Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-217-RC1, 2020

Earth System Science Data Discussions

Interactive comment on "Fine-grained, spatio-temporal datasets measuring 200 years of land development in the United States" *by* Johannes H. Uhl et al.

Tracy Kugler (Referee)

takugler@umn.edu

Received and published: 1 October 2020

General Comments The data described in this manuscript are an extension of the existing Historical Settlement Data Compilation for the United States (HISDAC-US) data collection. The new data layers include 250m resolution gridded time series (every 5 years from 1890-2015) of the number of built-up records (BUPR), the number of distinct built-up locations (BUPL), and a binary layer indicating the presence of any built-up records (BUA) in the cell. The BUPR and BUPL layers represent new ways of compiling information from the underlying ZTRAX data, while the BUA layer appears to be a refinement/revision of a previously published BUA layer (cf. Leyk et al. 2020 cited

in this manuscript). All three new layers describe dimensions not previously available in the data collection, and I expect they will prove valuable in a variety of research applications investigating the trajectory of the built environment in the U.S.

The data and methods are described clearly and in sufficient detail. In addition, the authors clearly describe sources and implications of potential uncertainties in the data, as well as a thorough series of validation procedures. The data are easily accessible via the Harvard Dataverse, as described in the manuscript, and include useful meta-data. The layer files accompanying the data are particularly useful for visualizing the data.

Specific Comments I have a few questions about the underlying ZTRAX data that would further clarify the development of the data: 1) Lines 139-140 (p. 5) state that the ZTRAX database contains more than 400 million data records, out of which around 150 million contain spatial information. What do the remaining 250 million data records (without spatial information) represent? In other words, what is missing from the final data by not including those records? 2) More generally, what is the universe of the ZTRAX database? Specifically, what, if any, information does it include for structures that were present historically but not in 2016? The conclusion alludes to the "absence of information on building teardowns or replacements" (lines 461-2, p. 15), but I don't believe this absence of information is mentioned earlier. Is the absence complete, or are there some instances where information about non-contemporary structures is present? This should be clarified in the description of the source data in section 3.1. 3) Were lat/long coordinates present in the ZTRAX database for the 150 million records with spatial information? Or did the authors conduct geocoding based on addresses in the ZTRAX database? The manuscript seems to imply the former, but it would help if it were explicitly stated. If I am mis-reading and it is the latter, information about the overall quality of the geocoding should be provided. For example, what proportion of records were successfully geocoded to a address point or parcel feature in the geocoding reference data?

I also have a question/request regarding the accompanying uncertainty surfaces, specifically the no built year (NBY) layer. This binary layer flags grid cells without any built year information, which is important data quality information for users. Would it be possible to create a layer indicating the proportion of records in each grid cell that lack built year attributes? Such a layer would enable users to select alternative thresholds for the level of missingness appropriate to their analysis.

Finally, kudos on the quasi-spatial organization of the thumbnail images in figure 3 (p. 24). I find this organization makes the figure much easier to follow than a more "conventional" organization, such as an alphabetical ordering of the cities.

Technical Corrections In discussing the incomplete geographic coverage of the ZTRAX data, the manuscript contains a potentially confusing parenthetical, "(i.e., RUC codes 4 to 9, inhabited by only 15% of the U.S. population in 2010)" (line 228, p. 8). I believe this means that 15% of the total U.S. population lives in all counties with RUC codes 4 to 9, not that 15% of the U.S. population lives in the 82 counties missing from the ZTRAX data, correct?. It would be more helpful to know how much of the U.S. population lives in those specific 82 counties.

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2020-217, 2020.