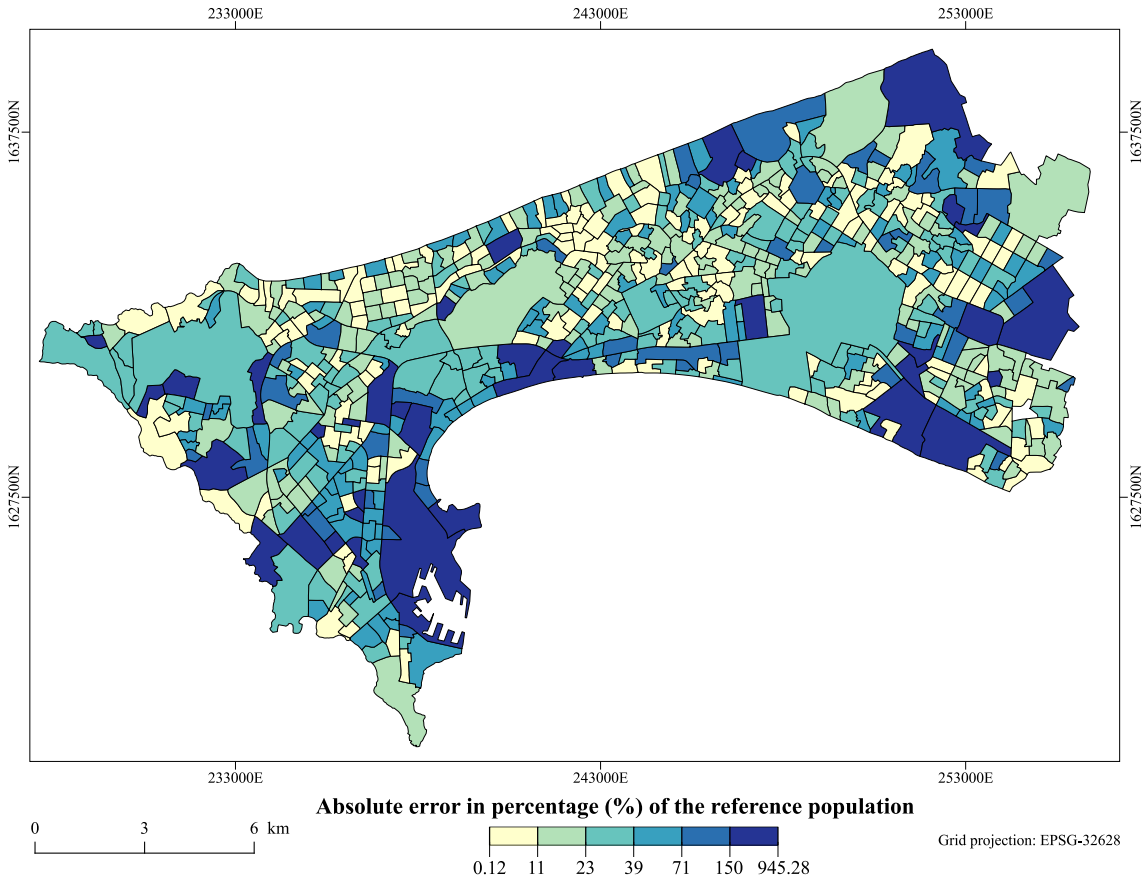

Figure 10. Errors of prediction, aggregated to level 1 in absolute percentage of the reference population.

Another interesting point, that can be highlighted from our analysis, is related to the validity of the Out-Of-Bag (OOB) score as a validation metric in dasymetric mapping. The OOB-score (or error) is an accuracy assessment metric, computed during the fitting of the Random Forest model. It is computed from the internal cross-validation procedure, and can be interpreted as an average goodness measure of the ability of the model to predict on unseen data in the training set. Since the training set is composed of administrative units of level 0, this metric could be seen as a measure of the ability of the model to predict on unseen units at the same specific level. Inversely, the external validation used here, as described in the "methods" section, is designed to assess the ability of the dasymetric mapping procedure to accurately redistribute population counts from one geographic scale to a finer one. When studies suffer from the lack of spatially detailed population data, external validations can not be systematically performed. In such contexts, it may be tempting to consider the OOB-score as a measure of the performance of the dasymetric reallocation. Nevertheless, our results show that there is no straightforward relationship between the external validation metrics and the internal OOB-score. Indeed, as visible in Table 3, the best-performing combination of input data (covariates) appears to be in the case of test H (OOB of 0.85 for a RF model built at level 0). However, when considering the RTAE or %RMSE, test J is identified to be the best-performing one. Furthermore, fitting RF models on both administrative levels revealed that the best-performing set of covariates identified at one scale is not obviously the one that performs best at another scale, which could be interpreted as a result of the MAUP effect.

The gridded population layer resulting of the dasymetric mapping procedure presented in this paper is available for anyone interested and for any purposes. The reference is provided in Appendix A.

4. Discussion

Medium resolution built-up settlement layers have been commonly used in population modeling [15,32]. They present the advantage of being free and providing global coverage, even though their spatial resolution is limited. However, high or very-high spatial resolution data are usually preferred for applications covering a relatively small geographic extent, as they allow the counting of dwelling units or the interpretation of residential land-use types, despite their expensive costs. To date, little research has explored the potential of built-settlement layers at different spatial resolutions for urban population mapping.

Even if the price of VHR remote sensing data tends to drop slowly, it is still an important limiting factor and reduces the merits of large-scale applications. Usually, the acquisition of VHR imagery is firstly dedicated to the production of detailed LC and LU maps, which are useful pieces of information by themselves. The gains that these VHR-LC and VHR-LU information could provide to the performance of a dasymetric mapping approach is important information regarding the cost effectiveness of these data. In this regard, our results show that there is a clear positive impact of using VHR products for population modeling, as well as a complementarity between VHR-LC and VHR-LU products. Future research could involve integrating low-cost imagery, such as SPOT-6/7 (1.5 m of spatial resolution for pan-sharpened images), which could provide an interesting cost-efficient compromise between MR and VHR.

Regarding the performance of the dasymetric mapping procedure presented here, it is important to mention that the accuracy of the predictions at the grid layer are probably lower than the one presented here (with a validation at level 1, after reaggregation of grid level estimates). Since grid population products tend to be commonly used in different fields of research, it is essential to inform end-users about the confidence and limitations of such products, as much as about their advantages. When using such population models, the end-user should always keep in mind that *"The most that can be expected from any model is that it can supply a useful approximation to reality: All models are wrong; some models are useful"* [33] (p. 440).

With regard to top-down dasymetric approaches, such as the one presented here, we should mention that they are completely dependent on official censuses which are frequently criticized regarding their reliability [6]. Furthermore, in the best case, official censuses are usually organized once in a decade, but in developing countries this rate is not systematically respected. This often forces studies to deal with population data that are asynchronous with ancillary geoinformation, used as covariates. When the official population data are not available or are outdated, remotely-sensed data can be used to support bottom-up approach [6,7]. The latter uses population counts coming from micro-census surveys—i.e., the collection of census information through field surveys on a limited portion of the territory—to extrapolate the population count on the rest of the territory, thus allowing implementation with limited human and financial capacity, compared to a regular census [6].

Additionally, we highlighted that, in a RF-based dasymetric mapping procedure, the OOB-score is not guaranteed to effectively help in the identification of the best-performing combination of input layers, in terms of reallocation accuracy. Therefore, a recommendation is made for future studies to exercise caution when using the OOB-score as an indicator of the performance of dasymetric mapping. More generally, the sensitivity of RF-based top-down population models to the scale factor and the MAUP effect is poorly explored in the literature. It would be beneficial for the field to further investigate the impact of the quantity and the spatial resolution of the administrative units used to train the RF model, and the impact of the difference of spatial resolutions between training and prediction (grid) levels.

5. Conclusions

Dasymetric mapping has been used to provide estimation of population densities at a finer scale than the available official data released in administrative units. To this end, this method relies on the use of ancillary data—such as settlement layers, land cover, or land use maps—that are used

as proxies of the real spatial distribution of the population within the administrative units. On one hand, MR satellite imagery can be used to derive such ancillary data at no cost. Unfortunately, when working at the intra-urban level, these data often fail to provide sufficient details in the population model. On the other hand, VHR satellite imagery can provide very detailed information and thus improve the quality of the population models, but their acquisition is much more expensive. In this research, we assessed the added value of VHR remote-sensing derived products, compared to MR products, when used as ancillary data in a dasymetric mapping procedure. When using a simple binary built-up/non-built-up mask, we showed that the use of VHR resulted in a drop of 13% in the error (RTAE). Moreover, our results showed that the use of the spatially and thematically detailed information, which can be derived from VHR land cover and land use maps, which are useful pieces of information by themselves, enabled significant improvement in the dasymetric reallocation accuracy, compared to what can be achieved using a MR built-up mask alone.

Author Contributions: Conceptualization, methodology, T.G. and C.L.; formal analysis and validation, T.G.; resources, A.G.; data curation, S.G. and T.G.; writing and editing, T.G., N.M., S.G., S.V., and M.L.; supervision, M.L. and E.W.; project administration, funding acquisition, C.L., S.V., and E.W.

Funding: This research was funded by BELSPO (Belgian Federal Science Policy Office) in the frame of the STEREO III program, as part of the MAUPP (http://maupp.ulb.ac.be) (SR/00/304) and REACT (http://react.ulb.be/) (SR/00/337) projects.

Acknowledgments: The authors gratefully thanks the ASSESS project (http://assess-sn.org/), funded by the ARES-CDD (https://www.ares-ac.be), which provided the access to the population data. The authors sincerely acknowledge the reviewers for their relevant comments which helped to improve this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

All the computed code that supported this research is available from the following repository: https://github.com/tgrippa/Dasymetric_mapping_using_GRASSGIS.

The gridded population layer produced in this study is accessible in a dedicated repository (CC-BY license): https://doi.org/10.5281/zenodo.2525671.

The data sets used as input for the analysis presented in this paper are as follows. These are all available under CC-BY license.

- The MR built-up layer is available from https://doi.org/10.5281/zenodo.1450931.
- The VHR land-cover map is available from https://doi.org/10.5281/zenodo.1290799.
- The VHR land-use map is available from https://doi.org/10.5281/zenodo.1291388.

References

1. UN DESA. *World Urbanization Prospects: The 2018 Revision, Online Edition*; United Nations Department of Economic and Social Affairs: New York, NY, USA, 2018.
2. United Nations Human Settlements Programme (UN-Habitat). *The State of African Cities, 2014: Re-Imagining Sustainable Urban Transitions*; UN-Habitat: Nairobi, Kenya, 2014.
3. Steiner, P.; Paulus, G. Dasymetric mapping for public health planning. In Proceedings of the 10th AGILE International Conference on Geographic Information Science, Aalborg, Denmark, 8–11 May 2007.
4. Ehrlich, D.; Lang, S.; Laneve, G.; Mubareka, S.; Schneiderbauer, S.; Tiede, D. Can Earth Observation help to improve information on population? Indirect Population Estimations from EO Derived Geo-Spatial Data: Contribution from GMOSS. In *Remote Sensing from Space: Supporting International Peace and Security*; Springer: Dordrecht, The Netherlands, 2009; pp. 211–237.
5. Su, M.D.; Lin, M.C.; Hsieh, H.I.; Tsai, B.W.; Lin, C.H. Multi-layer multi-class dasymetric mapping to estimate population distribution. *Sci. Total Environ.* **2010**, *408*, 4807–4816. [CrossRef] [PubMed]
6. Wardrop, N.A.; Jochem, W.C.; Bird, T.J.; Chamberlain, H.R.; Clarke, D.; Kerr, D.; Bengtsson, L.; Juran, S.; Seaman, V.; Tatem, A.J. Spatially disaggregated population estimates in the absence of national population and housing census data. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 3529–3537. [CrossRef] [PubMed]

7. Weber, E.M.; Seaman, V.Y.; Stewart, R.N.; Bird, T.J.; Tatem, A.J.; McKee, J.J.; Bhaduri, B.L.; Moehl, J.J.; Reith, A.E. Census-independent population mapping in northern Nigeria. *Remote Sens. Environ.* **2018**, *204*, 786–798. [CrossRef] [PubMed]
8. United Nations Economic and Social Council. *Report of the Inter-Agency and Expert Group on Sustainable Development Goal Indicators*; United Nations Economic and Social Council: New York, NY, USA, 2016.
9. Langford, M. Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps. *Comput. Environ. Urban Syst.* **2007**, *31*, 19–32. [CrossRef]
10. Langford, M.; Unwin, D.J. Generating and mapping population density surfaces within a geographical information system. *Cartogr. J.* **1994**, *31*, 21–26. [CrossRef]
11. Openshaw, S. The modifiable areal unit problem. *Concepts Tech. Mod. Geogr.* **1984**, *38*.
12. Mennis, J. Generating Surface Models of Population Using Dasymetric Mapping. *Prof. Geogr.* **2003**, *55*, 31–42.
13. Wu, S.; Qiu, X.; Wang, L. Population Estimation Methods in GIS and Remote Sensing: A Review. *GISci. Remote Sens.* **2005**, *42*, 80–96. [CrossRef]
14. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
15. Stevens, F.R.; Gaughan, A.E.; Linard, C.; Tatem, A.J. Disaggregating census data for population mapping using random forests with remotely-sensed and other ancillary data. *PLoS ONE* **2015**. [CrossRef]
16. Nieves, J.J.; Stevens, F.R.; Gaughan, A.E.; Linard, C.; Sorichetta, A.; Hornby, G.; Patel, N.N.; Tatem, A.J. Examining the correlates and drivers of human population distributions across low- and middle-income countries. *J. R. Soc. Interface* **2017**, *14*, 20170401. [CrossRef] [PubMed]
17. Tobler, W.R. Smooth pycnophylactic interpolation for geographical regions. *J. Am. Stat. Assoc.* **1979**, *74*, 519–530. [CrossRef] [PubMed]
18. Forget, Y.; Shimoni, M.; Gilbert, M.; Linard, C. Complementarity Between Sentinel-1 and Landsat 8 Imagery for Built-Up Mapping in Sub-Saharan Africa. *Preprints* **2018**. [CrossRef]
19. Forget, Y.; Linard, C.; Gilbert, M. Automated supervised classification of Ouagadougou built-up areas in Landsat scenes using OpenStreetMap. In Proceedings of the 2017 Joint Urban Remote Sensing Event (JURSE), Dubai, UAE, 6–8 March 2017; pp. 1–4.
20. Forget, Y.; Linard, C.; Gilbert, M.; Shimoni, M.; Lopez, J. Fusion Scheme for Automatic and Large-Scaled Built-up Mapping. In Proceedings of the IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, Spain, 22–27 July 2018; pp. 2072–2075.
21. Grippa, T.; Lennert, M.; Beaumont, B.; Vanhuysse, S.; Stephenne, N.; Wolff, E. An Open-Source Semi-Automated Processing Chain for Urban Object-Based Classification. *Remote Sens.* **2017**, *9*, 358. [CrossRef]
22. Grippa, T.; Georganos, S.; Vanhuysse, S.; Lennert, M.; Wolff, E. A local segmentation parameter optimization approach for mapping heterogeneous urban environments using VHR imagery. In *Remote Sensing Technologies and Applications in Urban Environments II*; International Society for Optics and Photonics: Bellingham, WA, USA, 2017.
23. Georganos, S.; Grippa, T.; Lennert, M.; Vanhuysse, S.; Wolff, E. SPUSPO: Spatially Partitioned Unsupervised Segmentation Parameter Optimization for efficiently segmenting large heterogeneous areas. In Proceedings of the 2017 Conference on Big Data from Space (BiDS'17), Toulouse, France, 28–30 November 2017.
24. Brousse, O.; Georganos, S.; Demuzere, M.; Vanhuysse, S.; Wouters, H.; Wolff, E.; Linard, C.; van Lipzig, N.P.M.; Dujardin, S. Using Local Climate Zones in Sub-Saharan Africa to tackle urban health issues. *Urban Clim.* **2019**, *27*, 227–242. [CrossRef]
25. Grippa, T.; Georganos, S.; Zarougui, S.; Bognounou, P.; Diboulo, E.; Forget, Y.; Lennert, M.; Vanhuysse, S.; Mboga, N.; Wolff, E. Mapping Urban Land Use at Street Block Level Using OpenStreetMap, Remote Sensing Data, and Spatial Metrics. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 246. [CrossRef]
26. Agence Nationale de la Statistique et de la Démographie (ANSD). *Rapport définitf du RGPHAE 2013*; ANSD: Dakar, Senegal, 2013.
27. Reed, F.; Gaughan, A.; Stevens, F.; Yetman, G.; Sorichetta, A.; Tatem, A. Gridded Population Maps Informed by Different Built Settlement Products. *Data* **2018**, *3*, 33. [CrossRef]
28. Batista e Silva, F.; Gallego, J.; Lavalle, C. A high-resolution population grid map for Europe. *J. Maps* **2013**, *9*, 16–28. [CrossRef]

29. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
30. Neteler, M.; Bowman, M.H.; Landa, M.; Metz, M. GRASS GIS: A multi-purpose open source GIS. *Environ. Model. Softw.* **2012**, *31*, 124–130. [CrossRef]
31. Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J.; Grout, J.; Corlay, S.; et al. Jupyter Notebooks—A publishing format for reproducible computational workflows. In Proceedings of the 20th International Conference on Electronic Publishing, Göttingen, Germany, 7–9 June 2016; pp. 87–90.
32. Linard, C.; Tatem, A.J.; Gilbert, M. Modelling spatial patterns of urban growth in Africa. *Appl. Geogr.* **2013**, *44*, 23–32. [CrossRef] [PubMed]
33. Box, G.E.P.; Hunter, J.S.; Hunter, W.G. *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd ed.; Wiley Series in Probability and Statistics; Wiley-Interscience: Hoboken, NJ, USA, 2005.