Reviewer’s report

Title: Population allocation at the housing unit level: Estimates around underground natural gas storage wells in PA, OH, NY, WV, MI, and CA

Version: 0 Date: 05 Apr 2019

Reviewer: Mathew Hauer

Reviewer’s report:

This article presents an interesting method to allocate sub-county populations based on residential housing units. The authors apply their method to an example of proximity to underground natural gas wells.

The method seems appropriate and the results seem interesting, however I do have some questions and concerns.

Major:

1) The use of address data from OpenAddresses.io and the NAD could be problematic, I'm not sure, and I would like a better description of the data. Not everyone has a mail delivery address and rural routes are extremely common. I'm assuming that underground gas wells are more likely to be located in rural areas than in suburban or urban areas, making the use of addresses alone potentially problematic. I say potentially because a) I don't know how many of the wells in the study area are in rural areas, b) I'm unsure of the accuracy of the OpenAddresses.io and the NAD data. The only mention of "rural" appears on line 125 and simply describes the wells located in "predominately suburban-rural residential areas." I would like to see a better description of this data -- including the advantages, disadvantages, and general coverage of housing units (as opposed to addresses) -- and a better description of the locations of the wells (how many area in urban, suburban, and rural areas?). We know the Census data has virtually 100% coverage but I don't know how well the address data covers rural areas.

2) The authors compare the PPA versus ABODE "accuracy" of the data but there is a temporal disconnect between the two datasets. The 2010 Census is a count of HU’s in year 2010, but the address data seems to be a "living" database and likely most representative of 2017 or even more recent. This can cause a mismatch in the # of people living within 200m of a well. The authors state on lines 296-297 that the "overestimates were due to inclusion of non-residential domiciles (i.e., empty parcels), rather than new RHU development following the 2010 census..." I *strongly* doubt this assertion and am curious as to how this assumption impacted the results of the authors analysis. 7 years is a long time for potential development.
3) On lines 248 the authors incorporate person per household or PPH into their analysis to convert housing units to population. As one moves from coarser scales to more fine scales, group quarters populations (those in barracks, college dorms, nursing homes, prisons, etc.) become increasingly important. The authors pay only minimal attention to this "special" population in line 254 when they write "For j areas that intersect with census group quarters populations, a similar location allocation and term summation can be added to equation (2)." This needs to be included more explicitly and can be accomplished very easily. The total population of a given area is the sum of the residential population and the group quarters population \((P = \text{Residential} + \text{GQ})\). The authors are calculating PPH by dividing the total population by the total housing units \((\text{PPH} = P / H)\). By using this definition of PPH, the authors will erroneously include group quarters populations inside their estimates of the residential population. They can easily correct for this by simply subtracting the GQ population from the Total Population and then calculating PPH based only on the residential population. As the authors have presented it, simply saying that such a calculation "can" (line 255) be added is insufficient. The authors need to be much more explicit. This seems trivial, but I can assure you, GQ at small geographies is VITALLY important and for some blocks and even block groups the GQ population could approach the entire enumerated population.

3b) I do not believe the Census "provides" an estimate of PPH but rather the authors are deriving the estimate.

4) The authors report most of their results using r-squared, which might tell us a degree of similarity between the estimates but is quite poor. Just looking at Figure 1 shows r-squared's are a poor metric (It would be nice if the individual facets were given letters or numbers to reference them). The Ohio facet on the bottom row for instance, has an r-squared of 0.6 but MANY of the estimates look to represent considerable deviation from the expected value \((y=x)\). I'd like the authors to quantify the differences using a better metric, something similar to a MAPE or MALPE, which is considerably more intuitive. Absolute Percent Errors (or differences in this case) can be subject of extreme value manipulation, but quantiles solve most of those problems.

Minor:

1) There were a number of typos throughout the manuscript.

2) The authors seem to use "nighttime population" and "residential population" interchangeably. I suggest, for readability, to pick one term and use that.

3) The sentence starting on line 109 is not entirely true. CIESIN's GPW data, Facebook's gridded Pop data, and WorldPop's gridded pop data all use various forms of daysmetric mapping in combination with other information -- including residential housing location data -- to produce small area population estimates. It would be interesting (thought not necessary for this paper) to compare the results of the authors here with the other freely available gridded datasets.
4) It doesn't look like it is "OpenAddress.io" but rather "OpenAddresses.io." The former led to a 404.

**Level of interest**
Please indicate how interesting you found the manuscript:

An article whose findings are important to those with closely related research interests

**Quality of written English**
Please indicate the quality of language in the manuscript:

Needs some language corrections before being published

**Declaration of competing interests**
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license ([http://creativecommons.org/licenses/by/4.0/](http://creativecommons.org/licenses/by/4.0/)). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.
I agree to the open peer review policy of the journal