August 30, 2022

Responses to specific review comments about Alexander and Genadek's "Using Administrative Records to Support the Linkage of Census Data"

Editor's comments

As well as addressing reviewers' comments, please explain what this work adds to the knowledge base e.g. is the planned strategy informative to others? It needs to be badged and structured it more clearly as a protocol paper, with anticipated challenges and benefits, etc.
We have fundamentally restructured the paper as a protocol piece, including revamping the organization and content.

Also, is there an indicative timeline? The results given as are of earlier studies – these need to be moved as per reviewer's comments.
We have moved the results section and added a timeline for the work described.

Are any of the datasets non-routine? If not, although the manuscript would not fit in the special issue, it could go in a standard issue, providing it's accepted. I hope this is helpful.
The census data itself and the Panel Study of Income Dynamics are non-routine data.

Reviewer A's comments

What is the most important contribution of this manuscript to the scientific literature? I think a paper describing the process after the process is done could be more relevant. Then, specific estimates about linkage rates, etc. could be included (and not only expectations as they are now). For example, specific subgroups for which linkage might be more difficult are not described, but a proper analysis could be performed after the linkage has taken place.
Thank you for raising this issues. We restructured the paper, and especially the introduction, in order to highlight the contribution of this work. We describe our novel plan for using administrative records and survey data to assign linkages and assess the quality of those linkages. Those are the paper's primary contributions.

I think a final paragraph describing the potential of such dataset could be useful. What type of research questions could this data answer? Or how could it be used?
That you for the suggestion. We have added more information on how the linked data has already been used and will be able to be used when the linkages are complete in the introduction of the paper.

In page 2, in the middle of 2nd paragraph, there is a sentence starting with "Our approach". Please check the sentence, as I think it is not correct.
We corrected this.

Reviewer B's comments

No "non-routine" database is actually considered in this manuscript.
The decennial census is non-routine data, as is the Panel Study of Income Dynamics.

It is not clear to me whether this is a protocol paper for project, or a description of an aspirational project.
We are sorry this was not clear. To address this, we have restructured the paper in an attempt to make it clear that this is a protocol paper for an ongoing project.

There is no research question or a goal for achievement of a degree to linkage that would show that this method is indeed better than those used before and overcomes the challenges described very briefly in the introduction. Even title is misleading, as the paper does not describe how the administrative records *are* used but how they *may* be used in the future.
You are correct here, we are not showing that the methods described are improving linkage. We have changed our title and the structure of the paper so that it should now be clear that this is a protocol paper describing the planned use of data to support the linkage of census records.

The paper structure is somewhat confusing as elements are somewhat arbitrarily placed in sections - especially baffling is inclusin of Results, considering that this appears to be a future project so there are no results as yet?
Thank you for noting this. We have reorganized the content and sections of the paper to address this.

The authors refer to success linking, reaching "even greater proportion" (p9), but do not specify what are their expectations in terms of what these proportions will be, and how will they judge that their method is a success. How low is it currently, and what this new method will bring, and will it be cost-effective in terms of labour involved and data potential?
We have expanded the section on estimating quality and error to describe how we will judge the success of our methods.

It would also be helpful to explain whether the Census is (and has been) mandatory in the US, and what is the % of population that completes it; and who might be missing.
The U.S. Census is a mandatory survey with estimated coverage rates ranging from 93.5% in 1880 to 98.6% in 1980. See Miriam King and Diana Magnusson, "Perspectives on Historical U.S. Census Undercounts", Social Science History 1995 (19:4) 455-466.

The first paragraph of Results really belongs in the Introduction as it sets the study goal – with judicious use of the existing references (as authors say "work that we and others have done" but there are no accompanying references). And many following paragraphs belong to Methods – see below.

We no longer have a results section, and we have now updated the paper to include the appropriate references.

Introduction, 1st paragraph – is this issue (low rates of coverages due to changes in names etc.) specifically a challenge in the US Census data, or is it a pervasive issue in other countries as well? The Introduction could be expanded to fuller illustrate the magnitude of the problem or how it was dealt with elsewhere.
While record linkage experts in many countries face similar problems, the exact circumstances differ enough by time and place that an overview of these challenges is beyond the scope of this protocol paper.

P2 “Our approach will be to creating linkages between censuses will be to...” please revise “Our work builds on methods…” should this be supported by a reference if those methods have already been established?
Thank you for the suggestions, we revised this paragraph to be grammatically clear and include citations.

P3 Last introduction paragraph – was there no research or at least a validation question?
Correct – we hope the paper restructure helps with this and it is not surprising when you get to the end of the introduction.

P4 first paragraph – I am a linked datasets’ user rather than creator, but the method described here seems to me a very standard one not necessarily novel?
We hope the flow of the paper now addresses this issue. We describe the background – not necessarily novel – linkage techniques, and how we build on these.

Second paragraph – could the authors describe briefly the records they mention (such as SSA Master Beneficiary Record etc.) for international audiences? Are SSA essentially tax data? I can see that they are described later – maybe add section titles so that this information could be easily found
We have added text to our methods section to describe these administrative records for an international audience.

“Though the SSN links are expected to be high quality…” why?
We have added more descriptive information on the routine data that we use, including explaining the importance of an SSN in the U.S. context.

P7 “The resulting file…” if the file will have one row per person, how will family members be cross-referenced?
That line now reads: “The resulting file will contain one row for each person ever observed in the administrative records, with all available information about that person and any observed family members included as well.” We hop it is now clear that we place at append family member information on the individual record.

P8 …" even though it will only be available for some cases" – should this protocol not include a detailed record of matches using each type of variables, especially for those that will only be available for some cases – should they not be used to evaluate the effectiveness of the procedure?
We have now linked a sentence following this line, "Since this variable provides an additional linkage key, we will expect slightly more accuracy in the linkage of long form census records to the composite." We do not only plan to use these for evaluation, because we want to make the best and most links possible for researchers. We hope this is clear now.

"we would have success linking…" Please define criteria for "success" of the linkage using this method – is there a percentage that you are striving for?
We are sorry for the confusion here, we are just stating that one could directly link the censuses together, so we have changed this sentence. There is not a percentage we are striving for, but rather we are looking for the most complete and accurate links across the period. We believe this is now clear in the text.

P8 – does it follow that the success rates will be better for those US citizens who completed the long form census?
Yes. We have made this more clear in the paper.

P9 "…used to produce research in top-tier journals across the social sciences" could you specify what kind of research exactly? Highlighting a few specific topics would certainly increase the reader's engagement.
We have added information about how these data have been used.

"For a large majority of the population…" such as? Do you have any evidence what particularly makes this "large majority" different and why it is so warranted to seek this additional "even greater proportion" match?
Thank you for bringing this up. Yes, we have now included specific information on who is linked and why it is important to try to expand linkage to the harder to reach populations.

P9/10 How many such additional passes the authors expect to make? Why would it be efficient?
We describe two approaches to improving the coverage of the linked data. It is not that this is efficient, but rather these are two approaches we think will be possible and useful.

P11/12 Pilot studies are described in Results, even though they appear to have been already published – if that is the case then they should have been described in the Introduction not Results. As far as I can gather the activities are to take place in the future, so there really aren't any results per se?
We have removed the results section and moved some of this information to the introduction and the methods section.

P12 – I finally found the information on people who are likely to be missed using the standard methods… (2nd paragraph) again this is something that should be up front as a rationale for the work. Will they not be also missed in this more detailed approach though?
Thank you for the suggestion. We have restructured our presentation of this information – and this moved up in the paper some.

P13 2nd paragraph – also finally there is some information on linkage quality – I would have expected that in the Methods – and at least mention at the end of Introduction. These paragraphs describe methodology yet are placed in results…
We now include this in the methods section.

P16 “when the digitization is complete…” so there really is no concrete timeline for this work?
Thank you for this suggestion. This work is ongoing and there is a timeline for it. We have now included it in the introduction.