Reviewer’s report

Title: Examining the validity and utility of two secondary sources of food environment data against street audits in England

Version: 0 Date: 24 Aug 2017

Reviewer: Madeleine Daeppe

Reviewer's report:

This paper assesses the validity of retail food outlet listings obtained from commercial and local authority datasets in comparison with street audits in four English Local Authorities. The research has a number of strengths; in particular, it fills a critical gap in the UK data validation literature, which (to my knowledge) lacks research comparing POI data with gold standard ground-truthed data. Given the large number of studies that rely on POI data, an assessment of data quality has important implications for the validity of prior findings and the methods appropriate for future research. Furthermore, the results allow for a comparison of the relative validity of Food Standards Agency versus Points of Interest data, which will help researchers decide which dataset to use in future food environments research.

The study has one major weakness, which is the purposive nature of the geographic sample. First, it seems the authors first selected four audit areas. The authors argue that these LAs were selected based on different ranges of urbanicity and affluence within each place, but because the methods do not seem to account for a multilevel sampling approach, the administrative nesting may confound the authors' results regarding urbanicity and deprivation. Furthermore, these LAs are geographically clustered and thus results may not be generalizable to other parts of England. This first sampling issue has to some extent been addressed by in the limitations section. However, the authors go on to select Lower Super Output Areas (LSOAs) purposively. The authors first stratify the sample by urbanicity and deprivation (but not by LA?) but then to select LSOAs based on (1) ease of access for the sampling team and (2) the number of expected outlets in the POI data. It is this second choice that introduces bias: POI data will be inflated in places with more false positives, so the authors may have artificially increased their estimate of the false positives in the data. Similarly, this choice as well as the choice to exclude LSOAs with fewer than 100 expected outlets could artificially reduce the number of false negatives in the results.

The construction of basic validity measures (PPV and sensitivity) is standard in this literature, but it is not clear to me what these scores actually mean. Suppose, for example, that the POI has 5 false positives but also 5 false negatives - you could imagine a food environment measure (e.g. counts of stores) that would be robust to the inaccuracies that authors have identified here. I would recommend that the authors also construct a measure of "representativity" (Clary & Kestens, 2013) or assess the correlations between audit-constructed and POI- or FSA-constructed
measures of the food environment they hope to use in future research (see e.g. Daeppe & Black, 2017; Lebel et al., 2017).

Specific changes the authors should make include:

1. Page 5, line 9: Many studies have relied on city council registers (Cummins & Macintyre, 2009; Lake, Burgoine, Greenhalgh, Stamp, & Tyrrell, 2010), so the authors should either cite a systematic review showing that POI and food hygiene data are the "most commonly used data sources" or delete the word "most".

2. On page 5, line 16, the motivation for validating FSA data remains somewhat weak. It seems that previous studies only validated data from specific LAs, whereas this study will validate a new data set of aggregated LA data. Even given prior research with FSA data, quality can vary considerably from place to place; the researchers could thus comment on the importance of local context in data validation research.

3. On page 5, the authors could also strengthen their case for the importance of validating POI data by commenting on the benefits of a study that allows for the comparison of both commercial and municipal datasets.

4. Page 5: While the introduction clearly builds to the proposal of research objective (i), it remains unclear why objectives (ii) and (iii) are warranted. The authors should add a paragraph explaining the importance of and introducing prior research on differences in validity by environment or outlet types.

5. Page 6: How many LAs are there in England in total?

6. Page 7 Line 5: What proportion of rural LSOAs are excluded by the authors' requirement that 100+ FOs were expected within each environment type? What is the reasoning for this cutoff?

7. Page 8 line 12: Was it not possible to download FSA data for the same time as POI data i.e. spring or summer 2016? Nevertheless, I am impressed by how closely the authors were able to temporally match the three datasets.

8. Page 10: Figure 2 should be moved to the supplementary material.
9. Page 13 line 2: I believe it's the version of R that is relevant here, not the version of R Studio.

10. Page 13 line 3: The authors generally present results for strict rather than relaxed criteria, but as noted in the discussion (page 18 lines 23-26), the relaxed criteria are actually more relevant to most researchers. Thus, the paper might be more relevant to other researchers if the authors reported the "relaxed" criteria as the default.

11. Page 15 Table 4: Because the outlets identified in the POI and FSA data may be different from those identified in the audits, I am not convinced that that a percentages column is appropriate here. Perhaps show correlations of the counts across the LSOAs (as well as N's for the LSOAs?)

12. Page 16 Table 5: This table is quite confusing. What is being compared with what? The authors should separate the urbanicity results from those reported for store classifications, include a clearer note on the reference level chosen and explain how significance was tested for rural and urban overall versus for rural and urban stratified by deprivation. The authors noted in their methods that Bonferroni corrections were used for multiple comparisons, so why is significance by outlet type still reported at the p = 0.05 level?

13. Page 17 lines 2-4: Why did the authors examine classification agreement only for POI but not for FSA data?

14. Page 20 line 11: what makes a dataset "sufficiently" valid?

15. Page 21 line 7: might this be a policy implication for the FSA data? Perhaps this difference supports a need for LAs to adopt a better classification scheme.

16. Page 21 last paragraph: this conclusion seems too broad because the methods did not actually examine geocoding accuracy e.g. the authors did not look to see how many outlets were coded as FNs because they were geolocated outside the LSOA and how many were FPs because they were inaccurately geolocated inside the LSOA.

17. Page 22 line 12: what is meant by "utility of the data"?
18. Page 23 line 16-17: One can imagine some scenarios in which a policymaker would need to know about specific outlets (e.g. if a local authority bans convenience stores within some distance of schools). While the results of this paper suggest adequate quality for research related to food environments, it's not clear that the results are strong enough to support complete confidence for policy tools using FSA or POI data. The authors have also not mentioned the cost of POI data - are these data something a standard local government with a limited budget would be able to access? Would paying for access to POI data make sense if the government already had FSA data at no extra cost?

19. Page 23 final paragraph: The authors could strengthen the conclusion by deleting the last two sentences. It does not make sense to end on a point about the role of methods in data quality, because that wasn't the main point of the paper. If the authors wish to retain a comment on the importance of reporting methods, I would recommend including a reference to Jones et al. (2017).

20. Anything the authors could do to reduce the number of abbreviations would improve readability.

References


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:

Acceptable

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/). 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