Reply to RC C2045: 'Comments on validation and technical corrections'

We would like to thank the referee for the time and effort put into reviewing the manuscript. We greatly appreciate the comments, suggestions and very constructive criticisms, with which we in general agree. With this response, we aim to address the issues raised in the review and propose changes to the manuscript accordingly. We believe that the implementation of such changes will considerably improve the overall quality of the revised version of the manuscript. Following the guidelines of the NHESS Editorial Board, the revised manuscript was not prepared at this point.

I. General and specific comments

I.1 Discussion of limitations of the methodology / Validation and accuracy assessment

General comments

The presented method for disaggregating census data for exposure datasets takes an innovative approach and has some potential with more and more building vector data being openly available. The method itself is outlined in a clear and understandable way, additionally supported by a flowchart visualizing the steps taken. However the manuscript is not convincing in the critical assessment of the model results, as well as the limitations of the presented method. The validation of the aggregation results is currently limited to a relative comparison of a not further described method.

(...)

Specific comments

Accuracy assessment:

Currently the performance of the presented method is evaluated by comparing the disaggregated floor area per storey for reinforced concrete buildings with the results of a disaggregation approach using the 2011 GEOSTAT dataset. In my opinion this approach has several weaknesses:

- *It does not quantify the improved accuracy over the population distribution method*
- *It does not validate the accuracy in regard of real world data*
- *It does not quantify the uncertainties of the presented method*

Therefore I would recommend taking data from field surveys or other real world data into account for validating the presented method. Validating disaggregation results with field data or other real world data is frequently treated in literature (e.g. Thieken et al. 2006, Stevens et al. 2015). According to your height classification method using field survey data, I would assume that there is already field data available.

Sample areas:

Regarding the sample areas, I would recommend to take a third grid cell from a more densely populated area of Pavia into account, or give more detailed reasons why you have decided to choose GC1 and GC2 as sample areas (see Fig. 7 in the manuscript). Currently one could speculate that the presented method only out-performs existing approaches in areas with low urbanization rates and therefore very uneven population distributions.

We fully agree that in the original version of the manuscript, the analysis of the limitations of the methodology was done in an incomplete way. First and foremost, we believe it is useful to synthetize what those limitations are, in order to allow addressing the comments as clearly as possible.

The assumptions present in the model are related with data limitations. These can be divided into two main groups:

1. Data that is, in a vast majority of cases, not possible to obtain at all:

   1.1. Data on some on the building variables of interest are only available at municipality level (e.g. Material, Year of Construction). To tackle this issue, the methodology assumes that the relative area distributions at grid cell level are equal to the one at municipality level (5055-L7f);

   1.2. The building occupancy types are not available on a building-by-building basis. Since risk models are generally occupancy-specific (in the example used in the manuscript, residential), the issue is tackled by incorporating additional data from other sources into the model (CORINE land cover maps together with an area threshold, 5052-L28ff).

2. Obtainable data, which is not exactly the required but has a direct relation with it (resulting in unavoidable assumptions when using building datasets; uncertainties are negligible):

   2.1. Building datasets contain information on building height, while census areas are divided by number of storeys. To tackle this issue, a correspondence must established between the two variables. In the case of this study, a sample of around 1000 buildings was used and the result is shown in Table 1 of the manuscript. The accuracy of the approach is very high, and very few buildings are wrongly classified in terms of number of storeys (5052-17f);

   2.2. Building datasets generally contain the footprints of the buildings, providing information on gross areas. Census data refers to internal net areas of dwellings (excluding walls and common areas). Given the practical impossibility of gathering this type of information on a building-by-building basis, it is assumed that the municipality ratio between the two can be applied at grid cell level. It is widely known that this relation is relatively similar among buildings of the same occupancy types; thus, the associated uncertainty is low (5053-26ff);

   2.3. Building datasets contain footprint areas, from which the gross area can be calculated by multiplying it by the number of storeys. However, we are interested in a specific occupancy type (residential in this case), and often residential buildings contain more than one (e.g. shops on the ground floor). The issue is tackled by assuming that the fraction of residential dwelling areas in relation to the total area of residential buildings is the same at municipality and grid cell level (5053-26ff).

The line of thinking behind the performance-evaluation in the original manuscript was that, since the building dataset is an accurate representation of reality in terms of building locations, footprints and heights (i.e. it is ‘real world’ data regarding these variables), the results of the ‘vector-based model’, shown in Figures 9 and 10 and Table 3, were the closest to reality we could hope to get. However, the assumptions listed in 1.1 (by using RC buildings) and 1.2 were present behind the model results, and they certainly are sources of uncertainty, to some extent. Therefore, we can say that in original manuscript, the proposed model’s areas were compared against the proxy-based ones, qualitatively showing its superior performance; however, a validation against observed data was missing, and the accuracies were not quantified.

When preparing the original manuscript, the observed data we had consisted in the number of floors for a sample of around 1000 buildings, collected within the scope of another previous work. This data was used as described in 2.1. No information on other variables was available, and certainly not for the entire municipality, which would be required for a complete validation. Following the reviewer’s suggestions and recognizing the need for measuring model accuracy in a more robust manner, we set out to collect it. However, for the vast majority countries/regions, the necessary data is not publically available at the required resolutions, and Pavia is not an exception. On the other hand, performing a building-by-building survey of the thousands of buildings in Pavia is simply not practicable. Thus, we decided to contact the department of urban planning and territorial management of the municipality of Pavia. They were kind enough to provide us with their own dataset of the city, containing occupancy types and year of construction classes on a building-by-building basis. Using this data, we were able to perform a more rigorous validation of the model. Accordingly, we have added this third observed dataset to the analysis. As suggested by the reviewer, an additional cell GC3, located in a more densely populated grid cell (in the historical centre of Pavia) was added. The results are shown in Figure 1. The total areas are shown in Table 1.

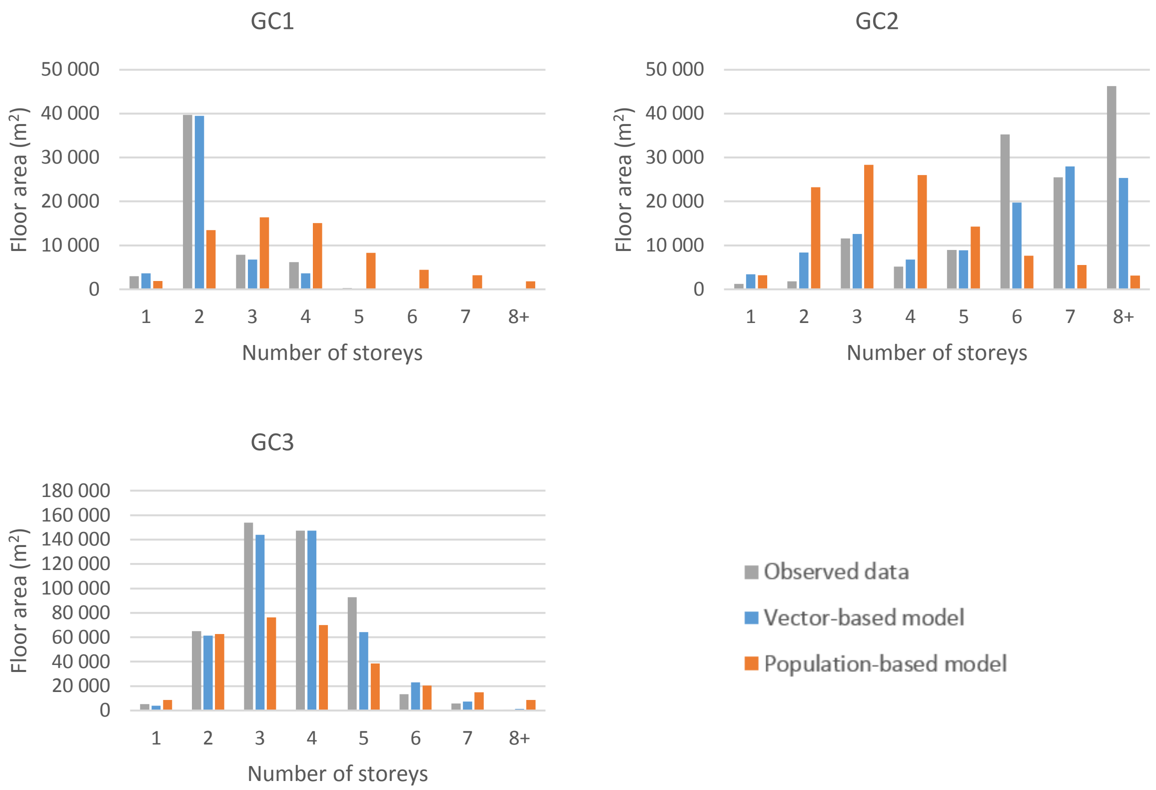

Figure 1. Residential building floor areas per height class, from the vector- and population- based exposure models, as well as the real world dataset, for three grid cells with different predominant building typologies. *[to replace Figures 9 and 10 in the original manuscript]*

Table 1. Total residential building floor areas for cells GC1, GC2 and GC3. *[to replace Table 3 in the original manuscript]*

	Total floor areas in grid cell (m^2)		
	GC1	GC2	GC3
Observed data	57 076.3	135 789.3	484 032.6
Vector-based model	53 573.1	113 125.8	463 185.9
Population-based model	64 544.2	111 469.2	299 838.3

In order to quantify the accuracy of the models, the root mean square deviations between each of them and the 'real world' dataset were computed (Table 2).

Table 2. Root Mean Square Deviations (RMSD) of the two models. *[new table]*

Height class	Population-based model		Vector-based model	
	RMSD (m^2)	NRMSD (%)	RMSD (m^2)	NRMSD (%)
1	1 213.1	16.30	1 031.8	13.92
2	9 930.0	15.28	6 531.8	10.13
3	12 472.7	8.10	4 244.4	2.91
4	13 289.1	9.03	5 020.3	3.47
5	9 349.6	10.06	7 389.5	7.99
6	8 283.8	11.73	3 136.0	4.48
7	5 202.2	13.96	1 177.1	3.19
8+	5 980.8	12.94	3 235.0	7.01
Overall	9 021.6	5.86	4 517.2	2.93

As expected, the model clearly outperforms the proxy-based model, when compared against not only aerial imagery, but now with real data as well (enabling the quantification of the accuracy, which will be included in the revised manuscript).

Regarding the population-based model, we agree with the reviewer's comment that it is not described in a clear manner. This proxy-based method consists in distributing the buildings among the grid cells proportionally to the population, by using the fraction of the population in each cell in relation to the population at municipality level. This is mentioned in 5051-L8f, but not very clearly, while in 5055-L25f it is not mentioned at all. The revised manuscript will be improved accordingly.

In the procedure mentioned above, the 'Age' variable is marginalized, as the age bands between the census and municipality datasets are different and cannot be directly associated. For this reason, the error introduced by assumption 1.1 needs to be measured separately. To do this, we compared, at grid cell level, the real areas with the ones that would be obtained by assuming the fraction distribution at municipality level. Since the dataset is compared against itself, the RMSDs computed in this way reflect assumption 1.1 individually. No data on the 'Material' variable is available, but it is reasonable to assume a similar degree of accuracy.

Table 3. Root Mean Square Deviations (RMSD) of the areas per grid cell compared with the areas obtained by assuming the fraction distribution at municipality level. *[new table]*

Age class	RMSD (m^2)	NRMSD (%)
1880	58 430.8	10.71
1913	8 638.4	13.52
1935	13 001.6	12.99
1963	15 177.9	10.35
1975	38 895.7	14.14
1986	15 714.9	24.19
2007	7 591.2	29.53
Overall	28 549.8	5.23

Finally, we would like to thank the reviewer for pointing out the two very interesting references. In these papers, two approaches for disaggregating population were used. The one in Thieken et al. (2006) is actually similar to the one in Thieken et al. (2008), which we mentioned in 5048-L11. The validation procedure consists in comparing the sum of the disaggregated population counts relative to grid cells inside administrative units, with population counts for those same units from other datasets, computing the respective error. In the case of our work, since we were able to obtain building-by-building observed data, we adopted a standard, cell-by-cell comparison between model and observations, computing the RMSD and NRMSD, as shown above.

The scope of the work by Thieken et al. (2006) is close to the one of the present work, and some very interesting considerations are made on the issue of resolution mismatch between datasets. We will be adding this reference to the revised version of the manuscript.

I.2. Time shift between data sets:

The four data sets used in this study (census data, building vector data, CORINE land cover data and GEOSTAT population dataset) are all from different time steps, spanning gaps of over 10 years. Although urban structures are not changing very dynamically, data sets with differences of 10 years and more can be an important source of uncertainty in model results. Therefore I would recommend to include this issue in the discussion, regarding limitations and uncertainties of the presented method.

We agree that a brief discussion on the time gap between datasets is missing in the original manuscript, since it will very rarely be possible to obtain datasets from the exact same year, making this an unavoidable procedure. However, the introduced uncertainty is not significant, especially when the other assumptions are taken into account.

First, by comparing the adopted 2003 building dataset with the up-to-date data from the municipality of Pavia, it is clear that very few buildings were constructed in this period, especially in relation to the total number of buildings in the city. As the reviewer pointed out, urban structures do not normally change very quickly, and this is very clear when looking at the Pavia datasets. Additionally, since the building datasets are used to distribute census data across grid cells rather than used directly, missing a few buildings has a negligible impact in the final exposure model. Even if some grid cells, where buildings were constructed more recently (i.e. postdating the building dataset), would be slightly more affected, overall the model would be largely insensitive to the issue.

On the other hand, since the model is mass-preserving and, in this case, the census data is from 2001, the building areas will be slightly underestimated in the exposure model, but this is known limitation with the input data itself. We know that if the data to be considered is from 2001, so will the final exposure model reflect that. A possible way to improve this could be, for example, to use population growth to estimate updated building areas, but we believe this is a consideration outside the scope of the manuscript.

II. Technical corrections

We would like to thank the reviewer for the specific technical corrections. They will be implemented in the revised version of the manuscript. Below, we have included replies to some of the comments, that either: 1) refer to issues that have not yet been addressed in part I; 2) refer to grammatical corrections that involve rewriting the sentences. Simpler comments are not included herein, as we fully agree with the suggested changes and have nothing to add.

5047-L27ff: The length and the structure of this sentence makes it very difficult to read.

We agree and propose the following: "Such an approach would thus introduce error in the exposure model and, consequently, in the results of the risk model. It is important to note that the impact on the latter also depends on the properties of the hazard itself. In fact, losses estimated for events with typically large, regularly shaped footprints, such as earthquakes, are less sensitive to the resolution of the exposure model than events with narrower and more irregular footprints, such as hailstorms or floods (Chen et al., 2004)."

5050-L1: "the aforementioned type of data": I would recommend to be more specific here, as you mention at least three different types of data before.

We agree that this choice of words made the sentence more confusing than it needed to be. We propose "open building data" instead.

5051-L9ff: Sentence is very long and difficult to read.

We agree and propose the following: "When using population as a proxy for disaggregation, the buildings can be distributed proportionally to the estimated population in each grid cell, but the relative frequency of the building classes at municipality level cannot accurately be changed at grid cell level without considering additional data."

5052-L1ff: Please add source and year of the data. I would also recommend to give more details about the building vector dataset.

Regarding the source, the reason why we had not included the URL here was that the address of the Geoportale Nazionale website had already been presented in 5049-L18. From that URL, accessing the data is straightforward. Having said this, we do agree that directly providing the address to the data on the WFS server can make things simpler for the reader. Therefore, the following will be added at the end of 5052-L3: http://wms.pcn.minambiente.it/ogc?map=/ms_ogc/wfs/Edifici.map

According to the metadata, the building dataset refers to 2003. This information will be included as well.

5052-L21: "large number of buildings": I would recommend to be more specific about the number of buildings, to maintain the traceability of results.

We agree: "a large number of buildings" will be replaced with "around 1000 buildings".

5053-L6f: Sentence structure unclear. Please rephrase.

We agree and propose the following: "In order to improve the accuracy in the estimation of residential building footprint areas from vector data without knowledge about the occupancy type of each building individually, the adopted approach consists in considering all the buildings in the dataset except those that fall within certain criteria in terms of area and location."

5071: I would recommend to add all information coming from the census data or at least the variables used in the study such as type and year of the building

With this figure, the idea was to represent the methodology conceptually, in a generic way, not only for the particular case of this study. For this reason, we suggest to keep "Other characteristics", adding "(in the case of this study, Material and Year of construction)".

References

Thieken, a. H., Müller, M., Kleist, L., Seifert, I., Borst, D., & Werner, U. (2006). Regionalisation of asset values for risk analyses. *Natural Hazards and Earth System Science*, *6*(2), 167–178. doi:10.5194/nhess-6-167-2006

Thieken, a. H., Olschewski, A., Kreibich, H., Kobsch, S., & Merz, B. (2008). Development and evaluation of FLEMOps – a new Flood Loss Estimation MOdel for the private sector. *Flood Recovery, Innovation and Response*, *I*, 315–324. doi:10.2495/FRIAR080301