Spatial Disaggregation of Population Subgroups Leveraging Self-Trained Multi-Output Gradient Boosting Regression Trees

Georgati, Marina; Monteiro, João; Martins, Bruno ; Keßler, Carsten

Published in:
25th AGILE Conference on Geographic Information Science

DOI (link to publication from Publisher):
10.5194/agile-giss-3-5-2022

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

AGILE: GIScience Series, 3, 5, 2022. https://doi.org/10.5194/agile-giss-3-5-2022
Proceedings of the 25th AGILE Conference on Geographic Information Science, 2022.
Editors: E. Parseliunas, A. Mansourian, P. Partsinevelos, and J. Suziedelyte-Visockiene.
This contribution underwent peer review based on a full paper submission.

Spatial Disaggregation of Population Subgroups Leveraging Self-Trained Multi-Output Gradient Boosting Regression Trees

Marina Georgati [1], João Monteiro [2], Bruno Martins [2], and Carsten Keßler [3,1]

[1]Department of Planning, Aalborg University, Copenhagen, Denmark
[2]Instituto Superior Técnico and INESC-ID, Universidade de Lisboa, Lisboa, Portugal
[3]Department of Geodesy, Bochum University of Applied Sciences, Bochum, Germany

Correspondence: Marina Georgati (marinag@plan.aau.dk)

Abstract. Accurate and consistent estimations on the present and future population distribution, at fine spatial resolution, are fundamental to support a variety of activities. However, the sampling regime, sample size, and methods used to collect census data are heterogeneous across temporal periods and/or geographic regions. Moreover, the data is usually only made available in aggregated form, to ensure privacy. In an attempt to address these issues, several previous initiatives have addressed the use of spatial disaggregation methods to produce high-resolution gridded datasets describing the human population distribution, although these projects have usually not addressed specific population subgroups. This paper describes a spatial disaggregation method based on self-training regression models, innovating over previous studies in the simultaneous prediction of disaggregated counts for multiple inter-related variables, by leveraging multi-output models based on gradient tree boosting. We report on experiments for two case studies, using high-resolution data (i.e., counts for different subgroups available at a resolution of 100 meters) for the municipality of Amsterdam and the region of Greater Copenhagen. Results show that the proposed approach can capture spatial heterogeneity and the dependency on local factors, outperforming alternatives (e.g., seminal disaggregation algorithms, or approaches leveraging individual regression models for each variable) in terms of averaged error metrics, and also upon visual inspection of spatial variation in the resulting maps.

Keywords. spatial disaggregation, gridded population datasets, gradient tree boosting, self-supervised learning

1 Introduction

Accessing socio-demographic data at high resolution is still challenging in many parts of the world, despite the wide availability of vast amounts of population data. This is usually due to limitations on the means and infrastructure for a regular census in developing countries, or privacy restrictions in developed ones. Nonetheless, accurate and consistent estimations on present and future population distribution, at fine spatial resolution, are fundamental to support public administrative functions in various sectors (Mennis, 2009; Lloyd et al., 2019; Qiu et al., 2022). Example applications relate to the environment (e.g., for environmental impact assessment), healthcare (e.g., for modelling epidemics (Hay et al., 2005)), economy (e.g., for studying inequality and segregation (Catney and Lloyd, 2020)), or urban development (e.g., for service planning and delivery (Langford et al., 2008)).

In order to address the lack fine resolution data, numerous attempts have been made at downscaling spatial population datasets, either by using traditional or novel computational approaches. Since access to aggregated population counts according to districts, municipalities, or even census tracts is very common, spatial disaggregation procedures have often been considered to support the production of harmonised, reliable, temporally regular, and spatially detailed datasets on population distribution. These methods have ranged in complexity, from simpler mass preserving areal weighting (Goodchild et al., 1993), to pycnophylactic interpolation capable of producing '*smooth maps*' (Tobler, 1979), or dasymetric weighting schemes that recognise that locations and population density are not homogeneous across space, depending on local factors. Recent methods for spatial data downscaling leverage regression modelling and machine learning (Stevens et al., 2015; Monteiro et al., 2018), combining ancillary information from different sources (e.g., satellite imagery or mobile phone data) in the definition of the dasymetric weights supporting the disaggregation. Still, and although modern procedures are reported to achieve accurate results, most previous studies have not addressed the disaggregation of specific population subgroups (e.g., age groups), instead focusing on the application to entire populations.

This study reports on an adjustable and reproducible methodology for disaggregating interrelated data consisting of multiple socio-demographic variables (e.g., population counts corresponding to different age groups, or different geographic regions of origin). Our approach is useful for situations where fine resolution data, concerning multiple variables, are required, but only aggregate data is available. Going beyond simple population density, towards more detailed socio-demographic variations and their distributions in the urban fabric, can contribute to applications that need to account with socially and geographically balanced allocation. Our contributions relate to improving spatial disaggregation results in the particular setting of multiple interrelated variables.

The core component of the proposed methodology is a spatial disaggregation approach based on the self-training of a regression model combining multiple ancillary variables, considering the simultaneous prediction of disaggregated counts for multiple inter-related variables, by leveraging multi-output models based on gradient tree boosting. Our main research contributions are as follows:

1. We propose a spatial downscaling method based on self-training, relying on gradient tree boosting to disaggregate counts associated to polygonal regions, into high resolution grid cells. The method advances over previous work by simultaneously predicting multiple inter-related values, corresponding to different socio-demographic variables of interest.
2. The proposed method was examined in two case studies involving densely populated urban areas, where land uses can be distinguished in broad categories. We specifically used high-resolution data (i.e., counts for different population subgroups, available at a resolution of 100 meters) for the the municipality of Amsterdam and the region of Greater Copenhagen.
3. We report on a quality assessment of the downscaled results, based on ground truth data at the same target resolution. Ours is thus one of the first studies that directly evaluates spatial disaggregation results at the level of high-resolution grid cells.
4. We show that our methodology is convenient for studies involving multiple socio-demographic variables, compared to separate single-output models. High quality results can be obtained with a single model, simplifying the overall disaggregation procedure and lowering the computational requirements.

The remainder of this document is structured as follows: Section 2 presents related work in the task of geospatial data disaggregation. Section 3 describes the proposed disaggregation approach, followed by the introduction of the used datasets and tools in Section 4. Section 5 presents the produced results for our two case study areas, and Section 6 summarises our conclusions and points out directions for future research. Lastly, Section 7 describes the data and software availability.

2 Related Work

This section describes traditional methods for spatial data disaggregation, followed by a survey of recent approaches.

2.1 Seminal Spatial Disaggregation Methods

The simplest spatial disaggregation method is perhaps mass-preserving areal weighting (Goodchild et al., 1993), in which the known counts associated to source administrative regions (e.g., the population associated to coarse administrative districts) are divided uniformly across their area, in order to produce estimates at target regions of higher spatial resolution. The total estimated value for a target zone is thus a weighted sum of the fractional count values from all source zones falling within the target zone.

Pycnophylactic interpolation (Tobler, 1979) can be seen as an extension of simple areal weighting, breaking the homogeneity assumption and assuming a degree of spatial auto-correlation in the variable being downscaled (i.e., areas that are close to one another should have similar values). The method starts by applying mass-preserving areal weighting, afterwards smoothing the resulting values by replacing them with the average of their neighbours (e.g., the adjacent cells in a raster grid). The aggregation of the predicted values for all zones within a source region is then compared with the original value, and adjusted to keep the consistency within the source regions. The method continues iteratively until there is either no significant difference between predicted values and actual values within the source regions, or until there have been no significant changes from the previous iteration.

In turn, dasymetric schemes use a weighted surface to distribute the source counts, instead of considering a uniform (or a smooth) distribution of the target values, as in the previous methods. The weighted surface can reflect ancillary spatial data such as land coverage, masks corresponding to building footprints, or night-time light emissions, to facilitate the disaggregation. The general idea is to apply weights for different source area types, which represent the percentage of the target variable that is likely to be contained within that area type. The main challenge in dasymetric disaggregation involves finding an appropriate set of weights to accurately reflect the distribution of the variable that is to be disaggregated. While some schemes use simple binary masks built from land coverage data (i.e., using data on water bodies or building footprints, to denote regions that should receive a fraction of the total value), other approaches rely on expert knowledge and manually-defined rules to define fractional dasymetric weights. More recent methods leverage machine learning to improve upon the heuristic definition of weights.

2.2 Machine Learning for Spatial Disaggregation

Several previous research initiatives have resulted in the production of openly available high-resolution gridded

datasets that describe the human population distribution, leveraging machine learning methods to combine ancillary information from different sources (e.g., satellite imagery or mobile phone data) in the definition of weights supporting the disaggregation. Well-known examples include the datasets made available in the context of the Gridded Population of the World (GPW), Global Human Settlement Layer (GHSL), or WorldPop projects. Still, it should be noted that most previous efforts have not considered the creation of datasets focusing on different population subgroups (e.g., according to age or gender).

In the context of the WorldPop project, Stevens et al. (2015) developed a technique for creating gridded predictions of population density with a resolution of approximately 100×100 meters, through a dasymetric approach that leverages random forest regression models. Among others, ancillary datasets which incorporate information on land coverage, digital elevation, the road network, and water bodies, were taken into account when estimating a weighting surface to perform dasymetric redistribution of census counts, originally at a country level, into the raster grid cells. The proposed approach relies on a multi-stage estimation technique, which tunes the number of covariates within the random forest model that estimates density from the ancillary variables. The density maps are computed from aggregated data available at coarse regions (e.g., provinces), and they are then used in a standard dasymetric approach for obtaining the population counts at each raster cell. For evaluation, the cells within each of the finer census units (i.e., villages or sub-locations) were summed and compared with the corresponding known counts, through metrics such as the Mean Absolute Error (MAE) or the Root Mean Squared Error (RMSE).

In another recent study exploring the use of regression analysis to infer dasymetric weights, Cheng et al. (2020) reported on the disaggregation of census data for China into a raster grid with a resolution of 1×1 km per cell for each month in 2015. The authors combined environmental information and mobile phone positioning data as the ancillary variables used to infer the dasymetric weights. The disaggregation method corresponds to a hybrid inference approach, combining random forests with area-to-point kriging. The random forest model is trained with data at the town level, aggregating the ancillary data (i.e., taking the mean values per town as the independent variables) and using the population density as the target variable. The model is then used to produce population estimates for the target cells, which are re-aggregated to the town level for computing the areal residuals for each town. The area-to-point kriging model finally uses this information to adjust the random forest predictions under the assumption that the sum of the encompassing residuals at the pixel level should match the town's residual.

Instead of training regression models with aggregated data, other studies have instead proposed to estimate models directly with data at the target resolution, e.g. through self-training procedures. For instance Vargas-Munõz et al. (2022) briefly described a method based on Markov Random Fields (MRF) that iteratively improves the initial estimations of a dasymetric disaggregation method. During the iterations, the MRF-based method minimises an energy function that encourages 100 meter cells with similar features to have similar population predictions, while at the same time ensuring that the predictions sum up to a value close to the available regional census data. Monteiro et al. (2018, 2019, 2021) proposed spatial disaggregation methods based on iteratively refining initial estimates produced by seminal methods (e.g., pycnophylactic interpolation or dasymetric mapping leveraging heuristic weights), by self-training different types of regression models. At each step, the previous predictions are used as the targets for training a regression model, which is then used to produce new estimates. The regression estimates are adjusted in order to ensure consistency with the source region counts (i.e., to enforce the pycnophylactic property), and the process is repeated for a fixed number of steps, or until no relevant changes are detected. Given the good results reported by Monteiro et al., this general method was also considered in the present study, adapting the procedure to the simultaneous downscaling of multiple variables.

3 The Proposed Method

The spatial disaggregation method used in our experiments relies on a self-training approach that combines weighted interpolation and regression-based dasymetric mapping. We specifically extended the method proposed by Monteiro et al. (2018, 2019, 2021), which in turn is adapted from a method described by Malone et al. (2012) for general spatial data downscaling. The approach is said to rely on self-training, in the sense that initial estimates are first computed through a simple disaggregation heuristic (e.g., weighted interpolation leveraging pre-existing population datasets), that can be seen as a teacher model. These results are then used to train a regression model (i.e., a student model), whose predictions are iteratively refined (i.e., the student from one iteration is then used as the teacher model for the next). Even though we have access to the ground truth disaggregated values, we do not use them during model training to simulate scenarios in which the high-resolution data is not available (not even for a part of the study region, or for a similar region). The ground-truth counts are used only for evaluation purposes, in order to assess the quality of the proposed methodology. In the experiments reported in this paper, we execute a fixed number of self-training iterations (10) and retain the disaggregated values computed at that iteration. The general idea is presented in Figure 1 and can be described as follows:

1. We start from a vector polygon layer with the aggregated (source) population counts, as provided by statistical offices for relatively coarse regions.
2. Based on the aforementioned layer, we create a raster representation of disaggregated estimates, through a

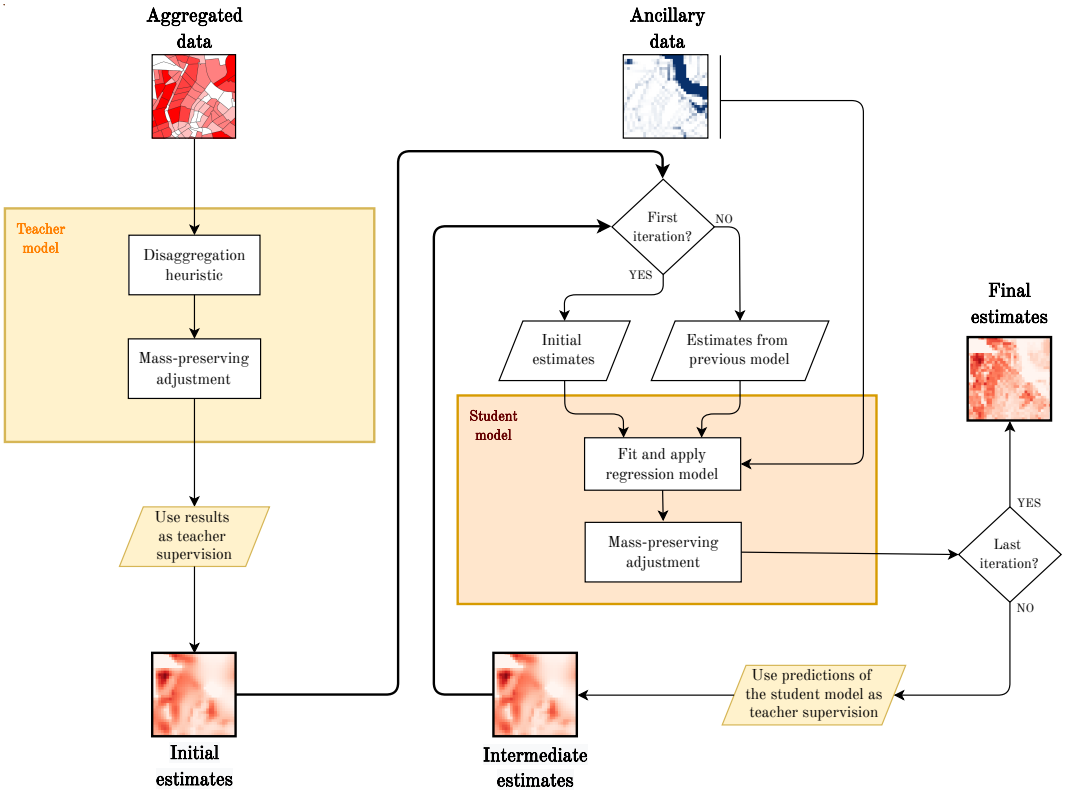

Figure 1. The different steps involved in the proposed spatial disaggregation method.

simple heuristic such as pycnophylactic interpolation (Tobler, 1979), or dasymetric mapping leveraging some heuristic weights.

3. We train a regression model to infer the results produced in Step 2 from ancillary information available as gridded rasters at the target resolution. After training, the regression model is used to produce new disaggregated values, leveraging and generalising patterns in the ancillary data to improve the predictions.

4. The values returned by the regression model in the previous step are proportionally adjusted to retain the original counts in the source zones.

5. Steps 3 and 4 are repeated, in order to adjust the disaggregated estimates and until reaching a maximum number of iterations.

The following subsections detail the heuristics that we tested for producing the initial estimates (Section 3.1), the regression algorithms that we considered for combining the different sources of ancillary data (Section 3.2), and the use of an appropriate loss function for model training, dealing with the characteristics of our data (Section 3.3).

3.1 Initial Estimates

We experimented with two different heuristics for producing the initial estimates: pycnophylactic interpolation (PI) and a dasymetric approach with heuristically defined weights for interpolation (WI). Spatial auto-correlation is the main driving factor in PI, with estimates computed on the assumption that regions close to each other tend to have similar values. However, no other properties are taken into account for estimating the distribution of the target variable and, consequently, results tend to be over-smoothed. This issue is addressed by our second approach, which disaggregates the data proportionally to weights derived from external information (in our case, derived from a pre-existing high-resolution population dataset). The method can be formalised through the following equation (Eq. 1), where W_t is the estimated count in a target zone t, S_s is the count in source zone s, P_t is the population count in target zone t, and P_s is the count in source zone s.

$$W_t = \sum_s \left(\frac{P_t}{P_s} \times S_s \right). \quad (1)$$

3.2 Regression Algorithms

We tested two different regression algorithms, based on ensembles of decision trees, to estimate the disaggregated values: random forests and gradient tree boosting. Given our objective of disaggregating counts for different population subgroups, we tested both algorithms on single- and multi-output scenarios.

Random forests (Breiman, 2001) correspond to one of the most popular regression algorithms, having been ex-

tensively used in a wide range of applications within the Geospatial Information Sciences (GIS) and demography (Georganos et al., 2021; Qiu et al., 2020; Stevens et al., 2015). It is an example of a bagging strategy with tolerance to overfitting, that improves accuracy by averaging several simple decision tree models (Verdhan, 2020).

In turn, gradient boosting is another ensemble method where different learners are trained sequentially, based on the results of previous ones. Learners improve in every step, by focusing more on the errors of the last iteration, while bias is simultaneously reduced (Verdhan, 2020). The foundations of gradient boosting rely on the frameworks of Freund and Schapire (1999) and Friedman (2001), and this type of model was selected in this study due to the high flexibility and customisability, in pair with very good results across a variety of tasks. Given the use of gradients for model training, this strategy also easily allows for the incorporation of customised loss functions (Natekin and Knoll, 2013). In our case, we used a task-specific loss function that combines linear and quadratic penalties, described in the following subsection.

3.3 A Task-Specific Loss Function for Training Multi-Output Gradient Boosting Models

In our experiments with multi-output gradient boosting models, we combined two of the most typical regression loss functions, namely the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE). Considering the objective of simultaneously predicting disaggregated counts for different population subgroups (e.g., the population according to different age groups, or according to different places of origin, noting that by summing the different groups we get the total population), we added separate penalties for errors in the predictions for the different groups, and errors in the total population.

The MAE and the RMSE respectively penalise the errors linearly and quadratically, making the RMSE more sensitive to large differences. Our customised loss function combines the RMSE between ground truth and predicted values for all the different subgroups being disaggregated, together with the MAE between the sums of the values within the groups. The loss function is presented in Equation (2), where y_i corresponds to the ground truth value of the variable i being disaggregated (i.e., for an individual population subgroup), while $\hat{y}_i$ corresponds to the predicted value of the same variable. We denote by G the set of groupings for the different variables.

$$\mathcal{L}\text{Combined}(y,\hat{y}) = \sum_{j=0}^{|G|} \sum_{i \in G_j} \text{RMSE}(y_i, \hat{y}_i) + \sum_{j=0}^{|G|} \text{MAE}\left(\sum_{i \in G_j} y_i, \sum_{i \in G_j} \hat{y}_i \right). \quad (2)$$

4 Data Sources

This section describes the pre-existing high-resolution population dataset used for producing the initial estimates, as well as the population datasets used for evaluating the produced results. The section also introduces the ancillary datasets used in the proposed approach. The data pre-processing and disaggregation experiments were performed with an open-source software architecture that primarily uses non-proprietary data formats and Python tools like GDAL, scikit-learn, and CatBoost. A few of the considered datasets are nonetheless of restricted use, due to confidentiality reasons. All spatial data used the ETRS89 Lambert Azimuthal Equal-Area (EPSG: 3035) geospatial coordinate system.

4.1 Population Data

Our spatial disaggregation experiments used a selection of socio-demographic groups corresponding to the data availability in two case study areas, corresponding to Amsterdam and the region of Greater Copenhagen.

For Amsterdam, the source units for disaggregation correspond to neighbourhoods, and the population is divided in 5 age groups – children (0-19 years), students (20-29 years), mobile adults (30-44 years), non mobile adults (45-65 years), and elderly (65+ years) – and 7 areas of origin – Natives, Western, Non-Western, and the 4 largest migrant groups: Suriname, Turkey, Antilles, and Morocco. Two datasets are provided by the Municipality (Onderzoek, Informatie en Statistiek – OIS), respectively at the neighbourhood level (OIS_n, https://data.amsterdam.nl) and at a 100 meters grid cell level (OIS_g).

In the case of Copenhagen, the data are provided by Statistics Denmark (DST) at the aggregated administrative level of municipalities (DST_m, https://www.statbank.dk). The age groups are the same from the Amsterdam case study, and two categories of migration background are explored: the general category includes three classes – Natives, EU, and non-EU – while the detailed category divides the population into seventeen geographic regions of origin (Statistics Division, United Nations).

Table 1 presents an overview of the population datasets. Counts are shown by demographic group as reference for each case study area, indicating the differences between the aggregated- and the grid cell datasets that are made available (our tests used exclusively the data made available at the level of grid cells, although small discrepancies exist in the official data reported at different aggregation levels). Groups with population lower than 1.000 people are omitted from the tables, but taken into consideration in the analysis. Lastly, it is worth mentioning that the source and target zones deviate significantly between the 2 cases. The mean area of the municipalities in Copenhagen is 60 times larger than the mean area of the neighbourhoods in Amsterdam (A_cph = 30.82km^2, A_ams =

Table 1. The population datasets used in our experiments.

Amsterdam			Copenhagen		
	OIS_g 7.513 cells	OIS_n 482 neighbourhoods		DST_g 21.615 cells	DST_m 17 municipalities
Total Population	862.973	862.987	Total Population	1.331.816	1.332.152
Age Groups					
Children	166.465	166.470	Children	286.975	286.995
Students	165.889	165.890	Students	244.584	244.672
Mobile Adults	212.799	212.802	Mobile Adults	298.115	298.263
Not Mobile Adults	209.843	209.846	Not Mobile Adults	307.503	307.577
Elderly	107.977	107.979	Elderly	194.635	194.645
Migration Status					
Natives	393.180	393.185	Natives	1.026.361	1.026.564
Western	160.565	160.566	EU	210.190	210.263
Non-Western	111.730	111.733	Not EU	209.379	210.306
Suriname	64.588	64.588	Australia and New Zealand	1.900	1.900
Turkey	44.054	44.054	Eastern Asia	10.122	10.123
Antilles	12.173	12.173	Eastern Europe	32.583	32.610
Morocco	76.683	76.688	Lat. Am. and the Caribbean	8.634	8.643
			Northern Africa	12.466	12.468
			Northern America	7.015	7.019
			Northern Europe	35.138	35.157
			South-eastern Asia	14.394	14.400
			Southern Asia	51.430	51.451
			Southern Europe	32.730	33.162
			Sub-Saharan Africa	17.989	18.005
			Western Asia	61.143	61.388
			Western Europe	18.527	18.535

$0.48km^2$), and Copenhagen covers almost 3 times more disaggregated units (E_{cph} = 407x281, E_{ams} = 236x179). In both cases, only approximately 18% of the cells in the examined areas are inhabited.

4.2 Preparation of the Ancillary Data

A wide range of ancillary layers were produced for this study, with a primary concern in the use of open data for facilitating reproducibility.

Specifically, the information on the distribution of the population, expressed in number of people per grid cell and used for an initial disaggregation heuristic, is provided by Schiavina et al. (2019) in the context of the Global Human Settlement Layer (GHSL) project, for the target year of 2015 and at the spatial resolution of 250 meters. This dataset (GHS-POP) is itself a product of disaggregation from census or administrative units into 250 meters grid cells, informed by the distribution and density of built-up areas, as mapped in the GHSL global layer. The original raster was re-projected and re-sampled with GDAL, using a combination of algorithms. The selection of the most accurate pre-processing strategy was based on the comparison of the produced rasters to the ground truth layers (OIS_g, DST_g). The lowest error was obtained when using nearest neighbours and cubic spline algorithms, respectively for re-projection and re-sampling. Despite its wide-spread use and recognition, the GHS-POP layer deviates significantly to the gridded ground truth datasets in our case study areas. For example, the GHS-POP population in the municipality of Amsterdam is revealed to be 830.352 people, 32.635 people lower than the ground truth in the OIS_n dataset, showing a mean error of 65 people per neighbourhood and 132 people among the grid cells.

A set of additional layers were considered as ancillary data. For instance the European Settlement Map (ESM) represents the human settlements in Europe for the year of 2015 (Sabo et al., 2019), classifying the built-up areas into residential and non-residential, at a spatial resolution of 10 meters. We used these data to represent the percentage of residential coverage at 100 meter grid cells. Five binary layers were also prepared at a resolution of 100 meters, obtained from the land coverage dataset of the Copernicus Land Monitoring Service (European Environment Agency (EEA)). These layers divide initially the artificial surfaces in urban fabric, transportation, and industrial areas; and secondly the natural environment in agricultural areas, forests and green spaces, and water bodies and wetlands. Due to the poor representation of the inner city water bodies, green spaces, and industrial areas, the produced binary layers were further processed and combined with corresponding Open Street Map datasets (OSM), resulting in scaled datasets of percentage coverage.

Apart from the aforementioned pan-European datasets, case specific layers were produced at high resolution, considering the major factors that affect the residential choices

Table 2. MAE and SD for the different approaches and different variables of interest, in Amsterdam and using 12 input layers (ams12).

	PI	WI	RFs100	RFm100	GB100	GB250	GB500
Total Population	12.4±36.5	11.1±33.9	9.6±30.9	11.1±33.9	12.1±41.4	11.4±38.5	**9.4±31.5**
Age Groups							
Children	2.8±8.6	2.6±8.3	2.3±7.6	2.6±8.3	2.7±9.7	2.5±9.0	**2.2**±7.6
Students	3.0±11.9	2.8±11.6	2.6±11.4	2.8±11.6	2.8±12.8	2.8±12.5	**2.5**±11.6
Mobile Adults	3.2±10.4	3.0±9.8	**2.6**±9.1	3.0±9.8	3.2±11.8	3.0±11.1	**2.6**±9.3
Not Mobile Adults	3.1±9.3	2.8±8.6	**2.4**±7.9	2.8±8.6	3.0±10.5	2.8±9.6	**2.4**±7.9
Elderly	1.8±6.5	1.7±6.2	1.6±6.2	1.7±6.2	1.8±7.0	1.7±6.8	**1.5**±6.1
Migration Background							
Natives	5.6±16.5	5.1±15.3	**4.4**±14.1	5.1±15.3	5.5±18.5	5.2±17.5	**4.4±14.5**
Western	2.5±8.1	2.3±7.7	**2.1**±7.4	2.3±7.7	2.4±8.8	2.3±8.5	**2.1**±7.4
Non-Western	2.0±7.5	1.9±7.4	1.8±7.0	1.9±7.4	1.9±7.9	1.9±7.6	**1.7**±7.0
Suriname	1.2±5.0	1.1±4.8	1.1±4.7	1.1±4.8	1.2±5.4	1.1±5.0	**1.0**±4.5
Turkey	0.9±3.8	**0.8**±3.6	**0.8**±3.5	0.8±3.6	0.9±4.6	0.9±4.3	**0.8**±3.5
Antilles	**0.3**±1.2	**0.3**±1.2	**0.3**±1.2	**0.3**±1.2	**0.3**±1.3	**0.3**±1.4	**0.3**±1.2
Morocco	1.6±7.1	1.5±6.8	1.4±6.5	1.5±6.8	1.6±8.5	1.6±7.9	**1.3**±6.3

of different demographic groups. The majority of these layers are common for both case study areas. The layers, their sources, and processing details, are as follows:

1. Proximity to railway stations, schools, universities, and cultural spaces like cinemas or theatres. They show the total number of accessible services in a biking distance of 15' with average biking speed of 15 km/h (Amsterdam: OSM, Copenhagen: Kortforsyningen, Bygnings- og Boligregistret (BBR)).
2. Proximity to bus stops. It shows the total number of accessible bus stops in a walking distance of 10' with average walking speed of 5 km/h (Amsterdam: OSM, Copenhagen: Movia Trafik).
3. Building height, volume, and construction year. They represent the average height, volume, and construction year of the buildings in the corresponding grid cell (Amsterdam: PDOK, Copenhagen: Bygnings- og Boligregistret (BBR)).
4. Real estate prices. It shows the average purchase price of the sold buildings from 2010 to 2020 (Copenhagen: Bygnings- og Boligregistret (BBR)).

5 Experimental Results

Our analysis was developed in various steps, starting from experiments around the combinations of the ancillary input layers, and leading to the comparison of the self-trained multi-output approach against baselines.

The predicted results are directly evaluated against the ground truth data, with Sections 5.1 and 5.2 presenting the numerical assessment and visual inspection, respectively.

The most suitable combination of ancillary data was selected after a series of small experiments with a restricted collection of variables. We report here the results of the best configuration for each case. For Amsterdam, we used 12 input layers (ams12) – residential coverage (ESM); green spaces (Corine/OSM); water coverage (Corine/OSM); industrial areas (Corine/OSM); transportation (Corine); proximity to bus stops, railway stations, schools, and universities; construction year; building volume and height. For Copenhagen, only 3 detailed layers were used (cph3) – construction year, building height, and real estate prices. In the end, we further report an additional combination of ancillary data for the case of Copenhagen, using 12 input layers.

5.1 Evaluation with Error Metrics

The error metrics considered in our analysis are the Mean Absolute Error (MAE), the Standard Deviation of the error (SD), and the Percentage Accuracy (PE), as described in Equations 3, 4, and 5.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}. \tag{3}$$

$$\mathrm{SD} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - x_i\right)^2}{n}}. \tag{4}$$

$$\mathrm{PE} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - x_i}{x_i} \right) \times 100. \tag{5}$$

Tables 2 and 3 present result quality scores for different disaggregation methods, highlighting the best output for each particular case study area. Each table includes the main demographic groups row-wise, and the MAE/SD of the explored models in each column – Pycnophylactic Interpolation (PI), Weighted Interpolation (WI), single-output Random Forest Regression (RFs), multi-output

Table 3. MAE and SD for the different approaches and different variables of interest, in Copenhagen and using 3 input layers (cph3).

	PI	WI	RFs100	RFm100	GB100	GB250	GB500
Total Population	13.1±36.8	9.2±30.5	7.9±27.6	8.8±29.1	8.7±30.1	7.7±28.2	**7.4**±27.4
Age Groups							
Children	2.9±7.6	2.1±6.5	**1.9**±6.1	2.0±6.3	2.1±6.5	**1.9**±6.2	**1.9**±6.1
Students	2.8±11.8	2.2±10.5	1.9±10.2	2.0±10.2	2.0±10.8	**1.8**±10.3	**1.8**±10.4
Mobile Adults	3.1±10.1	2.3±8.5	2.0±7.9	2.1±8.1	2.2±8.6	2.0±8.1	**1.9**±8.0
Not Mobile Adults	3.0±8.0	2.2±6.8	1.9±6.2	2.1±6.6	2.1±6.5	1.9±6.2	**1.8**±6.0
Elderly	2.1±5.8	1.6±5.3	1.5±5.1	1.6±5.3	1.6±5.1	**1.4**±4.9	**1.4**±4.9
Migration Background							
EU	2.6±10.0	2.2±9.4	2.0±9.3	2.1±9.5	2.1±9.4	**1.9**±9.0	**1.9**±9.1
Not EU	2.6±9.9	2.2±9.4	2.0±9.3	2.1±9.4	2.1±9.3	**1.9**±9.0	**1.9**±9.0
Natives	9.9±27.2	6.9±22.1	5.9±19.4	6.6±20.7	6.3±21.3	5.9±20.2	**5.6**±19.0
Australia and New Zealand	**0.0**±0.3	**0.0**±0.3	**0.0**±0.3	**0.0**±0.3	**0.0**±0.3	**0.0**±0.3	**0.0**±0.3
Eastern Asia	0.2±0.8	**0.1**±0.8	**0.1**±0.7	**0.1**±0.7	**0.1**±0.7	**0.1**±0.7	**0.1**±0.8
Eastern Europe	0.4±1.7	0.4±1.6	0.3±1.6	0.4±1.6	0.4±1.6	**0.3**±1.5	0.3±1.6
Latin America and the Caribbean	**0.1**±0.5	**0.1**±0.5	**0.1**±0.5	**0.1**±0.5	**0.1**±0.5	**0.1**±0.5	**0.1**±0.5
Northern Africa	**0.2**±1.0	**0.2**±1.0	**0.2**±1.0	**0.2**±1.0	**0.2**±1.0	**0.2**±1.0	**0.2**±1.0
Northern America	0.1±0.6	0.1±0.6	0.1±0.6	0.1±0.6	0.1±0.6	0.1±0.6	0.1±0.6
Northern Europe	0.4±1.8	0.3±1.7	0.3±1.6	0.3±1.7	0.3±1.6	0.3±1.6	0.3±1.6
South-eastern Asia	**0.2**±0.7	**0.2**±0.7	**0.2**±0.7	**0.2**±0.7	**0.2**±0.7	**0.2**±0.7	**0.2**±0.7
Southern Asia	0.7±2.9	**0.6**±2.8	0.6±2.7	**0.6**±2.8	**0.6**±2.8	**0.6**±2.7	**0.6**±2.7
Southern Europe	0.4±1.8	0.4±1.7	**0.3**±1.7	0.4±1.7	0.4±1.7	**0.3**±1.7	**0.3**±1.7
Sub-Saharan Africa	0.3±1.5	**0.2**±1.5	**0.2**±1.5	**0.2**±1.5	**0.2**±1.5	**0.2**±1.5	**0.2**±1.6
Western Asia	0.9±4.0	0.8±3.9	0.7±3.7	0.8±3.9	0.8±3.8	**0.7**±3.6	**0.7**±3.6
Western Europe	**0.2**±1.1	**0.2**±1.0	**0.2**±1.0	**0.2**±1.0	**0.2**±1.0	**0.2**±1.0	**0.2**±1.0

Random Forest Regression (RFm), and multi-output Gradient Boosting (GB). Each regression model name is accompanied by the selected number of estimators, indicating the number of trees used to build the ensemble. We include the results of 100 trees for random forests, and 100, 250, and 500 trees for gradient boosting. For RFs, the predictions for each variable are estimated by separate models (≈ 1.4'/iteration), while for RFm, the predictions for all groups are estimated simultaneously by one single regression model (≈ 10'/iteration). Similarly, for the case of GB, all groups are predicted together (≈ 600'/iteration, although we did not use Graphics Processing Units (GPUs) for model training), except for the total population which is later calculated as the sum of the corresponding age groups. The loss function in this last model counts also for the difference between the predictions and the real values in terms of the sum of two or three population subgroups, according to the examined area.

In terms of the MAE, the multi-output gradient boosting regressor with increased estimators performs better than the rest of the models, both in Amsterdam and Copenhagen. It slightly outperforms RFs for most of the features, or achieves the same numerical scores. According to Tables 2 and 3, the MAE of RFs100 and GB500 for the total population is less than 10 persons in both cases, with a standard deviation of error up to 30 persons. The MAE in Copenhagen is even lower, at 7.4 persons, despite the fact that the aggregated units are larger than the ones in Amsterdam. Error is also significantly decreased compared to the heuristic estimates of both the PI and WI methods, for all the age groups and the largest groups by migration background. However, all models produce similar scores for the smaller migrant groups. The small number of persons in these groups, and their high concentration in neighbouring areas, are well represented in the initial estimates of the WI heuristic, explaining these similarities. Lastly, the high percentage of non-inhabited cells for these small groups has a great impact on the results. With the MAE and SD only, it is difficult to estimate the spatial accuracy.

In spite of combining multiple target variables at once, the performance of RFm is poor and can only be compared to PI, WI, and GB100. The quality of its results does not improve even if the model's capacity and training time increase significantly, as Table 4 indicates. Moreover, a reasonable decrease is noticed in the errors for the GB models, as their capacity increases. Even though the first experiments with 100 estimators showed a low performance, they presented interesting spatial patterns with richer variability (further discussed in Section 5.2) than RFs. The error falls significantly at 500 estimators, at the cost of a much higher training time (although this can be significantly reduced if training the GB models on a GPU).

5.2 Visual Inspection of Maps and Errors

The disaggregation predictions can be mapped and compared directly to the ground-truth gridded data, giving us the opportunity to evaluate the performance of the models based on their spatial accuracy at high resolution.

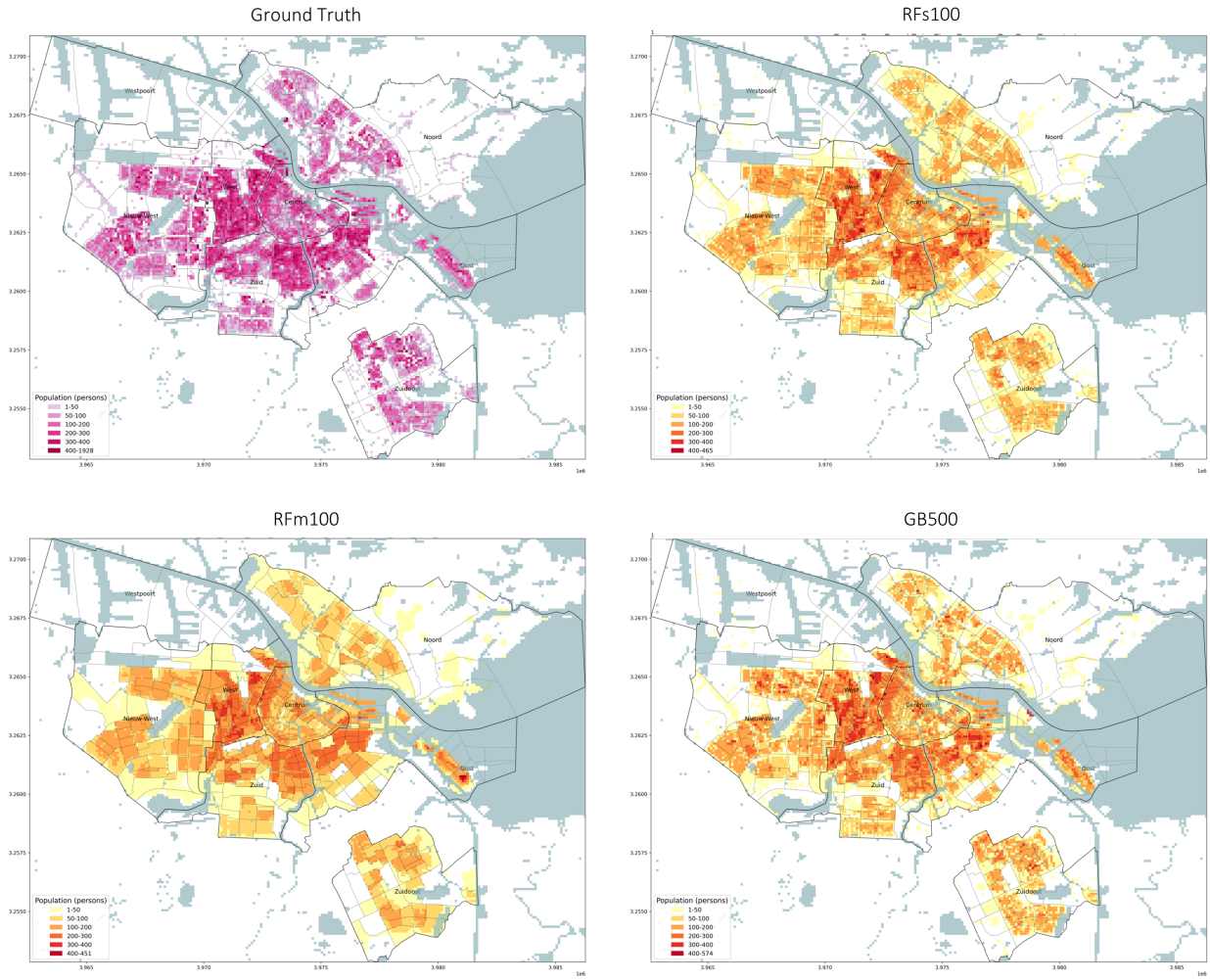

Figure 2. Distribution for the total population in Amsterdam, comparing ground truth counts (top-left) versus predictions: RFs (top-right), RFm (lower-left), GB500 (lower-right).

Due to the fact that each case study includes many variables, corresponding to different population subgroups, we choose to focus our detailed analysis on the total population (Fig. 2 and 3), although also presenting some indicative variables based on the size of the subgroups (Fig. 5).

Specifically, Figures 2 and 3 illustrate the distribution of the total population predicted by the 3 main models, in tandem to the ground truth distribution as reported by the statistical offices of the case study areas. It is evident in both cases that the RFs100 and the GB500 predictions show high spatial heterogeneity that approximates the real values. On the other hand, as expected from the MAE results, the population distribution predicted by RFm100 remains homogenised in each of the aggregated zones, similar to its original input layer, which was produced by the weighted interpolation method. Table 4 reports that this deficiency is not enhanced by increasing the capacity of the multivariate random forest and, regardless of its high flexibility and short training period, RFm is deemed unsuitable for accounting spatial heterogeneity at high resolution.

Considering the 2 best models – RFs100 and GB500 – in particular, their differences are relatively small and vary between the cases. Figure 4 represents the percentage error by grid cell for these 2 models, for each case, to facilitate their comparison and highlight their strengths and weaknesses. Starting from Amsterdam, RFs100 recognises more accurately the densely populated cells of West, but fails to distinguish the non-inhabited areas in the western part of Noord (industrial area of Nieuwendam) and in Zuidoost. Both models have poor performance in Rembrandtpark and by the lake of Sloteplas in Nieuw-West, in the industrial area of Oost, and in the port of Oosterlijke (Fig. 4, upper). In Copenhagen, the predicted population distribution is continuous in both models, including low values in all cells of the study area. Nonetheless, GB500 predicts more accurately the densely populated areas, especially in the northern part of the municipality of Copenhagen (Brønshøj, Vanløse), on Amager and in the western part of the city (municipalities of Ishøj and Vallenbæk), with the general pattern of low error values being reduced from 4 to 1 person per grid cell, as the estimators increase (Fig. 4, lower). The highest percentage error is noticed in

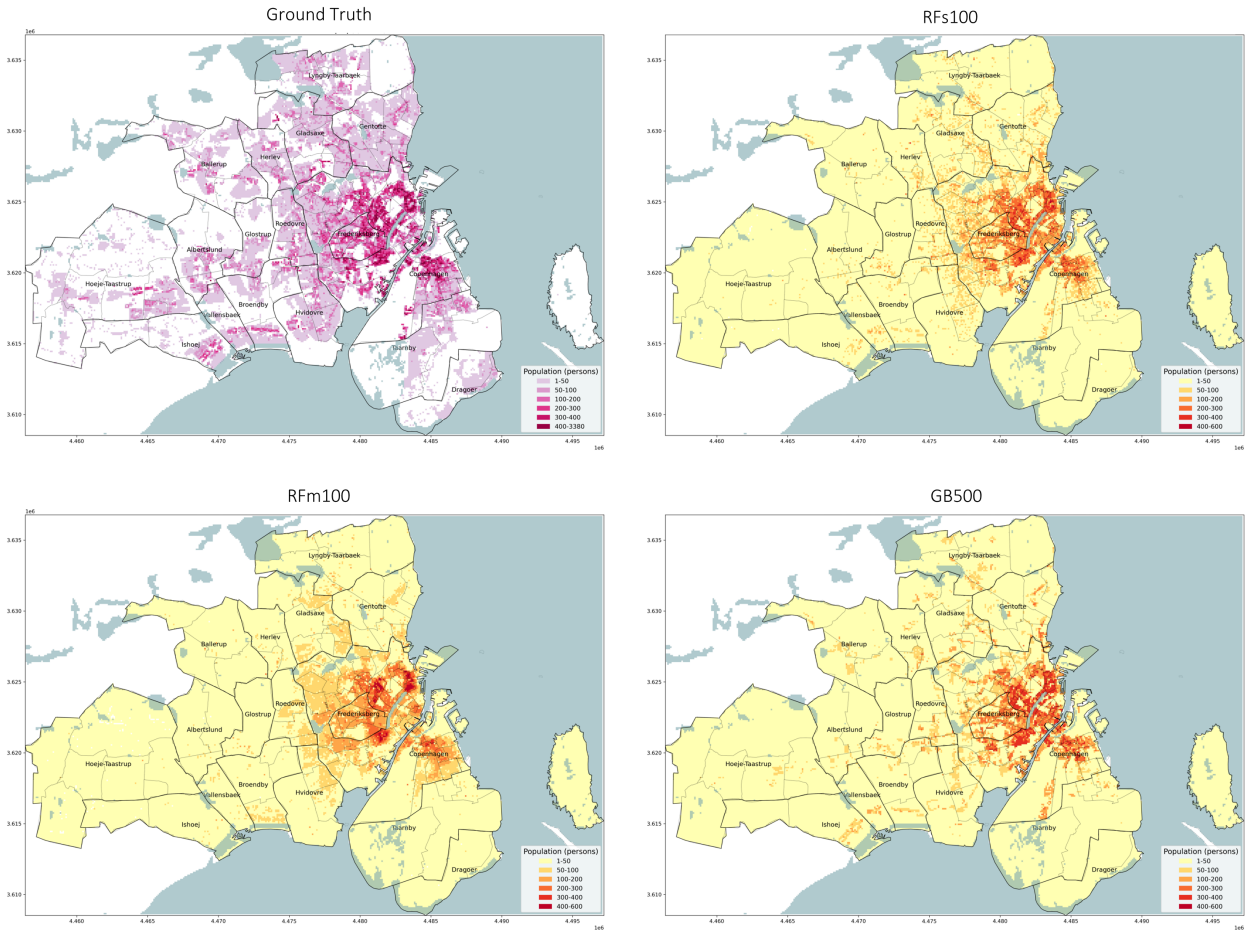

Figure 3. Distribution for the total population in Copenhagen, comparing ground truth counts (top-left) versus predictions: RFs (top-right), RFm (lower-left), GB500 (lower-right).

natural and industrial areas outside the urban fabric, as noticed in the scatter plot of Figure 6.

In order to try explaining the above observations, we assess the role of the input layers by examining the weights of importance for the included variables, for RFs100 and GB500 and for each of the case studies. The importance weights are compared against the distribution of the errors by land coverage. The bar chart for Amsterdam in Figure 6 (upper-left) shows high variations between the models, with the weights of GB500 being distributed in a more balanced way. The ancillary variable corresponding to residential coverage gains the highest weight, while the layers of green spaces, water coverage, and transportation are inadequately weighted in RFs100, rising the error in the corresponding areas. Apparently, neither the residential coverage nor the height of the buildings are adequate to increase the accuracy in the inhabited areas of the urban fabric, where our primary concern is focused.

Nevertheless, information on the age of housing in combination with the building height enhances the performance of the model significantly, as shown in the case of Copenhagen where the predictions are slightly lower than the ground truth for cells with medium-high values (50-100 persons). This is in accordance to both examined models, even if their weights are switched. In contemplation of improving the results of Copenhagen outside the urban fabric, and for comparability reasons to Amsterdam, we performed an additional experiment with RFs100 using 12 ancillary input layers, including the original 3 ones along with the layer for residential coverage, 4 binary layers for land coverage, and 4 layers for describing proximity to transportation, cultural spaces and schools. According to Table 4, the use of more layers leads to inadequate results, which do not meet the scores of the other models. The higher accuracy observed in the experiments with detailed datasets entails that the targeted selection of training input achieves more satisfactory outputs, at least in inhabited urban areas. It should lastly be mentioned that none of the examined models succeeds on predicting the higher real values noticed in either of the cases, showing smoother distributions among the cells of each source zone.

Figure 5 illustrates the distribution of mobile adults and migrants outside EU, as predicted by GB500, along with the ground truth counts for Amsterdam in the top row and for Copenhagen in the lower one. The mobile adults are chosen as an indicative age group with high population,

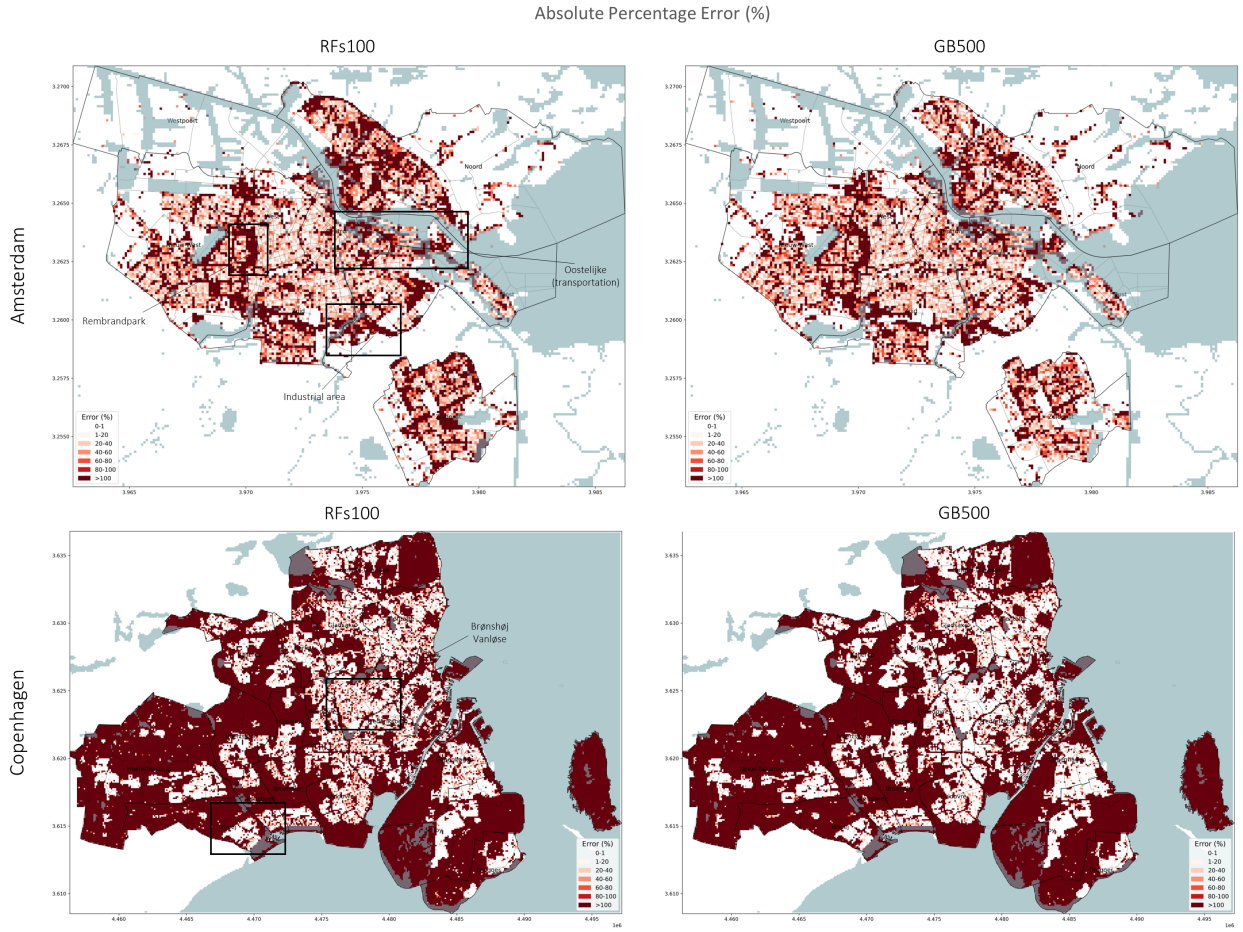

Figure 4. Absolute percentage error in the distribution of the total population for the regions of Amsterdam (upper) and Copenhagen (lower), for RFs100 (left) and GB500 (right).

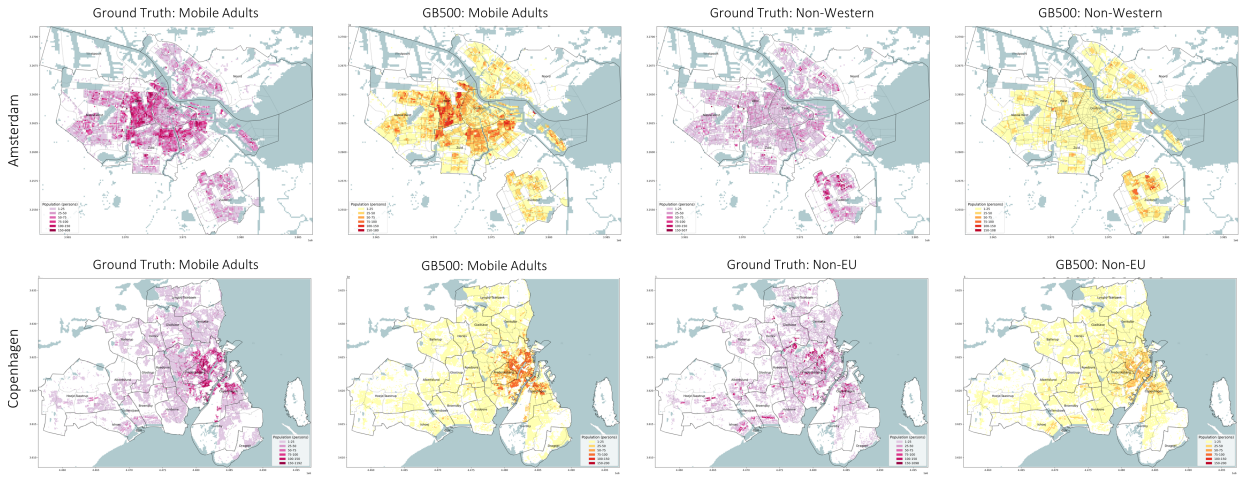

Figure 5. Distributions for two indicative population subgroups in Amsterdam (mobile adults and non-Western migrants) and Copenhagen (mobile adults and non-EU migrants), comparing ground truth counts versus GB500 predictions.

and the non-western migrants, or migrants outside EU, are choosen as population subgroups currently under public discussion due to the recent growing migration flows to Europe. It can be noticed that the model is able to pick up patterns related to these specific groups to some extent, reaching a percentage accuracy of 55-65%. This percentage might not seem high enough, but we need to take into consideration the fine resolution of the examined target zones and that the results are immediately compared to ground truth data at the same detail.

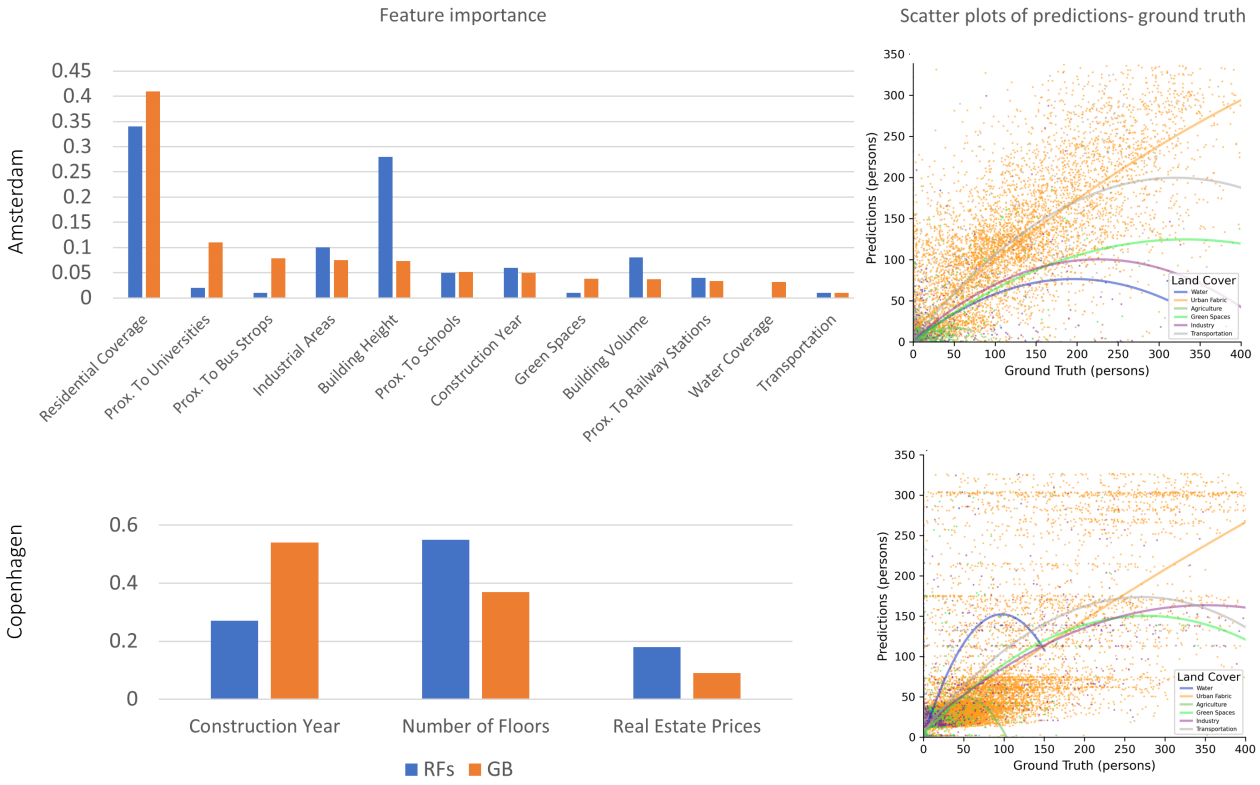

Figure 6. Importance of features for RFs100 and GB500 predictions (left), together with scatter plots comparing ground truth counts versus GB500 predictions, by land coverage (right) and for total population.

Table 4. Percentage errors associated to different regression models, for each case study region.

Amsterdam				Copenhagen				
Model	RFs100	RFs500	GB500	Model	RFs100		RFs500	GB500
Ancillary Data	ams12	ams12	ams12	Ancillary Data	cph3	cph12	cph3	cph3
Total Population	171.66	171.75	**135.09**	Total Population	160.77	262.6	160.68	**147.72**
Age Groups								
Children	48.3	48.31	**38.94**	Children	32.58	55.86	**32.57**	37.31
Students	65.12	65.09	**56.63**	Students	34.32	63.13	34.44	**30.02**
Mobile Adults	53.0	53.18	**44.16**	Mobile Adults	30.56	59.93	**30.55**	32.91
Not Mobile Adults	44.67	44.73	**36.93**	Not Mobile Adults	30.85	57.04	**30.81**	34.5
Elderly	32.1	32.02	**28.22**	Elderly	**28.96**	40.87	**28.96**	31.38
Migration Status								
Natives	83.63	83.57	**66.32**	EU	44.43	64.78	44.44	**41.61**
Western	46.92	46.94	**38.81**	Not EU	44.45	64.81	44.48	**41.67**
Non-Western	47.73	47.64	**42.18**	Natives	118.46	199.76	118.39	**108.76**
Suriname	30.38	30.39	**26.53**	Australia and New Zealand	-0.17	-0.18	-0.17	**-0.16**
Turkey	21.04	21.11	**19.58**	Eastern Asia	**0.1**	- 0.57	**0.1**	0.7
Antilles	9.57	9.58	**9.08**	Eastern Europe	3.77	**2.62**	3.77	3.68
Morocco	34.06	34.17	**32.42**	Lat. Am. and the Caribbean	**0.36**	-0.46	**0.36**	0.89
				Northern Africa	2.07	**0.41**	2.07	2.56
				Northern America	0.39	**-0.32**	0.39	0.9
				Northern Europe	4.0	4.1	3.97	**3.51**
				South-eastern Asia	**-0.05**	-0.78	**-0.05**	0.61
				Southern Asia	11.39	10.05	11.38	**9.93**
				Southern Europe	4.68	3.46	4.67	5.07
				Sub-Saharan Africa	3.39	**1.68**	3.39	4.04
				Western Asia	16.7	15.89	16.68	**14.46**
				Western Europe	2.37	**1.46**	2.37	2.63

6 Conclusions and Future Work

In this work, we discussed the potential of multi-output regression models for spatially disaggregating population groups, and performed a comparative study with a methodology based on self-training random forests or gradient tree boosting regression models. We evaluated the results on ground truth data at the target resolution of 100 meters, and demonstrated that gradient boosting with a customised loss function outperforms random forest regression. This method is not only more convenient for multi-output predictions, but it also achieves high accuracy results in densely populated areas, with a small amount of input layers that characterise the building features and can be easily accessed. Initialising the disaggregation procedure with heuristics based on seminal approaches, such as pycnophylactic interpolation or an heurisitic dasymetric weighting approach, we compared single- and multi-output random forest models to a multi-output gradient boosting regressor. Through tests with two case study areas, we concluded that the multi-output gradient boosting regressor obtains similar or higher quality to the single-output random forests, producing results with higher spatial heterogeneity. The results produced by gradient boosting are also more interesting in densely populated areas, which are the major areas of interest.

Despite the interesting results, there are still many open challenges to work on in the future. Decreasing the long training period (3-5 days based on the extent of the study area, with a single laptop CPU) is one of our priorities, and different software libraries (e.g., Tensorflow) can perhaps be used to train gradient boosting decision trees on a GPU.

Furthermore, our experimental results in Copenhagen showed that a small amount of targeted ancillary datasets can achieve high quality predictions in densely populated areas, but low accuracy in non-populated cells in suburban areas. Taking this observation into account, it may be interesting to explore other sources of ancillary information to address the poor performance in agricultural and industrial areas, or include Earth observation data in a more efficient combination of training input.

7 Data and Software Availability

All the used source code, together with a selection of visualisations for the results, is available on a GitHub repository[1], and can be openly re-used for similar analyses. The datasets used in the experiments are described in Section 4, and most of them are publicly available. The workflow underlying this paper was partially reproduced by an independent reviewer during the AGILE reproducibility review and a reproducibility report was published at https://doi.org/10.17605/osf.io/10.17605/OSF.IO/CDFAH.

[1] https://github.com/mgeorgati/spDisag

Acknowledgements. We certify that there is no actual or potential conflict of interest in relation to this article. This work is part of the FUME project, funded by the EU Horizon 2020 Programme (Grant agreement ID: 870649). The researchers from INESC-ID were partially funded by Fundação para a Ciência e Tecnologia (FCT), through the MIMU project with reference PTDC/CCI-CIF/32607/2017, and also through the INESC-ID multi-annual funding from the PIDDAC program (UIDB/50021/2020).

References

OpenStreetMap, https://www.openstreetmap.org/copyright.

Breiman, L.: Random Forests, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Bygnings- og Boligregistret (BBR): Forside - BBR Teknik, https://teknik.bbr.dk/forside.

Catney, G. and Lloyd, C. D.: Population Grids for Analysing Long-Term Change in Ethnic Diversity and Segregation, vol. 8, Springer International Publishing, https://doi.org/10.1007/s40980-020-00071-6, 2020.

Cheng, Z., Wang, J., and Ge, Y.: Mapping monthly population distribution and variation at 1-km resolution across China, https://doi.org/10.1080/13658816.2020.1854767, 2020.

European Environment Agency (EEA): European Union, Copernicus Land Monitoring Service, https://land.copernicus.eu/pan-european/corine-land-cover.

Freund, Y. and Schapire, R. E.: A Short Introduction to Boosting, Journal of Japanese Society for Artificial Intelligence, 14, 771–780, www.research.att.com/fyoav,, 1999.

Friedman, J. H.: Greedy function approximation: A gradient boosting machine., The Annals of Statistics, 29, 1189–1232, https://doi.org/10.1214/AOS/1013203451, 2001.

Georganos, S., Grippa, T., Gadiaga, A. N., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., and Kalogirou, S.: Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling, Geocarto International, 36, 121–136, https://doi.org/10.1080/10106049.2019.1595177, 2021.

Goodchild, M. F., Anselin, L., and Deichmann, U.: A framework for the areal interpolation of socioeconomic data, Environment Planning A, 25, 383–397, https://doi.org/10.1068/a250383, 1993.

Hay, S. I., Guerra, C. A., Tatem, A. J., Atkinson, P. M., and Snow, R. W.: Urbanization, malaria transmission and disease burden in Africa, Nature Reviews Microbiology 2005 3:1, 3, 81–90, https://doi.org/10.1038/nrmicro1069, 2005.

Kortforsyningen: Kortforsyningen | Kortforsyningen, https://www.kortforsyningen.dk/.

Langford, M., Higgs, G., Radcliffe, J., and White, S.: Urban population distribution models and service accessibility estimation, Computers, Environment and Urban Systems, 32, 66–80, https://doi.org/10.1016/j.compenvurbsys.2007.06.001, 2008.

Lloyd, C. T., Chamberlain, H., Kerr, D., Yetman, G., Pistolesi, L., Stevens, F. R., Gaughan, A. E., Nieves, J. J., Hornby, G., MacManus, K., Sinha, P., Bondarenko, M., Sorichetta, A., and Tatem, A. J.: Global spatio-temporally harmonised datasets

for producing high-resolution gridded population distribution datasets, https://doi.org/10.1080/20964471.2019.1625151, 2019.

Malone, B. P., McBratney, A. B., Minasny, B., and Wheeler, I.: A general method for downscaling earth resource information, Computers & Geosciences, 41, https://doi.org/10.1016/j.cageo.2011.08.021, 2012.

Mennis, J.: Dasymetric mapping for estimating population in small areas, Geography Compass, 3, 727–745, https://doi.org/10.1111/j.1749-8198.2009.00220.x, 2009.

Monteiro, Martins, Murrieta-Flores, and Moura Pires: Spatial Disaggregation of Historical Census Data Leveraging Multiple Sources of Ancillary Information, ISPRS International Journal of Geo-Information, 8, 327, https://doi.org/10.3390/ijgi8080327, 2019.

Monteiro, J., Martins, B., and Pires, J. M.: A hybrid approach for the spatial disaggregation of socio-economic indicators, International Journal of Data Science and Analytics, 5, 189–211, https://doi.org/10.1007/s41060-017-0080-z, 2018.

Monteiro, J., Martins, B., Costa, M., and Pires, J. M.: Geospatial data disaggregation through self-trained encoder–decoder convolutional models, ISPRS International Journal of Geo-Information, 10, 1–28, https://doi.org/10.3390/ijgi10090619, 2021.

Movia Trafik: Movia, https://www.moviatrafik.dk/.

Natekin, A. and Knoll, A.: Gradient boosting machines, a tutorial, Frontiers in Neurorobotics, 7, https://doi.org/10.3389/FNBOT.2013.00021, 2013.

PDOK: PDOK - 3D downloads, https://3d.kadaster.nl/basisvoorziening-3d/.

Qiu, G., Bao, Y., Yang, X., Wang, C., Ye, T., Stein, A., and Jia, P.: Local Population Mapping Using a Random Forest Model Based on Remote and Social Sensing Data: A Case Study in Zhengzhou, China, Remote Sensing 2020, Vol. 12, Page 1618, 12, 1618, https://doi.org/10.3390/RS12101618, 2020.

Qiu, Y., Zhao, X., Fan, D., Li, S., and Zhao, Y.: Disaggregating population data for assessing progress of SDGs: methods and applications, International Journal of Digital Earth, 15, 2–29, https://doi.org/10.1080/17538947.2021.2013553, 2022.

Sabo, F., Corban, C., Politis, P., and Kemper, T.: The European Settlement Map 2019 release : Application of the Symbolic Machine Learning to Copernicus VHR imagery, Tech. Rep. EUR 29886 EN / JRC118076, Publications Office of the European Union, https://doi.org/10.2760/979189, 2019.

Schiavina, M., Freire, S., and MacManus, K.: GHS-POP R2019A - GHS population grid multitemporal (1975, 1990, 2000, 2015), European Commission, Joint Research Centre, https://doi.org/10.2905/0C6B9751-A71F-4062-830B-43C9F432370F, 2019.

Statistics Division, United Nations: UNSD — Methodology, https://unstats.un.org/unsd/methodology/m49/.

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

Tobler, W. R.: Smooth Pycnophylactic Interpolation for Geographical Regions, Journal of the American Statistical Association, 74, 519–530, https://doi.org/10.1080/01621459.1979.10481647, 1979.

Vargas-Munõz, J. E., Metzger, N., Daudt, R. C., Kellenberger, B., Whelan, T. T.-T., Ayo, B., Imran, M., Ofli, F., Schindler, K., and Tuia, D.: Fine-grained population mapping using coarse census data and satellite imagery, https://ethz.ch/content/dam/ethz/special-interest/baug/igp/photogrammetry-remote-sensing-dam/documents/pdf/Papers/Metzger_LivingPlanetHAC.pdf, 2022.

Verdhan, V.: Supervised Learning for Regression Analysis, Supervised Learning with Python, pp. 47–116, https://doi.org/10.1007/978-1-4842-6156-9_2, 2020.