Nat. Hazards Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/nhess-2019-313-RC1, 2019

Natural Hazards and Earth System Sciences
Discussions

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

Anonymous Referee #1

Received and published: 14 October 2019

This paper presents an effort to estimate exposure of residential assets using publicly available or open source data. The rationale behind the choice of data sets is sound and the effort in preparing these different data sets is appreciated. Overall, the undertaking is commendable, since this approach (and resulting estimates) might be of importance for risk and vulnerability analysis in a broad context.

While I think that the approach itself as well as the data sets used are very interesting and promising, I have some concerns and suggestions with respect to the methodology. In general, the methodology section should be reworked to contain more precise, in-depth information on how the authors solved the given task from a methodological point of view. While section 2 is quite long, the methodology is sometimes not very clear, and some parts are quite verbose. Also, I have the impression that the full po-

tential of the data set is not exploited. Results are not ultimately convincing given the mediocre model quality as shown in the validation.

Please find detailed comments on the respective sections and subsections below.

2.1 Identification of residential buildings

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?

2.2 Building size estimation

- 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.113ff), the variable selection process is not clearly described.
- 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.

- 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.
- 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.
- 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.
- 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.
- 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*.

2.3 Valuation of buildings and household contents

- I suggest to include a supplementary table to show which formula for deriving S_t was used for each country.

2.4 Validation

- 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.
- 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.
- 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.

3 Results

- An overall R^2 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).
- 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?
- 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.

4 Discussion

- 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.

Figures

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

- 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 differences when compared to the JRC values, this could be explored/discussed further. Again, I suggest to consider using points instead of bars.
- Figure 9: 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.

Tables

- Table 1: I am wondering why two different sources for 'Population per area' were used. If both are based on the 2011 census, why not using the one with higher resolution if the model is fitted at a building level?
- Table 3: '% of correctly predicted floors' is confusing. Only the diagonal values indicates the percentage of *correctly* predicted floors, all other number are simply the percentage of predicted floors?

Formal aspects

- p.7: L_mean should read L_{mean}, or simply $\bar{L}$.
- Please check consistency regarding capitalization (e.g.: 'Eq.' vs 'eq.'). NHESS manuscript preparation instructions suggest 'Eq.'.
- Please format the supplement according to the journal's standards.

Final remark

To conclude: Given the mediocre performance, I suggest (1) trying to use the full data set instead of the 2% sample only, and (2) taking a look at alternative modelling approaches for comparative purposes. If this does not improve the results, the input data might indeed be of limited use for predicting building height, but it does not leave the reader wondering the potential of the full data set has not been explored. Since the effort of gathering and preparing this interesting data set has already been undertaken, I suggest to have another look on the readily compiled data set. I am in full support of publishing these findings, but I propose reconsideration after major revision.

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