scientific reports

OPEN

Fine-grained population mapping from coarse census counts and open geodata

Nando Metzger[1✉], John E. Vargas-Muñoz[2], Rodrigo C. Daudt[1], Benjamin Kellenberger[2,3], Thao Ton-That Whelan[4], Ferda Ofli[5], Muhammad Imran[5], Konrad Schindler[1] & Devis Tuia[2]

Fine-grained population maps are needed in several domains, like urban planning, environmental monitoring, public health, and humanitarian operations. Unfortunately, in many countries only aggregate census counts over large spatial units are collected, moreover, these are not always up-to-date. We present Pomelo, a deep learning model that employs coarse census counts and open geodata to estimate fine-grained population maps with 100m ground sampling distance. Moreover, the model can also estimate population numbers when no census counts at all are available, by generalizing across countries. In a series of experiments for several countries in sub-Saharan Africa, the maps produced with Pomelo are in good agreement with the most detailed available reference counts: disaggregation of coarse census counts reaches R^2 values of 85–89%; unconstrained prediction in the absence of any counts reaches 48–69%.

High-resolution population maps are crucial for many planning tasks, from urban planning[1] to preparing humanitarian actions[2] and effective disaster response[3]. Given the rapid population growth in many regions of the world[4] and the increasing rate at which populations shift in response to environmental and social changes, it is important to maintain accurate, up-to-date maps. Unfortunately, census data are often only available at very coarse spatial resolution (e.g., one aggregate number for a district with hundreds or even thousands of km^2) and therefore not suitable as a basis for local planning: whether for sustainable land use and infrastructure management or for targeted disaster relief, planners need to know in more detail where the people are. The problem is especially prevalent in developing countries in the global south, where humanitarian actions are more often needed yet census data availability and quality are limited.

Remote sensing products and other openly available geographical datasets like OpenStreetMap (OSM) can serve as auxiliary, high-resolution evidence to create fine-grained population density maps[5]. Yet, the design of effective population density models[6] that combine such data sources with low-resolution census counts remains a challenge. Generally speaking, two different approaches have been employed for population mapping[7]: bottom-up and top-down. Bottom-up methods[7,8] start from local surveys of population density, collected at a number of sample locations, and attempt to generalize from detailed but sparse samples to the unobserved regions to cover larger areas. Researchers have proposed different ways to locally measure population density, such as counting the (average) number of people per rooftop area[7,8] or, if more resources are available for the local survey, specific average densities for different types of residential zones (urban-, rural-, and non-residential)[7,9]. A main drawback of bottom-up methods is that local surveys will necessarily remain extremely sparse and can hardly provide enough data points to scale population mapping up to the country level. On the contrary, top-down approaches[6,10] rely on census data, which ensures complete coverage at the expense of much lower spatial resolution, in some cases down to a single head count per large district. The task then becomes to disaggregate that data to a much finer resolution, often a regular grid.

Top-down approaches[6,11] commonly use dasymetric disaggregation to redistribute the known, spatially coarse population counts for census areas on the order of many km^2 across smaller spatial units[12]—for instance square blocks of size 100×100 m—with the help of auxiliary variables that covary with population density. Examples of such covariates for population disaggregation are the presence of buildings[13], building counts[10], building volumes[14], and cumulative road lengths[15]. Dobson et al.[16] propose a weighted combination of several covariates, such as land cover and proximity to roads, to compute what portion of the population to assign to each target unit. Stevens et al.[6] resort to machine learning and train a random forest model to predict population density

[1]ETH Zurich, Zurich, Switzerland. [2]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. [3]Yale University, New Haven, USA. [4]International Committee of the Red Cross, Geneva, Switzerland. [5]Hamad Bin Khalifa University, Ar-Rayyan, Qatar. ✉email: nando.metzger@geod.baug.ethz.ch

1

from a set of covariates like building maps and night light images. A limitation of the method proposed by Stevens et al.[6] is that administrative regions with known census counts are used as training units, which means that the input features (covariates) must be aggregated across all pixels in such a region; whereas at prediction (inference) time, density values are predicted for each individual pixel, respectively feature vector. Consequently, the method's efficacy is reduced in countries with coarse census units, as aggregation over large spatial regions induces distribution shifts between the training and test data.

Recent work has employed deep learning, in particular convolutional neural networks (CNNs[17]) for population estimation, in an attempt to better account for spatial context. Gervasoni et al.[18] used a CNN to map data extracted from OSM to disaggregation weights per pixel with 200×200 m ground sampling distance (GSD). This allows one to super-resolve census counts to a finer granularity; however, since the predicted weights are relative, the method cannot be used to predict population numbers for regions without census data. Jacobs et al.[19] trained a CNN to predict fine-grained population maps from very high resolution (VHR) optical satellite imagery and census data. This is viable for individual cities, but difficult to scale up to larger geographic contexts, due to the limited availability and high cost of VHR images. Both CNN-based methods[18,19] depend on fairly fine-grained initial counts with census units of few km^2, e.g., urban census blocks in France, respectively the USA. For comparison, the administrative units for which census data are available in much of Africa have sizes on the order of several hundred to several thousand km^2, which precludes the use of such methods.

Here, we propose a methodology to estimate fine-grained population maps from such very coarse census data by fusing them with covariate maps of higher resolution. As inputs, we use publicly available products derived from remote sensing images and other open data sources, e.g., building footprint maps, night light images, and OSM road layers. They are used to train a model that predicts population on a spatial grid with 100×100 m GSD. Our method, which we call POMELO (short for "population mapping by estimation of local occupancy rates"), is inspired by work on *guided super-resolution* of low-resolution images[20–22]. In an experimental evaluation with data from three different countries in Sub-Saharan Africa (Tanzania, Zambia, Mozambique) POMELO delivers significantly more accurate population maps compared to several baseline methods, including the pioneering work of Stevens et al.[6] with which the widely used WorldPop maps[23] are created. Moreover, POMELO can not only disaggregate existing census data, but can also predict population maps in the absence of census counts. Consequently, it can be deployed to regions or countries where no suitable census information is available—of course these estimates, inevitably, have higher uncertainties, as the total population is no longer constrained by a known aggregate number. In summary, POMELO provides knowledge about where people are at the hectometer scale, even if census data are not available.

Results

We compare POMELO with other state-of-the-art baseline models using three different evaluation strategies, and three performance metrics (see "Methods" section), namely R^2, mean absolute error (MAE), and mean absolute percentage error (MAPE):

Coarse supervision.   In this strategy, used by several works in the literature[6,24,25], the census data at the finest available level are reserved exclusively for performance evaluation. First, the available census data are artificially coarsened by aggregating them to the next-higher ("coarse") level of administrative regions (e.g., counts at district level are aggregated into a single count per province). Then, the coarse level regions are divided into five random folds, of which three serve as training set, one as validation set for hyper-parameter tuning and checkpointing the model on the best MAPE, and one as test set to measure the model's predictive skill. Finally, the model is trained to predict gridded population numbers at 100 m GSD, such that they add up to the coarse level counts. To measure performance, the gridded population estimates are aggregated to obtain population numbers for each *fine* level administrative region in the test set and are then compared to the corresponding census data.

This scheme corresponds to an application scenario where fine-level census data are not available: all model fitting and mapping is only based on the coarse level. The fine level counts are employed exclusively to evaluate prediction performance.

The entire training and evaluation procedure is run as a fivefold cross-validation, rotating the validation and test folds. Figure 1 depicts the coarse census data used for training, the fine level census data used for evaluation, and the fine-grained population maps obtained by our model, for the surroundings of Zanzibar, Tanzania.

Fine supervision.   In the second evaluation scheme, census data at the finest available level are used for supervision. The data are *not* coarsened, rather the original fine-level data are divided into five random folds. These folds are split along the same coarse-level administrative boundaries as above, such that the training, validation and test portions during cross-validation are identical to the coarse supervision scenario. The model is trained to predict gridded population numbers at 100 m GSD, such that they add up to the fine-level counts. Just like above, the gridded population estimates are aggregated to obtain population numbers for each fine-level administrative region in the held-out test fold (e.g., Fig. 1b), which are then compared to the corresponding census data to compute performance metrics.

This scheme corresponds to the actual, realistic population mapping task with the available census granularity. The finest available resolution is used to learn the best possible model, geographically separating training and testing regions to ensure an unbiased evaluation. Note that the combined test set, after holding out each fold once for testing, is the same for *coarse supervision* and *fine supervision*, such that the error metrics are directly comparable.

(a) Coarse (district level)   **(b)** Fine (ward level)   **(c)** Pixel (prediction level);
100 m resolution
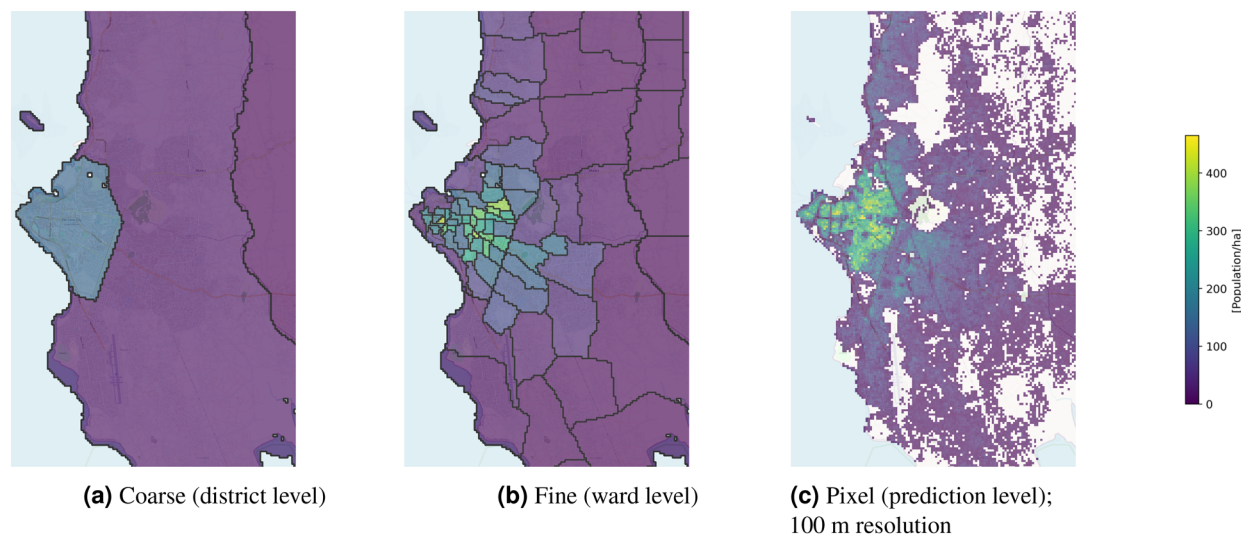
Figure 1. Overview of the different level of population data. Shown location: Zanzibar City, Tanzania. The map visualizations were created in QGIS 3.14[26].

Evaluation set	Method	R^2 ↑ [%]	MAE↓	MAPE↓ [%]
Tanzania	Building disaggregation	65.2	3700	23.1
	MRF	78.6	3300	21.7
	RF per region[6]	49.0 ± 3.0	4380 ± 50	26.9 ± 0.2
	CNN	69.5 ± 2.0	4200 ± 100	28.8 ± 0.5
	Pomelo	85.7 ± 0.9	3100 ± 40	21.6 ± 0.2
Zambia	Building disaggregation	82.1	4300	53.9
	MRF	87.0	3900	51.3
	RF per region[6]	76.4 ± 0.0	4750 ± 90	60.9 ± 1.2
	CNN	83.9 ± 0.7	4250 ± 40	51.4 ± 1.7
	Pomelo	87.9 ± 0.1	3730 ± 30	50.4 ± 0.4

Table 1. Performance with coarse supervision for Tanzania and Zambia.

Transfer task. Besides census disaggregation, Pomelo can also be used to estimate population density for a given country in the absence of census numbers. To evaluate the performance on such more challenging task, we use data from seven different countries. One of the countries is held out as test set, and the data from the remaining six countries are randomly split into 80% for training and 20% for validation (i.e., hyper-parameter tuning). Then, the model is trained to predict gridded population numbers at 100 m GSD, using the *fine-level* strategy. Finally, the model is deployed for the held-out country, and evaluated in the same way as above.

For each test country we run five models with different random parameter initialization and train/validation splits and report the average and standard deviation of each performance metric. We compare our proposed method with other four baseline methods, described in the "Methods" section: Building count disaggregation, the random forest model used by WorldPop[6], a Markov Random Field based method for population disaggregation, and a Convolutional Neural Network model.

For evaluation, we use data from Tanzania and Zambia with the three aforementioned evaluation strategies. Mozambique is used only in the *transfer task* evaluation. The data available for that country is not suitable for the other two evaluation schemes: in Mozambique there are only 413 fine-level regions grouped into 156 coarse-level regions. The low aggregation factor of only 2.6× causes the performance metrics to saturate at a similar value of 83% R^2 for all methods, while giving little indication about the correctness of the pixel-level maps.

Results—Coarse supervision. Table 1 presents the performance metrics for the coarse supervision strategy in Tanzania and Zambia. The first two rows show learning-free methods that are only capable of pure disaggregation, i.e., they cannot estimate population numbers for regions without known (aggregate) census counts. The MRF formulation brings a marked improvement compared to the simple disaggregation scheme based only on building counts. For the three learning-based methods, we first predict population numbers without using the target regions' census counts, then normalize those counts to relative fractions and use those as weights for dasymetric disaggregation. Pomelo achieves the best performance in all three metrics, with the closest competitor being the MRF. An interesting observation is that the region-based RF method[6], when trained with such

Evaluation set	Method	$R^2 \uparrow$ [%]	MAE↓	MAPE↓ [%]
Tanzania	RF per region[6]	79.1 ± 2.6	3320 ± 70	22.1 ± 0.3
	CNN	80.8 ± 2.1	3290 ± 80	21.7 ± 0.4
	Pomelo	87.6 ± 0.2	2890 ± 20	20.4 ± 0.3
Zambia	RF per region[6]	87.6 ± 0.0	3740 ± 40	46.5 ± 0.4
	CNN	88.2 ± 0.4	3680 ± 50	44.9 ± 0.4
	Pomelo	88.7 ± 0.3	3650 ± 20	48.0 ± 0.6

Table 2. Performance with fine supervision for Tanzania and Zambia.

(a) Tanzania   **(b)** Zambia
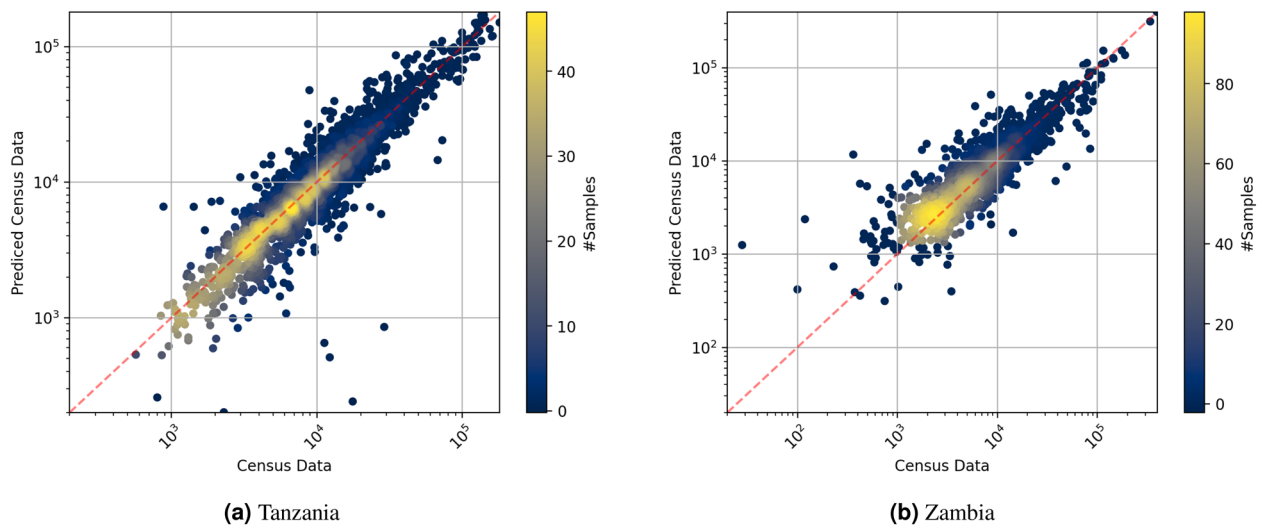
Figure 2. Comparison between the estimates predicted by Pomelo and the respective ground truth census in the fine level training task.

coarse census data, does not even reach the performance of simple building count disaggregation. The evaluation results for Zambia largely mirror those for Tanzania, where the MRF is the closest to Pomelo in terms of R^2 and MAE. Next follows the CNN, which handles Zambia much better than Tanzania.

Results—Fine supervision. Table 2 shows quantitative performance when the models are trained and evaluated on the finer census level. Note that we cannot use the learning-free building disaggregation and MRF baselines with this evaluation strategy, since they are only designed for disaggregation of regions with known overall population numbers (which makes no sense in the fine supervision setting, as the target quantity would have to be known in advance). For Tanzania, we again observe the best performance with Pomelo, although the advantage is smaller than in the case of coarse supervision (Table 1). The performance of the region-based RF improves a lot when trained with fine supervision, showing that the coarse supervision does not offer enough supervision signal for training at the region level. The second section of Table 2 shows results for Zambia. Pomelo maintains a slight edge in terms of R^2 and absolute error. It does not fare as well in terms of MAPE, which is due to large relative errors in few, sparsely populated regions (that nonetheless translate to small errors of the absolute population count): just by removing Zambia's smallest census region with only 28 inhabitants from the evaluation, the MAPE of Pomelo drops to 45%.

Figure 2 shows scatter plots (census counts versus our predicted counts) for the fine-level administrative regions of both Tanzania and Zambia. In both cases, the data are, with few exceptions, close to the (red, dashed) diagonal that corresponds to the ideal result, where predictions and ground truth coincide.

Results - Transfer task. We also evaluate Pomelo in a scenario where no census data are available for the target country. Here, the models must instead be trained on other countries and should be able to generalize to a potentially different unknown geographical context. From the seven countries considered (Tanzania, Zambia, Mozambique, Rwanda, Uganda, Democratic Republic of Congo, and Nigeria), we select six for training and validation and report the evaluation metrics on the respective held out country (Tanzania, Zambia, and Mozambique) in Table 3.

The first rows of each section in the table ("average occupancy rate") show the results of a naïve baseline that computes the average occupancy rate in the six training countries and multiplies it with the building count map of the target country to obtain the population map. One can see that this simple approach does not perform well

Evaluation set	Method	$R^2 \uparrow$ [%]	MAE ↓	MAPE ↓ [%]
Tanzania	Average occupancy rate	− 52.9	13,200	98.0
	RF per region[6]	22.1 ± 10.1	8000 ± 700	66.5 ± 6.2
	CNN	34.6 ± 6.2	6800 ± 300	49.6 ± 3.0
	POMELO	68.1 ± 0.9	5200 ± 200	35.6 ± 1.0
	POMELO—7 countries	81.7 ± 0.6	3170 ± 20	21.7 ± 0.6
	POMELO—Tanzania only	82.0 ± 1.5	3800 ± 70	26.4 ± 0.2
Zambia	Average occupancy rate	14.4	12,300	150.2
	RF per region[6]	− 193.4 ± 361.3	19,000 ± 8000	271.9 ± 139.0
	CNN	59.0 ± 1.6	6600 ± 300	82.7 ± 5.6
	POMELO	68.7 ± 3.1	6100 ± 200	75.6 ± 1.6
	POMELO—7 countries	85.4 ± 0.5	3810 ± 10	49.6 ± 0.2
	POMELO—Zambia only	74.0 ± 1.4	4820 ± 70	47.6 ± 0.1
Mozambique	Average occupancy rate	− 88.1	82,500	153.0
	RF per region[6]	− 49.5 ± 21.1	59,000 ± 3000	102.0 ± 3.6
	CNN	43.6 ± 5.2	35,200 ± 1300	63.0 ± 1.9
	POMELO	48.3 ± 3.2	36,900 ± 900	47.5 ± 0.6
	POMELO—7 countries	83.0 ± 0.5	16,800 ± 200	34.1 ± 0.2
	POMELO—Mozambique only	61.0 ± 3.0	30,600 ± 900	45.0 ± 1.0

Table 3. Performance on the transfer task for all countries.

in any of the countries, barely reaching a positive R^2 score only in Zambia. The region-based RF model does not generalize well either, to the point that the R^2 metric becomes negative and the MAPE exceeds 100% for Zambia and Mozambique. The performance of the CNN is still rather good across all evaluation settings, despite the more challenging scenario. POMELO exhibits even better overall performance, with the exception of a slightly higher MAE in Mozambique.

As a proxy for a "general" model that is valid for an entire geographic area, we also create a training set with data from all seven countries (including the respective target country), and apply the resulting model to the (unseen) test set of the target country, using fivefold cross-validation to cover the whole country. For comparison, we also run models trained specifically for each country, but without dasymetric rescaling, to keep the comparison fair. That scenario would normally not be relevant in practice, since the training implies that aggregate counts are available except in the special situation that no current census counts are available and a previously trained model is reused. However, by removing the influence of the postprocessing it shows the ability of the country-specific model to predict unconstrained, absolute numbers. As expected, the country-specific models perform better than the one learned only with data from four other countries, but not as well as a model trained on all seven countries, presumably due to the larger amount and variability of training samples.

Visual results. The finest census level available for validation are administrative regions with areas of a few km² to several hundred km², and ranging from very low to very high population densities. To visually examine the predicted population maps at 100 m resolution, we select regions with known low/high densities and resort to OSM for verification. Figure 3 presents example population estimates, shown as heat maps overlaid on OSM for regions around the city of Dar-es-Salaam in Tanzania. Each example covers an area of 400×400 m, corresponding to 16 cells in our gridded map. Figure 3a shows an example of a very low population estimate in a rural area. The estimates in low-density residential areas (Fig. 3b) are somewhat higher and more variable, due to varying numbers of buildings. In high-density residential areas (top part of Fig. 3c), the estimates are even higher. For non-residential areas with mostly commercial buildings (bottom part of Fig. 3c), the model predicts a relatively low and uniform density. These variations can occur over a small spatial distance: as an example, Fig. 3c show a sudden drop in population density between immediately adjacent city areas.

Figure 4 depicts the predicted building occupancy rates as heat maps for two neighborhoods in the city of Dar-es-Salaam, that are located close to each other but feature very different occupancy rates. The densely populated area of the Mtoni region in Fig. 4a has an estimated (mean) building occupancy rate of 4.2, whereas the area in the Kijichi region in Fig. 4b has a much lower estimated building occupancy rate of 2.8.

Figure 5 visualizes population numbers and building occupancy rates per administrative region ("ward") in the surroundings of Dar-es-Salaam. One can see that wards have rather heterogeneous sizes and populations. Figure 5a depicts our model's population estimates, obtained by aggregating the grid cells within each ward. Figure 5b depicts the reference values from the national census. Figure 5c are our model's estimated building occupancy rates, computed by dividing the estimated population by the number of buildings—or, equivalently, by averaging the predicted densities, with weights proportional to the building counts in their corresponding grid cells. Figure 5d are the true occupancy rates according to the census (i.e., the population count divided by the number of buildings). The relative error of the per-ward population estimates, as a percentage of the true numbers, is shown in Fig. 5e. In general, the predictions are in good agreement with the actual numbers. We do observe a trend that POMELO overestimates the populations in built-up low-density regions, while it tends to
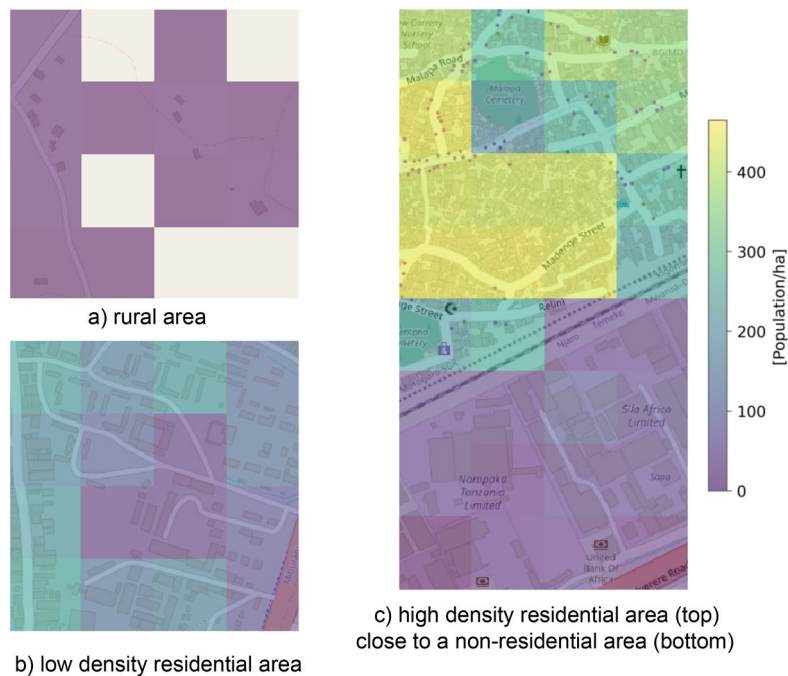
Figure 3. Visual examples of population estimates on a 100 m grid, represented as heat maps on OSM background[27]. The color scale for the heat maps is shown in the top right in absolute population counts.
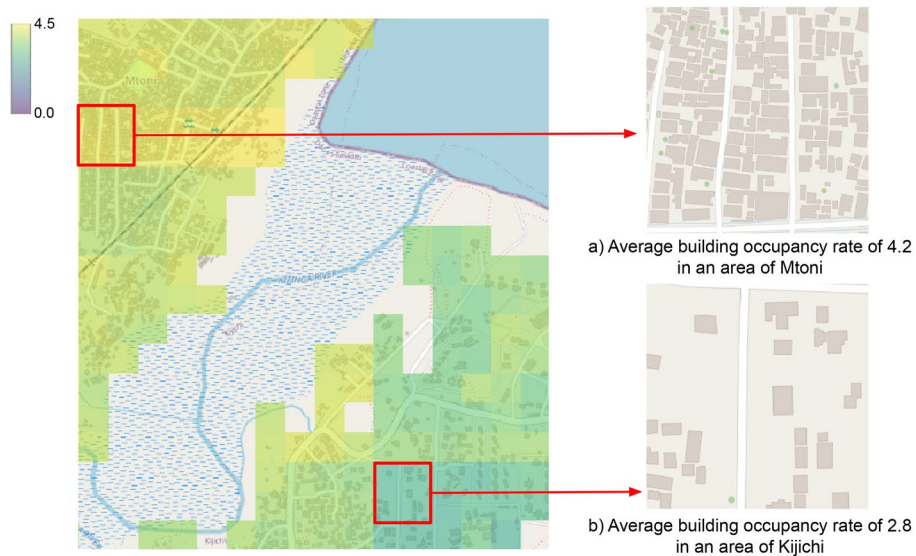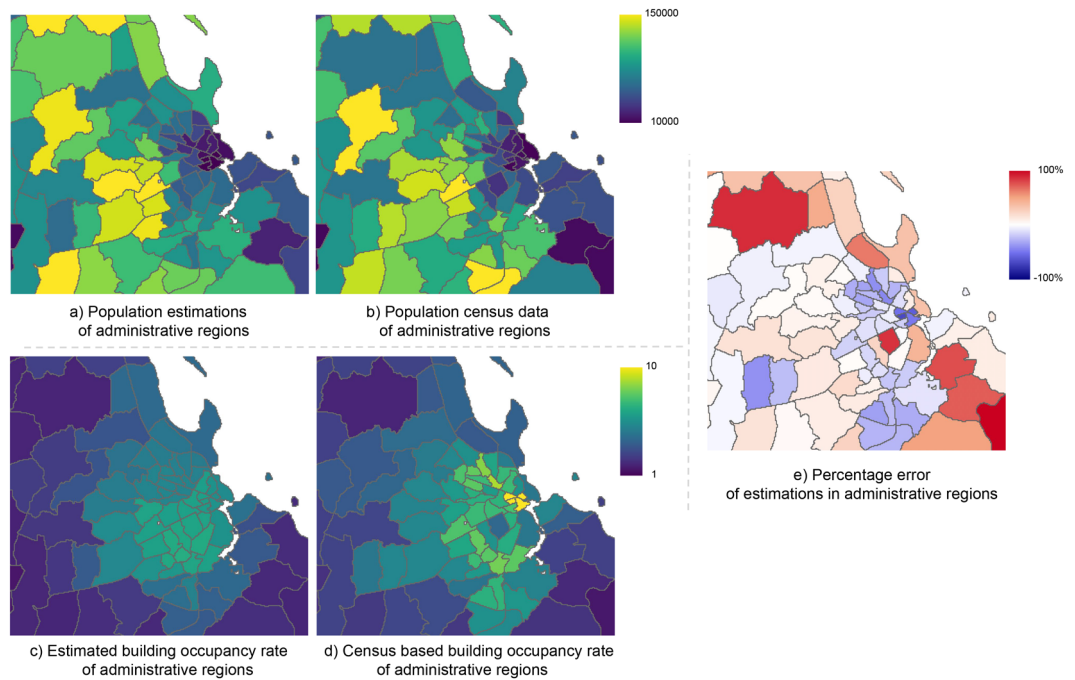

Figure 4. Estimated average building occupancy rate, depicted as heat map on OSM background, for the Mtoni and Kijichi regions of Dar-es-Salaam. Note the significantly different occupancy rates in the marked areas.

underestimate the extremely high occupancies in the centers of large cities. Finally, Fig. 6 shows a visual comparison of recent population maps of Zanzibar City, at 100 m resolution; Including the *High Resolution Population Density Maps* (HRPDM) project[29], Random Forest disaggregation per census region following[6] (but trained with the same data as our method), and Pomelo. It is apparent that Pomelo recovers the population distribution in more detail (i.e., with a higher effective resolution), whereas the HRPDM maps as well as the region-wise disaggregation result appear overly smooth and do not recover high-frequency variations of the population density.

Discussion

We evaluate the performance of our method using three different scenario: *coarse supervision*, *fine supervision*, *transfer task*. In the past few years, the first method has been the *de facto* standard for top-down population estimation[6,25] and connects our evaluations to related literature. However, we argue that the *fine supervision*

Figure 5. Population and building occupancy rate per ward around Dar-es-Salaam, Tanzania: (**a**) estimated population counts, (**b**) true population counts, (**c**) estimated building occupancy rate, (**d**) true occupancy rate, (**e**) signed relative errors of population estimates. The map visualizations were created in QGIS 3.14[26].
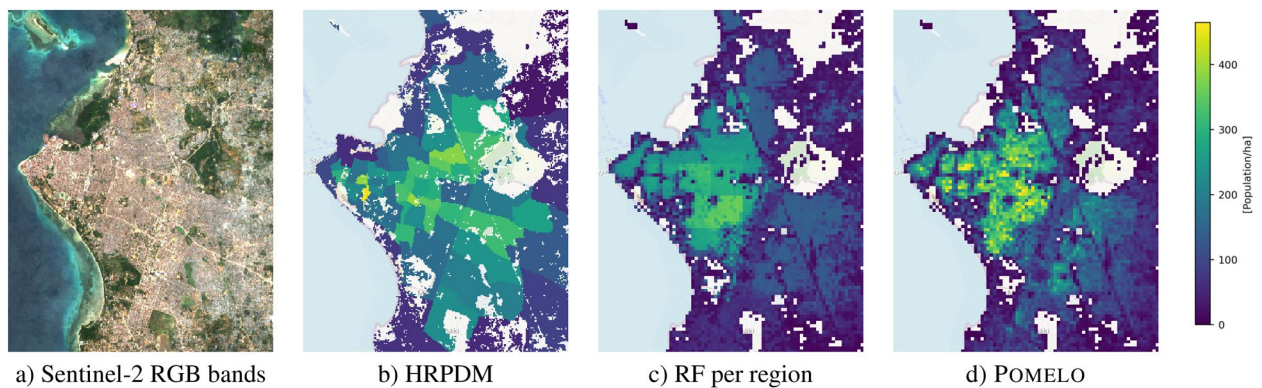

Figure 6. Visual comparison of the Sentinel-2 imagery[28] *High Resolution Population Density Maps* (HRPDM)[29] per-region RF disaggregation[6] and POMELO. Shown location: Zanzibar City, Tanzania. The Sentinel-2 mosaic was created with Google Earth Engine[30] and visualized in QGIS[26].

scenario is actually more representative of the procedure one would use in practice: On the one hand, it would seem unnatural to not use the finest available spatial units if the goal is to produce the best possible population maps from a given census dataset. On the other hand, in the *coarse supervision* setting one predicts population maps at a level that is already known from the census. The coarse scenario seems to have arisen largely from the desire to evaluate pure disaggregation methods, which requires access to ground truth counts at two different levels of the spatial hierarchy.

It is worth noting that the region-level RF method of Stevens et al.[6] does not work well if the census is only available for spatially coarse units, which is a rather frequent situation in developing countries. We see several possible reasons for this: first, there simply are fewer regions at coarser hierarchy levels and one may be left with too few training examples to learn a good model. Second, performing feature aggregation leads to a domain shift between coarse-level training data and fine-level test data—in particular at the pixel level.

We found it advantageous to first estimate the spatial distribution of the building occupancy rate, and then compute population numbers by multiplying the occupancy rate and the building count at a given location. This is in contrast to methods that directly estimate absolute population numbers[6] or relative population fractions (which are the weights for dasymetric disaggregation[18]). The finding that factoring into building counts and occupancy reduces the estimation errors lends support to two assumptions implicit in our approach: First, the

Feature importance
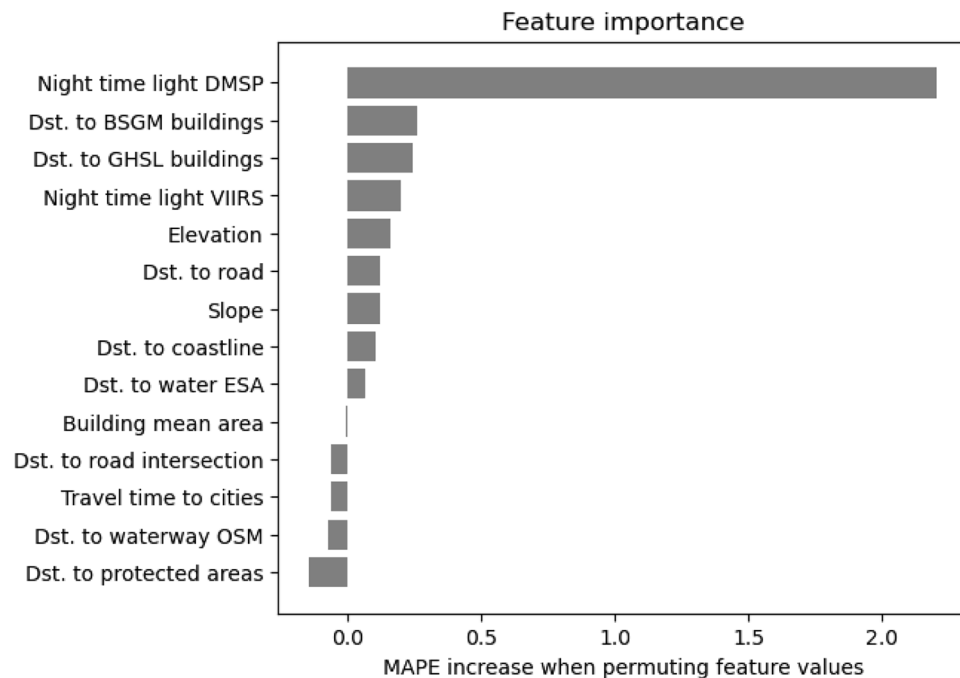
Figure 7. Feature importance analysis for the fine supervision setting in Tanzania.

available maps of building counts (respectively, building footprints) are apparently rather accurate, so that using them directly rather than as one of many "soft" covariates reduces the estimation error. Second, the occupancy rate appears to have lower spatial variability than the population density, making it easier to estimate from the same covariates.

The important role of country-scale building counts (respectively, building footprint maps) also merits some discussion. For our work we obtain them from free sources, namely the Open Buildings dataset[31] and the Grid Maps of Building Patterns[32]. Both these datasets are created with computer vision-based building detectors on the basis of high-resolution satellite imagery (GSD < 1m). Although the datasets are of high quality, such large-scale maps inevitably contain errors and data gaps, especially in rural areas[33]. We have tried to maximize completeness by fusing the two building datasets, but note that missing buildings (and to a lesser degree perhaps also spurious buildings) may cause errors in our population maps, mainly in scarcely populated areas.

Although the mentioned building footprint datasets are available for free, the underlying high-resolution images used to produce them are not. The processing of high-resolution data at country scale also requires considerable computational resources. Our method thus critically depends on data that are oftentimes primarily produced for a different purpose or for philanthropic reasons, and over whose production we have no control. If at some point in the future no up-to-date building dataset is at hand, one could try to resort to autoregressive settlement growth models[34,35]. Alternatively, one may keep the building detection implicit and also feed high-resolution images to the population estimator. Such approaches have been developed[7,36], but since they rely on (commercial) high-resolution satellite imagery they can hardly be scaled up to entire countries, except by entities who could equally supply the building counts. Crowdsourced data (e.g., OSM roads) have their own challenges due to incompleteness and temporal inaccuracies. However, in several regions of developing countries, commercial products and official data are not available, making crowdsourced data the only source of information with sufficient coverage in those regions[37].

In addition to the building counts, Pomelo is driven by several other geospatial data layers that are publicly available and correlate with population. We have analyzed feature importance with the permutation method[38], which essentially measures the performance drop caused by randomly shuffling a single input layer, so as to render that covariate uninformative. Exemplary feature importance scores for Tanzania (fine supervision) are shown in Fig. 7. Although the ordering of the covariates fluctuates depending on the chosen performance measure, we noted one commonality, namely the nightlight and settlement layers are among the most predictive inputs. Moreover, also features with low scores appear to carry some relevant information, as explicit feature selection based on the importance scores tends to harm model performance. For more details please refer to the "Methods" section.

Apart from the covariates used in the present work, there are others that could potentially be useful, and that one could obtain from open geospatial data sources. For instance, online land use maps as available in OSM could help to identify non-residential buildings such as schools or shopping centers, and building height estimates from Esch et al.[39], once publicly available, could help to predict more precise building occupancy rates. Furthermore, data extracted from social media could possibly also support population mapping in certain areas. For example, the Facebook Marketing API[40] allows one to access (anonymized and aggregated) information about the platform's users at a certain location[41,42], such as the number of active users, the type of internet connection

Type		Description
Buildings	1	Building counts from Google Open Buildings[31]
	2	Building counts from Gridded Maps of Building Patterns[32]
	3	Mean building areas from Google Open Buildings[31]
	4	Mean building areas from Gridded Maps of Building Patterns[32]
Accessibility	5	Travel time to city with more than 50k inhabitants[23]
Nightlight	6	Cloud-free DMSP nightlight composite[43]
	7	Cloud-free VIIRS nightlight composite[43]
Settlement	8	Distance to built-up area from Global Human Settlement layer (GHSL)[44]
	9	Distance to built-up area from Built-Settlement Growth Model (BSGM)[23]
Topography	10	Elevation model from the Shuttle Radar Topography Mission (SRTM)[45]
	11	Slope from the Shuttle Radar Topography Mission (SRTM)[45]
Water	12	Distance to Waterbody from ESA CCI Water[46]
	13	Distance to Waterway from OpenStreetMap[27]
	14	Distance to Coastline[23]
Land cover	15	Distance to Nature Reserves from World Database on Protected Areas (WDPA)[47]
Roads	16	Distance to Road from OpenStreetMap[27]
	17	Distance to Road Intersection derived from OpenStreetMap[27]

Table 4. Summary of the used covariates.

	# census regions		# pixels		Avg. fine census region	
	Fine	Coarse	Total	Built-up (%)	Area (km²)	Population
Tanzania	3654	170	111M	7.5	333	15,449
Zambia	1421	150	90M	3.4	633	13,035
Mozambique	413	156	96M	5.6	2324	75,095
Others	2779	295	412M	5.5	9975	139,171

Table 5. Overview of census data used and data extent in # of pixels. Each pixel covers a region of 100 × 100 m. The Category *Others* refers to the additional countries of Uganda, Rwanda, Nigeria and the Democratic Republic of Congo.

used, etc. It is quite possible that such information is to some degree predictive of population density. However, social media and OSM data are typically incomplete and biased, which is a challenge and would require computational methods that can exploit the important features without relying on their completeness. Finally, it would be important to develop techniques to handle the case where covariates are missing for certain regions or time periods in a given country.

Methods

Data.
To validate our proposed methodology, we use covariates that are related to population, and that can be derived from remote sensing imagery, open geo-spatial data (e.g., OSM) or governmental sources. For the present study we rely on data that have been preprocessed by the WorldPop[23] project. We collect the covariates listed in Table 4, with a resolution of 100 × 100 m. A selection of variables is visualized in the left part of Fig. 8. Moreover, we obtain census data for the countries of Tanzania (ward-level, $n = 3654$) Zambia (ward-level, $n = 1421$) and Mozambique (*postos administrativos*-level, $n = 413$) to evaluate the performance of our proposed POMELO method.

We also aggregate the census counts to the second-finest administrative level for the countries of Tanzania (district-level), Zambia (constituency-level), and Mozambique (district-level), using administrative boundaries from the Humanitarian Data Exchange[48], to simulate scenarios where the census counts are coarser. We will later refer to these aggregated data as *coarse census* counts. Finally, to assess the generalization performance of POMELO across countries, we collect the same source data for four further countries, Uganda (UGA), Rwanda (RWA), Nigeria (NGA), and Democratic Republic of Congo (COD), to be used as additional training data. The dataset is summarized in Table 5.

Proposed method: population mapping by estimating local occupancy rates.
Our goal is to produce population maps with a finer resolution than the underlying census counts. In machine learning terms we are thus faced with an instance of *weakly supervised* learning[49]: we do not have access to ground truth values for the individual grid cells, rather we only have one target value (the census count for a region) as supervision signal for a whole set of outputs (all grid cells within the region).
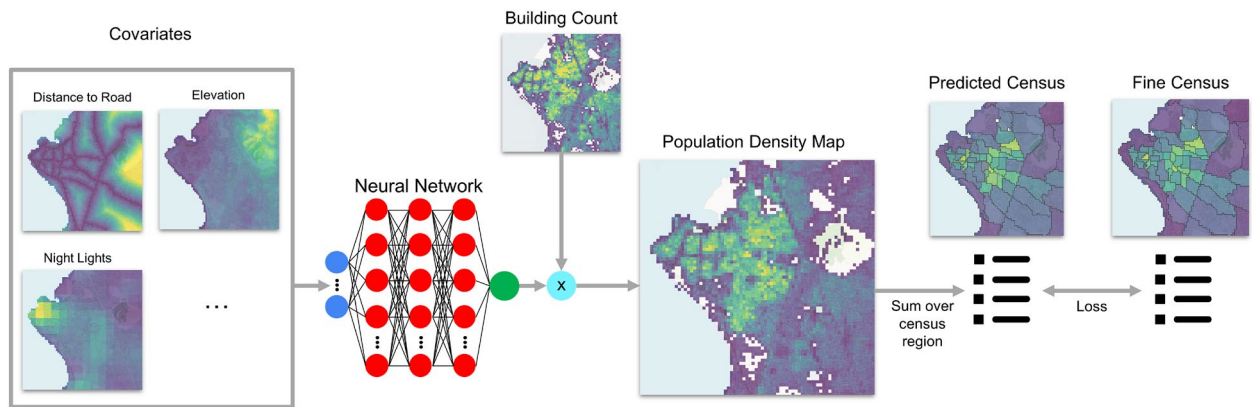
Figure 8. Overview of the proposed method: We use a neural network to create a mapping between covariates (blue circles) and population density. The direct output of the neural network are building occupancy rates (green circle) which we multiply with the building count map to obtain population counts. We receive the supervision signal for training by aggregating the population maps to the respective census regions through summation, and comparing our estimates with the available census data.

This cumulative sum per region is the supervision signal used by our model. More specifically, we employ a neural network, c.f. Fig. 8, that maps the covariates to population density estimates and that can be trained with lower-resolution ground truth: rather than per-pixel supervision, which is not available, the training procedure compares *aggregated* estimates per administrative region to the available region-level census data.

We denote the covariates at grid location l_i as $X(l_i)$, the population estimate at the same location as $\hat{p}(l_i)$, and c_j as the true census data for region A_j. We observe that the correlation between the number of buildings at a given location and its population turns out to be so strong that even dasymetric disaggregation based *only* on building counts produces reasonable (albeit less precise) population density maps, as shown in the "Results" section. We exploit this strong correlation for the design of the neural network's architecture as follows: To obtain even more accurate estimates, we predict a spatially varying *building occupancy rate* $f_\theta(X(l_i))$, rather than the population count directly. The population estimate $\hat{p}(l_i)$ is then computed from it as

$$\hat{p}(l_i) = f_\theta\big(X(l_i)\big) \cdot b(l_i), \tag{1}$$

where $b(l_i)$ is the number of buildings at location l_i (obtained from existing building products[31,32]) and θ are the parameters of the neural network. We use a neural network composed of a sequence of convolutional layers with kernel sizes 1×1, i.e., its architecture is equivalent to a per-pixel multi-layer perceptron (MLP). Empirically, we observe that adding more spatial context (e.g., via convolutional layers with kernel sizes > 1) does not significantly improve the model's prediction performance (see results in Table 7). At first this may seem surprising, but at the low GSD of our maps most covariates are smooth. Consequently, the disadvantages of a large receptive field outweigh the potential benefits: feature values of neighboring pixels tend to be very similar and hence do not add much information. Moreover, the spatial averaging effect still causes some loss of high-frequency detail (a similar effect has been observed by other authors, e.g., de Lutio et al.[22]). Also, the increased number of parameters is more prone to overfitting. A practical benefit of using 1×1 kernels is that, in the absence of spatial interactions, there is no spatial diffusion of the input information. All computations can be restricted to the small fraction of grid cells with at least one building. This makes the network memory-efficient, an important feature when processing large countries with coarse census regions, as GPU memory is often the computational bottleneck for modern neural networks.

The fundamental insight behind POMELO (and any other top-down disaggregation scheme) is that, given enough data, the weak supervision is sufficient: the sum over many grid cells is only a weak constraint per census region, but it accumulates over different regions because similar covariate features at different locations must yield similar population densities; whereas the input covariates have (at least) the target resolution and inject the missing high-frequency information. Formally, measuring the loss $\mathcal{L}$ over all grid cells l_i in each region A_j gives rise to the following optimization problem:

$$\hat{\theta} = \underset{\theta}{\mathrm{argmin}} \sum_j \mathcal{L}\big(\hat{c}_j, c_j\big),$$
$$\text{with } \hat{c}_j = \sum_{i \in A_j} \hat{p}(l_i). \tag{2}$$

Empirically, the loss function that achieves the best results is the L_1 distance between the log-transformed predictions and targets:

$$\mathcal{L}\big(\hat{c}_j, c_j\big) = |\log \hat{c}_j - \log c_j|. \tag{3}$$

Method	LogL1	Occ.	Aug.	$R^2 \uparrow$ [%]	MAE $\downarrow$	MAPE $\downarrow$ [%]
Baseline				71.4 ± 0.3	4130 ± 80	27.5 ± 0.9
+ Log L_1 loss	✓			73.2 ± 0.3	3580 ± 30	23.0 ± 0.2
+ Occupancy rate		✓		81.4 ± 1.2	3820 ± 180	26.8 ± 1.2
+ Augmentation			✓	73.0 ± 1.2	3820 ± 30	25.0 ± 0.4
POMELO	✓	✓	✓	85.8 ± 0.7	3100 ± 30	21.7 ± 0.3

Table 6. Effect of modeling choices (Tanzania data, coarse supervision).

Training setting	Kernel size	#params (k)	$R^2 \uparrow$ [%]	MAE $\downarrow$	MAPE $\downarrow$ [%]
Coarse	1 × 1 (ours)	34	85.7 ± 0.9	3100 ± 40	21.6 ± 0.2
	3 × 3	313	83.5 ± 1.3	3700 ± 110	26.6 ± 0.7
	5 × 5	870	80.9 ± 1.0	3900 ± 100	28.4 ± 0.8
Fine	1 × 1 (ours)	34	87.6 ± 0.2	2890 ± 20	20.4 ± 0.3
	3 × 3	313	86.0 ± 2.1	3010 ± 100	20.9 ± 0.5
	5 × 5	870	86.8 ± 1.0	3030 ± 50	21.5 ± 0.4

Table 7. Performance with varying kernel sizes on the dataset of Tanzania.

If census data are available at inference time, we can leverage the associated, additional constraints to perform disaggregation (dasymetric mapping) as a form of post-processing. In other words, we treat the final outputs of our model $\hat{p}(l_i)$ not as absolute quantities but as relative proportions. By linearly rescaling the predictions within a census region j such that they add up to 1, we obtain weights that indicate what fraction of the total census count c_j should be assigned to each location l_i. By multiplying those weights with the c_j, we then obtain the disaggregated population estimates $\hat{p}_{adj}$, adjusted to exactly match the coarse census counts:

$$\hat{p}_{adj}(l_i) = \frac{\hat{p}(l_i)}{\sum_{k \in A_j} \hat{p}(l_k)} \cdot c_j. \tag{4}$$

As long as the census data are correct, that adjustment can be expected to improve the population density map compared to the raw, absolute estimates. If no census counts are available at all, we just keep the raw estimates $\hat{p}(l_i)$.

Model setup. Our neural network is composed of four convolutional layers with kernel sizes 1×1 or, equivalently, of a four-layer MLP that is applied per pixel. The first three layers each have 128 filters, and the last layer has one filter to output a scalar density value per location. We use dropout[50] (with probability 0.4), and rectified linear unit (*ReLU*) layers after each convolutional layer. We apply a *softplus* function to the output of the last convolutional layer, constraining the occupancy rates to positive numbers. The model is trained using the Adam optimizer[51] with a base learning rate of 0.0001. The weight decay parameter for regularization is optimized via grid-search on the validation set. To mitigate the low number of training samples, we propose a data augmentation strategy specific to our task, namely we create artificial "pseudo-regions" from two real administrative regions, by merging their pixels and summing their population counts.

The importance of the main components of our model design is ablated in Table 6. It shows quantitative results obtained with different model setups for the scenario with *coarse supervision in Tanzania*. We start from a baseline that has the same model architecture, but

1. employs the standard L_1 distance between predictions and target values as loss function (no log L_1 loss);
2. directly predicts population counts per grid cell, and
3. is trained only with the actual administrative regions (no region-based augmentation).

We then gradually add the components used in our proposed setting. First, we find that computing the loss with log-transformed outputs significantly improves over the standard L_1 loss. Second, predicting the building occupancy rate instead of the population and transforming it to a population count in postprocessing also clearly reduces the error, especially the R^2 metric. Third, data augmentation by synthetically merging census regions also has a positive effect, confirming the beneficial effect of problem-specific augmentation when training data is scarce. Finally, combining all three measures, as in our proposed model, yields greatly improved predictive skill, reaching $\approx 25\%$ lower mean absolute error, respectively $> 14\%$ points higher R^2.

Table 7 show the effect of larger kernel widths on POMELO's performance for Tanzania, in both the coarse and fine supervision settings. It can be seen that larger kernel sizes are consistently detrimental, it appears that the they increase model complexity beyond the level that can be learned with the available supervision. On the one hand the bigger kernels inflate the number of weights that must be trained (9× more, respectively 25× more).

Method	$R^2 \uparrow$ [%]	MAE $\downarrow$	MAPE $\downarrow$ [%]
All features (ours)	87.6 ± 0.2	2890 ± 20	20.4 ± 0.3
Top 5 features	85.7 ± 0.4	2980 ± 10	20.7 ± 0.0

Table 8. Performance on Tanzania data with reduced covariate set (fine supervision setting).

These additional degrees of freedom increase the risk of overfitting and make learning harder—particularly in our weakly supervised setting, which becomes apparent with coarse supervision from only 170 regions. While on the other hand the covariate maps in general exhibit a high degree of spatial smoothness and do thus do not contain high-frequency context that larger kernels could capture.

We also empirically study the choice of input covariates. To that end, we only keep the five input layers with the highest individual importance according to the permutation method[38] and discard the remaining features. The results with this reduced input set for Tanzania (fine supervision setting) are reported in Table 8. Empirically, explicit variable selection noticeably harms performance. Apparently features with lower importance scores still carry valuable information that complements the most important features. In this context, note that in the proposed scheme the number of input layers is not critical: with the POMELO architecture, an additional covariate only introduces 128 learnable parameters. The computational savings achievable with feature selection are negligible.

Methods used for numerical comparisons.   We compare POMELO with four other methods:

Building count disaggregation Dasymetric disaggregation using only the available building counts per pixel as weights.

Random Forest (RF) at region level Our own re-implementation of the random forest model used by WorldPop[6]. The model is fed the same covariates (features) as the POMELO network. The training units for that scheme are not grid cells but administrative regions, with features aggregated over all pixels within a region. The map visualizations were created in QGIS 3.14[26].

Markov random field (MRF) Here, disaggregation is based directly on the assumption that locations with similar features have similar population densities. To enforce such prior knowledge explicitly, we resort to an MRF[52] that compares pixel-wise predictions to each other and to their encompassing region. This involves finding the combination of population values that minimizes the following (negative log-likelihood) energy function:

$$\hat{\mathbf{p}}_{MRF} = \arg\min_{\mathbf{p} \in \mathbb{R}^N} \sum_i^N \sum_{k \in Q_i} |p_i - p_k| + \lambda \sum_j |c_j - \sum_{k \in A_j} p_k| \,, \tag{5}$$

where N is the number of grid cells and Q_i the set of nearest neighbors to cell l_i in the (normalized) feature space. The first term encourages locations with similar features (i.e., the k nearest neighbors to pixel i in feature space) to have similar population counts; the second term pushes the aggregate count over an administrative region towards the region's total census. Parameter λ determines the balance between the two constraints; for all the experiments we set $\lambda = 1$. We create the (approximate) nearest-neighbor graph with the fast ANN method[53], initialize the population counts with the *building count disaggregation* described above, and find an optimal configuration by minimizing the energy function with the Iterated Conditional Modes algorithm[52], using update steps of ±1%. Empirically, we found that using too many covariates as features harms the performance of the MRF model, likely due to the higher dimension of the associated feature space. The best performance, reported in the "Results" section, are obtained with only three covariates: building count, average building size, and night lights.

Convolutional neural network (CNN) It has been proposed to estimate population counts with a CNN with spatial context[19], trained with the same form of aggregated supervision as POMELO. We instead advocate for a shared per-pixel model. Hence, we further include a baseline that has the same network architecture as POMELO, with two exceptions: the kernel size is set to 3 × 3 for all convolutional layers, and the network is trained to directly output population as suggested by[19] (rather than densities, as in our system). We point out that, while our CNN baseline is inspired by that study, the two are not directly comparable: Jacobs et al. target local population mapping in urban areas and base their estimates on high-resolution Planet imagery with 3m GSD, and on city block-level population and housing counts from the US census. Consequently, they can also afford to train a much larger (U-net[54]) model.

Evaluation metrics.   For all three evaluation strategies, shown in the "Results" section, the estimated maps are evaluated by aggregating the per-pixel counts back to a list of n_c population numbers Θ at finest available census level and comparing them to the actual census counts **c** in terms of R^2, mean absolute error (MAE), and mean absolute percentage error (MAPE):

$$R^2(\hat{\mathbf{c}}, \mathbf{c}) = 1 - \frac{\sum_{j=1}^{n_c}(c_j - \hat{c}_j)^2}{\sum_{j=1}^{n_c}(c_j - \bar{c})^2}$$

$$\text{MAE}(\hat{\mathbf{c}}, \mathbf{c}) = \frac{1}{n_c}\sum_{j=1}^{n_c}|c_j - \hat{c}_j| \qquad (6)$$

$$\text{MAPE}(\hat{\mathbf{c}}, \mathbf{c}) = 100 \cdot \frac{1}{n_c}\sum_{j=1}^{n_c}\left|\frac{c_j - \hat{c}_j}{c_j}\right|$$

Data availability

References

1. Naess, P. Urban planning and sustainable development. *Eur. Plan. Stud.* **9**, 503–524 (2001).
2. Lang, S. *et al.* Earth observation tools and services to increase the effectiveness of humanitarian assistance. *Eur. J. Remote Sens.* **53**, 67–85 (2019).
3. Maas, P. *et al.* Facebook disaster maps: Aggregate insights for crisis response & recovery. In *ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Vol. 19, 3173 (2019).
4. Thomson, D. R., Rhoda, D. A., Tatem, A. J. & Castro, M. C. Gridded population survey sampling: A systematic scoping review of the field and strategic research agenda. *Int. J. Health Geograph.* **19**, 1–16 (2020).
5. Vargas-Muñoz, J. E., Srivastava, S., Tuia, D. & Falcão, A. X. OpenStreetMap: Challenges and opportunities in machine learning and remote sensing. *IEEE Geosci. Remote Sens. Mag.* **9**, 184–199 (2020).
6. Stevens, F. R., Gaughan, A. E., Linard, C. & Tatem, A. J. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS ONE* **10**, e0107042 (2015).
7. Weber, E. M. *et al.* Census-independent population mapping in northern Nigeria. *Remote Sens. Environ.* **204**, 786–798 (2018).
8. Hillson, R. *et al.* Methods for determining the uncertainty of population estimates derived from satellite imagery and limited survey data: a case study of Bo City, Sierra Leone. *PLoS ONE* **9**, e112241 (2014).
9. Leasure, D. R., Jochem, W. C., Weber, E. M., Seaman, V. & Tatem, A. J. National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty. *Proc. Natl. Acad. Sci.* **117**, 24173–24179 (2020).
10. Huang, X., Wang, C., Li, Z. & Ning, H. A 100 m population grid in the CONUS by disaggregating census data with open-source microsoft building footprints. *Big Earth Data* **5**, 112–133 (2021).
11. Freire, S., MacManus, K., Pesaresi, M., Doxsey-Whitfield, E. & Mills, J. Development of new open and free multi-temporal global population grids at 250 m resolution. *Association of Geographic Information Laboratories in Europe (AGILE)* (2016).
12. Sapena, M. *et al.* Empiric recommendations for population disaggregation under different data scenarios. *PLoS ONE* **17**, e0274504 (2022).
13. Tiecke, T. G. *et al.* Mapping the world population one building at a time. *arXiv preprint* arXiv:1712.05839 (2017).
14. Sridharan, H. & Qiu, F. A spatially disaggregated areal interpolation model using light detection and ranging-derived building volumes. *Geograph. Anal.* **45**, 238–258 (2013).
15. Reibel, M. & Bufalino, M. E. Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environ. Plan. A* **37**, 127–139 (2005).
16. Dobson, J. E., Bright, E. A., Coleman, P. R., Durfee, R. C. & Worley, B. A. LandScan: A global population database for estimating populations at risk. *Photogramm. Eng. Remote Sens.* **66**, 849–857 (2000).
17. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
18. Gervasoni, L., Fenet, S., Perrier, R. & Sturm, P. Convolutional neural networks for disaggregated population mapping using open data. In *IEEE Conference on Data Science and Advanced Analytics (DSAA)* 594–603 (2018).
19. Jacobs, N., Kraft, A., Rafique, M. U. & Sharma, R. D. A weakly supervised approach for estimating spatial density functions from high-resolution satellite imagery. In *ACM SIGSPATIAL Conference on Advances in Geographic Information Systems* 33–42 (2018).
20. Diebel, J. & Thrun, S. An application of Markov random fields to range sensing. In *Advances in Neural Information Processing Systems* (2006).
21. Hui, T.-W., Loy, C. C. & Tang, X. Depth map super-resolution by deep multi-scale guidance. In *European Conference on Computer Vision (ECCV)* 353–369 (2016).
22. de Lutio, R., D'Aronco, S., Wegner, J. D. & Schindler, K. Guided super-resolution as pixel-to-pixel transformation. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2019).
23. Worldpop. http://www.worldpop.org (2022) (Accessed 20 June 2022).
24. Mennis, J. Generating surface models of population using dasymetric mapping. *Prof. Geogr.* **55**, 31–42 (2003).
25. Grippa, T. *et al.* Improving urban population distribution models with very-high resolution satellite information. *Data* **4**, 13 (2019).
26. QGIS Development Team. *QGIS Geographic Information System, Version 3.14*. Open Source Geospatial Foundation (2009).
27. OpenStreetMap contributors. Data retrieved from https://planet.osm.org, https://www.openstreetmap.org (2022).
28. European Space Agency. *Sentinel-2.*
29. Meta. *Facebook data for good high resolution population density maps* (2022).
30. Gorelick, N. *et al.* Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**, 18–27 (2017).
31. Sirko, W. *et al.* Continental-scale building detection from high resolution satellite imagery. *arXiv preprint* arXiv:2107.12283 (2021).
32. Dooley, C., Leasure, D. & Tatem, A. Gridded maps of building patterns throughout sub-Saharan Africa, version 1.1. Tech. Rep., (University of Southampton, 2020).
33. Vargas-Muñoz, J. E., Lobry, S., Falcão, A. X. & Tuia, D. Correcting rural building annotations in OpenStreetMap using convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **147**, 283–293 (2019).
34. Nieves, J. J. *et al.* Annually modelling built-settlements between remotely-sensed observations using relative changes in subnational populations and lights at night. *Comput. Environ. Urban Syst.* **80**, 101444 (2020).
35. Nieves, J. J. *et al.* Predicting near-future built-settlement expansion using relative changes in small area populations. *Remote Sens.* **12**, 1545 (2020).

36. Neal, I., Seth, S., Watmough, G. & Diallo, M. S. Census-independent population estimation using representation learning. *Sci. Rep.* **12**, 1–12 (2022).
37. Barrington-Leigh, C. & Millard-Ball, A. The world's user-generated road map is more than 80% complete. *PLoS ONE* **12**, e0180698 (2017).
38. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
39. Esch, T. *et al.* World settlement footprint 3D—a first three-dimensional survey of the global building stock. *Remote Sens. Environ.* **270**, 112877 (2022).
40. Meta. *Facebook Marketing API* (2022).
41. Fatehkia, M., Coles, B., Ofli, F. & Weber, I. The relative value of Facebook advertising data for poverty mapping. In *AAAI Conference on Web and Social Media* 934–938 (2020).
42. Leasure, D. R. *et al.* Ukraine crisis: Monitoring population displacement through social media activity. *SocArXiv*, https://doi.org/10.31235/osf.io/6j9wq (2022).
43. Elvidge, C. D., Baugh, K. E., Zhizhin, M. & Hsu, F.-C. Why VIIRS data are superior to DMSP for mapping nighttime lights. *Proc. Asia-Pac. Adv. Netw.* **35**, 62 (2013).
44. Corbane, C., Florczyk, A., Pesaresi, M., Politis, P. & Syrris, V. An overview and comparison of smooth labeling methods for land-cover classification. *IEEE Trans. Geosci. Remote Sens.* https://doi.org/10.2905/jrc-ghsl-10007 (2018).
45. Farr, T. G. *et al.* The shuttle radar topography mission. *Rev. Geophys.* **45** (2007).
46. Lamarche, C. *et al.* Compilation and validation of SAR and optical data products for a complete and global map of inland/ocean water tailored to the climate modeling community. *Remote Sens.* **9**, 36 (2017).
47. UNEP-WCMC & IUCN. *Protected Planet: The World Database on Protected Areas (WDPA)* (2021).
48. Humanitarian data exchange. https://data.humdata.org/ (2022) (Accessed 20 June 2022).
49. Zhou, Z.-H. A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **5**, 44–53 (2018).
50. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
51. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)* (2014).
52. Besag, J. On the statistical analysis of dirty pictures. *J. R. Stat. Soc.* **68**, 259–302 (1986).
53. Johnson, J., Douze, M. & Jégou, H. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data* **7**, 535–547 (2019).
54. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention* 234–241 (2015).

Acknowledgements

Author contributions

D.T., K.S., T.W., and F.O. conceived the study, N.M., J.V., R.D., K.S., and D.T. designed the research, N.M., and J.V. implemented the methods. All the authors discussed the results and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to N.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.