Review Report

Journal: Natural Hazards and Earth System Sciences

Reference code: nhess-2020-213

Title: A glimpse into the future of exposure and vulnerabilities in cities? Modelling of residential location choice of urban population with random forest

Authors: Sebastian Scheuer, Dagmar Haase, Annegret Haase, Manuel Wolff, Thilo Wellmann

Reviewer: Georgia Papacharalampous

Date agreed to review: 2020-08-15

Date review submitted: 2020-08-31

Recommendation: Minor revisions

<u>Summary</u>

The paper proposes residential choice modelling for the indirect assessment of disaster exposure, vulnerability and risk. The investigations are carried out for the city of Leipzig, Germany by building on the work by Scheuer et al. (2018). In this respect, random forests by Breiman (2001) are used to predict the probability *p* for a positive residential choice. The predictor variables include:

- Spatial housing attributes, i.e., location and neighbourhood amenities.
- Non-spatial housing attributes, i.e., size, rooms, rent, furnishing features and house type (including information on the apartment's condition).
- Household attributes, i.e., employment status, qualification, income and age.

The random forest model has been pre-trained in Scheuer et al. (2018) by using interactive interview data from the same city; therefore, the real-estate data used in the present paper are accordingly re-coded and geo-located to support the prediction process. A switch from the INSPIRE grid (i.e., a grid with cells of dimensions 1 000 m x 1 000 m) to the spatially homogeneous units (SHU) grid (i.e., a grid with cells of dimensions 250 m x 250 m) is also made. In the SHU grid, each cell is characterized by the following properties: (i) residential land-use; (ii) a predominant house type; and (iii) the presence (or absence) of each individual spatial housing attribute. The formed dataset encompasses information for the years 2008/2009, 2013/2014 and 2018/2019, and for

five socio-economic profiles (i.e., young adults in education, academic professionals, middle-aged workers, precarious unemployed persons and pensioners). Each of the latter is known to be characterized by a specific degree of vulnerability, which might also be different for flooding and heat stress.

The predicted likelihoods are summarized in the form of hot and cold spots by using local G* statistics (Ord and Getis 1995) through the R package spdep (Bivand and Wong 2018). This is made separately for each set {year, socio-economic profile}. The resulted hot and cold spots are presented in maps, which allow the inspection of the changes observed as the years pass (separately for each socio-economic profile). Overall, it is demonstrated that residential choice modelling can be informative in disaster risk assessment and management.

General comments

In general, I find that the paper is meaningful, interesting, and very well-formulated and -written. I have only a few minor comments that could be addressed for improving the presentation of the already conducted work.

Specific comments

(1) Since the paper uses random forests by Breiman (2001), this latter work should necessarily be cited, to my view.

(2) Moreover, some basic information on random forests (see e.g., the review paper by Tyralis et al. 2019) should be provided (e.g., in an Appendix). This could be made by emphasizing the appealing properties of the utilized variants for the application of interest (see again the review paper by Tyralis et al. 2019). More generally, I feel that it would be particularly relevant to answer key questions like the following ones: Why are random forests selected in Scheuer et al. (2018) and herein? Could they be replaced by other machine learning algorithms?

(3) It should also be noted that several references provided in Scheuer et al. (2018), such as Liaw and Wiener (2002), and Ishwaran et al. (2008, 2011), seem to be relevant in this paper as well. Currently, only the R package spdep is cited in the manuscript, while all the exploited R packages should be cited.

(4) A short summary (additionally to lines 110−113) of the experiments carried out by Scheuer et al. (2018) could built some extra confidence in the use of the pre-trained random forest model. This summary could again be given in an Appendix.

(5) Furthermore, basic information on selected machine learning concepts could be provided. This information could be particularly important, given the technical character of the manuscript. The reader could also be referred to several specialized books (e.g., Hastie et al. 2009; James et al. 2013; Witten et al. 2007), for further information.

(6) The abstract could be revised to better reflect the novelty of the work. For instance, it could start with lines similar to the following: "The most common approach to assessing natural hazard risk is by investigating the willingness to pay in the presence or absence of such risk. In this work, we propose a new (also indirect) approach to the problem, i.e., through residential choice modelling".

(7) Some hints on how the title should be perceived could also be provided in both the abstract and the introductory section. For instance, one could think that the paper is about forecasting (which is not the case).

(8) Finally, there are very few typos in the manuscript. For instance, in Figure 2(b) the right big box (including 16 cells in the INSPIRE grid and 256 cells in the SHU grid) is larger by four cells in the INSPIRE grid than the one marked in the middle sub-figure of Figure 2. Another example exists in Table 2, in which "pensioner" should be replaced with "pensioners".

References (not included in the manuscript)

[1] Breiman L (2001) Random Forests. Machine Learning 45(1):5–32. https://doi.org/10.1023/A:1010933404324

[2] Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: Data mining, inference and prediction, second edition. Springer, New York. https://doi.org/10.1007/978-0-387-84858-7

[3] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. The Annals of Applied Statistics 2(3):841–860. https://doi.org/10.1214/08-AOAS169

[4] Ishwaran H, Kogalur UB, Chen X, Minn AJ (2011) Random survival forests for high-dimensional data. Statistical Analysis and Data Mining 4(1):115–132. https://doi.org/10.1002/sam.10103

[5] James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer, New York. https://doi.org/10.1007/978-1-4614-7138-7

[6] Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 2(3):18–22

[7] Tyralis H, Papacharalampous G, Langousis A (2019) A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water 11(5):910. https://doi.org/10.3390/w11050910

[8] Witten IH, Frank E, Hall MA, Pal CJ (2017) Data Mining: Practical machine learning tools and techniques, fourth edition. Elsevier Inc. ISBN:978-0-12-804291-5