Geosci. Model Dev., 15, 7505–7532, 2022
https://doi.org/10.5194/gmd-15-7505-2022

Geoscientific Model Development
Open Access
EGU

Estimation of missing building height in OpenStreetMap data: a French case study using GeoClimate 0.0.1

Jérémy Bernard[1,3], Erwan Bocher[2], Elisabeth Le Saux Wiederhold[3], François Leconte[4], and Valéry Masson[5]

[1]University of Gothenburg, Department of Earth Sciences, Sweden
[2]CNRS, Lab-STICC, UMR 6285, Vannes, France
[3]Université Bretagne Sud, Lab-STICC, UMR 6285, Vannes, France
[4]Université de Lorraine, INRAE, LERMaB, 88000, Épinal, France
[5]Météo-France and CNRS, CNRM, UMR3589, Toulouse 31057, France

Correspondence: Jérémy Bernard (jeremy.bernard@zaclys.net)

Received: 21 December 2021 – Discussion started: 20 April 2022
Revised: 18 July 2022 – Accepted: 3 September 2022 – Published: 11 October 2022

Abstract. Information describing the elements of urban landscapes is required as input data to study numerous physical processes (e.g., climate, noise, air pollution). However, the accessibility and quality of urban data is heterogeneous across the world. As an example, a major open-source geographical data project (OpenStreetMap) demonstrates incomplete data regarding key urban properties such as building height. The present study implements and evaluates a statistical approach that models the missing values of building height in OpenStreetMap. A random forest method is applied to estimate building height based on a building's closest environment. A total of 62 geographical indicators are calculated with the GeoClimate tool and used as independent variables. A training dataset of 14 French communes is selected, and the reference building height is provided by the BDTopo IGN. An optimized random forest algorithm is proposed, and outputs are compared with an evaluation dataset. At building scale for all cities, at least 50 % of the buildings have their height estimated with an error of less than 4 m (the cities' median building heights range from 4.5 to 18 m). Two communes (Paris and Meudon) demonstrate building height results that deviate from the main trend due to their specific urban fabrics. Putting aside these two communes, when building height is averaged at a regular grid scale (100 m × 100 m), the median absolute error is 1.6 m, and at least 75 % of the cells of any city have an error lower than 3.2 m. This level of magnitude is quite reasonable when compared to the accuracy of the reference data (at least 50 % of the buildings have a height uncertainty equal to 5 m). This work offers insights about the estimation of missing urban data using statistical methods and contributes to the use of open-source datasets based on open-source software. The software used to produce the data is freely available at https://doi.org/10.5281/zenodo.6372337 (Bocher et al., 2021b), and the dataset can be freely accessed at https://doi.org/10.5281/zenodo.6855063 (Bernard et al., 2021).

1 Introduction

The topography – defined as the spatial distribution of natural and artificial land use features – has a significant influence on the microclimate. This is clearly visible in urban areas, where the great heterogeneity of forms, materials and land uses induces high variability in temperature, wind speed and humidity (Oke, 2002). Thus an in-depth knowledge about the topography of a location would lead to a better understanding and more accurate modeling of its climate as well as of other physical processes, such as noise propagation and air pollution (Tang and Wang, 2007; Bocher et al., 2019).

There are currently no standard geographical data to study the urban climate worldwide. However, urban data tend to increase under both closed license and open-source. A key open-source data approach is the OpenStreetMap (OSM)[1] project. Data from the latter have several features: they are

[1]https://www.openstreetmap.org (last access: 19 September 2022)

expected to be available worldwide, and the most important objects needed for urban climate studies (building footprints, isolated tree locations, water, vegetation and impervious patches) are available, well located and described using a great diversity of tags (Mocnik et al., 2017). Moreover, OSM has a free tagging system that allows users to improve the current tags (key, value) with their own information (an OSM user can describe a building object with the tags such as the following: "building"="house", "height"="10", "building:levels":"2").

However, information concerning building height is not available worldwide, neither in OSM nor in any other database (Masson et al., 2020). Lao et al. (2018) reported that less than 3 % of buildings globally have a height value, and less than 4 % have a number-of-levels value. For the city of Paris (where this information is known to be quite well informed), the values are only 0.1 % and 51.2 %, respectively. This is a major shortcoming, since the urban climate is often characterized by spatial indicators based on the third dimension:

- The sky view factor (SVF), which is calculated according to terrain level variations, building and tree locations and heights, is related to effective albedo (Bernabé et al., 2015) and is strongly correlated to temperature (Lindberg, 2007) and wind speed (Johansson et al., 2016).
- The building height variability within an area affects the vertical and horizontal wind speed (Hanna and Britter, 2010).
- The roughness length of an area, often calculated using facade density, is used to estimate the wind speed vertical profile of the urban canopy (Hanna and Britter, 2010).

The objective of this study is to develop a method to estimate the height of a building from its topographical context using only data available in OSM. Modeling building footprints and their height value has been largely covered by remote sensing. It can cover large areas at once quite efficiently, and the resulting datasets can be updated quite easily with a repetitive coverage. Different techniques for building height extraction have been developed based on photogrammetric processing (Fradkin et al., 1999; Zeng et al., 2014), analysis of point clouds from airborne light detection and ranging (Sohn et al., 2008; Shan and Toth, 2018), shadow detection (Song et al., 2013; Shao et al., 2011) and, more recently, a deep learning approach (Cao and Huang, 2021).

In the meantime, the recent and global movement regarding open data – specifically for vector topographic databases such as OSM – offers new opportunities to estimate building height. The geography of a territory and the pattern of the topographic elements are criteria that can be used as a proxy to identify the urban forms and therefore the distribution of building heights. Biljecki et al. (2017) have used building footprints and their corresponding attributes to derive building heights for the city of Rotterdam, the Netherlands. They have tested several random forest models using building properties characterizing its geometry footprint (size, shape and number of neighbors), other attributes (use, age and number of levels), and information concerning the inhabitants (the number and their levels of income) as independent variables. Milojevic-Dupont et al. (2020) have proposed a random forest approach to estimate building height using 152 features. These independent variables are related to buildings (e.g., footprint geometry), streets (e.g., closest street, closest intersection), street-based blocks (e.g., number of blocks in a given radius) and cities (e.g., total city area). For the base case, the training dataset includes the building heights of the Netherlands, the region of Friuli Venezia Giulia in Italy and five French urban areas, and the validation dataset includes the building heights of the state of Brandenburg in Germany.

A major limitation of these studies is the obstacle of reproducibility for experts and practitioners. Indeed, the algorithms are not fully available. Moreover, input datasets require many preprocessing steps, since the format and the access differ between city models (e.g., French 3D city models in Milojevic-Dupont et al., 2020). The method presented in the next sections uses a random forest approach and can be easily reproduced using the GeoClimate software without any preprocessing steps, since it is based exclusively on OSM data. The main spatial indicators used by the urban climate communities are also calculated using reference and estimated building heights. Compared to each other, they provide urban climate researchers with a good level of magnitude in terms of the impact of the estimated height on these indicators. OSM data do not contain as much detailed information about buildings as that seen in Biljecki et al. (2017) (number of levels, age, number of inhabitants), but other information describing the environment of the buildings will be used (roads, vegetation, rail, etc.).

In order to make OSM data available to urban climate researchers, the GeoClimate tool has been developed (E. Bocher et al., 2021; Bocher et al., 2021a). It is an easy way (i) to download most of the information needed for urban climate studies, (ii) to estimate building height from the topographical context, and (iii) to calculate spatial indicators (such as SVF, building height variability or roughness length) that are useful as input for parametric climate models. This paper focuses on the second item, namely how to estimate building height in OSM when the information is missing. First, the data and the methodology used to estimate the building height are presented (Sect. 2), and second, the accuracy of these estimations is analyzed (Sect. 3).

2 Data and method

This study presents a method to estimate values of building height when the information is missing in OSM. The height of a building is determined according to a regression-based statistical model (e.g., random forest model) using a set of spatial indicators – including the building's shape, the building's relation to its neighbors, and the organization and morphology of the building's environment – as independent variables. The true building height values come from the BD-Topo V2.2 (BDT) provided by the French National Geographic Institute (IGN). These values are defined as reference height. Two datasets have been considered for this study, namely a training dataset to build the random forest algorithm and a validation dataset to compare the outputs of the optimized random forest algorithm with reference heights. The overall methodology is illustrated in Fig. 1 and consists of the following steps:

1. *Building characterization*. Each building and its environment (limited to the topographical spatial units (TSU) it belongs to – defined in Sect. 2.2.1) are characterized by spatial indicators (building area, number of buildings neighbors, vegetation fraction, etc.). These indicators are the independent variables of the statistical model.

2. *Building height attribution*. The reference building heights (BDT building heights) are attributed to each OSM building according to their footprints. The resulting height in the OSM dataset is the dependent variable of the statistical model.

3. *Statistical analysis*. The random forest model is built based on the training dataset. In this step, parameters that maximize the performance of the random forest model are identified.

4. *Performance evaluation*. Outputs of the optimized random forest model are evaluated against the reference heights of the validation dataset.

Each step is described further in Sect. 2.2.

2.1 Study area

Building organization and height may differ a lot throughout the world, limiting the ability to model the height of a building based on the characteristics of its environment. Thus, although the method can be used to estimate the height of any building in the world, the application area of this preliminary work is limited to the French territory. The training and evaluation areas are selected to cover

- all types of communes (from small villages to large conurbations)
- a large part of France (Fig. 2), increasing the probability of having cultural and/or historical differences inducing urbanistic heterogeneities
- the main geographical constraints for construction (nearby mountains – Annecy, La Thuile, Corbonod – and nearby sea – La Rochelle).

To fulfill the first need, the communes have been chosen to cover each of the four French commune types defined by the French National Institute of Statistics and Economic Studies (INSEE, 2020): main urban area, secondary urban area, peripheral urban area, rural area. According to the French 2020 census data, the types are defined based on the following commune characteristics: number of inhabitants, density of population, number of employees, and population flow between households and workplaces. They are used to define the urban attraction cluster. The definitions of each type are given in Table 1.

Table 1. The four commune types defined by INSEE (2020).

Commune type	Definition
Main urban area	Commune centre of the urban attraction cluster
Secondary urban area	Other commune of the urban attraction cluster
Peripheral urban area	Commune in the attraction area of the urban cluster
Rural area	Commune outside of any urban attraction area

The training and the evaluation datasets contain 14 and 8 communes, respectively. The location of each commune is shown Fig. 2, while further information concerning each territory is in given Tables 2 and 3 (for training and evaluation datasets, respectively).

2.2 Methodology

2.2.1 Building characterization

Data from OSM are used to characterize the building and its environment: building footprint, vegetation footprint and type, water footprint, impervious footprint, rail and road footprint. The free and open-source GeoClimate software (E. Bocher et al., 2021; Bocher et al., 2021a) is used to compute the spatial indicators at three different scales:

- building scale
- block scale, defined as the aggregation of all buildings touching each other
- topographical spatial unit (TSU) scale, defined according to the central lines of roads and rails, commune boundaries, and water and vegetation boundaries when their area is higher than 2500 and 10 000 m^2, respectively (Fig. 3).

Each building is described by a total of 62 spatial indicators (note that all characteristics are calculated in two dimen-

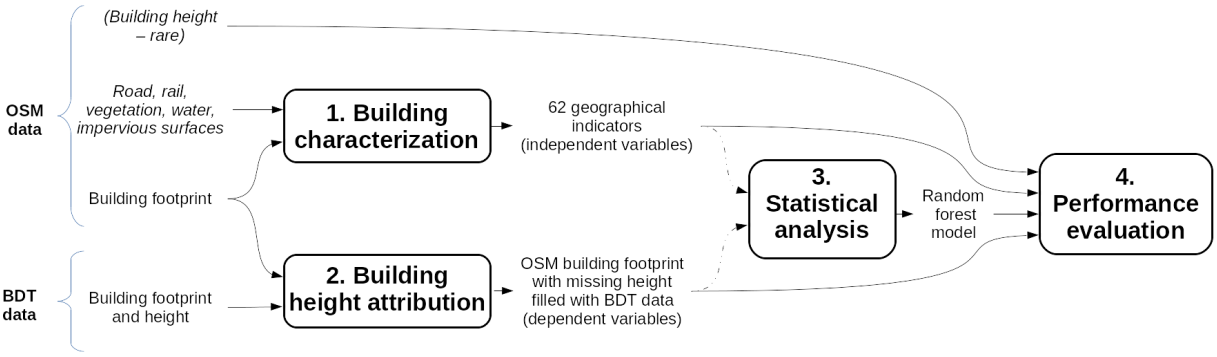

Figure 1. Overall methodology – the use of dashed arrows means that only the training dataset is used.

Table 2. Information and statistics about the training dataset.

Commune type	Commune name	Inhabitants (2017)	Number of buildings*	INSEE code
Main urban area	Paris (6th, 11th and 18th districts)	2 187 526 (40 525, 145 903, 193 665)	15 964	75 056 (75 106, 75 111, 75 118)
	Toulouse	479 553	103 368	31 555
	Nantes	309 346	57 550	44 109
	Annecy	126 924	21 153	74 010
	Avignon	91 921	29 113	84 007
	La Rochelle	75 735	31 194	17 300
Main peripheral urban area	Nanterre	95 105	10 851	92 050
	Meudon	45 352	5430	92 048
	Blagnac	24 517	9286	31 069
Secondary peripheral urban area	La Haie-Fouassière	4659	2323	44 070
	Gratentour	4158	2938	31 230
	Staffelfelden	3959	2254	68 321
	Bourgneuf	1275	782	17 059
	Lathuile	1016	732	74 147

* Only buildings that have a height value higher than 3 m in the BDT (cf. Sect. 2.2.2) have been conserved for this evaluation.

sions only, since most OSM data do not have height information). Four types of indicators are used (see Table 4) and the list of all indicators is given in Appendix A. Indicator values calculated at block and TSU scales are then attributed to each building (a building within a given block or TSU embeds the indicator values of the block and the TSU it belongs to).

2.2.2 Reference building heights attribution

In order to train and evaluate the statistical model, a reference height used as a true value should be assigned to each OSM building. Most of the OSM buildings do not have any information concerning their height. The few buildings with height information could have been considered as training and evaluation datasets; however, these buildings are sparse and not representative of common buildings (since most of them are filled out by OSM users because they are well-known buildings with specificities). Therefore, the training and evaluation datasets are created only with OSM buildings that have no height. The reference height is set using the BDT data. However, a single building in OSM may match with several buildings in the BDT (Fig. 4).

Thus, the height of an OSM building used as a reference ($H_{\text{osm,true}}$) is calculated from the height of all intersecting BDT buildings according to Eq. (1). This equation is applied for both the training and the validation datasets.

$$H_{\text{osm,true}} = \frac{\sum_{i=1}^{n} A_i \cdot H_{\text{BDT}_i}}{\sum_{i=1}^{n} A_i}, \quad (1)$$

with A_i being the area of the intersection between an OSM building and a BDT building i, and H_{BDT_i} being the height of the BDT building i intersecting the OSM building.

In the example presented in Fig. 4, if BDT building 1 is much taller than the others (BDT buildings 2 and 3), this information is lost (smoothed by the averaging) and could then lead to a bias in the learning process. To keep track of this potential bias, a simple index is proposed to characterize the proportion of the intersection between a BDT building and

Table 3. Information and statistics about the validation dataset.

Commune type	Commune name	Inhabitant (2017)	Number of buildings*	INSEE code
Main urban area	Rennes	216 815	30 527	35 238
	Dijon	156 920	23 044	21 231
Main peripheral urban area	Charnay-lès-Macon	7376	3574	71 105
	Saint-Nicolas de Redon	3179	2515	44 185
Secondary peripheral urban area	Allaire	3854	2744	56 001
	Pont-de-Veyle	1625	729	1306
Rural area	Corbonod	1264	1115	1118
	Saint-Ganton	424	469	35 268

* Only buildings that have a height value higher than 3 m in the BDT (cf Sect. 2.2.2) have been conserved for this evaluation.

Table 4. Types of spatial indicators used to define the main characteristics of each unit scale.

Indicators type	Scale of application			Examples of indicators
	Building	Block	TSU	
Type and use	x			Building type, building use
Form and size	x	x	x	Area, form factor, fraction of courtyard, etc.
Spatial relations	x			Minimum distance to another building, fraction of wall shared with other buildings, minimum distance to road, etc.
Planar density			x	Building fraction, vegetation fraction, etc.
Aggregated statistics from lower scale		x	x	Mean building area, standard deviation building form factor, etc.

an OSM building. This index, called uniqueness value (UV), is defined in Eq. (2):

$$\mathrm{UV} = \frac{\max_{i \in 1..n} A_i}{\sum_{i=1}^{n} A_i}. \quad (2)$$

The uniqueness value considers only the BDT building that demonstrates the largest intersection area with a given OSM building. The higher the UV, the more unique the BDT building intersecting the OSM building. UV is not impacted by the fraction of the OSM building shared with other BDT buildings. If only one BDT building overlaps only a small fraction of an OSM building, the uniqueness value will be 1.

2.2.3 Design and optimization of the random forest statistical model

For the statistical analysis, the OSM building height (reference height $H_{\mathrm{osm,true}}$) is defined as the dependent variable, while spatial indicators are defined as independent variables. Only the training dataset (see Table 2) is used for this step. To obtain an optimal model, the methodology illustrated in Fig. 5 is applied.

The random forest (RF) approach is chosen for several reasons: (i) it is simple to implement, (ii) it deals with quantitative and qualitative variables, and (iii) it is appropriate when using a large number of variables (Hastie et al., 2001). In order to limit overfitting and a high correlation between trees,

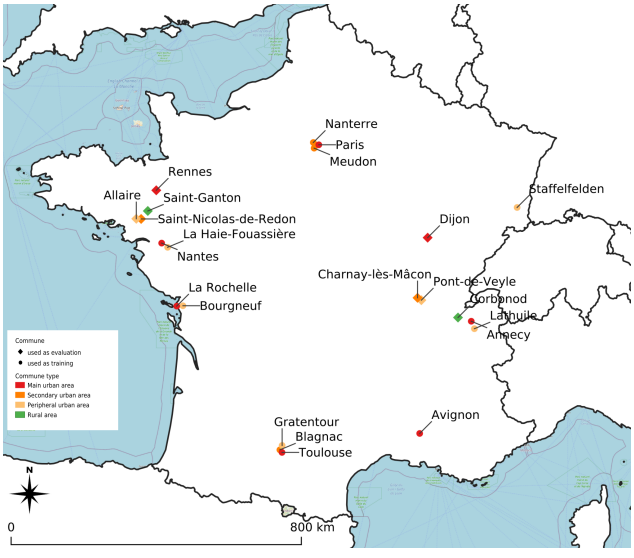

Figure 2. Location of the 22 communes used as training or evaluation data. ©OpenStreetMap contributors 2021. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

all combinations of the following RF regressor parameters are investigated:

– number of trees: 100, 350, 500, 650 (note that preliminary analysis showed lower accuracy when the number

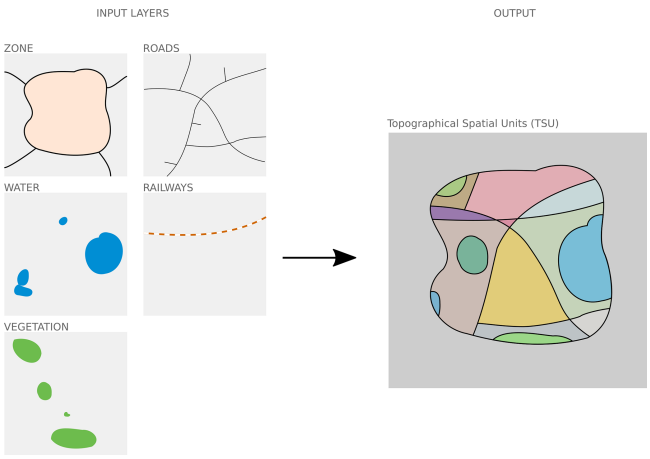

Figure 3. Example of topographical spatial unit calculation.

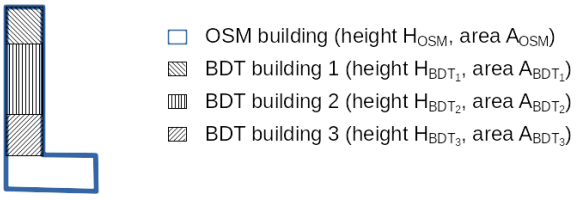

Figure 4. Example of the overlap between OSM and BDT buildings.

of trees was lower than 100 and no significant improvement when greater than 650);

- minimum node size (minimum fraction of the sample used to create a new node): 0.0001 %, 0.001 %, 0.005 %, 0.01 % (the whole sample size includes 345 418 individuals – note that preliminary analysis showed decreasing performance over 0.01 %);
- maximum variables per tree (maximum fraction of variables used in a tree): 20 %, 35 %, 50 % (of a total of 62 variables – note that preliminary analysis showed decreasing performance when the fraction was lower than 20 % and no significant improvement over 40 %);
- maximum leaf nodes (maximum number of leaves in a tree): 300, 500, 800, 1100 (note that preliminary analysis showed no significant improvement over 1100 while increasing the complexity and thus potentially the overfitting).

For a default combination of 500 trees, 0.001 % minimum node size and 30 % maximum variables per tree, the effect of UV on the accuracy is studied, keeping only buildings having a UV above 30 %, 70 %, 90 % and 95 %.

A total of 70 % of the training data are randomly drawn to construct the RF. The accuracy is calculated using the remaining 30 % of the data. This process is performed 10 times for each combination C_i and uniqueness value UVi. The scikit-learn Python algorithm is used for this investigation.

The optimized combination C_{opt} and uniqueness value UV_{opt} leading to the lowest mean absolute error (MAE) are used to construct the final RF model used in GeoClimate. For this purpose, the entire training dataset is used as input for the Smile library algorithm (since GeoClimate is Java-based).

2.2.4 Performance evaluation

The optimized RF model obtained in the previous step is run over the eight communes of the validation dataset to calculate the missing height values of the OSM buildings. For each building, the heights estimated with the optimized RF model ($\hat{H}_{OSM,model}$) are then compared to the reference height ($H_{OSM,true}$). The model error Err_{model} is defined for heights estimated by the random forest model:

$$Err_{model} = \hat{H}_{OSM,model} - H_{OSM,true}. \quad (3)$$

The building height values filled out by the OSM users ($\hat{H}_{OSM,user}$) are also compared to the reference height. If the user filled only the number of storeys, the building height is simply calculated by multiplying the storey number by 3 m. Even though the storey height may vary quite a lot between construction age and building type (see Biljecki et al., 2017 – Fig. 5), 3 m seems to be a reasonable value according to the one observed in the literature (ranging from 2.8 to 3.5 m – see Biljecki et al., 2017; Sect. 2.2.1). The user error Err_{user} is defined for heights filled by the users (Eq. 4).

$$Err_{user} = \hat{H}_{OSM,user} - H_{OSM,true} \quad (4)$$

In parametric urban climate models, parameters such as building height are aggregated within each square cell of a regular grid. Therefore, four indicators are calculated for a grid of 100 m width square: the mean building height and standard deviation, the roughness length (as defined by Hanna and Britter, 2010), and the SVF (as defined in Bernard et al., 2018).

3 Results and discussions

The dataset produced by the methodology described in Sect. 2.2 can be freely downloaded at https://doi.org/10.5281/zenodo.6855063 (Bernard et al., 2021). In this section, cells that have no building with an estimated height are not considered for the statistical calculations.

3.1 Optimized configuration of random forest characteristics

Very little accuracy difference is observed between all combinations described in Sect. 2.2.3. For all studied configurations, the median RMSE ranges between 2.05 and 2.2 m. The

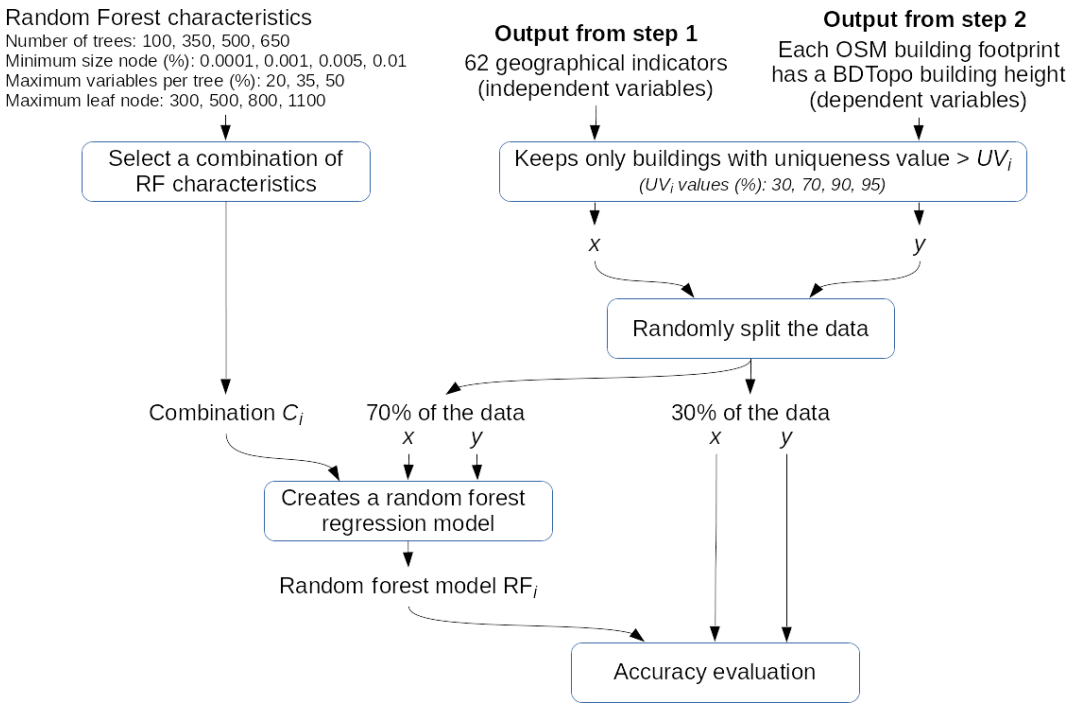

Figure 5. Method to train and optimize the random forest model. Only the training dataset is used at this step.

minimum node size and the number of trees (when greater than 100) have the least significant impact on the accuracy. The highest accuracy is reached when the maximum variables per tree is 50 % and the maximum leaf nodes is 1100. Thus, the RF scenario chosen for GeoClimate has 350 trees, 40 % maximum variables per tree (based on previous results showing little difference between 40 % and 50 %) to minimize the correlation between trees, 0.01 % minimum node size, and 1100 maximum leaf node to minimize the potential of overfitting (Hastie et al., 2001). Since the maximum tree depth obtained in Python for this configuration is 33, this value is also applied to the GeoClimate RF algorithm. The uniqueness value has an unexpected effect on the accuracy: the MAE decreases when UV value increases up to 70 % and increases for UV values above 90 %, while it could be expected to continue decreasing. However, the difference is slight (0.05 m, 2.5 %) and may be explained by the size of the sample, which is larger (+23 %) for the 70 % scenario than for the 95 % (having 345 418 and 281 081 individuals, respectively). Therefore, the data used to train the GeoClimate model are created with UV = 70 %.

3.2 General building height accuracy

For all cities, more than 50 % of the buildings have an estimated height within a ±3.97 m (3.22 m if the 18th district of Paris is excluded) interval around the true building height (Fig. 6a). At cell scale (Fig. 6b), the same statistic is ±4.61 m (2.74 m if the 18th district of Paris is excluded). If cities demonstrating a specific behavior are not considered (Paris and Meudon), the median absolute error at cell scale is always lower than 1.6 m, and 75 % of the buildings or cells of any city have an error lower than 3.2 m. This error is equivalent to the floor height of one building and could appear quite high. However, it seems quite reasonable when compared with the accuracy of the reference data (i.e., the height uncertainty of more than half of the BDT dataset is ±5 m). Note that we have tested a previous version of the model (independent variables have been updated since then) on several cities far from the ones presented in this manuscript, and the error was almost similar.

Surprisingly, the worst results are obtained with communes belonging to the training dataset (the median relative absolute error – MRAE – is 24 % for *main peripheral* communes – Table 5). Overall, there is almost no accuracy decrease when the model is applied to the validation dataset (Fig. 6). No city type shows a specific pattern; even the rural areas, which are not included in the training dataset, do not show a clearly higher error (MRAE = 23 %) than the overall trend (MRAE = 22 % on average). However, a specific behavior is observed for the city of Meudon and for the 18th district of Paris, which have a higher error than the main trend (Fig. 6b). A part of the explanation can be found in their uncommon urban fabrics, which makes them more difficult to estimate by the RF model:

- Meudon has quite a low median height (Fig. 6b) while also having a high building height variability within a 100 m square (Fig. 6c);

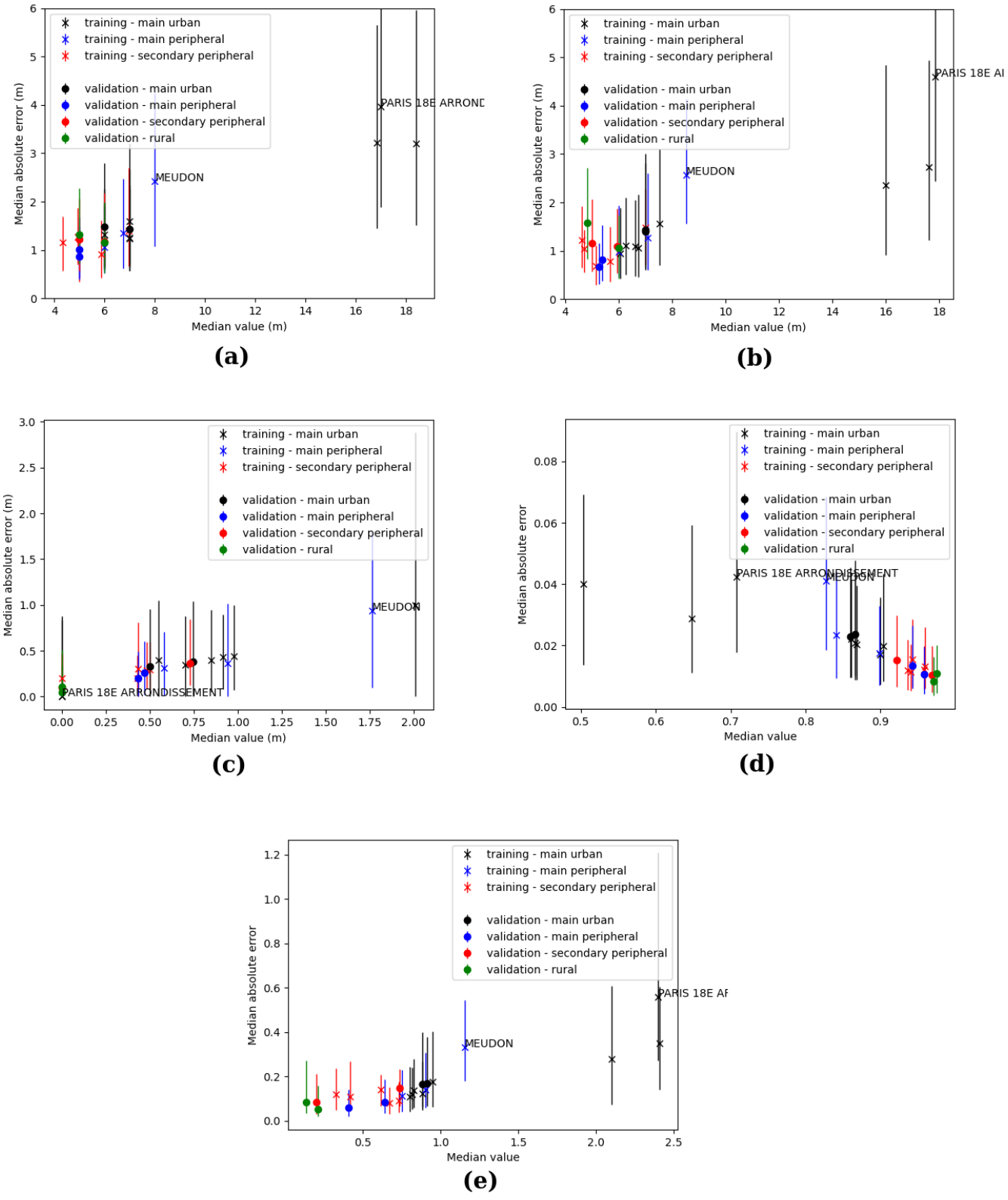

Figure 6. Median absolute error versus median true value for **(a)** building height, **(b)** RSU average building height, **(c)** RSU standard deviation building height, **(d)** RSU mean ground sky view factor, and **(e)** RSU effective terrain roughness length. The cross and the dot are the medians, while the whiskers are the first and third quartiles.

Table 5. Summary of the main statistics for each combination of {indicator, commune type, dataset type}. Bold values are the ones described in the text.

Indicator	Urban class	Training			Validation		
		Median absolute error	Median	Median relative absolute error*	Median absolute error	Median median	Median relative absolute error*
Building height (m)	Main urban	2.13	10.65	0.21	1.45	6.50	**0.23**
	Main peripheral	1.61	6.91	**0.24**	0.93	5.00	**0.17**
	Secondary peripheral	1.14	5.41	0.23	1.19	5.50	**0.23**
	Rural	–	–	–	1.24	5.50	**0.23**
RSU average building height (m)	Main urban	1.93	10.58	0.19	1.41	7.00	**0.20**
	Main peripheral	1.61	7.20	0.22	0.75	5.31	**0.13**
	Secondary peripheral	1.04	5.43	0.20	1.12	5.47	**0.21**
	Rural	–	–	–	1.32	5.42	**0.26**
RSU standard deviation building height (m)	Main urban	0.38	0.75	0.70	0.35	0.62	0.67
	Main peripheral	0.54	1.10	**0.62**	0.23	0.45	0.73
	Secondary peripheral	0.26	0.37	0.71	0.23	0.36	0.76
	Rural	–	–	–	0.08	0.00	0.76
RSU mean ground SVF	Main urban	0.03	0.78	0.17	**0.02**	0.86	0.20
	Main peripheral	0.03	0.86	0.21	0.01	0.95	0.26
	Secondary peripheral	0.01	0.95	0.26	0.01	0.94	0.28
	Rural	–	–	–	0.01	0.97	0.37
RSU effective terrain roughness length	Main urban	0.23	1.40	**0.20**	0.17	0.90	**0.23**
	Main peripheral	0.19	0.94	**0.24**	0.07	0.52	**0.20**
	Secondary peripheral	0.11	0.55	**0.27**	0.11	0.47	**0.32**
	Rural	–	–	–	0.07	0.17	**0.48**

* Contrarily to the other indicators, the reference value used for MRAE is not the SVF of the grid cell but 1-SVF of the grid cell.

– the 18th district of Paris has the highest average building roof height at grid scale (Fig. 6b) but also the highest sky view factor of the three Parisian districts (Fig. 6d) even though it would be expected to have the lowest.

3.3 Accuracy of standard spatial indicators

The building height estimation is slightly improved when averaged at RSU scale (MRAE decreases from about 3 % for all communes types except for rural areas). The variability of height within a cell is very roughly calculated using the estimated height: for more than 50 % of the cells, the MRAE on the standard deviation of the building height is higher than 62 % (Table 5). This behavior is quite understandable, since the RF model smooths the values of the estimated height; it cannot reproduce entirely the complexity of the initial dataset. The roughness length MRAE is about 25 %. This error is slightly higher for secondary peripheral and rural communes (29.5 % and 48 %, respectively – Table 5), while Meudon and the 18th district of Paris show higher values than the MAE trend (Fig. 6e). The sky view factor is quite accurately calculated: more than 50 % of the cells have an absolute error lower or equal to 0.02 (Table 5). The effect of the building height error is probably limited, because the sky view factor is a 3-dimensional indicator: it also accounts for the horizontal footprint of the building, which is the same between the estimated and observed data.

3.4 Limitations of the model for high-rise buildings

The accuracy differs a lot between low-rise and high-rise buildings. For all types of cities, the building height is often overestimated for buildings smaller than 5 m and often underestimated for the taller ones (Fig. 7). The bias for high-rise buildings can be quite high, but it does not affect the general accuracy of the model, since most of the buildings are low-rise (see Fig. 7: 80 % of the buildings are lower than 10 m – even in main cities). A better estimation of the high-rise buildings may be achieved using a training dataset containing an equal number of buildings for all levels. This would allow a better representation of the spatial heterogeneity of the third dimension. However, this would most probably affect the accuracy of the estimation for low-rise buildings. Indeed, in France, low-rise building are much more numerous than the high-rise ones.

As previously observed, there is almost no accuracy decrease between the training and the validation estimations, even for high buildings. Only a slight difference can be ob-

served for main urban cities: above 15 m, the training dataset performs better than the validation one (almost 3 m difference – Table 6).

This difference is attributed to the Paris buildings dataset: the two curves almost coincide if the latter is excluded. The reason is that Paris buildings are quite accurately calculated and represent a large part of the training dataset (43.2 % of the buildings higher than 15 m). The urban fabric (very dense block of buildings with courtyards) and the building heights are quite homogeneous in Paris, thus being well taken into account by the model. In most other cities, a large amount of the high-rise buildings are isolated buildings (see Fig. 8 for an example with the city of Nantes). These buildings probably have very little shape or environmental differences with smaller, isolated buildings and are probably less numerous. Therefore, most of these buildings are seen as low-rise by the model.

It is interesting to notice that the buildings that already have a building height value in OSM (or at least a number-of-floors value) are, most of the time, slightly higher than the BDT ones (Fig. 7). This is the case for low-rise buildings (lower than 5 m – Table 6) in particular, and it may be explained by the fact that the BDT heights are taken at the lowest part of the roof. The OSM data can take into account the roof height, which is, most of the time, equal to zero for tall buildings but non-negligible for small buildings. The difference between height derived from OSM user filling and the reference data (BDT) is quite low for any building height. This result may be used to improve the model performance – when estimating the height of a given building, the random forest may take into account the height of a nearby building filled by an OSM user as an extra independent variable.

3.5 Spatial distribution of the building height at city scale

While most of the buildings higher than 15 m are underestimated, the model allows one to represent the spatial patterns of the third dimension well; at grid scale, the average building height maps of estimated and reference values look quite similar (Fig. 9; Appendix B for other cities). While the model smooths the values slightly, the city center, first ring and second ring are quite easily distinguishable. Note that this is not the case for cities that have a more homogeneous spatial distribution of the building height values (e.g., Annecy, which is constrained by the topography – see Fig. B1 in Appendix B).

For most of the city pixels, the absolute error is under 2.5 m, and only a small proportion of cells have an error higher than 5 m (Fig. 9). This absolute error magnitude is within the accuracy of the reference building dataset. Indeed, according to the data supplier's (IGN) information (based on a sample of 7 299 422 buildings), 8.2 % of the buildings have an accuracy of 1 m, 13.5 % an accuracy of 2.5 m, and 68.8 % an accuracy of 5 m, while 9.5 % have no accuracy information.

4 Conclusions

There is a need for a world-wide database of morphological indicators that would be useful for many physical process interests (e.g., parametric urban climate models, noise modeling, urban planning). The GeoClimate tool aims to tackle this issue using the OpenStreetMap data. However, most of the OSM buildings do not have any information concerning their height, which is a crucial parameter for urban climate studies. A random forest model has been integrated within GeoClimate to estimate the height of a building based on spatial indicators describing its shape, its relations to other buildings and the 2D characteristics of its close environment.

This article presents the method for building and evaluating this model. The buildings from 14 French communes have been used to train the model, while the evaluation was based on 8 French communes. Attention was paid to having as many types of territories (based on the French definitions) in the samples as possible, including main urban, main peripheral, secondary peripheral and rural.

The random forest model was tuned according to four parameters: the number of trees (best 350), the minimum node size (best 0.01 %), the maximum variables per tree (best 40 %), and the maximum leaves per tree (best 1100). The reference heights used for the training of our OSM buildings were based on a dataset (French BDTopo – BDT) where buildings could not fit exactly with the OSM ones. Thus, the matching between each OSM building footprint and BDT building footprint has been quantified using the uniqueness value indicator. The latter equals 1 if only one building from the BDT was used to feed the OSM building height; it is otherwise lower than 1 and is best suited to the random forest model when values are higher than 0.7.

Two communes (Paris and Meudon) demonstrate a specific behavior within the analysis. Apart from these, the median absolute error at cell scale was always lower than 1.6 m, and 75 % of the buildings or cells of any city had an error lower than 3.2 m. This level of magnitude is similar to the BDT data used for the training: 68.8 % of the buildings heights demonstrated an uncertainty of 5 m.

Geographical indicators commonly used in urban climate studies have also been calculated at a 100 m grid cell according to the estimated building height. While the building height variability (standard deviation within a grid) is strongly affected by the building height estimation error (50 % of the cells have more than 50 % error in building height standard deviation value), the roughness length and sky view factor have a relative error of about 20 % for 50 % of the cells.

One of the major limitations of the model at the French scale is presented when applied to tall (> 15 m), isolated buildings. However, it does not affect the recognition of the general patterns of a city: most of the high-rise buildings located in the centers of the cities are quite well modeled, though slightly underestimated.

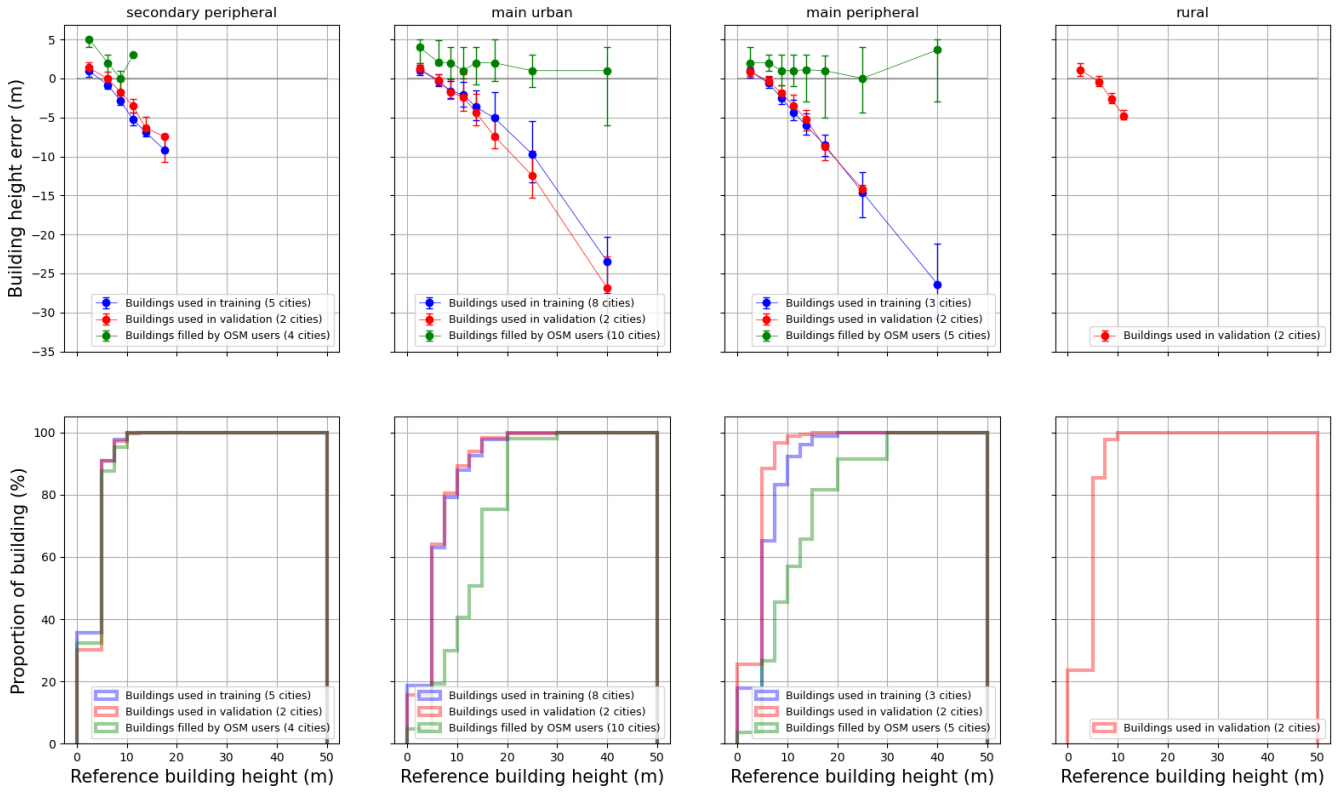

Figure 7. On the top: building height errors (Err_{model} and Err_{user}) versus reference building height ($H_{OSM,true}$) for each type of city. The dots represent the median, while the whiskers are the 1st and 3rd quartiles. On the bottom: cumulated distribution of reference building height ($H_{OSM,true}$) for each type of city. The intervals used for the reference building height (the abscissa) are based on the following values: 0, 5, 7.5, 10, 12.5, 15, 20, 30, 50 m (values above 50 m are not considered, since their number is negligible and they affect the reading).

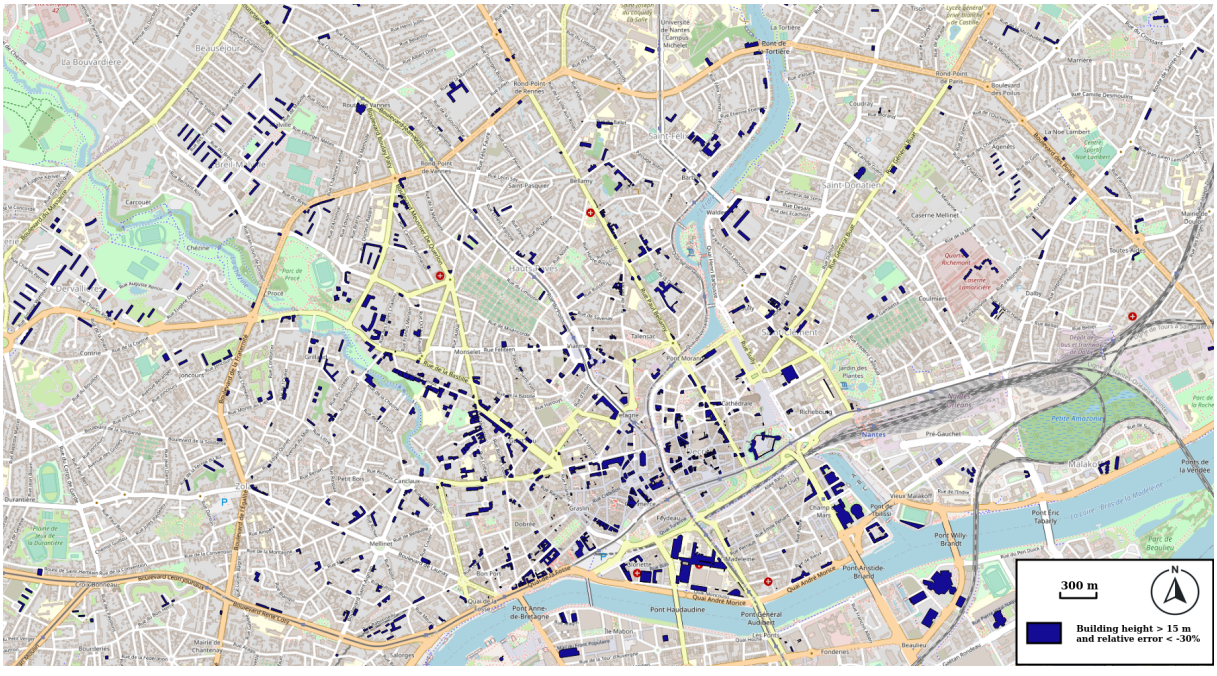

Figure 8. Buildings taller than 15 m for which the height underestimation is higher than 30 %. © OpenStreetMap contributors 2021. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

Table 6. Summary of the building height estimation error by commune type, building height range and dataset type.

Commune type	Reference building height range (m)	Training dataset			Validation dataset			OSM dataset		
		Proportions	Median error (m)	Median relative error	Proportions	Median error (m)	Median relative error	Proportions	Median error (m)	Median relative error
Main urban	0–5	0.19	1.10	0.25	0.16	1.28	0.28	0.05	**4.00**	0.80
	5–10	0.44	−0.30	−0.05	0.48	−0.23	−0.04	0.15	2.10	0.33
	10–12.5	0.16	−1.62	−0.19	0.17	−1.80	−0.22	0.11	2.00	0.20
	12.5–15	0.09	−2.08	−0.18	0.09	−2.37	−0.21	0.11	1.01	0.09
	15–20	0.05	**−3.66**	−0.27	0.05	**−4.35**	−0.32	0.10	2.00	0.15
	20–30	0.05	**−5.08**	−0.30	0.05	**−7.45**	−0.43	0.25	2.00	0.12
	30–40	0.02	**−9.69**	−0.42	0.01	**−12.43**	−0.53	0.23	1.00	0.04
	40–50	< 0.01	**−23.48**	−0.69	< 0.01	**−26.86**	−0.73	0.02	1.00	0.03
Main peripheral	0–5	0.18	0.93	0.21	0.26	0.90	0.19	0.04	**2.00**	0.50
	5–10	0.47	−0.58	−0.09	0.63	−0.32	−0.05	0.23	2.00	0.29
	10–12.5	0.18	−2.48	−0.29	0.08	−1.89	−0.22	0.19	1.00	0.12
	12.5–15	0.09	−4.33	−0.38	0.02	−3.46	−0.31	0.11	1.00	0.09
	15–20	0.04	−6.05	−0.44	0.01	−5.25	−0.40	0.09	1.11	0.08
	20–30	0.03	−8.55	−0.50	< 0.01	−8.79	−0.54	0.16	1.00	0.06
	30–40	0.01	−14.58	−0.63	< 0.01	−14.16	−0.66	0.10	0.01	0.00
	40–50	< 0.01	−26.37	−0.74	–	–	–	0.08	3.66	0.09
Secondary peripheral	0–5	0.36	0.94	0.22	0.30	1.36	0.31	0.32	**5.00**	1.25
	5–10	0.55	−0.84	−0.14	0.61	−0.06	−0.01	0.55	2.00	0.29
	10–12.5	0.07	−2.89	−0.35	0.06	−1.77	−0.21	0.08	0.00	0.00
	12.5–15	0.02	−5.25	−0.47	0.02	−3.48	−0.31	0.05	3.00	0.25
	15–20	< 0.01	−6.89	−0.53	< 0.01	−6.37	−0.42	–	–	–
	20–30	< 0.01	−9.12	−0.47	< 0.01	−7.43	−0.47	–	–	–
	30–40	–	–	–	–	–	–	–	–	–
	40–50	–	–	–	–	–	–	–	–	–
Rural	0–5	–	–	–	0.24	1.13	0.24	–	–	–
	5–10	–	–	–	0.62	−0.41	−0.07	–	–	–
	10–12.5	–	–	–	0.12	−2.61	−0.32	–	–	–
	12.5–15	–	–	–	0.02	−4.85	−0.47	–	–	–
	15–20	–	–	–	–	–	–	–	–	–
	20–30	–	–	–	–	–	–	–	–	–
	30–40	–	–	–	–	–	–	–	–	–
	40–50	–	–	–	–	–	–	–	–	–

Care should be taken for territories that have limited OSM data available (which is not the case in this study, since all cities used in this work have a higher building fraction in OSM than in BDT). In this case, the first step before applying our work would be to contribute to OSM and fill the gap in the study area. In our opinion, aside from this issue, the dataset resulting from the optimized random forest could be useful for climate analysis (even though the model is far from being perfect). We recommend a prior evaluation of what the effect of using the output of the RF model compared with the reference data usually employed by urban climate researchers could be. While we do not expect major differences when applied with parametric urban climate models at city scale, the spatial error might be quite high at neighborhood scale. Thus, for researchers and practitioners willing to use GeoClimate at a finer scale (for example, to automatically download land-type and land-use information for explicit modeling purpose), we recommend that they contribute to the OSM project first. Specifying the height of the most important buildings of their studying area in OSM can be done before running GeoClimate. At the end of the day, they can contribute to the improvement of the OSM data and also freely benefit from the GeoClimate tool. Concerning the building height modeling, the work may be continued by

- identifying the model's sensitivity to a lack of OSM information (for example, removing some or all of the roads, vegetation and water data);
- evaluating the accuracy of the estimations using other reference datasets – in France, it could be performed us-

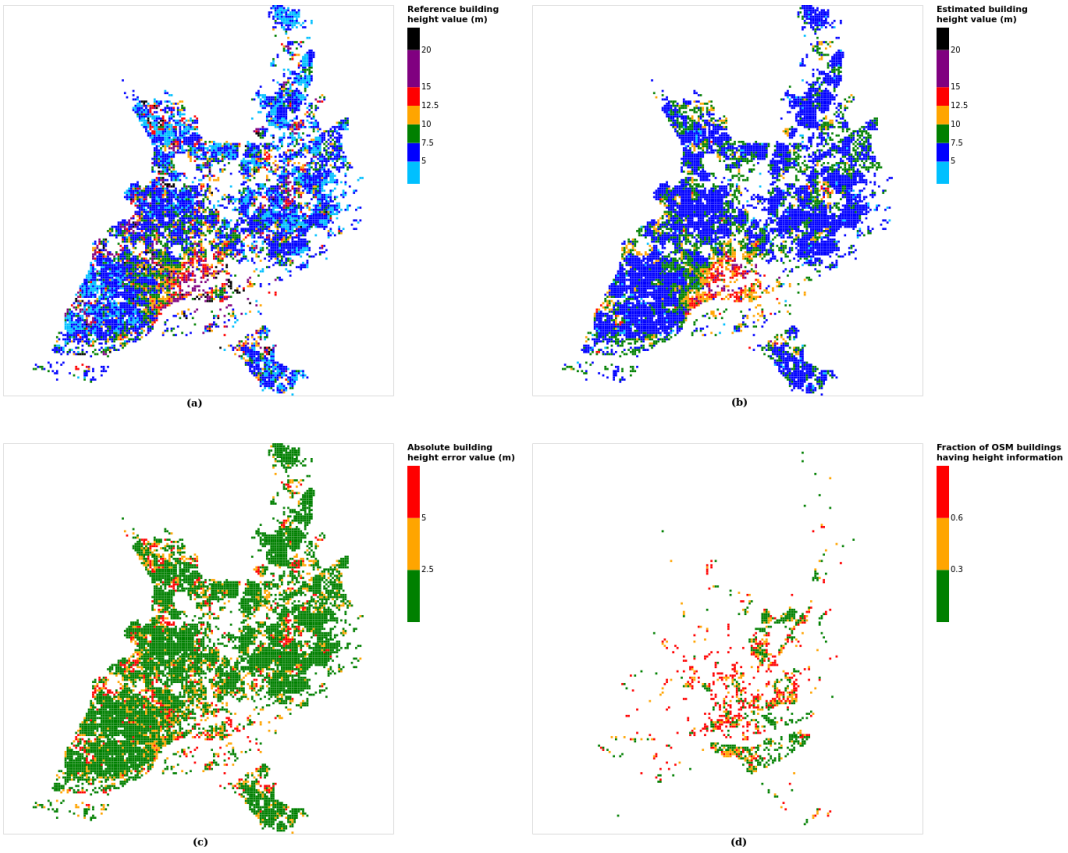

Figure 9. Results for the city of Nantes at grid cell: **(a)** reference building height, **(b)** estimated building height, **(c)** absolute building height error, and **(d)** fraction of OSM buildings that have height information. For cases **(a)**, **(b)** and **(c)**, only cells that have buildings with at least 90 % of their buildings having no height value in OSM are displayed.

ing more accurate reference data and in other countries with any existing reference data;

– improving the statistical modeling: (i) selecting a dataset that has a uniform distribution of building height, as described Sect. 3.4; (ii) using the height from OSM buildings that already have this information as an additional independent variable (e.g., the average building height at RSU scale may be used); (iii) investigate other supervised methods; (iv) in the training data, get rid of OSM buildings that have less than a certain fraction of BDT buildings covering them; (v) find more appropriate building properties that can be used as independent variables (e.g., the height of the nearby buildings filled by OSM users); and (vi) identify a subset of the most appropriate variables in order to limit the adverse effects of noisy variables.

Appendix A: List of all spatial indicators used as independent variables

A1 Building scale

Table A1. List of all building scale spatial indicators used as independent variables.

Indicator name	Indicator type			Name of the indicator in the GeoClimate documentation (version 0.0.1)
	Type and use	Form and size	Spatial relations	
BUILD_TYPE	x			None (this is an input of GeoClimate)
BUILD_MAIN_USE	x			None (this is an input of GeoClimate)
BUILD_PERIMETER		x		PERIMETER
BUILD_AREA		x		AREA
BUILD_TOTAL_FACADE_LENGTH		x		TOTAL_FACADE_LENGTH
BUILD_COMMON_WALL_FRACTION			x	COMMON_WALL_FRACTION
BUILD_NUMBER_BUILDING_NEIGHBOR			x	NUMBER_BUILDING_NEIGHBOR
BUILD_AREA_CONCAVITY		x		AREA_CONCAVITY
BUILD_FORM_FACTOR		x		FORM_FACTOR
BUILD_PERIMETER_CONVEXITY		x		PERIMETER_CONVEXITY
BUILD_MINIMUM_BUILDING_SPACING			x	MINIMUM_BUILDING_SPACING
BUILD_ROAD_DISTANCE			x	ROAD_DISTANCE
BUILD_LIKELIHOOD_LARGE_BUILDING		x		LIKELIHOOD_LARGE_BUILDING

A2 Block scale

Table A2. List of all block scale spatial indicators used as independent variables.

Indicator name	Indicator type		Name of the method in the GeoClimate documentation (version 0.0.1)
	Form and size	Aggregated statistics from lower scale	
BLOCK_BUILDING_DIRECTION_UNIQUENESS	x		BUILDING_DIRECTION_UNIQUENESS
BLOCK_AREA	x		AREA
BLOCK_BUILDING_DIRECTION_EQUALITY	x		BUILDING_DIRECTION_EQUALITY
BLOCK_HOLE_AREA_DENSITY	x		HOLE_AREA_DENSITY
BLOCK_CLOSINGNESS	x		CLOSINGNESS
BUILD_AVG_PERIMETER		x	None (average from lower scale)
BUILD_STD_PERIMETER		x	None (standard deviation from lower scale)
BUILD_AVG_AREA		x	None (average from lower scale)
BUILD_STD_AREA		x	None (standard deviation from lower scale)
BUILD_STD_TOTAL_FACADE_LENGTH		x	None (standard deviation from lower scale)
BUILD_STD_COMMON_WALL_FRACTION		x	None (standard deviation from lower scale)
BUILD_STD_NUMBER_BUILDING_NEIGHBOR		x	None (standard deviation from lower scale)
BUILD_AVG_AREA_CONCAVITY		x	None (average from lower scale)
BUILD_STD_AREA_CONCAVITY		x	None (standard deviation from lower scale)
BUILD_AVG_FORM_FACTOR		x	None (average from lower scale)
BUILD_STD_FORM_FACTOR		x	None (standard deviation from lower scale)
BUILD_AVG_PERIMETER_CONVEXITY		x	None (average from lower scale)
BUILD_STD_PERIMETER_CONVEXITY		x	None (standard deviation from lower scale)
BUILD_STD_MINIMUM_BUILDING_SPACING		x	None (standard deviation from lower scale)
BUILD_AVG_ROAD_DISTANCE		x	None (average from lower scale)
BUILD_STD_ROAD_DISTANCE		x	None (standard deviation from lower scale)
BUILD_AVG_LIKELIHOOD_LARGE_BUILDING		x	None (average from lower scale)
BUILD_STD_LIKELIHOOD_LARGE_BUILDING		x	None (standard deviation from lower scale)

A3 TSU scale

Table A3. List of all TSU scale spatial indicators used as independent variables.

Indicator name	Indicator type			Name of the method in the GeoClimate documentation (version 0.0.1)
	Form and size	Planar density	Aggregated statistics from lower scale	
RSU_HIGH_VEGETATION_FRACTION		x		AREA_FRACTION_x
RSU_HIGH_VEGETATION_WATER_FRACTION		x		AREA_FRACTION_x
RSU_HIGH_VEGETATION_BUILDING_FRACTION		x		AREA_FRACTION_x
RSU_HIGH_VEGETATION_LOW_VEGETATION_FRACTION		x		AREA_FRACTION_x
RSU_HIGH_VEGETATION_ROAD_FRACTION		x		AREA_FRACTION_x
RSU_HIGH_VEGETATION_IMPERVIOUS_FRACTION		x		AREA_FRACTION_x
RSU_WATER_FRACTION		x		AREA_FRACTION_x
RSU_BUILDING_FRACTION		x		AREA_FRACTION_x
RSU_LOW_VEGETATION_FRACTION		x		AREA_FRACTION_x
RSU_ROAD_FRACTION		x		AREA_FRACTION_x
RSU_IMPERVIOUS_FRACTION		x		AREA_FRACTION_x
RSU_VEGETATION_FRACTION_URB		x		VEGETATION_FRACTION_URB
RSU_LOW_VEGETATION_FRACTION_URB		x		LOW_VEGETATION_FRACTION_URB
RSU_HIGH_VEGETATION_IMPERVIOUS_FRACTION_URB		x		HIGH_VEGETATION_IMPERVIOUS_FRACTION_URB
RSU_HIGH_VEGETATION_PERVIOUS_FRACTION_URB		x		HIGH_VEGETATION_PERVIOUS_FRACTION_URB
RSU_ROAD_FRACTION_URB		x		ROAD_FRACTION_URB
RSU_IMPERVIOUS_FRACTION_URB		x		IMPERVIOUS_FRACTION_URB
RSU_AREA	x			AREA
RSU_GROUND_LINEAR_ROAD_DENSITY		x		GROUND_LINEAR_ROAD_DENSITY
RSU_AVG_NUMBER_BUILDING_NEIGHBOR			x	AVG_NUMBER_BUILDING_NEIGHBOR
RSU_AVG_MINIMUM_BUILDING_SPACING			x	AVG_MINIMUM_BUILDING_SPACING
RSU_BUILDING_NUMBER_DENSITY		x		BUILDING_NUMBER_DENSITY
RSU_BUILDING_TOTAL_FRACTION		x		BUILDING_TOTAL_FRACTION
RSU_BUILDING_DIRECTION_EQUALITY			x	BUILDING_DIRECTION_EQUALITY
RSU_BUILDING_DIRECTION_UNIQUENESS			x	BUILDING_DIRECTION_UNIQUENESS

Appendix B: Results for all cities

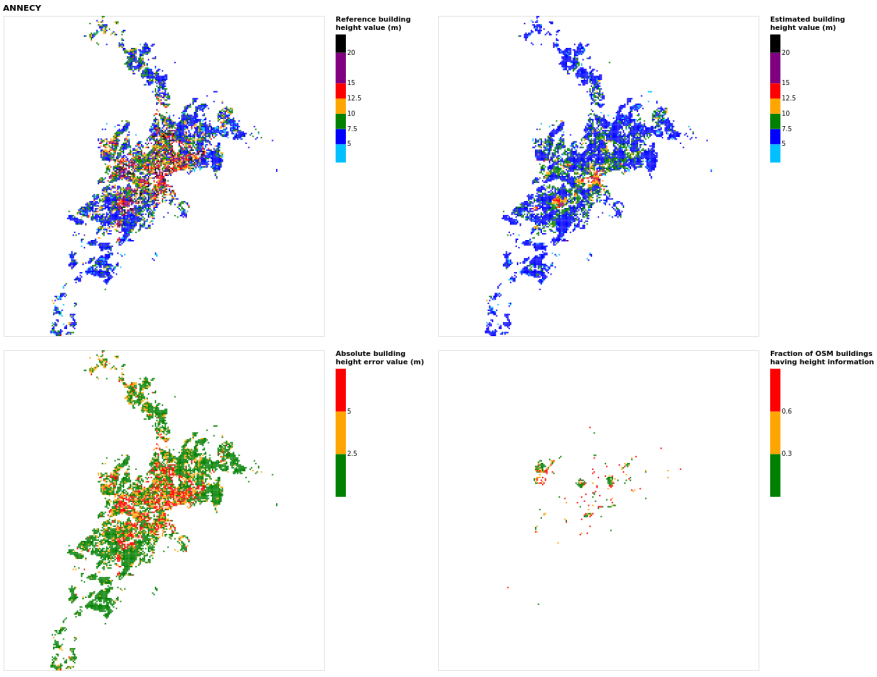

Figure B1. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

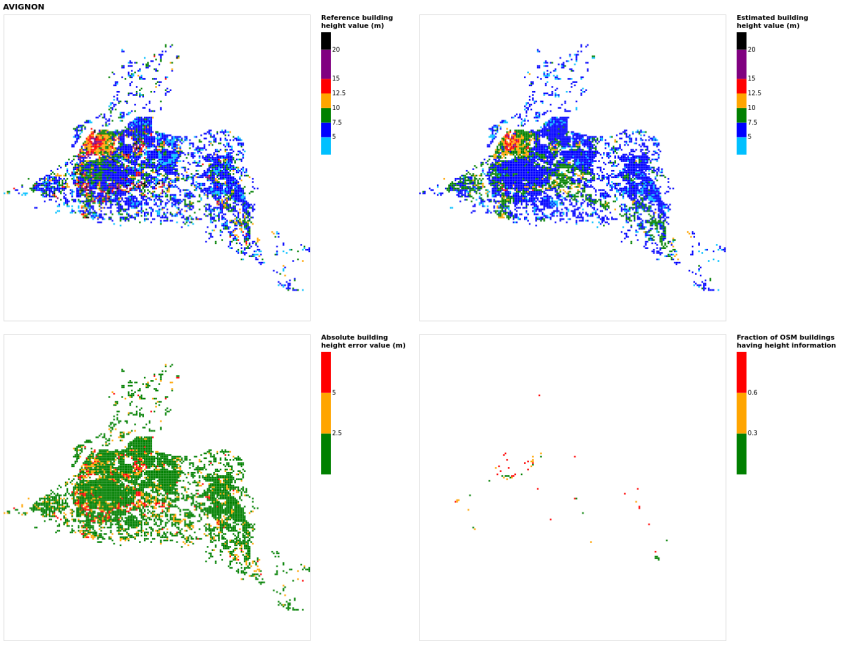

Figure B2. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings having height information. For all panels except the lower right, only cells that buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

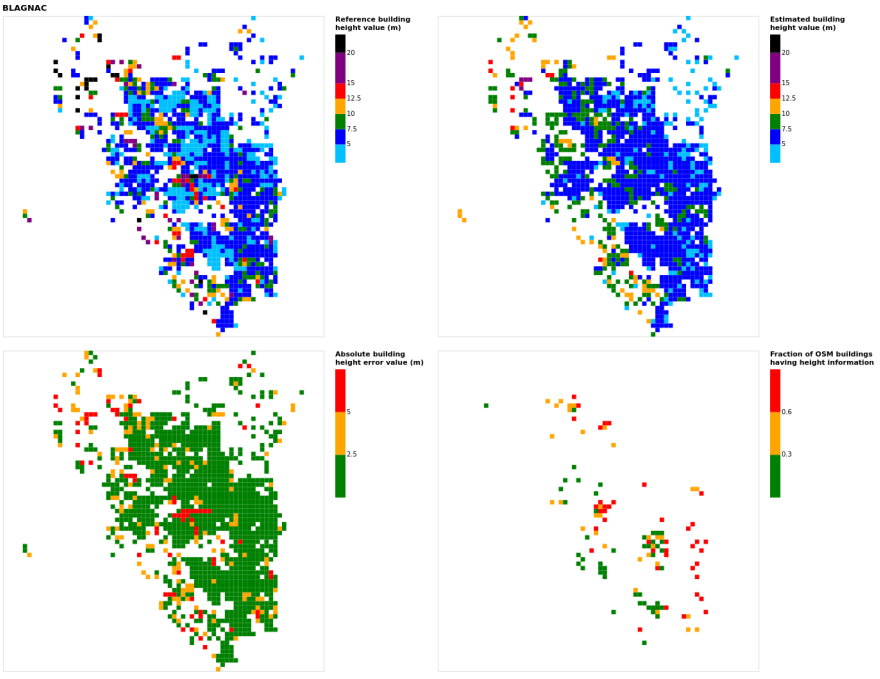

Figure B3. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

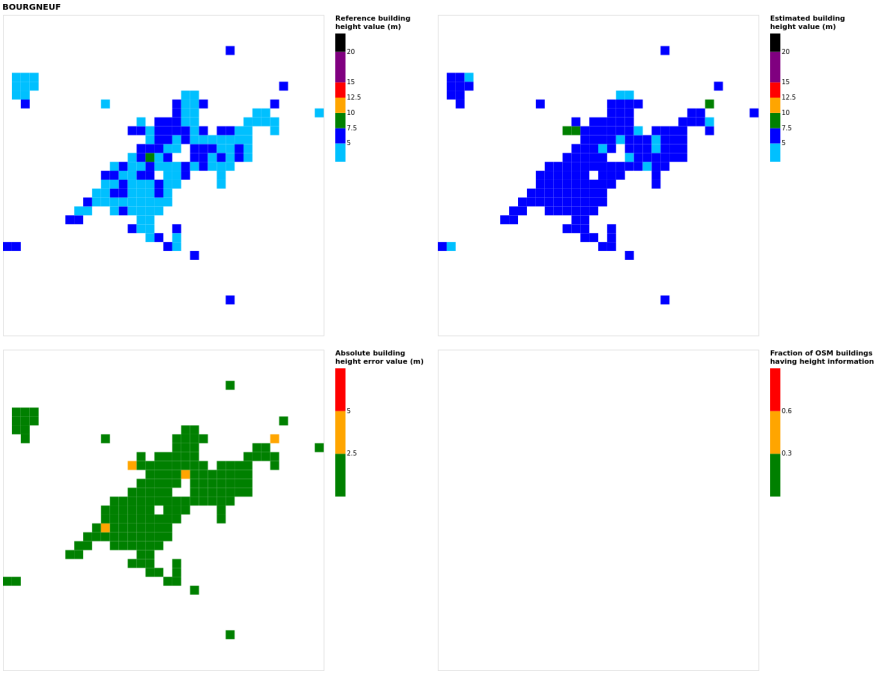

Figure B4. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

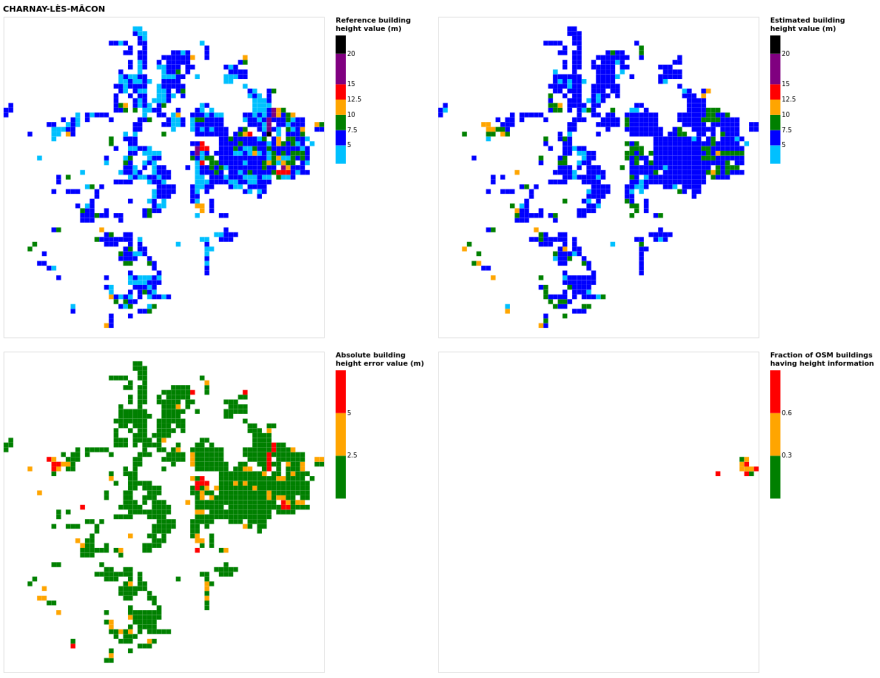

Figure B5. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

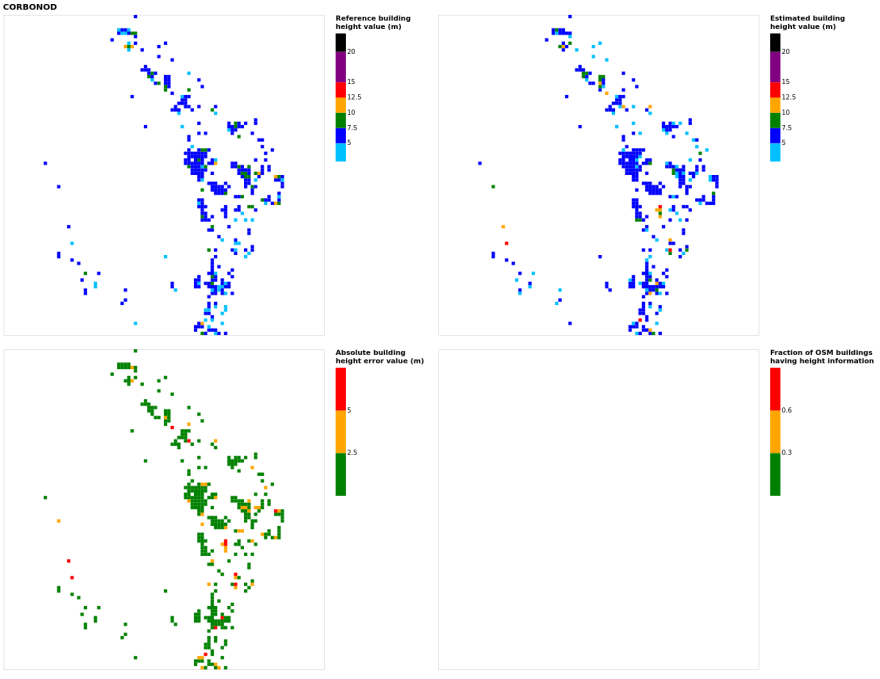

Figure B6. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

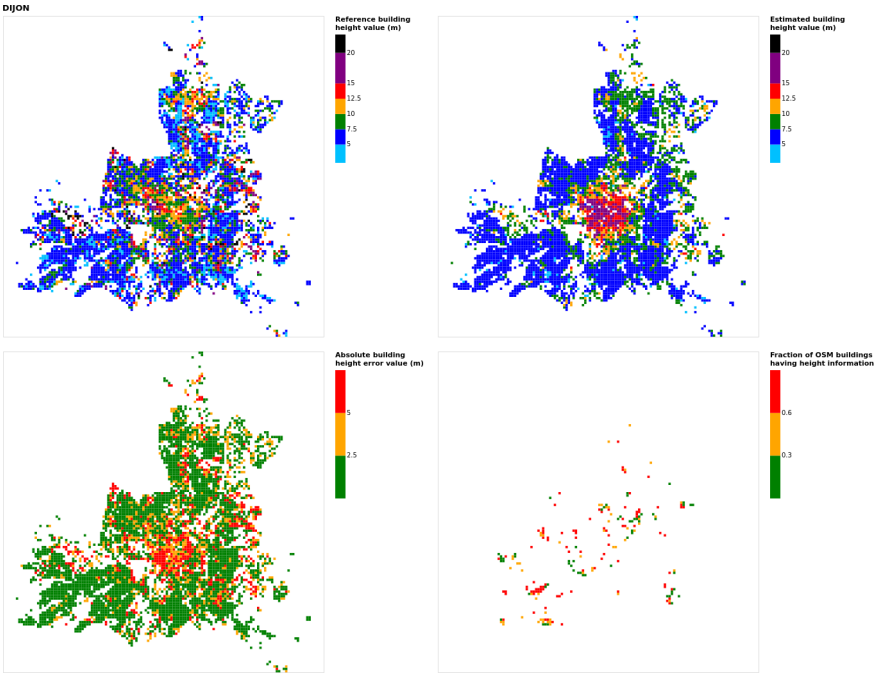

Figure B7. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

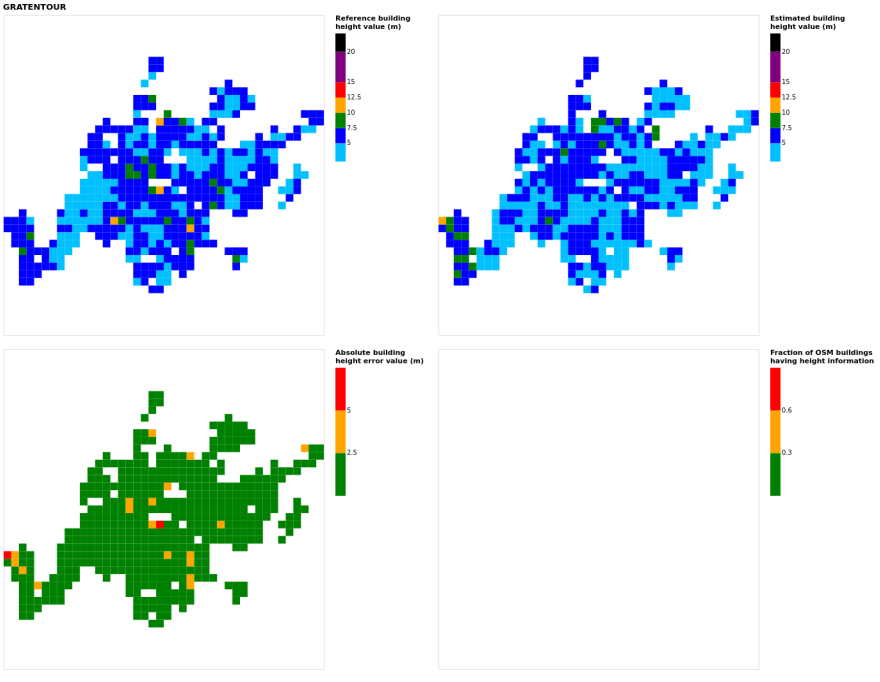

Figure B8. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

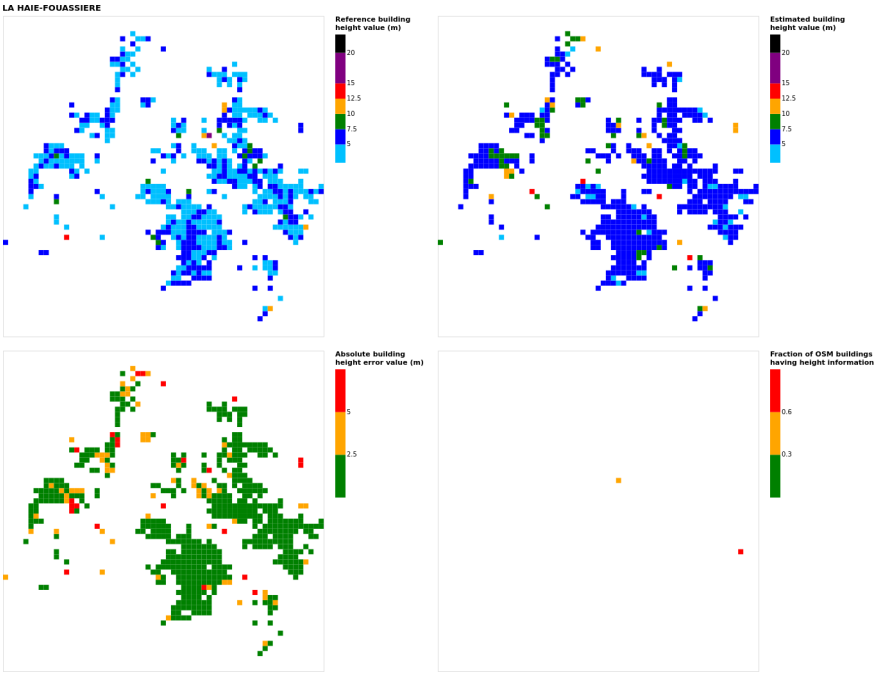

Figure B9. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

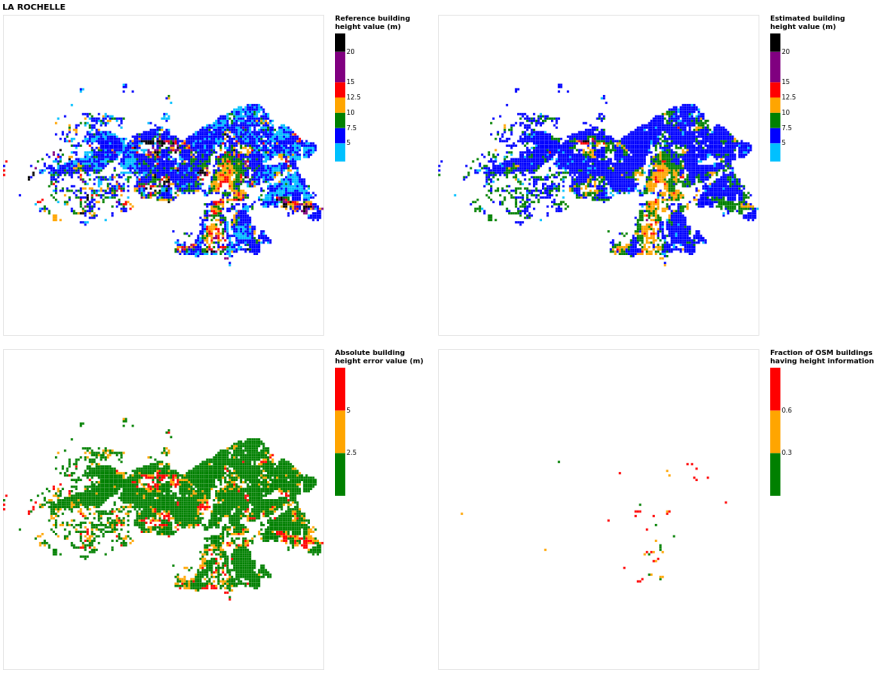

Figure B10. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

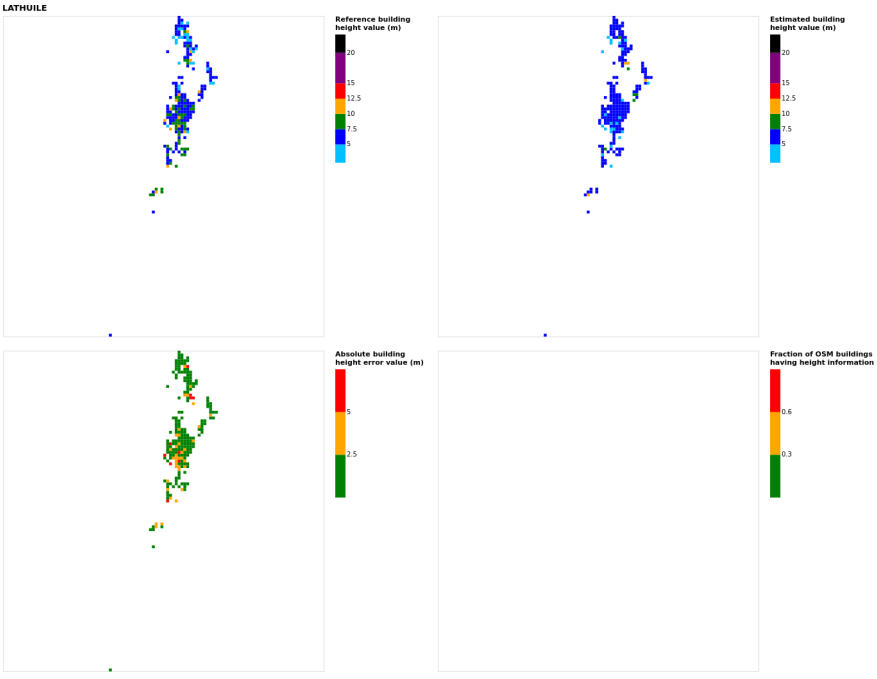

Figure B11. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

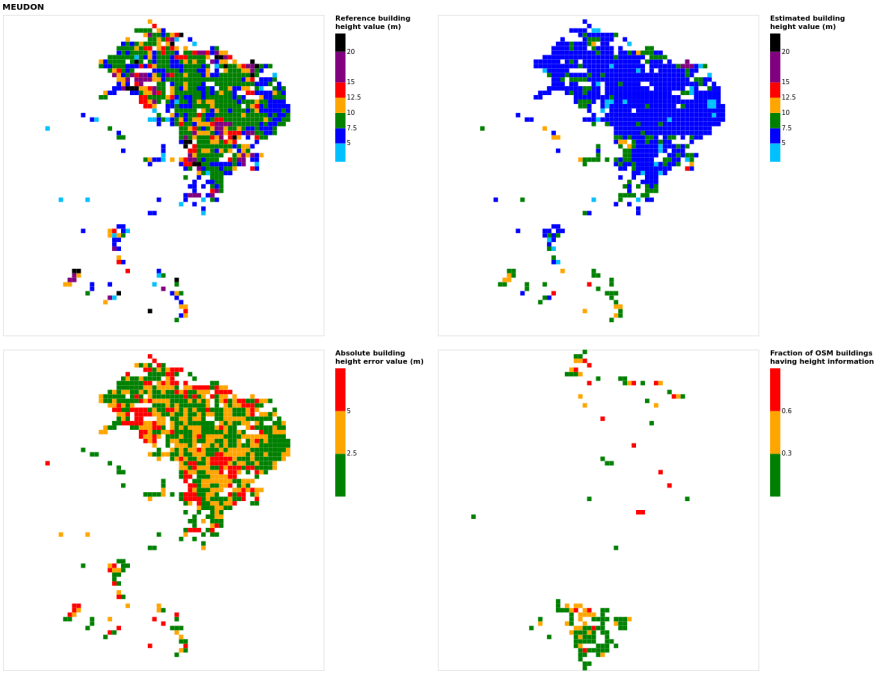

Figure B12. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

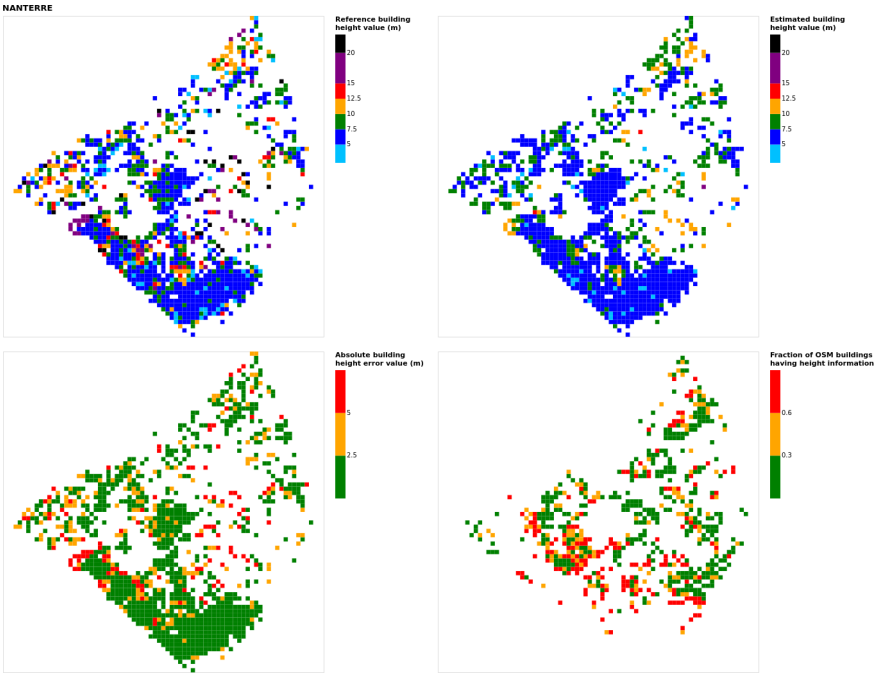

Figure B13. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

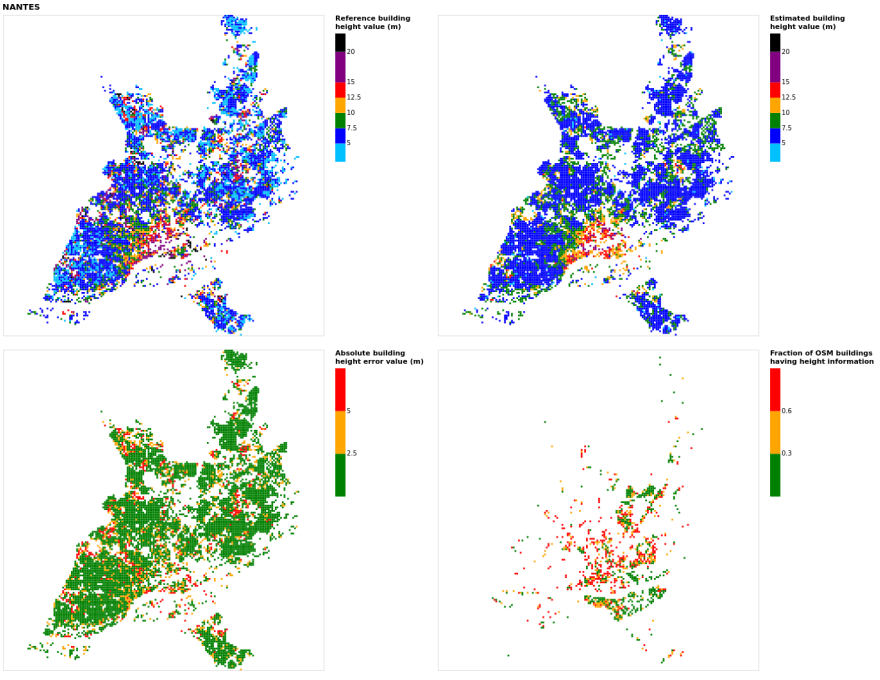

Figure B14. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

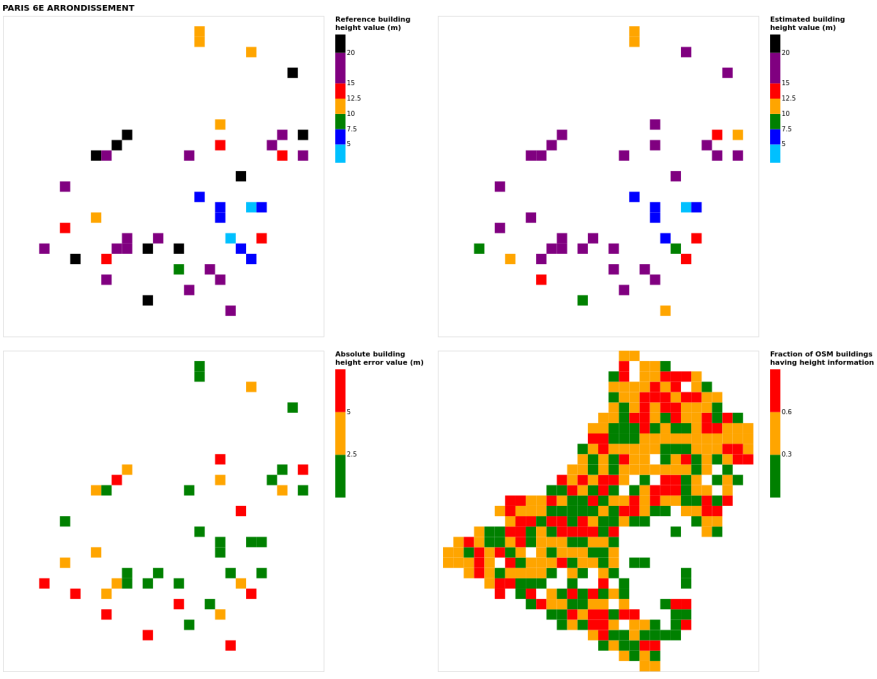

Figure B15. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

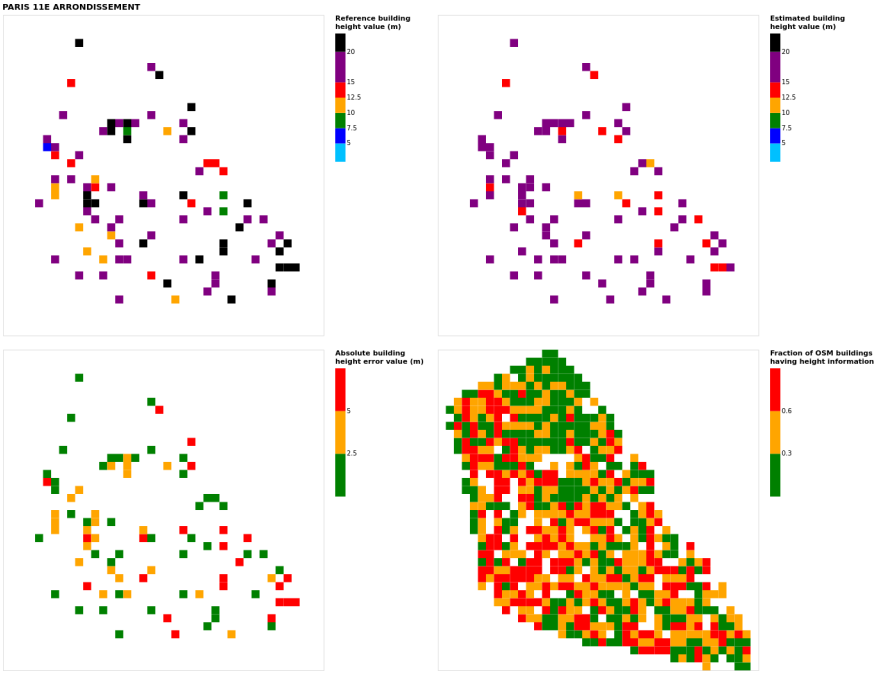

Figure B16. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

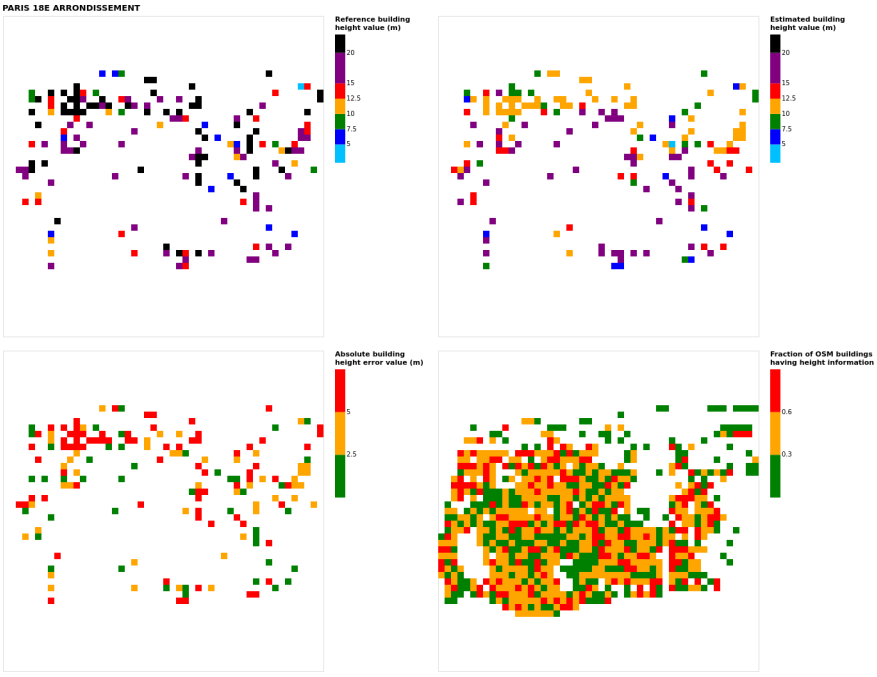

Figure B17. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

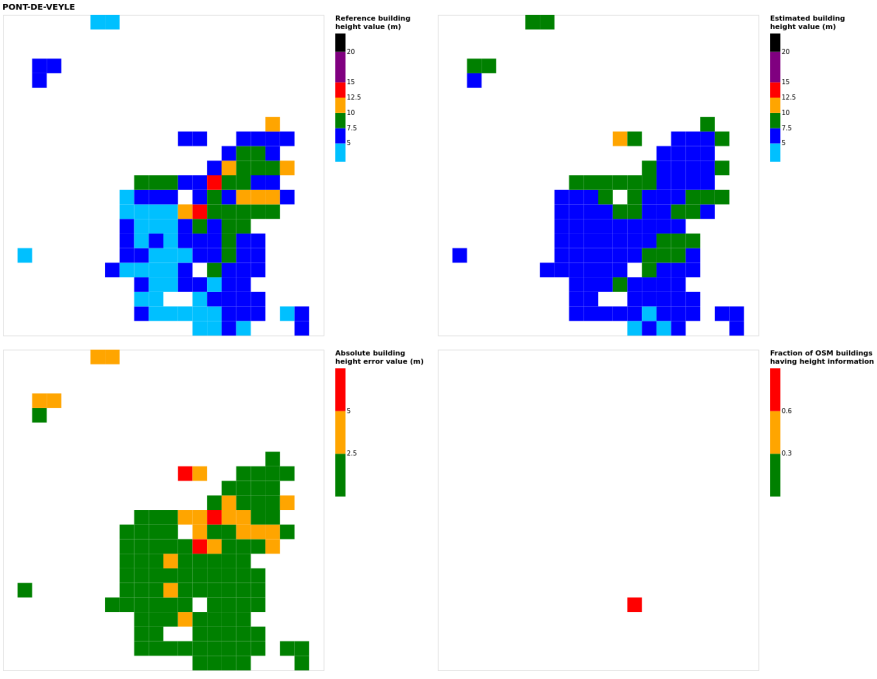

Figure B18. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

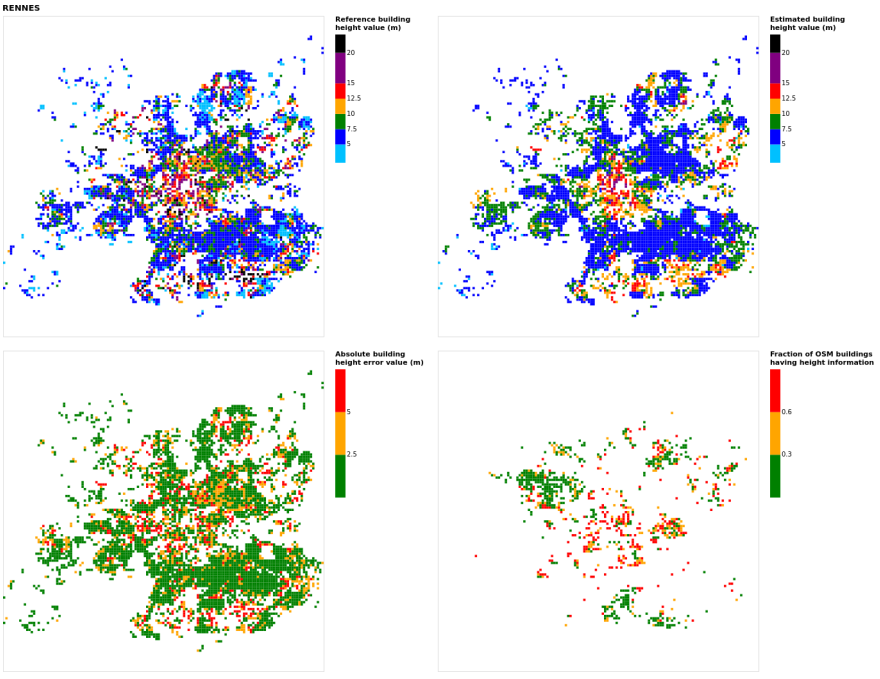

Figure B19. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

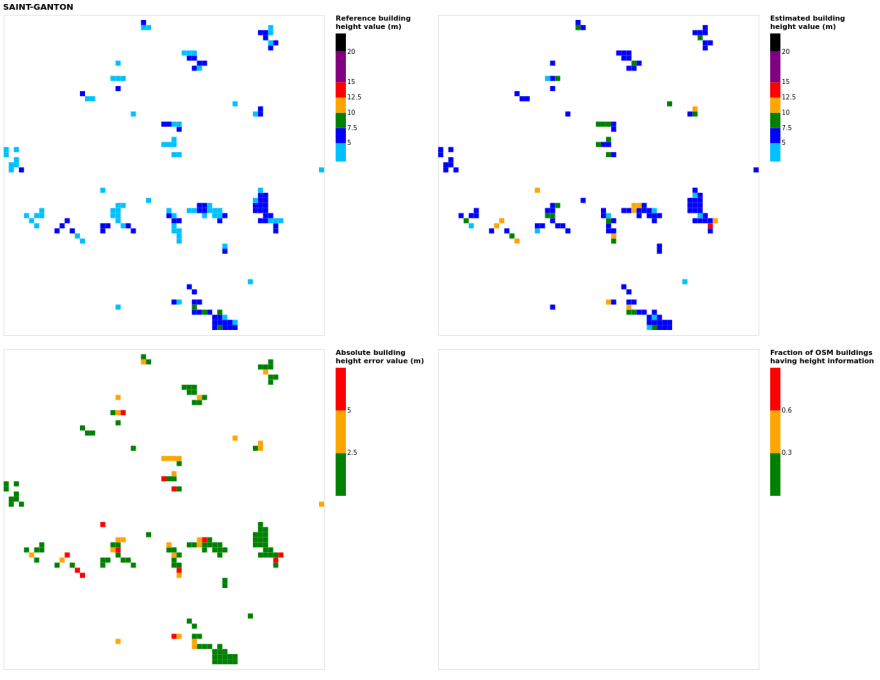

Figure B20. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

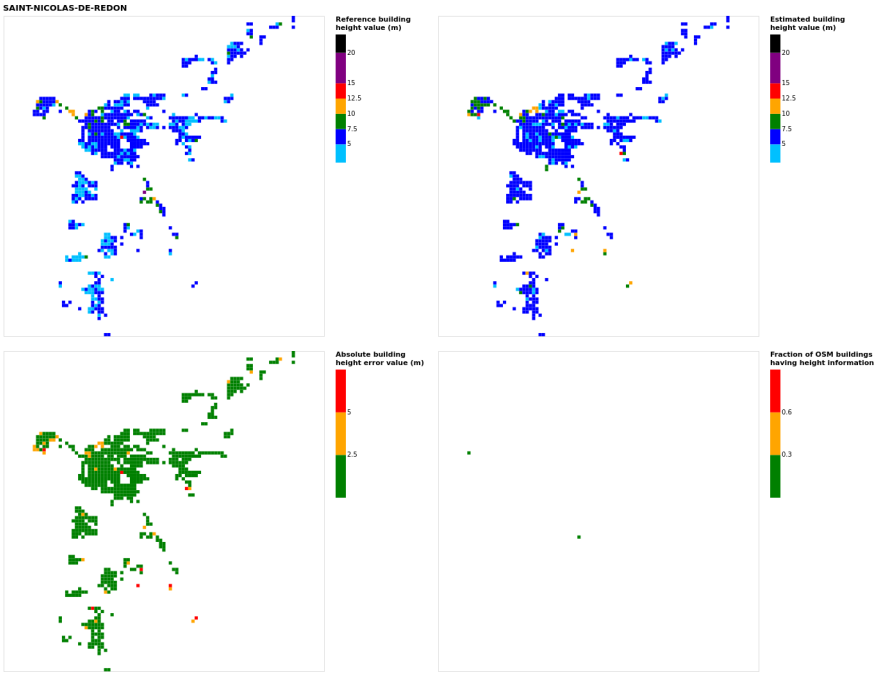

Figure B21. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

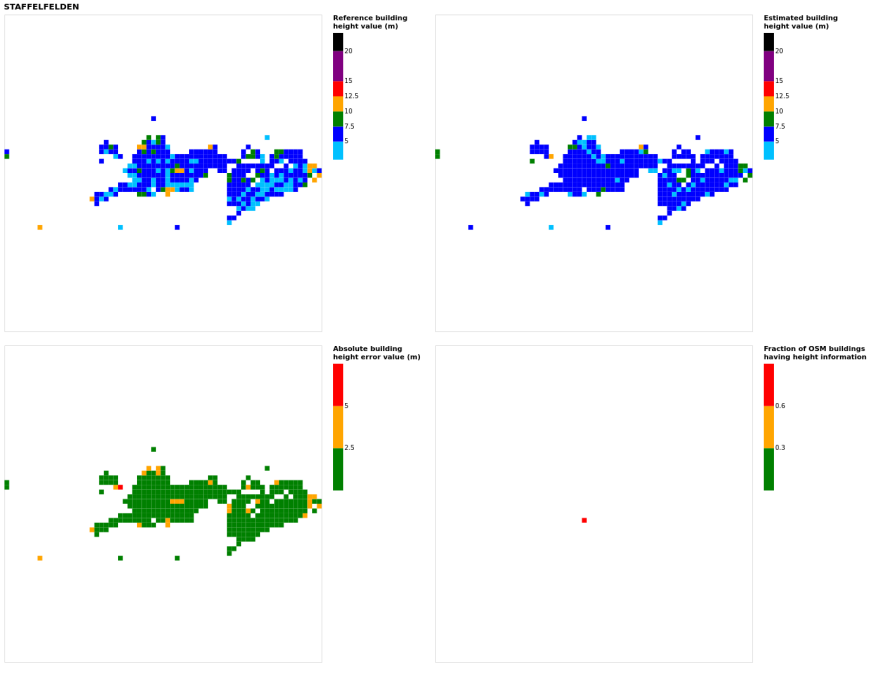

Figure B22. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

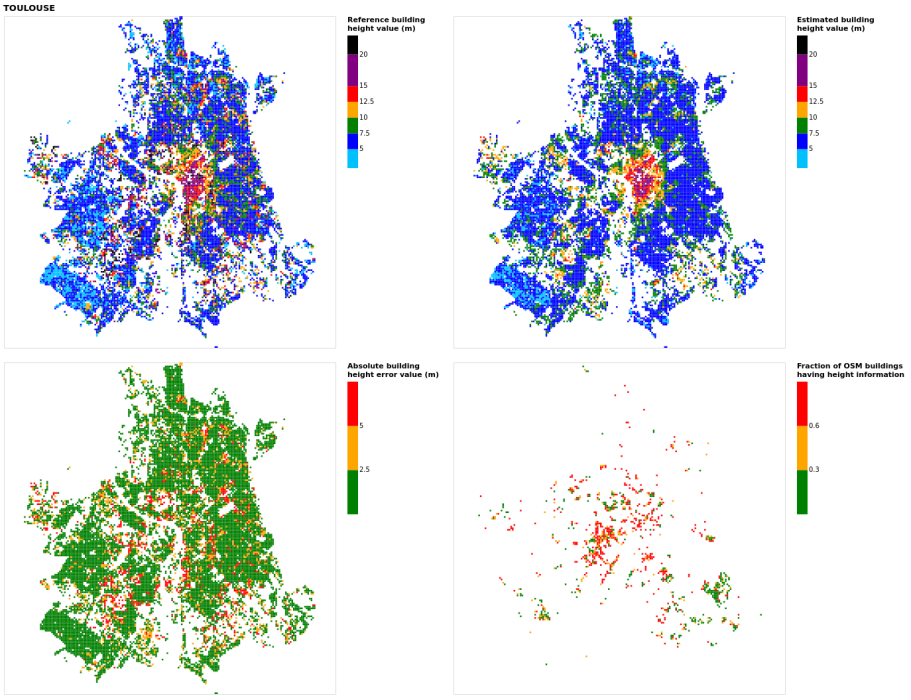

Figure B23. Results for the commune at grid cell (upper left panel) reference building height, (upper right panel) estimated building height, (lower left panel) absolute building height error, and (lower right panel) fraction of OSM buildings that have height information. For all panels except the lower right, only cells that have buildings – with at least 90 % of their buildings having no height value in OSM – are displayed.

Code and data availability. The major part of this work can be reproduced directly using the Software GeoClimate version 0.0.1 (the source code and executable file of this software version are permanently available on Zenodo at https://doi.org/10.5281/zenodo.6372337, Bocher et al., 2021b); the scripts and data are available on Zenodo at https://doi.org/10.5281/zenodo.6855063 (Bernard et al., 2021). GeoClimate downloads OpenStreetMap data using the overpass API from the end point https://overpass-api.de/ (last access: 28 September 2022), estimates building height when missing and calculates geographical indicators. The resulting datasets presented in this paper have been obtained using the OpenStreetMap data between June and September 2021. It can be freely accessed at https://doi.org/10.5281/zenodo.6855063 (Bernard et al., 2021). The French BDTopo (version 2.2) is used only for training and evaluation purposes. It is a proprietary dataset provided by the French National Geographic Institute (IGN) and is available upon request. Thus it is unfortunately not possible to make this dataset freely accessible. This was one of the major motivations for perform this work, i.e., to create a methodology to automatically create a topographic dataset containing buildings with estimated height.

Author contributions. The conceptualization was performed by JB, EB and VM, the data curation by JB, EB and ELS, the formal analysis by JB, EB, ELS and FL, the acquisition of the funding by EB and ELS, the investigation, the definition of the methodology and the project administration by JB and EB, the resources by EB, the software development by EB, JB, ELS and FL, the supervision of all tasks by JB and EB, the validation of the work by EB, JB, ELS and FL, the work dedicated to visualization by JB and EB, the original draft preparation by JB and EB, the review and editing by JB, EB, FL, ELS and VM.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. The method presented in this paper has been integrated in the GeoClimate tool and developed within the following research projects:

- URCLIM (2017–2021), part of ERA4CS, a project initiated by JPI Climate and co-funded by the European Union under grant agreement no. 690462
- CENSE (2017–2021), funded by the French National Research Agency (ANR) under grant agreement no. Projet-ANR-16-CE22-0012

– SLIM (2020–2021), a Copernicus project C3S_432 Provisions to Environmental Fore- casting Applications (Lot 2).

Financial support. The article processing charges for this open-access publication were covered by the Gothenburg University Library.

Review statement. This paper was edited by Richard Mills and reviewed by two anonymous referees.

References

Bernabé, A., Musy, M., Andrieu, H., and Calmet, I.: Radiative properties of the urban fabric derived from surface form analysis: A simplified solar balance model, Sol. Energy, 122, 156–168, 2015.

Bernard, J., Bocher, E., Petit, G., and Palominos, S.: Sky View Factor Calculation in Urban Context: Computational Performance and Accuracy Analysis of Two Open and Free GIS Tools, Climate, 6, 60, https://doi.org/10.3390/cli6030060, 2018.

Bernard, J., Bocher, E., Wiederhold, E. L. S., Leconte, F., Masson, V., and Noûs, C.: Estimated height of the OpenStreetMap buildings of 24 French communes using the GeoClimate Software (version 0.0.1), Zenodo, https://doi.org/10.5281/zenodo.6855063, 2021.

Biljecki, F., Ledoux, H., and Stoter, J.: Generating 3D city models without elevation data, Computers, Environment and Urban Systems, 64, 1–18, 2017.

Bocher E., Bernard J., Le Saux Wiederhold E., Leconte F., Petit G., Palominos S., and Noûs C.: GeoClimate: a Geospatial processing toolbox for environmental and climate studies, Zenodo, https://doi.org/10.5281/zenodo.5534680, 2021a.

Bocher E., Bernard J., Le Saux Wiederhold E., Leconte F., Petit G., Palominos S., and Noûs C.: GeoClimate: a Geospatial processing toolbox for environmental and climate studies (0.0.1), Zenodo, https://doi.org/10.5281/zenodo.6372337, 2021b.

Bocher, E., Guillaume, G., Picaut, J., Petit, G., and Fortin, N.: NoiseModelling: An Open Source GIS Based Tool to Produce Environmental Noise Maps, ISPRS Int. J. Geo-Inf., 8, 130, https://doi.org/10.3390/ijgi8030130, 2019.

Bocher, E., Bernard, J., Wiederhold, E. L. S., Leconte, F., Petit, G., Palominos, S., and Noûs, C.: GeoClimate: a Geospatial processing toolbox for environmental and climate studies, Journal of Open Source Software, 6, 3541, https://doi.org/10.21105/joss.03541, 2021.

Cao, Y. and Huang, X.: A deep learning method for building height estimation using high-resolution multi-view imagery over urban areas: A case study of 42 Chinese cities, Remote Sens. Environ., 264, 112590, https://doi.org/10.1016/j.rse.2021.112590, 2021.

Fradkin, M., Roux, M., Maître, H., and Leloglu, U. M.: Surface reconstruction from multiple aerial images in dense urban areas, in: Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), 2, 262–267, IEEE, 1999.

Hanna, S. R. and Britter, R. E.: Wind flow and vapor cloud dispersion at industrial and urban sites, vol. 7, John Wiley and Sons, https://doi.org/10.1002/9780470935613, 2010.

Hastie, T., Tibshirani, R., and Friedman, J.: Data mining, inference, and prediction, The elements of statistical learning Springer Series in Statistics, Springer-Verlag, New York, https://doi.org/10.1007/978-0-387-84858-7, 2001.

INSEE: Method for determining functional areas in 2020, 2020.

Johansson, L., Onomura, S., Lindberg, F., and Seaquist, J.: Towards the modelling of pedestrian wind speed using high-resolution digital surface models and statistical methods, Theor. Appl. Climatol., 124, 189–203, 2016.

Lao, J., Bocher, E., Petit, G., Palominos, S., Le Saux, E., and Masson, V.: Is OpenStreetMap suitable for urban climate studies?, in: OGRS2018, Open Source Geospatial Research and Education Symposium, 9–11 October 2018, Lugano, Switzerland, https://halshs.archives-ouvertes.fr/halshs-01898612 (last access: 19 Sptember 2022), 2018.

Lindberg, F.: Modelling the urban climate using a local governmental geo-database, Meteorol. Appl., 14, 263–273, https://doi.org/10.1002/met.29, 2007.

Masson, V., Heldens, W., Bocher, E., Bonhomme, M., Bucher, B., Burmeister, C., de Munck, C., Esch, T., Hidalgo, J., Kanani-Sühring, F., Kwok, Y.-T., Lemonsu, A., Lévy, J.-P., Maronga, B., Pavlik, D., Petit, G., See, L., Schoetter, R., Tornay, N., Votsis, A., and Zeidler, J.: City-descriptive input data for urban climate models: Model requirements, data sources and challenges, Urban Climate, 31, 100536, https://doi.org/10.1016/j.uclim.2019.100536, 2020.

Milojevic-Dupont, N., Hans, N., Kaack, L. H., Zumwald, M., Andrieux, F., de Barros Soares, D., Lohrey, S., Pichler, P.-P., and Creutzig, F.: Learning from urban form to predict building heights, PLOS ONE, 15, e0242010, https://doi.org/10.1371/journal.pone.0242010, 2020.

Mocnik, F.-B., Zipf, A., and Raifer, M.: The OpenStreetMap folksonomy and its evolution, Geo-spatial Information Science, 20, 219–230, 2017.

Oke, T. R.: Boundary layer climates, 2nd ed., Routledge, https://doi.org/10.4324/9780203407219, 2002.

Shan, J. and Toth, C. K.: Topographic laser ranging and scanning: principles and processing, 2nd ed., CRC press, https://doi.org/10.1201/9781315154381, 2018.

Shao, Y., Taff, G. N., and Walsh, S. J.: Shadow detection and building-height estimation using IKONOS data, Int. J. Remote, 32, 6929–6944, 2011.

Sohn, G., Huang, X., and Tao, V.: Using a binary space partitioning tree for reconstructing polyhedral building models from airborne lidar data, Photogramm. Eng. Rem. S., 74, 1425–1438, 2008.

Song, H., Huang, B., and Zhang, K.: Shadow detection and reconstruction in high-resolution satellite images via morphological filtering and example-based learning, IEEE T. Geosci. Remote, 52, 2545–2554, 2013.

Tang, U. and Wang, Z.: Influences of urban forms on traffic-induced noise and air pollution: Results from a modelling system, Environ. Modell. Softw., 22, 1750–1764, https://doi.org/10.1016/j.envsoft.2007.02.003, 2007.

Zeng, C., Wang, J., Zhan, W., Shi, P., and Gambles, A.: An elevation difference model for building height extraction from stereo-image-derived DSMs, Int. J. Remote, 35, 7614–7630, 2014.