Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-379-RC2, 2021

Earth System Science Data Discussions

Interactive comment on "Mid-19th-century building structure locations in Galicia and Austrian Silesia under the Habsburg Monarchy" *by* Dominik Kaim et al.

Anonymous Referee #2

Received and published: 17 January 2021

General comments:

This manuscript describes a vector dataset of reconstructed mid-19th-century building structure locations in former Galicia and Austrian Silesia covering an area of more than 80,000 km2 in present-day Czechia, Poland and Ukraine, derived from detailed Second Military Survey maps (at a scale of 1:28,800) that were built off of cadastral mapping (1:2,880) of the 19th century. The dataset includes two building categories, residential and outbuildings (mainly farming). The dataset is compared to census and cadastral data to evaluate local variations in differences between these and the extracted building data. The dataset is a useful resource that will be welcomed by researchers interested

in historical assessments of settlement, population and land use changes. The data represent the build structures in this regions at a very important point in time providing opportunities to better understand the evolution of the built environment and land use patterns over extended time periods. There are some concerns with this study and its design and the authors are encouraged to address them and add important detail and expand the scope of the research.

Specific comments:

There are three major issues. First, there is a significant lack of methods details. The authors dedicate no more than one sentence to the actual classification approach: "We used a semiautomatic, colour-based method involving the classification toolbar from ArcMap software." While the signature of the buildings might allow to use default tools to extract these symbols with high accuracy, the method underlying this ArcMap tool needs to be explained in detail. If there is not detail available it might not be the best idea to use a black box tool, to be frank. However, assuming, details can be found, the authors need to describe the underlying method/ type of classification run, parameters and any other aspects that might be relevant. The authors also need to ensure all details are included related to what they call "Data cleaning" in their workflow figure. Please make sure you include all the details necessary for any user to fully reproduce the methods and approaches and understand the choices made. Second, the validation of the classification results needs to be strengthened. It appears that the authors are validating the classification results for 1.3 Mio buildings using a sample of 1,500-1,600 objects. This is a 0.12% sample if this is all correctly understood. This represents a problem in terms of robustness and statistical power. This is true, especially as this validation is supposed to be valid across several dozens of map sheets that can be expected to have high levels of variation in their graphical properties and quality and thus, likely, the level of performance of the classification. The authors need to increase the sample size and based on underlying results from different map sheets show whether their validation statistics are representative and robust against underly-

ing variation of the map images. This will make this validation step more credible for the data user. Also, a relative error measure would be a valuable addition to better understand the nature and magnitude of existing errors. Third, the authors need to think about ways to integrate uncertainty-related information in the final data product, and provide respective metadata that users can refer to for any quality-related aspects. There is no description entry (metadata) provided with the shapefile posted online. Uncertainty details will improve the data usefulness and instruct users about the fitness of the data for the intended use. This could include summaries of deviation statistics between the created data and the information on the map frame or the census-based data. Releasing such uncertainty-related information will increase the usability of and confidence in the data. The authors are encouraged to be creative on how this kind of information could be provided. It could be included in additional map-level files or for different regions.

The existing variation in agreements between the building data and the map frame information as well as the census data are very interesting. The authors are encouraged to add more of this exploration into the analysis of underlying uncertainties as they might be able to pave the way for some interesting substantive research on historical aspects of mapping and settlement patterns in the 19th century. For example, variation in such agreements could illustrate the role of other ancillary variables such as topography, water, transportation and accessibility. Such aspects would make the analysis of local differences much more interesting and provide more detail that users of the data could refer to in their applications.

Finally, it would be a valuable addition in the concluding part to lay out more detailed potential applications of the data to illustrate possible directions where it could be useful and which research areas could benefit by exploring new questions. To enrich the study, the authors could even consider the calculation of settlement change estimates using respective contemporary building data (or data layers that offer similar enough data such as the GHSL or the GUF data).

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2020-379, 2020.