Interactive comment on "Estimating exposure of residential assets to natural hazards in Europe using open data" *by* Dominik Paprotny et al.

Dominik Paprotny et al.

d.paprotny@tudelft.nl

Received and published: 11 November 2019

We thank the reviewer for taking the time to analyse our manuscript. Below we list the comments (C) and our responses (R).

C: I agree that 'the problem of accurately identifying buildings and occupancy, especially with open data, is outside the scope of this paper'. However, it remains unclear how residential buildings were eventually defined in this study. This needs to be clearly stated for the sake of reproducibility. Apparently, two OSM layers (buildings and land use) were downloaded (On a sidenote: a date indicating the day of the download would be nice to reference the status/version of the data set used). Was information obtained from the buildings layer enhanced or modified based on the land use? If so, how?

R: Firstly, we downloaded two Map Features ("Buildings" and "Landuse") for the 30 study areas and the example application study from section 3.3. For the analysis, residential buildings were objects from "Buildings" layer which (1) had tags "residential", "apartments", "house", "detached" or "terrace", and (2) had tags "yes" (indicating that a building exists, but the function is not defined) and were located within an object from "Landuse" layer which had a tag "residential". OSM is updated continuously, and the data used here were downloaded between 22 and 25 January 2019. Data for the example application was downloaded from OSM on 18 July 2019. We will add this information to the text.

C: Seven potentially important variables were initially defined. Three of these variables were included in the final model. Even though it can be guessed how these variables were selected (p.4, l.113), the variable selection process is not clearly described.

R: The variables first variable (POP) was chosen as it had the highest unconditional rank correlation with H (Table S1). The second variable (B) had the highest conditional rank correlation with H among the remaining six variables, and then IMD had the highest conditional rank correlation with H among the remaining five variables. Further variables had very low (<0.1) correlation with H only, therefore only three variables were used to explain H. Remaining arc between POP and IMD was added due to high correlation between the two. All arcs further have a theoretical explanation, as described in lines 117-122. We will clarify the explanation in the text.

C: I think that the use of a 2% sample is somewhat critical, since a lot of information is dropped. Why were so many instances dropped, how was this number (2%) chosen, and how can the authors guarantee that this is a representative sample? The full data set should include roughly 2,373,300 records (2% correspond to 47,466 records). A data frame with 2 million rows and maybe 10 columns is definitely still manageable on local machines.

R: The principle reason for using a sample of the OSM buildings for all cities was reduc-

ing the time of data processing with GIS to arrive with table of data. Then, the efficiency of a non-parametric Bayesian Network calculation (or a Random Forest computation) drops significantly with large numbers of data points, which would make it less suitable for a possible operational pan-European application. Those aspects notwithstanding, the buildings of 30 capital regions are not the complete population of the European residential buildings, but a small fraction of them, and using all data would still mean using a sample of European buildings. We went back to the full dataset and extracted a larger share of the data, so that the influence of using different dataset size can be presented. The whole population of usable data (i.e. no data missing for any variable at a given location) results in a correlation matrix almost identical correlation matrix, differing by a rank correlation of 0.007 at most. This matrix (N = 2,375,058) will be used to revise Table S1.

C: In addition, the 2% sample was only used once. Results were then tested once on a 1% sample. This approach is not very robust. Proper k-fold cross-validation using the full data set would be desirable.

R: We made a new extraction of the data from the full dataset, containing 10% of the data. This sample was used for a 10-fold cross-validation, and the revision of other results. We have revised Tables 2 and 3 (see supplement to this response), and we will revise related text accordingly.

C: What was the reason to use a BN for predicting exposure? Was the BN the only model that was tested, or was it contrasted to other approaches? Once the full data set is created, model comparison is comparatively less time-consuming than data preparation. Since Bayesian approaches are often computationally demanding, a classical regression approach or simple machine learning model (e.g. random forest) might be worth trying. This would also allow to investigate more complex interactions between variables as well as non-linear effects.

R: There are several advantages of the class of BNs used here, compared with other

C3

approaches: (1) they are probabilistic, providing uncertainty bounds, (2) they can be applied when some of the input data is missing (which Random Forests can't), as e.g. the imperviousness dataset is not gap-free for Europe, which makes it more useful for applications, (3) the whole model can be presented graphically (in Random Forests, only singular trees out of the forest could be presented), (4) accuracy depends only on the configuration of nodes and arcs, and not on model parameters such as number of splits, leaves, trees etc. We applied Random Forests to our data (using Matlab functions 'TreeBagger' and 'predict'), with a 10-fold cross-validation on the basis of the 10% of the data – the same set-up as for the BN. The number of trees was set to 100, maximum number of leaves to 50, in-bag fraction to 1/3, and splits to two. The resulting R2 is slightly lower, RMSE higher, MBE strongly negative and only MAE is slightly lower than when using the BN. Overall, the performance is quite similar, though still worse. The comparative results were added to the supplement of this response.

C: The authors assume that there are no country-specific differences in H, apart from those that are implicitly modelled by including POP, IMD and B. The authors claim that they provide a 'universal method for estimating exposure of residential assets' (p.1, l.3f) across whole Europe. Since the method was only validated with data from Poland, Germany and the Netherlands, I am not sure if this statement is fully justified. Since the characteristics might be different in different countries, using a variable specifying geographical location (e.g. country or even broader geographical region) might be helpful to tackle unobserved heterogeneity.

R: The datasets for Poland, Germany and the Netherlands were the only independent datasets with information on floor space or number of floors for individual buildings that we were able to collect. If the reviewer knows other datasets like those for other countries, we would add them to the analysis. Regarding the heterogeneity, individual models for each country could be made, and they would improve performance at each, but this would still give only a good prediction for the capital city region of a given country, leaving uncertainty how well such a model would perform in other parts of the

country. We will move results for individual cities from the Supplement to the main text to make the variation in the model's quality more visible to the reader (see supplement to this response). We also note that we discovered and corrected an error related to the dataset German flood-affected households. In the submission, only those households affected by fluvial floods were included, and not those affected by pluvial floods as suggested by the years provided in Table 2. We have added those missing households, increasing the number of records from 2330 to 2868, though the validation results didn't change much.

C: I found the explanation for the empirical relationship given in Eq. (1) a little bit difficult to understand, since the numbers are scattered throughout the paragraph below the formula. I suggest to streamline this explanation.

R: We will clarify the explanation in the paragraph.

C: Also, I realized that within Eq. (1), B is used (1.) to derive H, and (2.) to compute F, which is based on H. I don't think that this is a problem, but I noticed that this puts quite a lot of weight on B.

R: The conditional correlation between B and H was considerable, therefore it had to be included in predicting H, and it is indispensable in calculating F, which strongly depends on both B and H.

C: I suggest to include a supplementary table to show which formula for deriving St was used for each country.

R: A supplementary table will be added (see table X in the supplement to this response).

C: Generally speaking, the coefficient of determination denotes the share of explained variance in the dependent variable that is predictable using independent variable. Note that $R^2 == r^2$ holds only in special cases such as simple linear regression if an intercept is included. While this is the case in the assessment of predicted vs observed

values presented in the paper, where the coefficient of determination equals the square of the correlation coefficient, the authors may want to clarify this.

R: The linear regression of revised Fig. 1 is approximately 1.02x − 0.3 (varies slightly within the 10-fold cross-validation).

C: Being a very common error metric, root mean squared error could be included as well, since it provides more information content with respect to outliers.

R: RMSE will be added to the results.

C: The first two sentences of Section 2.4.2 are unclear to me. The collective out-of-sample validation was done using an unseen 1% sample across all cities. How was the individual validation performed? By using stratified 1% samples of each city? The second sentence starting with 'Then' suggests that the procedure is different and that the samples are not the same. If the same stratified sample is used, validation results can be assessed both city-specific and at an aggregated European level.

R: We replaced the existing results with a 10-fold cross-validation (see attached revised tables).

C: An overall R2 of 0.36 is moderate, indeed. This means that only a third of the observed variance in building height can be explained using modelled building height (given that observed vs. predicted regression was used). The confusion matrix (Table 3) showing around 25% (and an increasingly lower amount as the number of floors increases) correctly classified outcomes for buildings with more than 2 floors is also slightly puzzling. Again, this might be a hint to try (1.) using more data and (2.) comparing different modelling approaches. Good results for average height are of rather limited explanatory power in terms of model quality assessment, since I would naturally assume that the differences in means are not too large when using any reasonable model. The problem of low variance might also be tackled by (1.) and (2.) mentioned in the previous sentence. That the model does not perform satisfactory at all for cities

like Nicosia and Reykjavik might indicate that there are country-specific differences. All cities that exhibit good performance are located in Central Europe (Vienna, Berlin, Amsterdam, Luxembourg, Warsaw, Zagreb).

R: We added more data (10% instead 2%), but this had only marginal effect on the data, as one might expect from a randomized sample of a very large dataset. Several cities have rather poor performance, but it is also partly due to variation in the quality of height data (which routinely shows errors of 1–3 m, according to the validation information contained in the dataset) and OSM buildings (which is particularly poor for Nicosia, for instance, using visual inspection). Some of the cities with good results are not located in Central, but Western and Northern Europe (Amsterdam, Berlin, Luxembourg, Stockholm, Vienna). We revisited the dataset using Random Forests, as described in a previous comment, which didn't lead to improvement. However, we tested the model for different urban-rural typologies using "Degrees of Urbanisation 2014" dataset from Eurostat. This allows classifying our data points according to whether they were located in "cities", "towns and suburbs" or "rural areas" (defined at the level of local administrative units - LAUs). The results presented in a table in the supplement to this response show that the R2 is lower for towns and rural areas than for cities, though this stems from the much lower variation in building height; MAE is lower in those cases, and in all three types of LAUs MAE has almost the exact same proportion to average height.

C: In the abstract, a validation with (1) buildings in Poland and (2) a sample of Dutch and German houses is mentioned. In the paper, (1) can be found in section 3.3, and (2) is described in the last paragraph of 3.1. I think the title of subsection 3.3 should be reworked, as 'Example application' is rather generic. Maybe a dedicated validation subsection for these new data sources could be helpful?

R: The example in section 3.3 is not based on the validation dataset for Poland. The Polish dataset is from a government-run national database (BDOT) and is used for validation in 3.1 as shown in Tables 2 and 3. The example in 3.3 uses an extraction of OSM data, as the BDOT dataset is not openly available in contrast to OSM. The

example is meant to present how the methods from the paper can be used in practice. We will highlight this better in the text.

C: In fact, there does seem to be a slight systematic bias in the results. Figure 2 shows overestimation for low building heights and underestimation of high building heights, with accurate results around 12 m. The regression line likely has a negative intercept and a slope larger than 1.

R: The linear regression of revised Fig. 1 is approximately 1.02x – 0.3 (varies slightly within the 10-fold cross-validation). We note that the model underestimates the height of tall buildings, but this is partly because few buildings are very tall.

C: The structure of the discussion is generally well thought through. However, the authors again solely focus on the BN model. Maybe the use of other models might lead to better results on the same data set? Limitations of the BN model itself and implications of using a comparatively small sample size (given available data) are not discussed.

R: We will add more information about the uncertainty related to the data analysis, and add results generated with Random Forests.

C: Figure 1: The histogram plots do not have any axis labels and units, which is a major limitation (in terms of information content) of this figure, since the histograms are essentially incomplete. Also - for the sake of consistency: the unit for population density is missing in the caption.

R: The graph and caption will be corrected.

C: Figure 2: Please use the same spacing for axis ticks (either steps of 5 or 10).

R: The ticks will be corrected.

C: Figure 3: I suggest to use points instead of bars. The information that needs to be transported is the value at the end of the bar, not the area of bar itself. There-

fore, information density is higher when using points. Also, the two colors of the bars are different (orange indicating building value in a and yellowish indicating household contents value in b), but the legend matches only the color in b.

R: The legend will be corrected and the graph will be reworked, so that points are used instead of bars.

C: Figures 4 & 5: I think it should be mentioned in the caption that values for each country are based on the respective capitals, since this is important when interpreting the results.

R: The values for each are not based on capitals, but on the economic statistics at national level (section 2.3).

C: Figure 7: Legend for a is missing, only legend for b is provided. Again, I suggest to consider using points instead of bars. If points overlap, you may slightly jitter them along the x axis or use some transparency.

R: The legend will be corrected and the graph will be reworked, so that points are used instead of bars.

C: Figure 8: It seems that this figure is not referenced in the text. If this is the case (I might have overlooked it), please add a reference in the text. Also, it reveals a substantial difference when compared to the JRC values, this could be explored/discussed further. Again, I suggest to consider using points instead of bars.

R: We incorrectly refer to Fig. 7 instead of Fig. 8 in line 410. The graph will be reworked, so that points are used instead of bars. The large difference between JRC and our estimates could be caused by the assumption of a single ratio between building and contents loss (which we show is far from uniform), transposition of this value from an American flood damage model, and possible differences in definition (JRC estimate including more items). We will describe this aspect better in the revision.

C: Figure 9: It seems that this figure is not referenced in the text. If this is the case (I

Please also note the supplement to this comment:
https://www.nat-hazards-earth-syst-sci-discuss.net/nhess-2019-313/nhess-2019-313-AC1-supplement.pdf

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess-2019-313, 2019.