Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., author comment AC1
https://doi.org/10.5194/gmd-2021-428-AC1, 2022

Reply on RC1

Jérémy Bernard et al.

Author comment on "Estimation of missing building height in OpenStreetMap data: a French case study using GeoClimate 0.0.1" by Jérémy Bernard et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-428-AC1, 2022

Thank you to *Anonymous Referee #1* for the time spent for the review. Below we have carefully answered each of your points. For most of them we modified the text which are highlighted in the enclosed pdf file. Some more modifications in the manuscript will be performed whenever agreement will be found for some of the other points.

1. Thanks to *Anonymous Referee #1* for this remark and the very interesting study he pointed out which is indeed quite similar to our work. Even thought we did not have knowledge of this study, we think that at least three points make our work useful anyway:

- The geographical indicators used as explicative variables are different between [NMD20] and our study, however results tend to be similar, which in a sense is worth of being investigated.
- While the geographical scale is larger and the investigation much deeper than ours, [NMD20] work seems hard to replicate (and thus hard to use) since their code seems only partially available (https://gitlab.pik-potsdam.de/nikolami/learning-from-urban-form-to-predict-building-heights). We think that the strength of our work (and this is why we chose GMD to publish our work) is its accessibility and its possibility of reuse.
- The building height estimation accuracy has direct consequences on spatial indicators used to model / analyze urban climate. Nothing is related to this topic in [NMD20] while these consequences are investigated in our work.

We have modified the introduction to better highlight our contribution to the field.

2. The main objective of our work is actually two-fold:

- Proposing a whole methodology to estimate building height from Free and Open Source data using a Free and Open-Source Software (FOSS) available online at https://github.com/orbisgis/geoclimate/wiki ,
- Evaluating the performance of this methodology in an area where we can easily obtain data as French researchers.

However, as *Anonymous Referee #1* guessed it, the mid-term objective of our work is to verify that the methodology applied in France is still valid in other countries. This is why we have designed GeoClimate to be simple to use for anyone in any country. Any

researcher with an access to local data is able to assess the performance of the current model for his country or to use GeoClimate to create a new model (might be the one developed by [NMD20]) which would be more appropriate for his country.

3. It seems that this point is viewed by *Anonymous Referee #1* as a major methodological concern. In the following, we try our best to answer each of these points.

We are not sure to understand well the differences between your second and your third points. We try to clarify what we have done in the training and validation stages:

- Concerning the training, we have actually trained the model on 70% of the training data and validate it on the 30% remaining (cf. section 2.2.3 and Figure 5). This has been done using all cities of the training dataset. The objective was to identify what was the best set of parameters for the RandomForest model.
- Then we have applied this "optimized" model on the validation dataset which consists of spatially distant cities from the ones used in the training. Thus if the trained model would have been poorly designed, it would have led to worse results using the validation cities. Your third point mentions the distance between training and validation cities which is not sufficient large. We do not think that using cities for validation "more remote from the training ones" will modify the performance of the model. Many reasons leads to this assumption:
- We have actually tested a previous version of the model on more remote cities and the performance was similar,
- We think that the city's topography (mainly water bodies and mountains), the city's historical heritage (e.g. periods of demographic expansion) and how cities are located within the attraction cluster such as defined by the French INSEE (types defined Table 1) plays an important role on the city's morphology. From this perspective, Corbonod (validation dataset) is located within a mountainous area (such as Annecy or La Thuile - training) but it is not at all affected the same way by the cluster attraction (rural area versus urban areas respectively) since its expansion is much less constrained by the mountains. The other cities in the region (Dijon, Charnay-lès-Mâcon and Pont-de-Veyle - validation) are located in much flater areas. Concerning West France, Nantes (training) is a main urban area built along a large river (the Loire) while the closest cities in the validation dataset are cities within or outside a much smaller attraction cluster (Redon).
- [NMD20] has obtained almost no improvement of their prediction when they add to the training dataset a city which is really close to the ones used for validation. In the meantime, [NMD20] has obtained a much better improvement adding randomly chosen local data: "*In Experiment 2, we added 2% of local data to the training set data which resulted in noticeable accuracy gains compared to Experiment 1 for both test sets. In contrast, Experiment 3 where we added Berlin to the training set for predicting Brandenburg did not noticeably improve the results.*" (p. 12). [NMD20] also mentions that "*the townhouses that are so typical for Berlin are not as common even in large cities in Brandenburg, despite the geographical vicinity."* (p.14).

4. The introduction has been modified for clarification purpose (cf. the "diff" version of the attached pdf file).

In the preprint, there is no RMSE value line 170, and the words "all cities" are not mentioned line 175. Could the reviewer cite the paragraph or the section that he suggests to improve ?

Concerning Fig. 6, training and validation were on purpose shown on the same axis to show how similar is the error for both of them. However, for more clarity we have also summarized Fig. 6 and 7 in tables as recommended by *Anonymous Referee #1.*

5. Concerning the debate about the use of OSM database instead of government data, there are several arguments in favor of the first:

- When we started the project, OSM was the unique open data set available on the whole French territory (the BDTopo V3 is free and open access but the V2.2 is not).
- OSM covers the whole planet.
- OSM gives free and unlimited access to the entire database, with a complete history of changes.
- OSM provides easy data access thanks to the Overpass API that permits to download data on demand for any part of the territory (using a bbox, a name for a commune...).
- OSM data model is flexible (thanks to the tags approach) and can quickly be updated by any people in the world
- And since it's open, anyone can also help improve the quality : edit the geometry or add new descriptors.

One of our main objectives is to provide a methodology and an open tool to produce climate and environmental indicators for any communities (e.g. geographers, urban stakeholders, environmental and climate specialists), therefore we believe that the OSM source was the best option. It is now possible for anyone having local government database to compare them to the building height estimated within GeoClimate.

Please also note the supplement to this comment:
https://gmd.copernicus.org/preprints/gmd-2021-428/gmd-2021-428-AC1-supplement.pdf