Application of deep-learning techniques to very-high-resolution satellite images supporting population censuses in developing countries

Gafarou Kpegouni,[a,*] Yacine Bouroubi,[a] Harolde Coulombe,[b] Damien Echevin,[b] Etienne Lauzier-Hudon,[a] and Mikaël Germain[a]
[a]Sherbrooke University, Department of Applied Geomatics, Faculty of Letters and Human Sciences, Sherbrooke, Quebec, Canada
[b]ApexMachina, Montreal, Quebec, Canada

Abstract. Knowledge of demographic data is valuable information for planning initiatives. Typically, census, survey, and population projection exercises provide this information. In some developing countries, these operations pose a variety of economic and logistical challenges, thereby depriving authorities of accurate and timely information on their populations. To provide approaches for solving this situation, our study evaluates a population estimation method that is based on detection of residential geo-objects (houses) on very-high-resolution (VHR) satellite images using convolutional neural networks (CNN). The approach would be applicable to countries where a complete census is difficult to perform due to resource constraints or political instability. A 2008 VHR satellite image of Sudan is annotated according to seven classes of buildings to create a dataset that was used to train an object detection model, faster region-based CNN, by transfer learning. The model obtained mean average precision of 79% and 99% during training and validation, respectively. This unusual difference is due to the dominance of well detected classes in the validation dataset. The model was fine-tuned to detect the same building classes on images in 2021. A link between residential geo-objects and population size was established using 2008 population data and available field data. Subsequent characterization of the current population should assist in preparation of the 2023 census. Limitations of this approach were raised, but it could be used to improve the framework for population data collection in developing countries. [DOI: 10.1117/1.JRS.17.024506]

Keywords: population estimates; very high resolution satellite images; deep-learning; residential geo-object detection; convolutional neural network; faster region-based convolutional neural networks.

Paper 220632G received Nov. 5, 2022; accepted for publication Apr. 5, 2023; published online Apr. 18, 2023.

1 Introduction

A population census is one of the most complex and costly statistical exercises that a nation can periodically undertake.[1] It provides information on the number and characteristics of a population in terms of its density and spatial distribution, sex and age structure, and other fundamental social and economic characteristics, including its housing conditions both nationwide and in each locality. This information is frequently used to assess, for example, future demand for food, water, energy, and services.[2] The use of such data in development projects has the merit of saving planners from the risks of poor targeting and poor design of development plans, for example.[3,4]

Conducting a census requires that a range of different challenges be overcome under diverse circumstances.[5] Although many industrialized countries establish population data collection frameworks with rigorous solutions,[6] censuses in many developing countries are not representative of the population, due to weak civil registration and vital statistics systems, for example. They are expensive to conduct, are rarely conducted (at best every 10 years), and provide only

*Address all correspondence to Gafarou Kpegouni, gafarou.kpegouni@usherbrooke.ca

population averages over large areas.[7] Furthermore, in some of these countries, conflict and violence affect population data collection operations. In this context, and in view of the difficulties faced by the census services of the various countries, it is imperative to explore alternative solutions for the collection of population data. Innovative modeling approaches, particularly those based on Earth observation data, can produce population estimates that are of growing interest in such situations.[5]

In low-income countries, census or polling managers are confronted by challenges, such as difficulties in mobilizing financial and material resources, technical shortcomings in data processing, socio-political crises, conflicts, and natural hazards, among others.[4] This affects the process of population data collection operations. In the absence of a census or population surveys, population projection models are used to provide an overview of a country's population landscape.[2] Yet, estimates are obtained by demographic models, which in some cases do not reflect relevant social indicators or they lack reliability.[8] When faced with challenges that are related to data collection in these problematic situations, the following question arises: how does one go about collecting alternative data? Important technical advances in geomatics can offer various solutions. Remote sensing data have been used in several studies to assess population density at multiple scales.[9–14]

The number of people living in an area cannot be observed directly from remote sensing data, but the latter can be used as a tool for detection and determining population density estimates. Currently, very high resolution (VHR, 30 to 50 cm) satellite images can provide very detailed ground information and, therefore, are well suited to the detection and analysis of human activity.[15,16] Nevertheless, extracting objects such as buildings from remotely sensed images is an important task with many applications, but it also remains a challenging task due to very large variations in the shape and features of buildings.[17] Image processing operations, such as filtering, edge detection, or segmentation, have been proposed for this purpose.[18,19] Yet, none of these methods provide an unambiguous answer as to the precise location of buildings, let alone their properties.[16] Due to the advantages of accurate building detection in VHR satellite images, work on building detection models has accelerated considerably since the introduction of neural networks. These research efforts have grown even more dynamically since 2013, when convolutional neural networks (CNNs) were first introduced into image processing. CNNs have improved several image analysis applications, including object detection[20–22] and semantic segmentation.[23–27] With the development of these efficient deep-learning algorithms, it has become possible to achieve high accuracy when performing remote sensing analysis on VHR images, especially in the context of building detection and classification.[16] This study aims to contribute to better estimation of the population by an approach using detection of building classes on VHR satellite images using CNN architecture, more specifically faster region-based CNN (R-CNN). The approach is applied to Sudan, a country where a complete census is difficult to conduct due to resource constraints and political instability. Specifically, the study aims: (a) to detect the types of residential geo-objects (houses) on VHR images by proposing an adequate taxonomy and an adapted CNN architecture; (b) to establish a link between each class of residential geo-objects and the average number of inhabitants, based upon an old census (2008); and (c) to estimate the population from recent VHR images.

2 Materials and Methods

2.1 *Study Site*

The study site is Sudan, which is the second largest country in Africa (1.86 million km^2). The 2008 census is the most recent source of population data. These data are old and not very useful for planning purposes. The ever-changing socio-economic and demographic structures of the country, which were brought on by war that has lasted for more than two decades, merit attention respecting the issue of population data collection for this vast nation. Sudan has a vast area, diverse topography, and varying climatic conditions across its land surface. Its socio-political structure is complex given its socio-cultural fabric, which is characterized by a multifaceted tribal, ethnic, and religious composition. Different levels of development between regions are

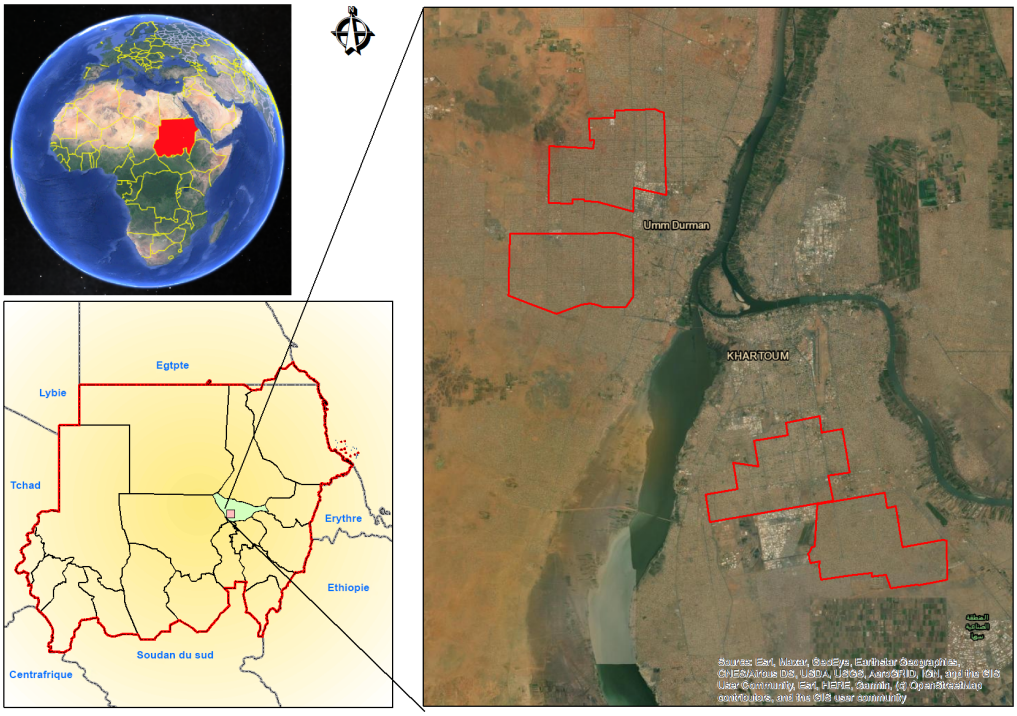


Fig. 1 Location of the study site in Greater Khartoum, Sudan.

easily discernable. All these aspects make the census a very difficult undertaking, especially under conflict or postconflict conditions. The high mobility of the population within the country and neighboring countries, the very high number of internally displaced persons, and the widely dispersed nomadic population add another dimension to the complexity of a comprehensive census project (source: Ref. 28).

In this study, we focus upon the principal urban area of the country, i.e., Greater Khartoum. This tripartite metropolis has more than five million inhabitants, who dwell in the capital Khartoum, which is joined by bridges spanning tributaries of the Nile to North Khartoum and Omdurman; Omdurman is the second largest city in the country in terms of population (Fig. 1). This municipality was chosen because of its representativeness of different urbanization densities, together with the availability of population data and VHR satellite images from 2008.

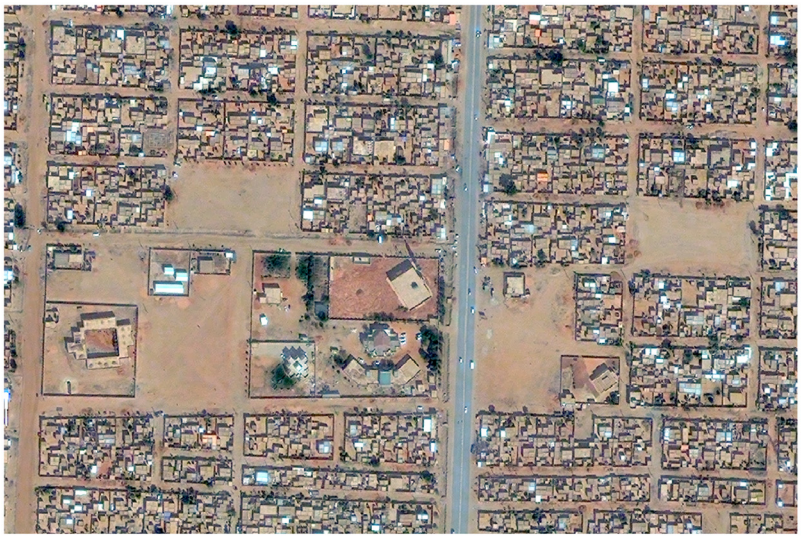
2.2 *Data*

2.2.1 *Population data: 2008 census*

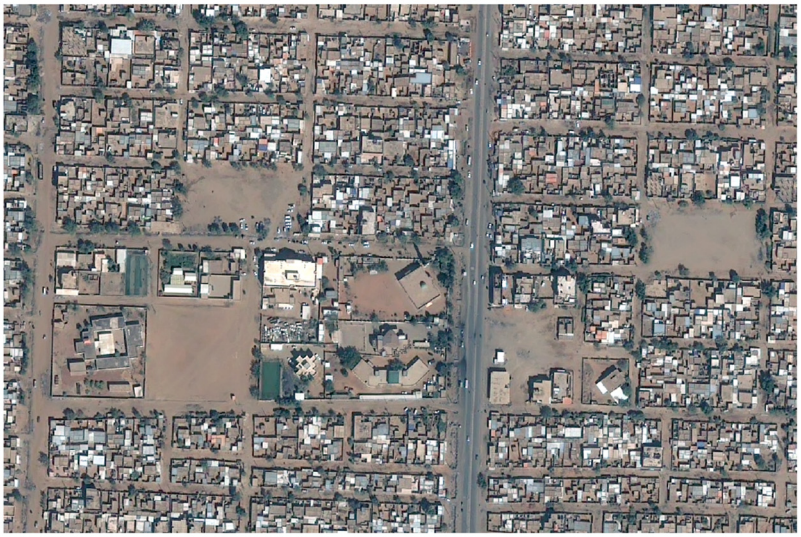
The main population data that were used is the 2008 census database. The file, which is in .shp format, contained the administrative divisions of the country, the division of certain large agglomerations into enumerated areas (EA), and the population sizes that were associated with these different divisions. Data on population size (per household) and housing characteristics were used, which had been collected during various surveys and polls that were conducted after the 2008 census.

2.2.2 *VHR satellite images*

The choice of search areas for VHR images corresponding to the 2008 census was based on the density of buildings/population, following verification of field data (photographs and videos), on Google Earth and on OpenStreetMap. Four subareas of interest were considered given the availability of good quality VHR images, which were provided at that time only by QuickBird (Fig. 1). This image (65 cm QuickBird with pan-sharpening) was acquired on January 16, 2009, which was close to the census year (hereafter, referred to as the 2008 image).



(a) **Extract of 2008 QuickBird image**



(b) **Extract of 2021 Pleiades image**

Fig. 2 Extracts of VHR images: (a) QuickBird from January 16, 2008 and (b) Pleiades from January 4, 2021.

The recent VHR image that was used to estimate the population and guide future censuses is a Pleiades image (50 cm with pan-sharpening) that was acquired on January 4, 2021. Figure 2 shows an example of built-up areas on the two types of VHR images that were used (QuickBird, 2009; Pleiades, 2021) in a residential area of Greater Khartoum.

2.2.3 *Taxonomy and constitution of the dataset*

To our knowledge, no semantically rich, annotated building detection dataset exists that takes into account the realities of our study area. Therefore, we constituted our own dataset for training the neural network.

Exploration of the 2008 census data allowed us to establish a taxonomy of residential geo-object classes (houses) according to the number of inhabitants per unit (Table 1). The data included a "household questionnaire" (housing characteristics), an examination of the VHR images, the state of the art, the images and photographs that were available on Google Earth, as well as field data (photographs taken on site).

Table 1 Description of different classes of residential geo-objects (houses).

Classes	Description
Class 1: Single-storey house	Structure without upper floors. May consist of several buildings forming an inhabited unit. Either a single- or multifamily dwelling.
Class 2: Two-story residential house	Dwelling with a ground floor and an upper level. Could consist of small, single-story buildings next to a large building with upper floor. Could be a single- or multifamily dwelling.
Class 3: Multilevel residential house	Ground floor plus 2 to 5 upper stories.
Class 4: Semi-commercial/residential house	Commercial buildings. Building generally divided into two: one part commercial, one part housing. It could be single-story or multilevel.
Class 5: Nonresidential house	Buildings not used for housing, e.g., schools, services, etc.
Class 6: Tall building	Large building exceeding five stories, approximate.
Class 7: Other	Places of worship and any residential structure, the membership of which cannot be established in any of the preceding classes.

Figure 3 illustrates the key to interpreting the taxonomy of different classes that were selected. Image preprocessing consisted of an enhanced QuickBird image in the RGB composite; this step was performed the match the RGB format of the datasets that were used in the CNN. Images were then sliced and converted to an unsigned 8-bit format, which was compatible with the libraries and faster R-CNN model that was used.

Manual annotation of residential geo-object classes, which were clipped to 10 QuickBird image areas of interest using a GIS tool (QGis), resulted in a .shp file with ground-truthed bounding boxes that mark terrain coordinates, while specifying class and probability of the object. The images of these areas of interest have been cut into thumbnails of 300×300 pixels, with 50% overlap between them. This overlap is applied so that information is not lost should a building be cut between two images. Due to the very limited quantity of annotated data, it was necessary to perform data augmentation (flips, and 90 deg and 270 deg rotations), which is a practice that is widely used with CNNs.[29] A total of 556 300×300 pixel thumbnails, covering 2413 annotations, were used for training (80% of the data) and validation (20% of the data) of the faster R-CNN model. They are distributed according to Table 2, which summarizes the class structure of the dataset. A large imbalance characterized the distribution of classes with an imbalance ratio (number of majority annotations/number of minority annotations) of 93.6. The distribution has been assigned in such a way that a thumbnail image and its versions resulting from the augmentation are found in the same dataset. The situation where an image seen in training is found during validation is thus avoided. Figure 4 shows examples of annotation of different classes of residential geo-objects.

2.3 *Methodology*

2.3.1 *Detection of residential geo-objects by deep learning*

The irregular nature of buildings in the study area and the difficulty of recognizing house classes in VHR images make the annotation task difficult. Judicious photo interpretation was necessary to provide a good dataset on which the detection model can be trained. The faster R-CNN model,[30] which has been shown to perform well in computer vision applications,[22,31] was used. With the generalization capabilities of pretrained models, for example, PASCALVOC-2012[32] in the case of faster R-CNN, it is possible to commence training from prior knowledge rather than starting from scratch.[29] Our residential geo-object detection approach is an application of this transfer learning technique.[33]

Fig. 3 Interpretation key for annotating house classes: (a) single-story house, (b) two-story house, (c) multistory house, (d) semicommercial house, (e) nonresidential house, (f) tall building, and (g) other.

Table 2 Distribution of annotations according to the class of residential geo-objects.

Residential geo-object class	Number of annotations	% annotation
1. Single-story house	1405	58.2
2. Two-story house	765	3.7
3. Multistory house	75	3.1
4. Semicommercial/residential house	90	3.7
5. Nonresidential house	15	0.6
6. Tall building	17	0.7
7. Other	46	1.9

The operation of faster R-CNN is based on three main components. (a) The first component is a pretrained model (i.e., backbone), the goal of which is to extract deep features from the input image. The convolutional neural network (ResNet50) was selected for the task after various tests of network depths; this particular network is 50-layers deep. (b) The second component is a region proposal network,[31] which generates bounding boxes in the input image, followed by a bounding box pooling layer that projects the attributes extracted by the backbone to a vector

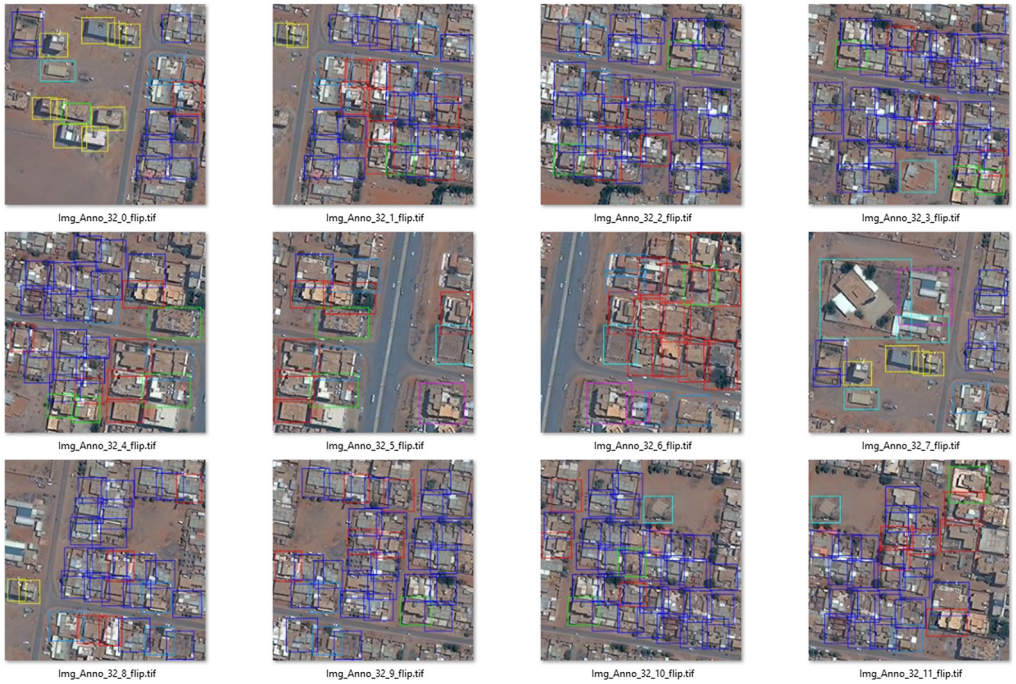


Fig. 4 Different types of annotations of residential geo-objects on VHR thumbnails.

and, finally, the application of classification and regression operators. (c) The third component is a classification operator that is combined with a regression operator to estimate the probability of each bounding box as belonging to a class (thanks to the classification operator) and to refine its final positioning (thanks to the regression operator). The cross-entropy loss function in Ref. 31 was used. Since our class distribution was not very homogeneous, the technique that was used by Ref. 34 is adopted, which consists of weighting the class loss function. The function was estimated as the total number of annotations per thumbnail, over all classes, divided by the total number of annotations per thumbnail of each class; 1 is subtracted from the result. We obtained the following weights: 1 (background); 0.41 (class 1: single-story house); 0.68 (class 2: two-story house); 0.96 (class 3: multistory residential house); 0.96 (class 4: semicommercial/semi-residential house); 0.99 (class 5: nonresidential house); 0.99 (class 6: tall building); and 0.98 (class 7: other). As is the case with hyper-parameters, we used the Adam stochastic gradient optimizer[35] with a learning rate of 0.0001 and a batch size of 10 thumbnails. Windows 11 and the Torchvision library from PyTorch were used as the computing framework. The computer that was used was equipped with an Intel Core i7 CPU and a Nvidia RTX GPU with 16 GB of memory.

2.3.2 *Detection evaluation*

The evaluation of the performance of our detection approach is achieved using mean average precision (mAP) in a multiclass detection system.[36] AP is a widely used metric in many object detection applications, where its values are calculated over recall values ranging from zero to one. Recall that "recall" is calculated as the number of true positives, divided by the sum of true positives plus false negatives. Computation of mAP thus involves important submetrics, such as recall, precision, intersection over union (IoU), the precision recall curve (PRC), and a confusion matrix.[36] The GitHub page by Rafael Padilla (https://github.com/rafaelpadilla/Object-Detection-Metrics) should be consulted regarding details of this metric and its concepts. Given the complexity of the built environment in the Greater Khartoum area, an IoU $\geq 50\%$ was used in our study. Furthermore, PRC evaluates classification tasks better with unbalanced data,[37] which justifies its selection for our study.

2.3.3 *Generalization of the model in the 2008 image*

Once the model is trained, inferential detection is performed over the entire 2008 image. This allows characterization of the buildings across the whole study area and uses these to establish the "house–population" relationship from the 2008 census.

2.3.4 *Characterization of residential geo-objects from the 2021 image*

The faster R-CNN model that was trained on the 2008 QuickBird image is used to detect houses in the 2021 image, after fine-tuning. Thus, annotations were performed on the 2021 images. In total, 2861 annotations were made, with proportions per class relatively similar to those of 2008. Unsurprisingly, they were made in the same areas as the 2008 image. The increase in the number of annotations on the 2021 image is explained primarily by the appearance of new buildings between 2008 and 2021 (see Fig. 2).

2.3.5 *Population estimates*

The 2008 census data provide a population count at the enumerated area (EA) level. These data were used to establish a "house–population" relationship. An average population factor is determined by residential geo-object class using the population data by EA from the 2008 Census according to the following equation

$$(N_1 \cdot \mu_1) + (N_2 \cdot \mu_2) + \ldots + (N_i \cdot \mu_i) \approx EA_{\text{pop}}, \tag{1}$$

where

N_i = number of geo − objects per class i, μ_i = average population size for class i, and

EA_{pop} = population size at the scale of the enumerated area.

For example, if EA_1 contains 100 residential geo-objects of class 1, 50 of class 2, and 0 of the other classes, and EA_1 has a population size of 600 according to the census, a possible solution might be that $\mu_1 = 5$ and $\mu_2 = 2$ ($600 = 100 \times 5 + 50 \times 2$). Choice of a likely solution is guided by the dominance of each house class in the EAs. Thus, all EAs are traversed to estimate μ that was found for each class while taking into account information from informal household surveys. We assumed that the household size per house class (μ_i) did not change between 2008 and 2021, having found no evidence for this.

3 Results and Discussion

3.1 *Detection of Residential Geo-Objects*

3.1.1 *On the 2008 image*

The faster R-CNN network was trained over 20 epochs. Figure 5 shows the mAP curve as a function of each epoch (i.e., one cycle or pass through the full training dataset). Following an increase in mAP during the first 10 epochs, the pattern achieved a stable plateau. The final mAP that was recorded at the end of the 20th epoch was about 79% during training and about 99% during validation. Loss reached 0.12 by the end of 20 epochs of training. Note that the constitution of the training and validation datasets was done randomly on the annotated thumbnails. However, the validation set contains almost instances of the majority classes (1 et 2) on which the model has very well performed in training. Since these are the instances that the model sees more in validation, the mAP in validation becomes higher than the mAP in training.

Given that training curves remain stable after a number of epochs reflects nonrepresentativeness in the dataset,[38] a nonrepresentative dataset may not capture statistical characteristics relative to another dataset that is drawn from the same domain, such as between training and validation data. This can usually happen if the number of samples in one dataset is too small

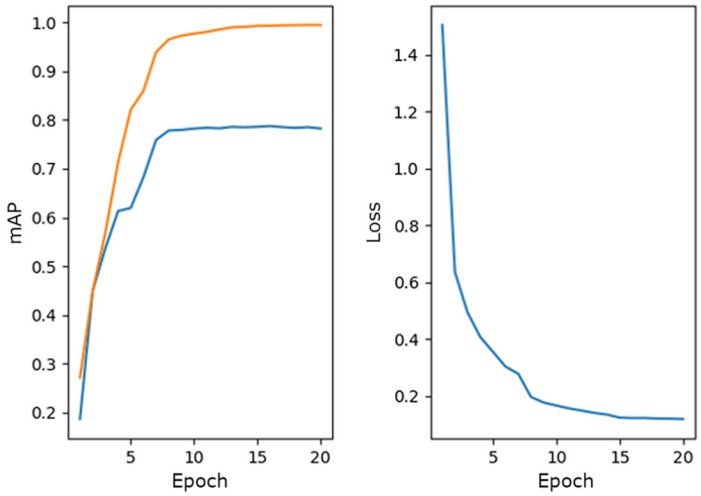


Fig. 5 Evolution of precision during training and validation, and loss in training for the faster R-CNN model.

compared with another dataset. Two common cases could be observed: (a) the training dataset is relatively unrepresentative; or (b) the validation dataset is relatively unrepresentative. In our particular case, it is the validation dataset that is poorly represented.

The precision–recall curve (Fig. 6) shows that six of the seven residential geo-object classes are well detected by the model, with precision ranging from 84% to 100% (Table 3). Detection of Class 5 ("nonresidential house") seems to be difficult, with an AP = 25%. Note that this is the least represented class in the dataset (< 1%). In general, the area under the precision–recall curve shows a large area between the curves and the baseline. This translates into perfect performance of the model and, therefore, appropriate functioning of the classifier, with a mAP of 79.46% for all seven classes.

In comparing the AP of each class to the distribution of class annotations (Table 2), the model clearly detects classes that appear more frequently in the training dataset and has difficulties for classes that appear less frequently. The detection performance of our approach is limited by the class imbalance.

Figure 7 shows the confusion matrix, which is a convenient and intuitive way of evaluating a model with respect to its detection performance on individual classes. In the case of a multiclass classification, a false negative in one class according to the prediction will be a false positive of

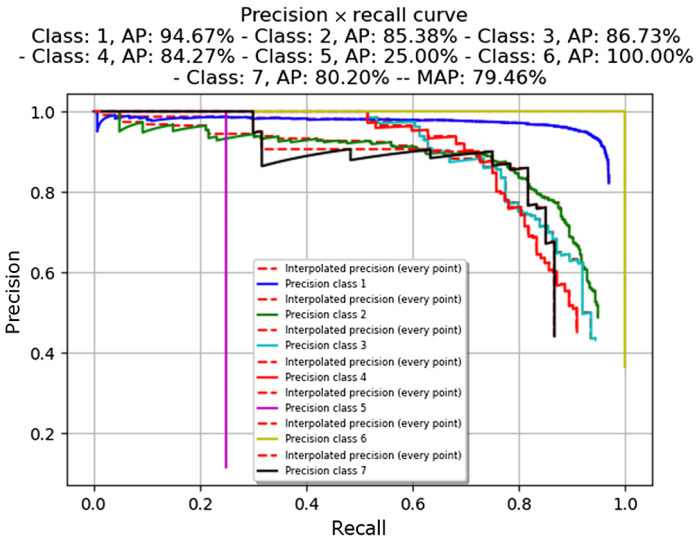


Fig. 6 Precision of detection for different house classes.

Table 3 Average precision (AP) among classes of residential geo-objects.

Residential geo-object class	Average precision (%)
1. Single-story house	94.67
2. Two-story house	85.38
3. Multistory house	86.73
4. Semicommercial/residential house	84.27
5. Nonresidential house	25.00
6. Tall building	100.00
7. Other	80.20

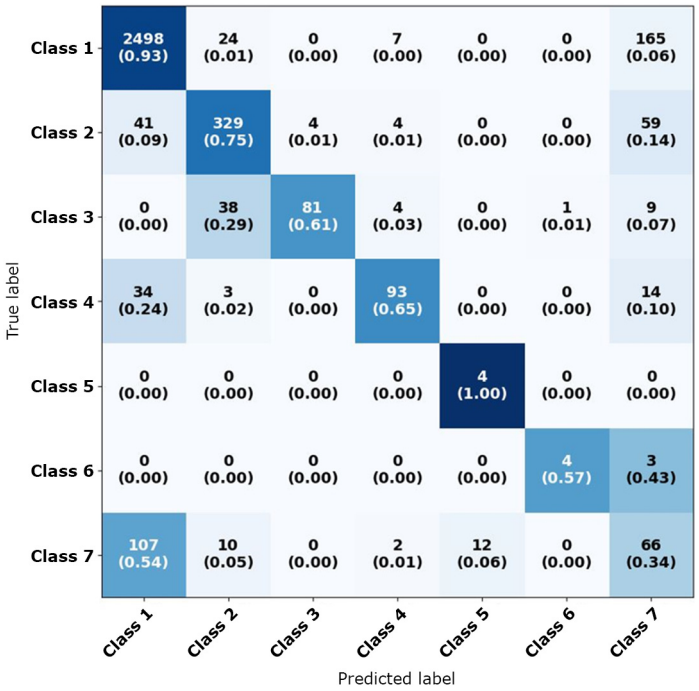


Fig. 7 Detection confusion matrix of the seven residential geo-object classes.

the other classes, i.e., false negatives are found in the rows (observed) that are associated with each class, and false positives are in the columns (predicted) that are associated with each class.

Classes 1, 2, 3, 4, and 5 are better detected than 6 or 7, with success rates exceeding 60%. Yet, it should be noted that some confusion in detection persists among classes. The most important ones are: (a) class 7 ("other") versus class 1 ("single-story house"), with 54% of class 7 confused with class 1; (b) class 6 ("tall building") versus class 7 ("other"), at 43%; and (c) between class 4 ("semicommercial/residential house") versus class 1, at 24%. Some classes have a detection rate below 60%. Other less important confusions are observed between several classes, reflecting the inability of the model to correctly detect certain types of buildings. Figure 8 shows examples of (a) and (b) good detections and (c) and (d) bad detections.

In general, large area geo-objects (e.g., nonresidential houses, schools, services, and mosques, among others) are not well detected. These geo-objects are often a collection of other classes. Occasionally, the model detects these classes individually in this case, rather than the

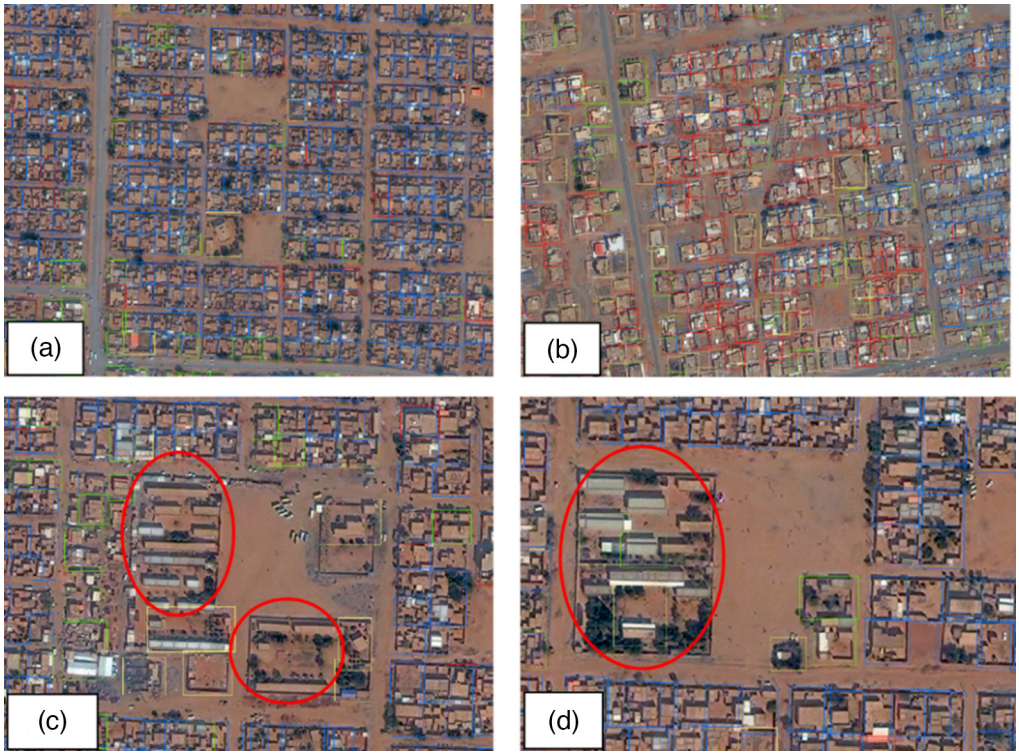


Fig. 8 Examples illustrating detections with faster-R-CNN on the 2008 QuickBird image: (a) and (b) good detection. (c) and (d) bad detections.

set forming a distinct geo-object [circled in Fig. 8(d)]. Another explanation could be that the area of these geo-objects is larger than the size of the patches (300 × 300 pixels) that were cut to train the model. The corresponding annotations are cut into several patches, thereby preventing the model from properly learning how to recognize these residential geo-objects.

Major confusion between class 1 and 7 (54%) could be explained by the very definition of the latter class. Indeed, class 7 is designated "other," given that it consists of a set of buildings for which identification was problematic during the annotation process. Class 1 geo-objects have characteristics in common with structures annotated in this seventh class; hence, the confusion that arose during detection. We also noted "empty fences" throughout the study area, i.e., parcels enclosed by a perimeter fence where the interior was empty, giving the appearance of a residential geo-object. The model detects these objects as constructed (either class 1 or class 7), but occasionally as background. This increases the confusion between classes 1 and 7.

3.1.2 *On the 2021 image*

Fine-tuning of the model on the Pleiades image resulted in detections of residential geo-objects in 2021 on all four subareas of the study site. Figure 9 illustrates detections, along with a zoomed inset of a residential area as an example.

3.2 *Population Estimates*

3.2.1 *Population estimated for 2008*

Analysis of the "house–population" relationship was performed on 87 EAs. Figure 10 shows the distribution of the numbers of each class detected by each EA and the average population/house ratio. The most dominant class among EAs is class 1, followed by classes 7 and 2. The average population/house ratio across the AEs varies greatly, ranging from 1.4 (likely an error in the census database) to 19.1 (likely a very special case).

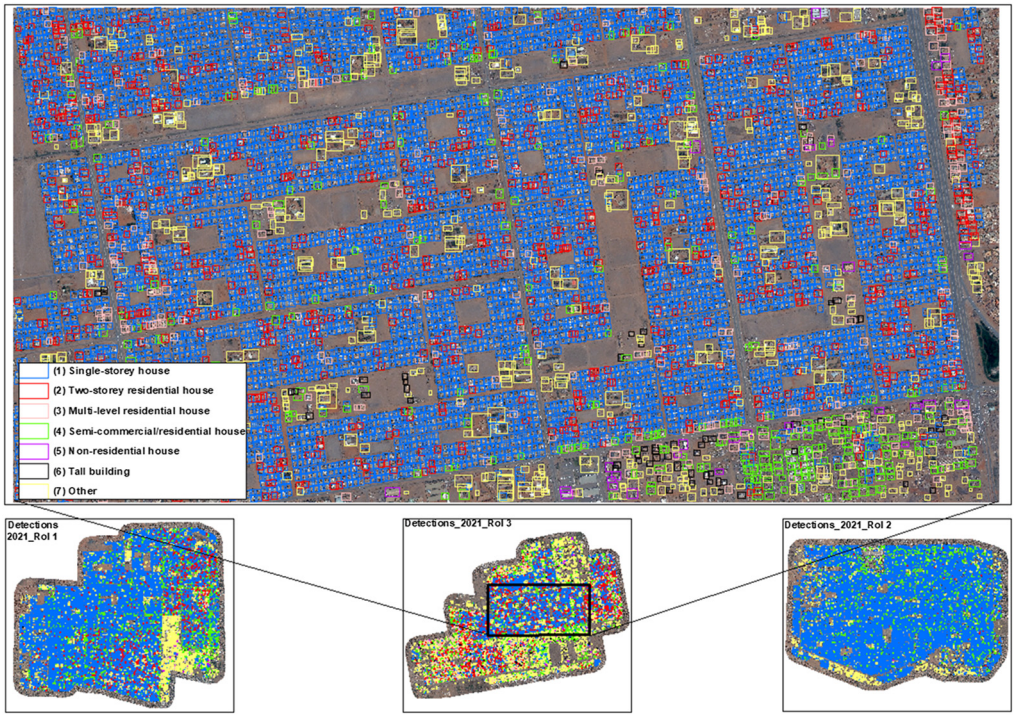


Fig. 9 Detection of residential geo-objects on Pleiades images from 2021 (on the ensemble of certain RoIs and extract of RoI).

We further note that variation in the average population/house ratio is related to the dominance of a class in an EA. The principle of the method for estimating this ratio is based on finding the EAs with a single dominant class and estimating its ratio, then looking for EAs with a second dominant class and evaluating its ratio, and so on.

The dominant classes that were detected are 1, 7, and 2, respectively. Their degree of dominance is related to variation in the population/house ratio. We note that for EAs with 90% dominance of class 1, the population/house ratio varies between 9 and 13. When class 7 is important (EAs with 35%), variation ranges between 4 and 9. When class 2 reaches 25% in the EAs, we observe a population/house ratio that varies between 4 and 7.

The different variations in the population/house ratio guided the choice of the μ factor (average population size per class), starting with the dominant classes and rejecting implausible extreme/erroneous values. For the other classes, estimation of μ is based on Eq. (1), while considering the data that were reported by field specialists. Thus, the μ-factors per residential geo-object class are as shown in Table 4.

Applying these different average sizes to all houses that were detected across all EAs yields a good correlation between the total population of EAs and the total population of all houses in that EA, as shown in Fig. 11, with the exception of some EAs that were investigated.

Analysis of Figs. 10 and 11 reveals three main sets (Table 5): (a) 6.9% of houses with $\mu > 2$; (b) 87.36% of houses with μ between 4 and 14; and (c) 5.75% of houses with $\mu < 15$. The set in case (b) corresponds to the range of population sizes per house that were generally reported in the field. The two possible outlier groups could likely be attributed to data errors, as discussed below.

We note the EAs that were of concern are all found in the same zone for each set (Fig. 12). EAs with a population/house size ratio < 2 are very close to an area where population data for some EAs are missing from the census database. Estimates for the population data that were captured in these EAs are likely erroneous (Fig. 12, in blue, EAs with missing data in the database and in black, EAs with probably erroneous data).

EAs with a population/house size ratio > 15 originate from areas where division of EAs is often difficult. In general, several enumerators are placed in these areas for systematic surveying, with the risk of double counting. Population estimates in these EAs are often high compared to

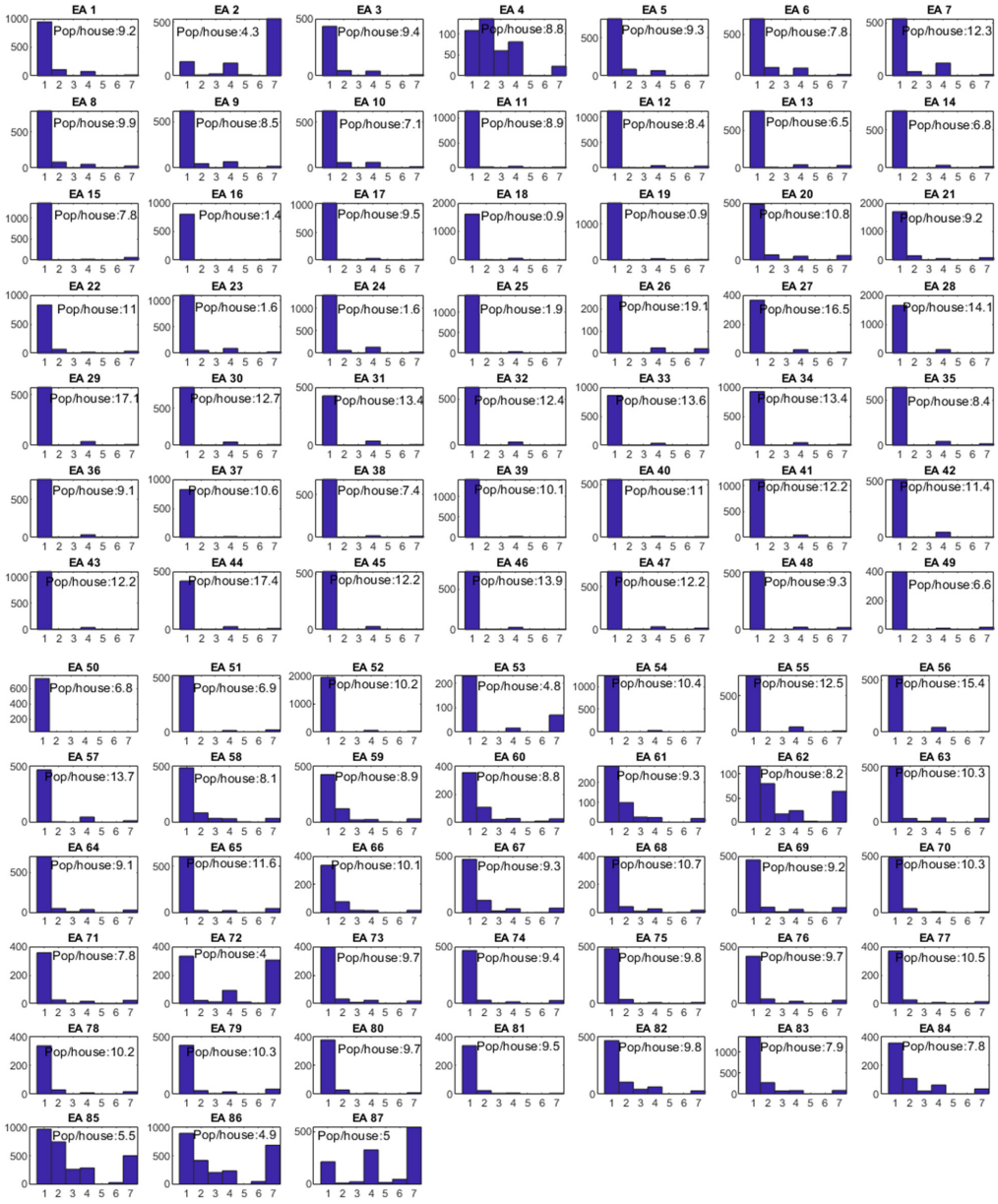


Fig. 10 Distribution of each class population/house ratio by EA.

Table 4 Estimated mean sizes by residential geo-object class.

Residential geo-object class	μ
1. Single-story house	13
2. Two-story house	9
3. Multistory house	7
4. Semicommercial/residential house	5
5. Nonresidential house	2
6. Tall building	5
7. Other	7

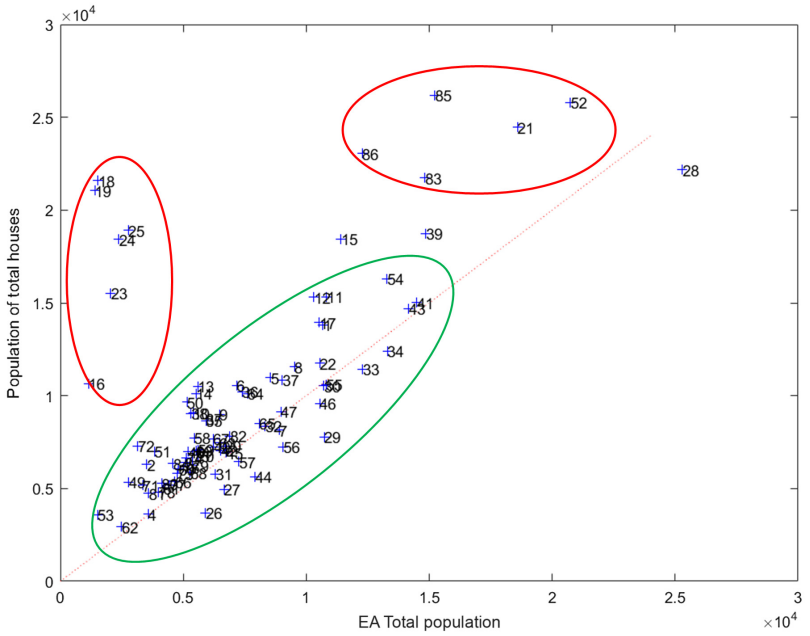


Fig. 11 Correlation of total EA population and total population per house.

Table 5 Variation in population/house ratios of EAs.

Population:house	Number of EAs	% EA
≤2	6	6.9
4 to 14	76	87.36
≥15	5	5.75

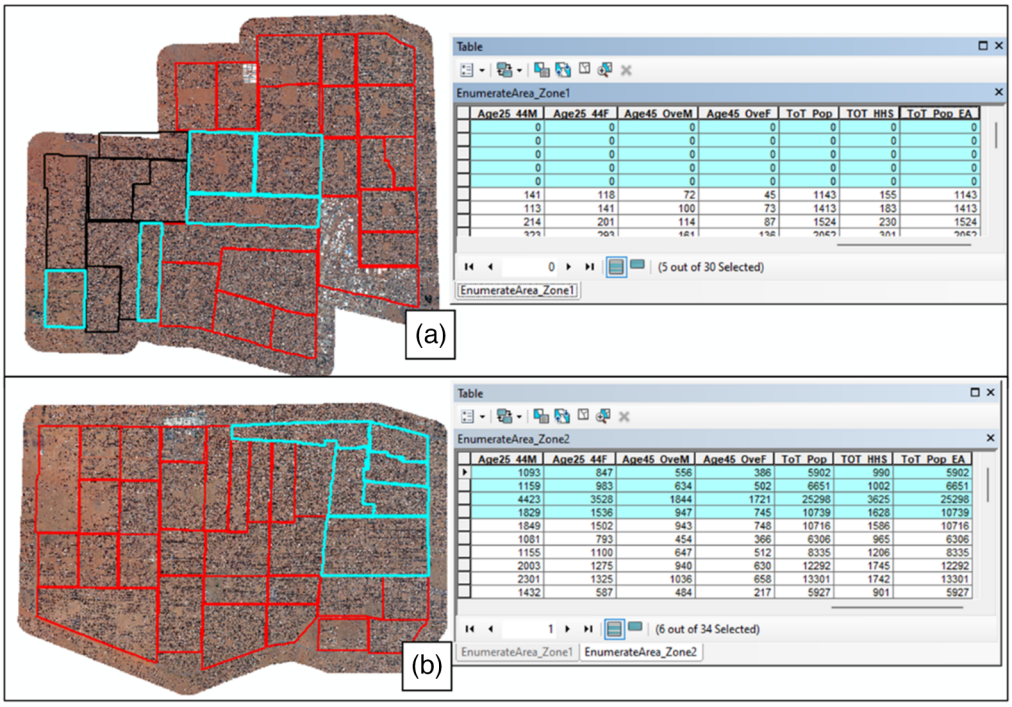


Fig. 12 Illustration explaining EAs with bad population data: (a) EAs with missing data and (b) EAs with probably erroneous data.

the average. Population sizes of 25,298 and 10,739 are observed in the EA of this zone [Fig. 12(b)].

3.2.2 *Population estimated for 2021*

Applying the model to the 2021 images resulted in detections of residential geo-objects in 2021 across the study site. Some examples at the scale of an EA are shown in Figs. 13 and 14. Tables 6 and 7 show an estimate of population in 2021 in these AEs by applying the average population sizes by house class determined above. The same value of the μ-factor (average population size per class) has been used (in the absence of a reliable update). The increase in population is therefore due to new residential developments that are revealed in the 2021 image.

Fig. 13 Example 1 of detection for 2021.

Fig. 14 Example 2 of detection for 2021.

Table 6 Example 1 of population estimation for an EA.

Classes	Bldg./EA	μ	2021 Pop_estimate
Class 1	337	13	4381
Class 2	77	9	693
Class 3	16	7	112
Class 4	14	5	70
Class 5	0	2	0
Class 6	0	5	0
Class 7	16	7	112
Total estimate_pop_2021			**5368**
Total pop_2008			**4663**

Table 7 Example 2 of population estimation for an EA.

Classes	Bldg./EA	μ	2021 Pop_estimate
Class 1	397	13	5161
Class 2	42	9	378
Class 3	10	7	70
Class 4	26	5	130
Class 5	0	2	0
Class 6	1	5	5
Class 7	17	7	119
Total estimate_pop_2021			**5863**
Total pop_2008			**5275**

4 Conclusion

Our study developed a method for estimating populations at the residential unit scale from VHR satellite images and a CNN. Residential geo-objects (houses) in our study were located in Khartoum, Sudan and were difficult to classify using a taxonomy related to residence type and number of inhabitants (large intraclass variation, little interclass variation). The CNN detection approach required manual annotation of all houses according to an established taxonomy that respected the observed built environment in the study setting. Use of a pretrained network helped to overcome the small size of the available dataset. To rectify the class imbalance, a weight adjustment was used. The model under these conditions obtained an mAP of 0.79. The determination of an average population size per detected geo-object class allowed for population estimation at different scales.

The methodology that was developed here offers an alternative, based upon Earth observations and artificial intelligence, to estimate the population in developing countries where census operations, surveys, and population projections pose various economic and logistical challenges. As another example of application, this methodology can be used for a better division of a territory into census units and the production of field sampling scenarios to guide population surveys and polls. Therefore, it could be adopted by international organizations in their sociodemographic information collection activities.

The study has limitations that must be raised and resolved for the implementation of a reliable tool in practice. First, the study faced the difficulty of defining consensus taxonomy of residential geo-objects. The difficulty lies in environmental heterogeneity with respect to habitat. The houses in the study environment exhibit great variability in shape, size, and neighborhood, among others, making their identification difficult. Further, the data that were used (50 cm resolution images) do not easily permit recognition of the type of buildings for a better annotation, whereas 30-cm images (WorldView-3 and Pleiades Neo) would likely have yielded better results. Second, the model was not able to distinguish among some classes, leading to confusion during detection. Finally, to improve the detection process, a larger sample size is needed to train and test the model. Therefore, it is possible to improve our method by: (a) examining the built environment of the study area in greater detail; (b) collecting more field data; (c) adding more annotations; (d) using a newer and better CNN for performing object detection; and (e) using 30-cm images, such as WorldView-3 and Pleiades Neo.

References

1. S. Juran and A. L. Pistiner, "The 2010 round of population and housing censuses (2005-2014)1," *Stat. J. IAOS* **33**(2), 399–406 (2017).
2. C. Robinson, F. Hohman, and B. Dilkina, "A deep learning approach for population estimation from satellite imagery," in *Proc. 1st ACM SIGSPATIAL Workshop on Geosp. Hum.*, Association for Computing Machinery, New York, United States, pp. 47–54 (2017).
3. A. J. Tatem, "WorldPop, open data for spatial demography," *Sci. Data* **4**(1), 170004 (2017).
4. J. Hoogeveen and U. Pape, *Data Collection in Fragile States: Innovations from Africa and Beyond*, Springer Nature (2020).
5. T. Jhamba et al., "UNFPA Strategy for the 2020 round of population and housing censuses (2015–2024)," *Stat. J. IAOS* **36**(1), 43–50 (2020).
6. Y. Wang et al., "Approaches to census mapping: Chinese solution in 2010 rounded census," *Chin. Geogr. Sci.* **22**(3), 356–366 (2012).
7. Vereinte Nationen, ed., *Principles and Recommendations for Population and Housing Censuses: 2020 Round*, Revision 3, United Nations, New York (2017).
8. A. Boumezoued, "Micro-macro analysis of heterogenous age-structured populations dynamics. Application to self-exciting processes and demography," p. 338 (2016).
9. C. M. Almeida et al., "Multilevel object-oriented classification of QuickBird images for urban population estimates," in *Proc. 15th Annu. ACM Int. Symp. Adv. in Geogr. Inf. Syst.*, pp. 1–8 (2007).
10. C. Deng, C. Wu, and L. Wang, "Improving the housing-unit method for small-area population estimation using remote-sensing and GIS information," *Int. J. Remote Sens.* **31**(21), 5673–5688 (2010).
11. P. Dong, S. Ramesh, and A. Nepali, "Evaluation of small-area population estimation using LiDAR, Landsat TM and parcel data," *Int. J. Remote Sens.* **31**(21), 5571–5586 (2010).
12. P. Füreder et al., "Monitoring displaced people in crisis situations using multi-temporal VHR satellite data during humanitarian operations in South Sudan," in *Proc. GI_Forum*, pp. 391–401 (2015).
13. D. Tiede et al., "Automated analysis of satellite imagery to provide information products for humanitarian relief operations in refugee camps–from scientific development towards operational services," *Photogramm. Fernerkundung Geoinf.* **2013**(3), 185–195 (2013).
14. L. Tomás et al., "Urban population estimation based on residential buildings volume using IKONOS-2 images and LiDAR data," *Int. J. Remote Sens.* **37**(sup1), 1–28 (2016).
15. R. Neuville et al., "3D viewpoint management and navigation in urban planning: application to the exploratory phase," *Remote Sens.* **11**(3), 236 (2019).
16. K. Reda and M. Kedzierski, "Detection, classification and boundary regularization of buildings in satellite imagery using faster edge region convolutional neural networks," *Remote Sens.* **12**(14), 2240 (2020).
17. J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(11), 2793–2798 (2017).

18. N. Audebert, B. Le Saux, and S. Lefèvre, "Segment-before-detect: vehicle detection and classification through semantic segmentation of aerial images," *Remote Sens.* **9**(4), 368 (2017).
19. W. Li et al., "Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data," *Remote Sens.* **11**(4), 403 (2019).
20. J. Deng et al., "Voxel R-CNN: towards high performance voxel-based 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, Vol. 35, pp. 1201–1209 (2021).
21. G. Han et al., "Meta faster R-CNN: towards accurate few-shot object detection with attentive feature alignment," in *Proc. AAAI Conf. Artif. Intell.*, Vol. 36, pp. 780–789 (2022).
22. H. Zhang et al., "Dynamic R-CNN: towards high quality object detection via dynamic training," *Lect. Notes Comput. Sci.* **12360**, 260–275 (2020).
23. B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., Vol. 34, pp. 17864–17875 (2021).
24. R. Fan et al., "SNE-RoadSeg: incorporating surface normal information into semantic segmentation for accurate freespace detection," *Lect. Notes Comput. Sci.* **12375**, 340–356 (2020).
25. P. Hu et al., "Real-time semantic segmentation with fast attention," *IEEE Rob. Autom. Lett*. **6**(1), 263–270 (2021).
26. R. Strudel et al., "Segmenter: transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 7262–7272 (2021).
27. K. Yang et al., "Omnisupervised omnidirectional semantic segmentation," *IEEE Trans. Intell. Transport. Syst.* **23**(2), 1184–1199 (2022).
28. United Nations Statistics Division, "Soudan demographics," Octobre 2016, http://www.unstats.un.org/unsd/demographic/meetings/wshops/Soudan (accessed January 2020).
29. Y. LeCun et al., "Deep learning," *Nature* **521**(7553), 436–444 (2015).
30. R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1440–1448 (2015).
31. S. Ren et al., "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6), 1137–1149 (2017).
32. M. Everingham and J. Winn, "*The Pascal visual object classes challenge 2012 (voc2012) development kit*," Pattern Anal. Stat. Model. Computat. Learn., Tech. Rep. 8(5) (2011).
33. F. Zhuang et al., "A comprehensive survey on transfer learning," *Proc. IEEE* **109**(1), 43–76 (2020).
34. B. Kellenberger, D. Marcos, and D. Tuia, "Detecting mammals in UAV images: best practices to address a substantially imbalanced dataset with deep learning," *Remote Sens. Environ.* **216**, 139–153 (2018).
35. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980 (2017).
36. M. Everingham et al., "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.* **88**(2), 303–338 (2010).
37. G.-H. Fu, L.-Z. Yi, and J. Pan, "Tuning model parameters in class-imbalanced learning with precision-recall curve," *Biometr. J.* **61**(3), 652–664 (2019).
38. J. Brownlee, "How to use learning curves to diagnose machine learning model performance," 2019, https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/ (accessed 13 February 2022).

Gafarou Kpegouni received his master's degree in applied geomatics and remote sensing from the University of Sherbrooke, Quebec. He received his first master's degree in remote sensing and GIS and the second in physical geography from the University of Felix Houphouet Boigny (Cote d'Ivoire) and the University of Lomé (Togo), respectively. He is currently pursuing a post-master's degree at CARTEL, University of Sherbrooke. His research interests include automated processing and analysis of earth observation data using computer vision techniques. His fields of application of these techniques and data are among others urban planning, demography, telecommunication and transportation infrastructures, and environment.

Biographies of the other authors are not available.