Response to Referee Comments

Editor Comments

Thank you very much for submitting manuscript ID JSSAM-2022-055 entitled "Bayesian data integration for small area estimation of pathogen prevalence dynamics from pooled and individual data" to the Journal of Survey Statistics and Methodology. The submission has been reviewed by four reviews, an associate editor, and me. Comments are enclosed/attached. The reviewers found points of interest in your paper, but also requested major revisions.

As I mentioned in an earlier communication, the recommendation from the guest editors of the upcoming Special Issue on Data Integration was to submit the manuscript through the regular peer review process as it was not a good fit for the intended issue. When I reassigned the paper, I strongly recommended that the associate editor consider the paper for the Applications sections, not the Survey Statistics section. This "request" is reflected in comments 5 and 6 of the associate editor's report, and I strongly encourage you to consider comment 6 in the revision. Along the same lines, I recommend revising the title to emphasize the case study application.

I invite you to respond to the reviewer(s)' comments and revise your manuscript. Eventual publication will depend on successfully addressing the reviewer criticisms. Please keep in mind that the revised manuscript should adhere to the 6,500 word count limit; the online submission system will not accept manuscripts over 6,900 words, excluding figures and tables. There is no word count limit for supplementary online material, and I encourage you to take advantage of that.

If you choose to submit a revision, please read the attached, Submitting a Revision, to guide you on the revision process. Please be sure to include a point-by-point response to all of the comments. Then you may follow the link to submit your paper.

We would like to have your revision within 90 days. If you have not submitted a revision within 80 days you will receive a reminder note, saying that you need to request a time extension if you do not submit your revision in the next two weeks. If you do not request and receive an extension it is likely that your paper will be treated as a new submission.

I look forward to receiving your revision, and please continue to consider Journal of Survey Statistics and Methodology as an outlet for your research.

Sincerely,

Katherine Thompson

Editor in Chief, Journal of Survey Statistics and Methodology

katherine.j.thompson@census.gov

Response: *Thank you for the transferring our manuscript to a standard submission. We have revised the manuscript to align with the formatting for an application type submission.*

Associate Editor Comments to the Author

The AE agrees with all four reviewers that the paper is well written and easy to follow. The proposed methods are clearly presented for the most part. The details of the simulation studies and case studies are presented and the author has provided detailed explanations of the results.

The AE also agrees with many of the concerns raised by the reviewers. The main ones are summarized below.

1. Model expression notations need improvement, e.g., consider including an individual level model in Section 3.2, check and edit expressions in Section 3.3 (specifically, “iid” or “ind” and conditional distribution).

response: *We have clarified the distributional assumptions from section 3.3. Additionally the individual level model is included after Section 3.2 in Section 3.3.1.*

2. Model assumptions need clarification, e.g., how to include covariates in Section 3.3.2, whether it makes sense to use time-specific covariates and regression coefficients in Equation (7), whether it makes sense to adopt a prior distribution for the standard deviation parameter in Equation (8), the reason to use a Gaussian process model with a covariance structure that effectively treats time as being continuous over an autoregressive model that would treat time as being discrete.

response: *Model assumptions have been clarified - see specific responses to reviewers below. We added additional details to the new Section 3.3.1 that details how to include covariates in these models; a description of time-varying coefficients to Equation (8), which is now Equation 8; a prior distribution for σ^2_μ in Equation (8), now Equation (9); a discussion of using a GP model for continuous time.*

3. Results discussion need clarification, e.g., the reason why Table 1 last column having notably wider intervals, discrepancy between text description and Figure 2.

response: *We have added additional clarification about Table 1 - namely that the last column corresponds to a method using a much smaller subset of the data. We've also adusted the axis in Figure 2.*

4. Model fits need more detail, e.g., consider including Stan fits detail such warmup, number of chains etc.

response: *Additional details about model implementation in Stan are added to Section 3.6. “For all of the analyses, four Markov chains were used. Generally, 2000 - 3000 iterations with 1000 - 1500 iterations for warmup were sufficient for parameters to converge. The analyses were conducted on a 3.1 GHz Dual-Core Intel Core i7 took less than about 20 minutes when running 4 chains in parallel.”*

5. The novelty of the work needs clarification. Since this manuscript is targeting JSSAM Applications, it might be worthwhile highlighting the novelty aspect of applying multiple existing methods in particular settings in the Introduction and also be careful using words when describing whether the methods are novel or not.

response: *While reworking the manuscript to highlight our datasets, we have been careful to describe whether methods are novel - they generally are not - or if they have been combined in novel ways to solve a specific problem associated with our applications.*

6. The structure of the manuscript can be improved. Since this manuscript is targeting JSSAM Applications, it might be a good idea to follow one reviewer's suggestion to place one case study front and center (specifically the Notre Dame COVID-19 data analysis) to motivate the manuscript while the simulation studies and other case studies can be used to compare / contrast how the pooled methods work under different types of assumptions. The AE welcomes the suggested restructuring and / or reasons of not making such a restructuring from the authors.

response: *We have restructured the manuscript to lead with our applications. Rather than just the Notre Dame COVID-19 data analysis, we opted to use both datasets to motivate the work. The bat study has obvious groups, species of bats, where the prevalence may vary and highlights an additional layer in our modeling framework.*

7. The use of “small area estimation” needs clarification.

response: *We have de-emphasized the "small area estimation" language in the article - including removing it from the title. We also added a sentence to the fifth paragraph in the introduction that explicitly defines how our approach compares to traditional small area estimation.*

8. The authors should carefully proofread the articles as many grammar mistakes and typos have been found.

response: *We have fixed all identified typos and given our final manuscript a close read to catch others.*

The AE has found the following typos / grammar mistakes in addition to the ones commented by the reviewers. The list is not supposed to be exhaustive.

1. Page 1 line 47, run-on sentence.
2. Page 1 line 54, consider adding a reference for group testing / pooled testing.
3. Page 1 line 57, what is "a perfect test"?
4. Page 2 line 10, should be "further reducing. . . ".
5. Page 2 line 15, "a prediction" or "the prediction".
6. Page 2 line 18, should be "the correlation".
7. Page 2 line 38, "and" before "2)" should be deleted.

response: *All typos*

Reviewer 1

Comments to the Author: This paper addresses pooling and jointly testing multiple samples to reduce testing costs, and data integration modeling to make inference about disease prevalence in subpopulations. The methods rely on non-parametric hierarchical Bayes inference and are illustrated using various applications. Minor comments are provided below.

1. In general,

   a) The manuscript seems rich in subject-specific language. This makes it less clear to follow by the statisticians readers.

   response: *While reshaping the manuscript to be application-motivated than methodological-motivated, we also attempted to avoid, or clearly define, any subject-specific language. In particular, we removed the second paragraph and de-emphasized language about zoonotic spillover.*

   b) It is not clear what assumptions are made for the survey design. For example, if it isn't simple random sampling, should any cautions be stated or the model specifications be expanded to include it (the survey design)? **response:** *The pools are selected with simple random sampling, comment added to the first paragraph of page 4. "Pools are constructed using random sampling without replacement."*

2. Page 5, equation (3). It would make sense to define i here, as seems to be used as both an index for time points and an index for samples.

response: *Reviewer 4 had a similar point. We added a more explicit definition of i, and t_i, to page 4 line 6. "For example, if 17 individuals were originally sampled at time t_i,* ***corresponding to the i^{th} time point in the study,***

3. Page 5, lines 56-57. It is not clear how the model specification would extend to include covariates, when available.

response: *We restructured the model specification section, based on suggestion 1 from reviewer 3, to first include a portion on individual tests. We also added a comment here about using a link function and covariates in the model. "Covariates can be included in this model framework, or those that follow, by using a link function where $g(p_{t_i}) = x_{t_i}\beta$."*

4. Page 6, lines 33-34. What do you mean by 80,000 data; are these points/individuals?

response: *We have replaced data with "individual tests.*

5. Page 6, equation (7). Would it make sense to use time-specific covariates and regression coefficients leading to a μ indexed by t_i, too?

response: *Yes, we added a sentence after Eq 7 stating that "Furthermore, time-varying coefficients could also be used where $\mu_{t_i} = X_{t_i}\beta$."*

6. Page 7, equation (8). Would one adopt a prior distribution for the standard deviation parameter σ_μ, too?

response: *Yes, good catch. We have added a prior distribution to Equation 8 for this parameter.*

7. Page 7, line 52. What does 'm=1 is informed by k x m individual data' mean?

response: *The text has been updated to be more specific, stating "the parameter estimates and predictions from model $m = 1$ are informed by all $k \times m$ individual data".*

8. Page 8, Table 1. Is there any insight as to why the intervals in the last column are notably wider than the intervals in the other columns? Also, any reason for the large bias in the estimate of l from the second study and m=1* scenario? The general observations in both of these questions are applicable to the results in Table 2 and Table 3, too – it seems that estimating l under the m=1* scenario isn't quite precise.

response: *The last column in these scenarios corresponds to the m=1* scenario where substantially less data is used for the analysis; hence the larger intervals. Text is included in the paragraph below Table 1 to describe why the m=1* scenario has wider intervals: The budgeted individual curve ($m = 1^*$) is both less precise and less accurate in general, "largely due to much less data used for the estimates."*

9. Check for typos. For example, see line 24 on page 3; lines 35-36 on page 7; last column in Table 1; lines 51-52 on page 10.

response: *Thanks for pointing out these typos. They have been fixed.*

Reviewer 2

Comments to the Author: Overall this is a very thorough study with convictive case study and simulation study with nice plots. However I believe it will be much more helpful if the variables and corresponding subscriptions could be specified within the model.

response: *Based on this, and specific suggestions from other reviewers, variables and subscripts have been clarified.*

Reviewer 3

Comments to the Author:

The paper describes an approach for estimating the prevalence of a disease using pooled data and analyzes data related to bats in the Congo Basin and COVID testing. While I have some minor quibbles with how the material was presented (as I'll describe in this review), overall the paper is well-written and was relatively easy to read and understand. On a brief side note before I get to my real comments, I'd like to say that I found the use of the phrase "small area estimation" — particularly in the title — to be a bit misleading in the sense that the data don't appear to come from "small areas". I understand that the methods used may be motivated by the small area estimation literature, but I think most readers will assume "small area estimation" will mean that the data come from a collection of small geographic areas rather than data collected from (or being treated as coming from) a single location.

response: *We de-emphasized the "small area estimation" language from the manuscript by removing this phrase from the title and deleting most occurrences. We did retain an paragraph in the introduction about small area estimation, but added a sentence clarifying that our situation is not concerned with "small geographic areas." Specifically, we now state "While small area estimation often refers to small geographic areas, in this work we are concerned with subgroups of the population."*

With that out of the way, the primary thought in my head while I was reading this paper was "How novel is this work?" My assumption is that analyzing pooled samples is not — in and of itself — particularly novel, nor is the use of a Gaussian process to account for temporal correlation, and the proposed "data integration" approach seems to be a special case of the pooled approach in which some pools have m > 1 and others (i.e., the individual- level data) have m = 1. Moreover, the inferiority of the subsampling approach seems a bit obvious (the statement beginning with "Notably..." on lines 45–46 of Page 8 speaks directly to this). I want to make it clear that this is not to say that the paper is bad by any means, but rather that I think its structure could be improved. For instance, rather than focus the paper around the "novelty" of the methods, I'd suggest putting the Notre Dame COVID-19 data analysis front-and-center (given the attention currently being paid to the pandemic) and using it motivate the methods.

response: *This is a good point and the work will be better position as an applied paper than one that is purely methodological. On their own, each individual element of the methodology is not novel, but in response to these situations, they have been combined in a novel manner. We have restructured the article to be motivated by estimating dynamics in pathogen prevalence.*

Other substantive comments:

1. Section 3: Perhaps I'm too comfortable/familiar with the methods being used, but I found the way the methods were presented a bit frustrating at times.
   - First and foremost, I think it would be more natural to write the individual level model first — e.g., $y_{t,j} \sim Bernoulli(p_t)$ where p_t has the structure described in Section 3.2 — and then proceed to discuss the extension to analyzing pooled data from Section 3.1. **response:** *Good section, we have added a section covering this - and how to include covariates in the model - before presenting any of the pooled models.* – On a minor but related note, I believe the authors have gone into a bit too much detail when describing the Gaussian process model in Section 3.2. Perhaps I'm mistaken, but I'd like to think that Section 3.2 could be replaced by a few sentences (or at most, a short paragraph) describing the particular structure used rather than providing such a broad overview of Gaussian processes. **response:** *We have kept Section 3.2 that details GP models, but have removed all, or parts, of 10 sentences to reduce the unnecessary information provided.*
   - I'd like the authors to expand a bit on their decision to use a Gaussian process model with a covariance structure that effectively treats time as being continuous rather than, say, an autoregressive model that would treat time as being discrete. More specifically, the authors describe the computational burden associated with fitting a full-rank Gaussian process model, yet those issues could largely be avoided with an alternative structure. **response:** *We didn't give a good justification for this choice, but we appreciate the flexibility to treat time in a more continuous manner with a GP. We added the following acknowledgement of AR processes and justification*

to the end of the 4th paragraph in the introduction: "Autoregressive models are common for time series modeling, going all the way back to the 60s (Akaike 1969), but we opt for a GP due to the ability to treat time in a continuous manner. Many of our applications, such as capturing and sampling bats do not necessarily occur in discrete time, but this modeling choice does result in additional computational burden."

2. Simulation study #3 and Case study #2: I feel like more detail needs to be provided for these examples. For instance, the authors say on Page 9 that each group will have a different prevalence driven by the different μ_i values, but it seems to only be implied that all groups share the same time trends, which seems like a major weakness/limitation of the approach (e.g., the curves in Figure 4 all have the same bumps, but these don't seem to be supported by the data (or at least they don't seem to be supported in each of the groups). To that end, I believe the authors should describe in more depth how their approach can be used to analyze pooled data from multiple groups when prevalence estimates from each group are desired.

response: *The point about groups sharing the same trend is a good one. We made this point explicit at the end of Section 3.5: "It is important to note that this framework does assume that all of the subpopulations share the same overall trend, but the trend can be shifted up or down with the μ_j values. If this is not reasonable, hierarchical structure would be needed in the W_t processes." and also added this comment to the last paragraph of the discussion: "One other specific extension would be a process to enable different temporal trends for subgroups, rather than subgroups sharing the same trend with a constant shift. This could be done by manipulating the W_t processes from Equation 8."*

Additional description is added to the last paragraph of page 9: ", which shift the latent prevalence curve and result in shared, but shifted prevalence curves across the groups"

For Case Study 2, we add the following additional text: "The overall prevalence is not as high as simulation study 3, but Microteropus pusillus, Epomops franqueti, Megaloglossus woermanni*, and the overall groups all have a clear increase in early 2014. While* Myonycteris torquota *and* Pipistrellus sp. *do not have obvious increases at this same time, less sampling effort is given for those species; nevertheless, the mean prevalence stays near zero during this time period, but uncertainty intervals do increase based on information from other species."*

- It is also not lost on me that in my main comment above, I suggested that the authors should focus the paper around the COVID-19 example (due in part to its timeliness) and yet this example does not appear to have multiple "groups". What's less clear to me is if this is much of a loss — i.e., do the authors need to consider an example that allows for multiple groups, or should they instead spend more time on "model selection" / "sensitivity analyses" to compare/contrast different sets of model assumptions?

response: *This is a fair point, and determining which portion of a continuing stream of research to include in a particular manuscript is challenging. Our intention with this manuscript is to use this work to publish methods that will enable estimating prevalence for multiple variants. So our inclination would be to leave this in the paper.*

Minor comments:

3. Middle of Page 4: Since it was clear that the authors would soon begin discussing prevalences and probabilities for different time periods, I found it strange that p and π were first introduced without t subscripts for time.

response: *The notation is updated, and text added, to index prevalences to time t*

4. Notation in Section 3: In Section 3.2, the authors use μ (bold μ) to denote a vector, but in Section 3.3.1, Y_{ti}, (not bold) denotes a vector.

response: *To be consistent with notation, we have removed bold notation.*

5. Page 5, lines 19–20: Stating that σ 2 and l control the "oscillation speed" and "amplitude" of the process rather than simply saying that σ 2 represents the variance of the process and that l is related

to the "(effective) range" of the covariance structure seemed strange to me. Perhaps this is simply a difference in terminology between different fields of statistics, but if the authors decide to keep their terminology, I think they should offer alternative interpretations of these parameters.

response: *We replaced oscillation speed and amplitude with variance and effective range, as suggested.*

6. Bottom of Page 6: In the equation/expression you have z_{ti} as the random variable, but the text below it you have y_{ti}.

response: *Thanks for catching this, we have have updated the text part to be z_{ti} rather than y_{ti}.*

7. Section 4: I'm a little confused by the use of the words "synthetic" and "simulated" seemingly being used interchangeably. In particular, I'm accustomed to seeing the phrase "synthetic data" used in privacy applications (e.g., the release of synthetic data in place of of sensitive data), and I don't see much gain in using both "synthetic" and "simulated" when "simulated data" and "simulation study" would suffice.

response: *We have replaced synthetic study / synthetic data with simulation study / simulated data.*

Reviewer 4

This article described a nonparametric, hierarchical Bayesian small area model to infer population prevalence from the pooled test results through time. The author(s) also proposed pooling and jointly testing multiple samples to reduce testing costs when estimating the prevalence of a disease. This approach is shown to reduce uncertainty compared to individual testing at the same budget and to produce similar estimates compared to individual testing at a much higher budget based on synthetic studies and case studies. The paper is well written but not very in detail. Some revisions can help the articles. Generally, I believe the paper needs some minor amendments and clarifications required, as listed below:

The things I would recommend for any revision are: - The paper talked a lot about the cost of the tests. One major comparison is between pooled estimates and estimates obtained from a subsample of the original data representing the same number of tests. I am not very clear about the cost definition here. Is it purely related to the number of tests? Will the single grouping test cost more than a single individual test cost? And It is very straightforward that few number of individual test results must be worse than the grouping ones. More explanation about the cost can help.

response: *We added the following clarifying comment about cost to the third paragraph in the introduction. "Testing costs are generally proportional to the number of laboratory test conducted rather than the number of individuals tested; hence, pooled testing can reduce overall costs relative to testing specimens individually."*

- More literature reviews about the small area estimations based on categorical data, time series can help to support this application. In addition, there is no literature review about the key model feature: Gaussian process. Any similar applications related to Gaussian process?

response: *With regard to GPs, we added the following statement to the second paragraph on page 2. After considering reviewer 3's comments about small area estimation, we have de-emphasized this phrase in our work.*

- In Section 3.3, those Bayesian models' expressions are not accurate. No "iid" or "ind" notation; no conditional distributions notation. Please check them and edit them correctly.

response: *We have changed $\sim$ to $\overset{\text{iid}}{\sim}$ for all of the binomial / Bernoulli random variables in Section 3. Additionally, conditional distributions have been included, for instance the first line of equation 3 is updated from $Y_{t_i,j} \sim Bernoulli(\pi_{t_i,j})$ to $Y_{t_i,j}|\pi_{t_i,j} \overset{\text{iid}}{\sim} Bernoulli(\pi_{t_i,j})$.*

- No definition and description on t_i , especially the "i" on page 4 line 6.

response: *The following text was added to page 4, line 6 to define t_i: "corresponding to the i^{th} time point in the study"*

- On page 4, what is the method for subsampling?

response: *Pools are constructed using random sampling without replacement. - comment added to first paragraph on page 4.*

- On page 7, in Section 3.6, it mentioned that the inference in Stan remains practical. What are the number of warmup, the number of Markov chains, and the number of thinning to make every parameter to converge? What is the computation environment for this research?

response: *The following statement is added to Section 3.6. "For all of the analyses, four Markov chains were used. Generally, 2000 - 3000 iterations with 1000 - 1500 iterations for warmup were sufficient for parameters to converge. The analyses were conducted on a 3.1 GHz Dual-Core Intel Core i7 took less than about 20 minutes when running 4 chains in parallel."*

- The figures show the estimated median curve and tables show all the posterior means. Please add some explanation about why showing the median curves. And what is the 95% credible regions based on, posterior median or posterior mean?

response: *To be consistent we have updated all figures to show posterior means rather than medians. Additionally, it is noted that credible regions are quantile-based.*

- On Page 9 Figure 2, the y axis limit is 0.5 but as stated in the paper, the prevalence is generally less than about 0.25. Why using the 0.5 as the upper bound for y-axis?

response: *Good point. We have adjusted the y-axis limit on Figure 2 to be closer to 0.3.*