nature research

Peer Review Information

Journal: Nature Computational Science
Manuscript Title: Quantifying Spatial Under-reporting Disparities in Resident Crowdsourcing
Corresponding author name(s): Professor Nikhil Garg

Editorial Notes:

Transferred manuscripts	This manuscript has been previously reviewed at another journal that is not operating a transparent peer review scheme. This document only contains reviewer comments, rebuttal and decision letters for versions considered at Nature Computational Science

Reviewer Comments & Decisions:

Decision Letter, initial version:

Date: 22nd August 23 17:51:40
Last Sent: 22nd August 23 17:51:40
Triggered By: Fernando Chirigati
From: fernando.chirigati@us.nature.com
To: ng343@cornell.edu
BCC: fernando.chirigati@us.nature.com
Subject: Decision on Nature Computational Science manuscript NATCOMPUTSCI-23-0304A-Z
Message: ** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your co-authors. **

Dear Professor Garg,

Your manuscript "Quantifying Spatial Under-reporting Disparities in Resident Crowdsourcing" has now been seen by 3 referees, whose comments are appended below. You will see that while they find your work of interest, they have raised points that need to be addressed before we can make a decision on publication.

The referees' reports seem to be quite clear. Naturally, we will need you to address *all* of the points raised.

While we ask you to address all of the points raised, the following points need to be substantially worked on:

- Referee #1 asks for empirical validation of the method with external data, which we definitely think would be important to strengthen the results of the paper.
- Please improve the proof of the main theorem and write all of the derivations clearly, as requested by Referee #3.
- The referees noted a number of missing discussions, including: generalization of the work to other datasets and contexts; inequity and the role of population density; and more context into how the NYC Department of Parks and Recreation uses reporting data to inform their inspections. Please make sure to add these discussions to the paper.
- Some issues of presentation were discussed by the referees. As an additional note, our Articles have the following structure: Introduction, Results (where a brief summary of the methods, enough for the readers to understand the results, can be added), Discussion, and Methods. If you think it's useful, you can start modifying the paper accordingly to improve its presentation.

In addition to these points, it would also be beneficial to address the following concerns:

- Referee #2 suggested creating a package to make it easier for less-technical users to use the proposed method. We strongly recommend doing so, as this will likely increase the potential impact of the paper.

Please use the following link to submit your revised manuscript and a point-by-point response to the referees' comments (which should be in a separate document to any cover letter):

[REDACTED]

** This url links to your confidential homepage and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this e-mail to co-authors, please delete this link to your homepage first. **

To aid in the review process, we would appreciate it if you could also provide a copy of your manuscript files that indicates your revisions by making use of Track Changes or similar mark-up tools. Please also ensure that all correspondence is marked with your Nature Computational Science reference number in the subject line.

In addition, please make sure to upload a Word Document or LaTeX version of your text, to assist us in the editorial stage.

If you have any issues when updating your Code Ocean capsule during the revision process, please email the Code Ocean support team Cc'ing me.

To improve transparency in authorship, we request that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and

Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit please visit www.springernature.com/orcid.

We hope to receive your revised paper within three weeks. If you cannot send it within this time, please let us know.

We look forward to hearing from you soon.

Best,
Fernando

--
Fernando Chirigati, PhD
Chief Editor, Nature Computational Science
Nature Portfolio

Reviewers comments:

Reviewer #1 (Remarks to the Author):

This reviews "Quantifying spatial under-reporting disparities in resident crowdsourcing," submitted for consideration for publication at Nature Computational Science (NATCOMPUTSCI-23-0304A-Z). Overall, I think this paper is very well done. It uses a clever statistical device to reverse engineer the tendencies of different communities to report (or not) issues requiring government intervention—specifically issues regarding tree maintenance. It has promise as a more generalizable technique, as well. It is well-analyzed, well-written, and a valuable contribution. Of course, I have a few suggestions that would strengthen the paper.
--The biggest outstanding question here is, "Does it work?" As the authors know, others have tried to solve this problem with internal devices (see O'Brien, Sampson, & Winship in Sociological Methodology) that were replicable without additional outside data. My sense is that this approach could be more durable and extensible but a major advantage of the O'Brien et al. paper is that they collected external data to validate their technique. That did not happen here. Thus, while I'm convinced of the theoretical argument that their technique works, I would appreciate an empirical validation of it using public audits. I know this is a heavy ask for a revision, but it really is necessary to substantiate the claims that the authors are making about their methodology. Otherwise, there will always be the caveat that its effectiveness is rooted solely in assumptions. Further, such a validation will reveal nuance and detail as to the level of precision this technique can achieve, contexts in which it is more or less accurate, and more.
--I found the Discussion / Conclusion a bit underwhelming. I think there is a need to describe the limitations of the current demonstration, but that also feeds into a missed opportunity to instruct readers on how to carry the work forward. The demonstration here was of a very narrow use case—not just 311, but a very small

slice of the types of issues that 311 supports. How does this generalize? Are there considerations that will need to be made for any generalization? What are the fundamental components that need to be adapted to other use cases, both within and beyond the 311 system? Could we use this logic in some way to adjust for other types of naturally occurring data sets, like Yelp ratings, Craigslist postings, 911 reports, restaurant code violations, and more? The authors don't need to answer these questions definitively for any of these or others examples, but I think it's important they provide a roadmap for people excited to work with those data sets.

Small points:
--The parenthetical reference to "co-production" in the first sentence seems either unnecessary or insufficient. If the authors want to incorporate that term, I think it deserves at least a definition or other contextualization.
--There is the claim that previous techniques only estimate whether an issue is reported or not. This is only partially true. O'Brien et al. use report delays to estimate a standardized "capacity for reporting."
--There are over 300 municipalities (pre-pandemic) using 311 systems. The number is probably far more at this point.
--I quibble over the "unknown" period between the last report and the closing of the case. Shouldn't that period be considered known if we are confident that the agency has not yet closed the case? Typically agencies close immediately after doing so. Thus, from report #1 to closure we know that there is a problem. The only period we don't really know about is the birth ◊ report #1 period.
--I'm not sure I'm comfortable with 90% confidence for statistical significance. I tend to think of 95% as more reliable.
--The posterior distributions for regression coefficients confused me. If these were all entered into a model, the model would be over-saturated. i.e., You can't have all 5 boroughs in the model and have it converge, unless you have a handful of cases mapped outside of boroughs. That, however, would generate a whole new set of issues as your comparison group is idiosyncratic and outside the jurisdiction of the program itself. If the authors ran the models correctly and then are inferring the reference group's response rate from the intercept, that's defensible but needs to be explained. The same issue arises for a variety of other categorical variables in the model.

Reviewer #1 (Remarks on code availability):

No, I focused on the substance of the paper.

Reviewer #2 (Remarks to the Author):

This paper addresses a well-known and longstanding challenge in using user-generated, crowdsourced data—--e.g., calls to 311 hotlines to report hazardous incidents like downed trees—--to inform the provision of municipal services, namely that the true incident rate is unobserved, and that different incidents are reported at different rates and with different (unobserved) initial delays. The paper provides a novel strategy based on Bayesian regression and Poisson rate estimation for estimating heterogeneous incident reporting rates and reporting delays, conditional on an incident having occurred. Rather than leverage external (approximate) ground

truth data, as other methods do, the strategy outlined in this paper relies on using duplicate reports for the same incident. The paper also includes compelling empirical illustrations of the method to data from NYC's Department of Parks and Recreation, as well as two municipal departments in Chicago.

To my knowledge, this approach is new, and a sensible strategy for tackling the underlying statistical challenges involved in measuring heterogeneous reporting rates when ground truth data are unknown/unavailable. The methodological choices (e.g., using a zero-inflated model, and a Bayesian approach, given the high dimensionality of incident characteristics) strike me as appropriate for the suggested applications. I really like the emphasis on real-world relevance for policymaking, and could see this approach being used by a much wider variety of municipal (and state and federal) agencies (e.g., police departments receiving 911 calls about publicly observable incidents, housing inspectors receiving complaints, etc) to inform and improve decision-making. At a purely technical level, my impression is that the novelty is limited; the contribution consists of combining fairly standard statistical methods (and a bit of theory) to convincingly address a real-world challenge in a new way.

The paper is fairly clearly written (see more on this below), and has a comprehensive set of appendices that include various robustness checks (the discussion of possible sources of bias in Appendix D.1 was appreciated). To the extent that they are included directly, the statistics presented seem appropriate and are given with uncertainty (e.g., the posterior summaries in Table 1a). The code and exact data used in most of the analysis (some data are private) are publicly available, easy to reproduce via CodeOcean, and it seems like the analysis would be straightforward to reproduce on one's own machine or to modify for a new problem.

Below, I list my main comments, questions, and suggestions first, followed by some minor corrections and typos and such. I hope these will be useful to the authors as they revise this paper.

General thoughts:
1. At a high level, the writing is clear, but there are numerous typos throughout (that could have been caught by a spellcheck, e.g., "Chicago" is spelled "Chciago" on p. 16, "statistically" is spelled "statisitically" on p.8, "occurrence" is spelled "occurence" on p. 12, and many others). I'd encourage the authors to take careful pass through the entire manuscript to polish it and make it publication-quality.

2. Similarly, the latter half of the paper is written in a noticeably less careful manner than the first half. The first part of section 4 stood out to me in this regard; it was quite hard to follow exactly what was going on without skipping back and forth to other parts of the paper (e.g., to section 6 and to various tables in the appendix) that weren't referenced. Again, I would encourage the authors to take a careful pass through the entire manuscript to make sure it contains enough information at the appropriate points (and is properly organized) for a reader to easily follow.

3. I would like to see a deeper discussion of *inequity* overall. The results given in the paper largely boil down to "estimated reporting rates are higher over here than over there, or for these incident types compared to those incident types, and that's likely to be inequitable." However, I think the method presented here is quite

powerful, and can be leveraged more than the simple descriptive statistics in sections 4 and 5 to provide some insight into inequity. Perhaps provide a definition or two (of some quantitative notions of inequity) and use that to contextualize the results? I'd also like to see a more thorough discussion of the role of population density, which is the naive first-order explanation for apparent spatial inequities. The authors do include population density as a covariate in some of their analyses, but why not include that and census tract indicators (that are used in the spatial analysis)? Maybe there are more reports (and faster service) in some places just because there are more people (and hence more people who would benefit from the problem being fixed)? More ambitiously, is there any way to get even an approximate measure of foot traffic, rather than just a residential measure?

4. I'd like to see some discussion, even if it's not extended, of how NYC DPR currently uses reporting data to inform their inspections and response more generally. Do they send inspectors out as some function of the number of reports? Presumably their process also involves the incident type, the location of their inspectors, things like that? Can more detail be provided about how specifically DPR would change their practices in response to more detailed knowledge of reporting delays?

5. Figure 3 on p. 12 is very interesting. Can more insight be provided into the nature of the heterogeneity in delays? My naive guess would be that the distribution of incident types is very different in Manhattan vs. the other boroughs, with incidents that require immediate response much more prevalent in Manhattan. But is that actually the case?

6. On P. 13, "Estimating relative potential of different interventions" subsection. I think this could be fleshed out more. A couple ideas: (1) it would be interesting to see reporting/inspection/work order delay times for one specific kind of incident (say, downed trees) where the intervention type and speed is clear; (2) it seems like it would be pretty simple to provide some quick estimates of some type of overall utility gain DPR would get from following one of the ideas laid out (e.g., what if DPR were to prioritize responding faster in certain neighborhoods? What would that look like? Which neighborhoods would get more resources? If one were going to deploy targeted advertising encouraging reporting, where should one do so?)

7. The authors could consider creating a package to help less-technical users (e.g., practitioners) use these methods on their own.

Minor comments:
1. (abstract) some inconsistencies in saying "white" (e.g., intro) vs. "White" (e.g., abstract).

2. (Section 4) reference section 6 as providing more detail on the data, etc, at the beginning of this section.

3. P. 8, beginning of section 4.1, "Base" covariates are mentioned but haven't been defined and I don't know where I can look those up (edit: they are defined in Section 6; please mention that "more detail can be found in section 6, including further details on our model specifications"). Are these the covariates in Table 1a, on p.9? What are the values of INRiskAssessment? What units is the Tree Diameter variable

measured in?

4. P. 9, Table 1a, give descriptive names for the covariates (not, e.g., "INSPcondition[T.Dead]"). Also, no need to give the 50th percentile of each posterior distribution, just say that they're all basically the same as the means.

5. P.9, Table 1b, how are these numbers calculated (say, for the Manhattan example incidents)? It might be nice to give one example calculation.

6. P.10, Figure 2, caption text "coefficients on spatial coefficients" doesn't make sense. Also, one shouldn't assume the reader knows, based just on the map in Figure 2(a), where "downtown Manhattan" is, Queens, etc. Perhaps annotate Fig. 2(a) accordingly?

7. p.12, very bottom, why were risk assessment scores discretized in this application but not in the model fit in Section 4 (i.e., the one presented in Table 1a)? Why not include the interaction between borough and risk category in that model as well?

8. (Section 6.1) what exactly are the incident-level covariates? More is said about this on p.15, but e.g., the values that the 'inspection results' variable can take on weren't provided, at least not that I saw. How is location recorded in the data?

9. (Appendix D) why is Table 5 so small? Please make it the same, standardized (and readable) size as other tables in the data.

Reviewer #2 (Remarks on code availability):

The CodeOcean capsule ran without any issues (it takes a while, though!). The code (including the README file) and data are generally well organized and it seems like the analysis (with public data) would be pretty easy to reproduce on one's own machine.

Reviewer #3 (Remarks to the Author):

Summary:
In this paper, the authors propose a Poisson estimation method to identify heterogeneous reporting delays using duplicate reports about the same incident. They also provide a theorem to justify their method. Then they apply their method to New York data and Chicago data. They find out that there are substantial spatial disparities in reporting rates after controlling for incident characteristics. Finally, they explain how their method can help people to come up with practical solutions and insights. The paper is mainly divided into the following parts: introduction, model and research question, empirical method, heterogeneity in NYC and Chicago, discussion of the application of findings, and data processing.

Strength:
The authors propose a reasonable model that takes the spatial disparities in reporting rates into account and estimates the reporting rates with duplicate reports. The

theory is easy to understand and the authors apply the method to two real datasets to support their statements. The experiments are clear and they also explain how the findings of their methods could potentially help with real-world problems. Overall, the method is reasonable in solving the problem that the authors aim at.

Critics:
The proof of the main theorem is not carefully written. The author skips some steps and does not define every variable and function clearly. Some indexes also seem to be wrong, which makes the proof a bit confusing to read. Considering that this is the main and only theoretical result in this paper, I think it needs to be written more carefully and clearly. In the paper, the authors use many words to describe something, but sometimes they skip some proofs and it would be better if they write all the derivations clearly.

Questions:
1. P6 2nd paragraph: why the stopping time is independent of the process parameter lambda?
2. P7 4th paragraph: why do you choose \bar{T} in this way, and is there a criterion to choose this?
3. P7 6th paragraph: why the reporting rate is defined as equation (4)? Can you explain the reason?
4. P8 3rd paragraph: the authors mention that there can be other specifications. Can you give some examples and explain the pros and cons?
5. P22 4th paragraph, the index is confusing, what is interval start, is it $\tilde{t}_i$? If t_i^{0} is the time between the interval start and the first report, why $\tilde{T}_i$ is equal to the sum of t_{i}^{m} without $m=0$?
6. P22 1st equation 2nd line: why the sum in the first bracket is over subscript j? Are you summing over different incidents?
7. P22 How do you marginalize out {t_i} to obtain the final result?
8. Would it be possible to deal with the case where the report rate changes over time?
9. What would be a future direction, and what's the limitation?

Reviewer #3 (Remarks on code availability):

Properly documented code.

Author Rebuttal to Initial comments

nature research

Round 1 Authors' response: "Quantifying Spatial Under-reporting Disparities in Resident Crowdsourcing"

AR.1 Summary of changes

Dear Editorial Team,

Thank you for the careful read of our paper. We are grateful to the editorial team and reviewers for their time and thoughtful comments. We have overhauled our model and work in response to the feedback, and believe that the draft is substantially improved – especially in how our results are validated and discussed – as a result. Based on the review team's feedback, this draft has several primary changes compared to the previously submitted version:

1. We conduct and report *four* substantial additional analyses using external data and various additional uses of existing reporting data, to validate, "does the method work:" (a) We use *storms* as a source of ground truth knowledge of when incidents occurred (large storms cause substantial tree-related damage), and validate that our model's estimates align with *true* reporting delays after storms. (b) We validate the socioeconomic/demographic coefficients specifically by comparing them to voter participation rates in NYC. (c) We perform out-of-sample test set validation for whether our model can correctly predict the delay between the *first* and *second* report. (d) We train a *new* model that *only* uses data after the second report, and then see if the model can predict the (unknown to the model, but known to researchers) time between the first and second report. Together, these results establish that our model works strikingly well in predicting average reporting delays based on incident characteristics, and especially is effective at predicting *differences* between reporting delays of incidents with different characteristics. These changes are detailed in the response to R1.2.

2. We have overhauled the theoretical writing, both in the main text and the appendix. We now carefully define all variables, stochastic processes, and assumptions, and provide more detailed proof than before. To maintain accessibility for more audiences, we give a correct but high-level theorem statement in the main text, and then the fully-defined-in-language-of-stochastic-processes version in the appendix. We detail these changes in the response to R3.

3. We have overhauled the writing and paper structure. We have shortened/combined the model and theoretical section, into a Theoretical Results section (now Section 2), including only enough of the model for the reader to understand the theoretical and empirical results. We then present the Empirical results (Section 3), Results on impact and applying the model to the NYC Parks setting (Section 4), Discussion and conclusion (Section 5), and Methods (Sections 6 and 7). We further expand various discussion points, including those requested

by the reviewers. In order to streamline the main results and analysis depth, we have moved more of the Chicago replication results to the Appendix.

4. There is a renewed focus (in both the writing and choosing what results/plots to include in the main text) on equity throughout the paper, including focusing on differences between Boroughs and highlighting through a map the cumulative association of demographic/socioeconomic differences in reporting delays across census tracts. We detail these changes in response to R2.4.

5. To be accessible to less technical readers, we now provide a separate GitHub repository[1] that has the minimal required code and clear functions in order for others to apply our methods in their applications. This code only uses standard Python packages (e.g., statsmodels to train the Poisson regression instead of Stan) and so should be more accessible than our other repository, which is primarily meant for replication of our results.

We detail our changes below, and further provide a separate pdf file that has track changes enabled[2] (alongside a clean pdf file for resubmission).

AR.2 Response to editor

Point E.1 — While we ask you to address all of the points raised, the following points need to be substantially worked on:

- Referee #1 asks for empirical validation of the method with external data, which we definitely think would be important to strengthen the results of the paper.
- Please improve the proof of the main theorem and write all of the derivations clearly, as requested by Referee #3.
- The referees noted a number of missing discussions, including: generalization of the work to other datasets and contexts; inequity and the role of population density; and more context into how the NYC Department of Parks and Recreation uses reporting data to inform their inspections. Please make sure to add these discussions to the paper.
- Some issues of presentation were discussed by the referees. As an additional note, our Articles have the following structure: Introduction, Results (where a brief summary of the methods, enough for the readers to understand the results, can be added), Discussion, and Methods. If you think it's useful, you can start modifying the paper accordingly to improve its presentation.

In addition to these points, it would also be beneficial to address the following concerns:

- Referee #2 suggested creating a package to make it easier for less-technical users to use the proposed method. We strongly recommend doing so, as this will likely increase the potential impact of the paper.

[1] `https://github.com/ZhiLiu724/reporting_rate_estimation`

[2] One thing to note: we use the pdf-diff latex tool to generate this pdf, which marks every change between the files. It generally works very well, but is also a bit confused with numbers in tables – it does not always correctly note what the old values were in the table.

Reply: Thank you for handling our paper and your helpful comments throughout the process, including providing this clear path to revision. As summarized above and detailed below in response to specific reviewer points, we believe that we have thoroughly addressed each of these points, and that the paper is much stronger as a result.

AR.3 Response to Reviewer 1

Point 1.1 — This reviews "Quantifying spatial under-reporting disparities in resident crowdsourcing," submitted for consideration for publication at Nature Computational Science (NATCOMPUTSCI-23-0304A-Z). Overall, I think this paper is very well done. It uses a clever statistical device to reverse engineer the tendencies of different communities to report (or not) issues requiring government intervention—specifically issues regarding tree maintenance. It has promise as a more generalizable technique, as well. It is well-analyzed, well-written, and a valuable contribution. Of course, I have a few suggestions that would strengthen the paper.

Reply: Thank you for the kind words and support. We believe that your suggestions, especially to find ways in which to validate our method, have made for a stronger paper.

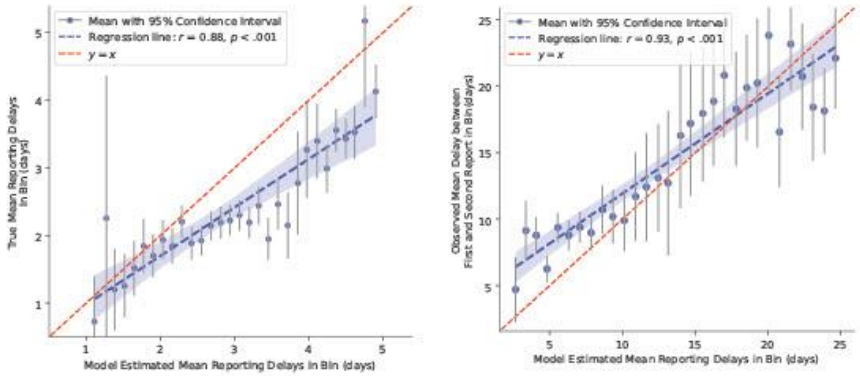
Point 1.2 — The biggest outstanding question here is, "Does it work?" As the authors know, others have tried to solve this problem with internal devices (see O'Brien, Sampson, & Winship in Sociological Methodology) that were replicable without additional outside data. My sense is that this approach could be more durable and extensible but a major advantage of the O'Brien et al. paper is that they collected external data to validate their technique. That did not happen here. Thus, while I'm convinced of the theoretical argument that their technique works, I would appreciate an empirical validation of it using public audits. I know this is a heavy ask for a revision, but it really is necessary to substantiate the claims that the authors are making about their methodology. Otherwise, there will always be the caveat that its effectiveness is rooted solely in assumptions. Further, such a validation will reveal nuance and detail as to the level of precision this technique can achieve, contexts in which it is more or less accurate, and more.

Reply: Thank you for this comment – pushing on this point has added substantial confidence in the method. As summarized above, we focused on this suggestion to find validation for our approach, and believe that we have done so. We detail our additional analyses here, which appear in the now Section 6.2 and Appendix D.8.

Addition of external verification in the main text

In Section 6.2, we now validate our presented model results (primarily results of the spatial model) by comparing the model estimated reporting delays from our methods to two sets of (out-of-sample) approximate ground truth.

Using *storms* as a cause for approximately *known* incident occurrence times. Storms cause substantial serious tree damage, including knocking down trees; crucially, *we can approximate the true incident time as the time that the storm in question arrived.* Thus for incidents reported in the days following a storm, we approximately know the true reporting delay. On August 4$^{\text{th}}$, 2020, Tropical Storm Isaias hit New York City, which caused major damage and triggered a travel



(a) Validation on incidents caused by Tropical Storm Isaias.

(b) Validation on time between observed first and second reports.

Figure 1: (a) Comparison of true reporting delays and model estimated reporting delays, based on data after Tropical Storm Isaias hit New York City on 8/4/2020 until 12 PM on 8/14/2020. True reporting delays for these incidents are calculated as the time between the first report of an incident and 12 PM on 8/4; model-estimated reporting delays are obtained using coefficients learned using the spatial model. (b) Comparison of observed delay between first and second reports and model estimated reporting delays, based on data between 9/1/2020 and 8/31/2022. Model-estimated delays are obtained using the spatial model. For both (a) and (b), all incidents are then categorized into 30 bins based on their estimated reporting delays, and we calculate the means of true and estimated delays within each bin. Model estimates remain correlated without binning. For the storm analysis, at the individual incident level, we observe Pearson $r = 0.20$ with $p < .001$. For true first-to-second report delays at the individual level, Pearson $r = 0.21$ with $p < .001$. Together, these results indicate that our model's estimates of reporting delay are accurate.

advisory [2], predicting that the strongest winds and rains would be from 12 PM to 2 PM on that day. We look at service requests submitted to NYC DPR from 12 PM on 8/4 to 12 PM on 8/14 and calculate the true reporting delay of an incident as the time between the submission of the first report associated with it, and 12 PM on 8/4. We then use results from our spatial model to calculate the model-estimated reporting delays for each incident.[3]

Figure 1a shows the relationship between model-estimated delays and true reporting delays. There is a strong correlation (Pearson $r = 0.88$, $p < .001$) between binned model estimated reporting delays and the true reporting delays, and predictions are approximately on the $y = x$ line suggesting that our model accurately recovers heterogeneity in incident-level reporting delays.[4] The Appendix

[3]To calculate the model estimated reporting delay, we take the reporting rate for each incident implied by the model and calculate the average reporting delay for that rate, *conditional* on the reporting delay being less than 10 days (to reflect our data filtering to only include incidents reported in that period).

[4]We note that our model slightly *over-predicts* reporting delays: reporting behavior likely changes after natural disasters, due to the amount of damage (e.g., DPR received 15,266 service requests on 8/4 alone, far exceeding their

Section D.8 contains similar results from analyses using different specifications of the end date, using data on service requests induced by Tropical Storm Ida, and without binning.

Comparing predicted delays to out-of-sample delays between *first* and *second* reports. Our method is designed to predict the time from the incident occurrence and the first report. While we cannot always validate that prediction (except in cases like storms, where incident occurrence times can be approximated), we can always validate that our approach is effective at predicting the time between first and *second* reports. We perform the following validation, which is approximately equivalent to test set validation in machine learning.

We analyze service requests submitted to DPR from 9/1/2020 to 8/31/2022 (after the model training data period). For each incident that received 2 or more service requests, we calculate the time between the first and second reports and estimate their reporting delay using the spatial model.

Figure 1b shows the relationship between model-estimated delays and observed delays between the first and second reports, when model-estimated delays are binned. As above, there is a strong correlation, and predictions are approximately on the $y = x$ line.

Additional external verification in the Supplementary materials

In the appendix, we provide more details for the above verification (such as for other storms and making different modeling choices). We further newly provide two additional analyses to verify that our method works.

***Training* a new model on data starting from the second report, and testing its predictions against *actual* delays between first and second reports.** Our theoretical analysis also holds if we instead use as the Poisson interval the time starting after the *second* report – if we split a Poisson process on the time of the first jump, and start counting subsequent jumps at that time, the resulting counting process would still be a Poisson process with the same rate. Thus, we can estimate the Poisson process rate using our methods, starting with the second report time. However, crucially, we as researchers know for the incident the *time between the first and second report*, which was not used to train the model (but which the model is trying to predict). Intuitively, this procedure evaluates our methods in a setting in which the ground truth is known.

In Appendix Section D.8.2, we carry out the idea; we train a model on the second report time onwards and compare its per-incident delay estimates to the (known) delay between the first and second report. We find that the estimates well-predict the ground truth.

Comparing census tract coefficients to voter participation rates The disparities in reporting behavior along socioeconomic variables highlight behavioral heterogeneity in resident crowdsourcing, and civic engagement at large. The level of civic engagement can be measured by many other means, chief among which is participation in political voting. One might expect that levels of civic engagement across avenues are correlated. We validate our socioeconomic coefficient estimates using this idea.

Voter-level public data on participation in voting is generally scarce in New York; however, the NYC Campaign Finance Board published a dataset containing participation scores of more than 4 million voters calculated based on their voting history from 2008 up to 2018, along with the census tract in which they resided in 2010.[5] We calculate cumulative association scores (the same

daily average of around 200) and city communication asking the public to report damage.

[5] `https://data.cityofnewyork.us/City-Government/Voter-Analysis-2008-2018/psx2-aqx3`

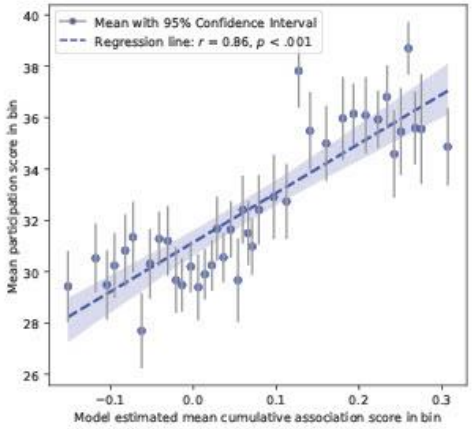


Figure 2: Relationship between voter participation rates in a census tract, and our model estimated cumulative association of socioeconomic variables with reporting rate. A significant positive correlation between these two measures affirms that our model estimates on disparities in reporting behavior are relating to other measures of civic engagement.

analysis as in the new main text Figure 2[6]) for each 2010 census tract, using the corresponding census data. We compare these scores with the average voter participation score in each tract. At the census tract level, our model-estimated mean engagement score is significantly correlated with the mean voter participation score (Pearson $r = 0.33$, $p < .001$). Response Figure 2 shows the full relationship, between binned cumulative association scores and voter participation rates. A significant positive correlation shows that our estimated disparities in 311 system usage relate to other forms of civic participation and representation.

Additional comments regarding whether and when the method works

Together, these validations suggest that our method works – model estimates correspond to actual delays, and *differences* in model estimates for different incident characteristics correspond to actual differences in reporting delays. We note that, as the analysis in Appendix D.8 explores, whether model estimates are *on the* $y = x$ line depends on the *maximum interval duration* (i.e., upper bound on the death time) that is chosen in the empirical method. This illustrates our theoretical requirement, that the interval end period must be *before* the actual incident death time – otherwise, the method will underestimate the reporting rate. However, we note that, as shown in Appendix Figure 16, even in this case, *differences* in estimates for different incident characteristics reflect real differences.

Furthermore, we note that in addition to the above new analyses that use external data to validate delay estimates, our work contains several other "validations" that may inspire confidence

[6]Since we are focusing on individual-level behavior, for that figure and this analysis we do not use the density variable, as that does not reflect individual-level characteristics, and only use the remaining five: Median age, Fraction white, Fraction college degree, Fraction renter, log of per capita income

that our method works, both in theory and practice. For instance, the coefficients that we find for incident-specific covariates (category, risk level) are all *face valid* (e.g., more urgent incidents are found to have delays of a few days, less urgent incidents up to hundreds of days), in both New York and Chicago. We further conduct simulations with known ground truth reporting rates and validate that our method correctly recovers them.

Finally, we note that the type of audit approach as in the O'Brien paper – in which researchers log on-the-ground conditions by themselves walking the streets – (already challenging at scale in that setting, and impressive that they did it!) is likely *impossible* for our setting: the estimand of our paper is reporting *delays* – so, to do a public audit of the kind of O'Brien, we would need to walk the streets enough to actually *observe incidents occur* in real-time, and then wait until these incidents are reported. We are unlikely to observe many/any incidents occur in real-time; if instead we log incidents that we see already having occurred sometime in the past, like O'Brien et al do (without knowing *when* they occurred) and then wait until they're first reported, that is conceptually the same as counting the time between the first report and the second report, as our method does (and is in fact identical to our validation approach of training the model using the second report onwards as described above). Further, especially for urgent/dangerous incidents, it may be unethical to not report incidents that we observe, to wait for them to be first reported by the public. For these reasons, we leverage existing external data instead of collecting our own data.

Point 1.3 — I found the Discussion / Conclusion a bit underwhelming. I think there is a need to describe the limitations of the current demonstration, but that also feeds into a missed opportunity to instruct readers on how to carry the work forward. The demonstration here was of a very narrow use case—not just 311, but a very small slice of the types of issues that 311 supports. How does this generalize? Are there considerations that will need to be made for any generalization? What are the fundamental components that need to be adapted to other use cases, both within and beyond the 311 system? Could we use this logic in some way to adjust for other types of naturally occurring data sets, like Yelp ratings, Craigslist postings, 911 reports, restaurant code violations, and more? The authors don't need to answer these questions definitively for any of these or others examples, but I think it's important they provide a roadmap for people excited to work with those data sets.

Reply: Thank you for this comment, it caused us to consolidate this related discussion (which was previously spread across the Model section and the conclusion) as well as expand it, specifically to emphasize applicability to other such types of systems. The relevant component of the conclusion now reads,

> Applying the method to other settings. In what other settings is the model applicable? First, while the incident occurrence process Λ_θ can be arbitrary, the reporting process must be Poisson with rates λ_θ – this means reports about the same incident must be independent; i.e., our method primarily works for *public* incidents such as potholes and downed trees that may be encountered and reported by separate people. However, the method does not easily handle *private* incidents, such as bed bugs within an apartment or food poisoning, in which potential reporters are likely to communicate with each other. Second, the analysis requires that the agency log information about duplicate reports and that this data be available to researchers – our empirical analyses were restricted to agencies that make this data publicly available via Open Data portals; applying the methods to other cities or agencies may require data sharing agreements

or automated methods for researchers to identify likely duplicate reports for the same incident. Third, it is important to note that our method does *not* require external estimates for how many incidents of each type θ one expects to see, which may be extremely difficult to estimate for high-dimensional types.

While our data application focuses on reports made in NYC and Chicago to specific city agencies, these requirements are widely met in other crowdsourcing systems in other cities, and in other domains such as software bug reporting (for example, users can report issues and bugs in open source repositories such as GitHub, and there may be heterogeneous reporting based on who the bug affects). Similarly, it is possible to apply our methods to 911 or other emergency systems, for types of incidents where there may be multiple independent witnesses (e.g., gunshots). Such broad applicability is especially important because reporting behavior (and its association with socioeconomic factors) may vary by context. For example, in India, marginalized communities tend to use *formal* channels (such as reporting systems) to access government resources, while dominant groups use informal ones [1].

Point 1.4 — The parenthetical reference to "co-production" in the first sentence seems either unnecessary or insufficient. If the authors want to incorporate that term, I think it deserves at least a definition or other contextualization.

Reply: Thank you for this suggestion. We have removed reference to 'co-production' in our manuscript, as 'crowdsourcing' is a term more widely used in the literature and more accessible to the readers.

Point 1.5 — There is the claim that previous techniques only estimate whether an issue is reported or not. This is only partially true. O'Brien et al. use report delays to estimate a standardized "capacity for reporting."

Reply: Thank you for pointing out this out to us – we definitely want to be accurate in such statements about prior work.

We have looked through the three O'Brien papers we cite [3–5] and we are unable to locate a mention to "capacity for reporting," or where reporting delays are estimated/how they're estimated. Would you please point us to which paper you mean/where it is? We'd love to learn about it and cite it appropriately.

In the meantime, we've softened the language to reflect that previous works *primarily* focus on whether a report happens or not, while leaving room for works that estimate delays. We're happy to update the language further as appropriate.

Point 1.6 — There are over 300 municipalities (pre-pandemic) using 311 systems. The number is probably far more at this point.

Reply: Thank you for pointing out this information. If you have one available, we'd love a citation that we can use for this fact – we were only able to find online web pages/blogs/non-academic sites that say close to 300 without an exhaustive list. For now, we omit the mentioning of the exact number of cities, and now mention that 311 systems are widely used, including in the four most populous cities in the U.S. – New York, Chicago, Los Angeles, and Houston.

Point 1.7 — I quibble over the "unknown" period between the last report and the closing of the case. Shouldn't that period be considered known if we are confident that the agency has not yet closed the case? Typically agencies close immediately after doing so. Thus, from report #1 to closure we know that there is a problem. The only period we don't really know about is the birth to report #1 period.

Reply: We mostly agree – the fundamental unknown is the time of incident birth, and that's what requires a method such as ours. Assuming that death times are exactly known would not materially change the method or its need to use duplicate report information, though it would of course simplify the application of our theorem, to finding an end interval for each incident that is before the incident close. The reason that we stated that death time is unknown is that it often is unreliably detected: many cases are not closed because the agency has not yet inspected it or addressed it, and many other incidents are resolved much before the agency actually closes the case – e.g., by community groups or a good Samaritan. For example, many of the agency's "resolution notes" are that they inspected the area and could not find the incident, suggesting that the incident disappeared at some point before they inspected and closed the case. In our empirical approach, we do take into account the resolution period (the death time as stated by the agency), choosing as the end of the interval that we measure as the minimum of inspection time, closing time if available, and a fixed time period.

We have updated the writing to reflect this nuance. The model section now reads:

> Crucially, birth times are unobserved in the model; death times may be unobserved or partially observed.

with footnote:

> In practice, death times are partially observed: agency actions such as inspections and work orders are observed by the agency. However, we model true death times as noisily observed as (1) such data may not be available to external researchers; (2) many incidents are resolved by outside community groups, and these times are unobserved even by the agency. Fully observed death times improve our method's accuracy but do not change the problem difficulty: crucially, birth times are always unobserved by both the agency and researchers, by definition.

Point 1.8 — I'm not sure I'm comfortable with 90% confidence for statistical significance. I tend to think of 95% as more reliable.

Reply: Thank you for this comment. We have updated all confidence intervals to 95%, with almost all statistical significance results staying the same. We note that we needed to rerun all of our models as we did not save all chain details, and so coefficients in tables have slightly changed as well.

Point 1.9 — The posterior distributions for regression coefficients confused me. If these were all entered into a model, the model would be over-saturated. i.e., You can't have all 5 boroughs in the model and have it converge, unless you have a handful of cases mapped outside of boroughs. That, however, would generate a whole new set of issues as your comparison group is idiosyncratic and outside the jurisdiction of the program itself. If the authors ran the models correctly and then are inferring the

reference group's response rate from the intercept, that's defensible but needs to be explained. The same issue arises for a variety of other categorical variables in the model.

Reply: Thank you for pointing out this clarity issue in our writing. The reason why our models still are identifiable (and still converge) with the full set of levels for a categorical variable is that we enforce a *zero-sum constraint* on their coefficient, effectively absorbing the mean of these estimates into the estimate for the intercept (note that the Borough coefficients sum up to 0 up to rounding) – there are still only K-1 degrees of freedom, as when a level is dropped. We do this for the Borough and category variables, which allows us to present, e.g., all coefficients for all 5 NYC boroughs in the tables, thus aiding interpretability. Such zero-sum constraints are standard in Bayesian modeling (see, e.g. `https://mc-stan.org/docs/stan-users-guide/parameterizing-centered-vectors.html`) for this reason.

To make this point clearer, we now mention this in the caption of Table 1 where the coefficients first appear:

> Note that for Borough and category, we enforce a *zero-sum* constraint on the coefficients and so can present coefficients for every level, without collinearity issues.

with more detail in the methods section:

> Our **Base** analysis includes just these covariates along with *Borough* level fixed effects. As the covariates are standardized, these coefficients all have a zero-mean Normal prior. *Borough* and *Category* are categorical variables; instead of dropping one level for each variable, to maintain identifiability, we enforce a sum-zero constraint[7] on their coefficients to ease interpretability. For the other categorical covariates, we drop one level each.

Point 1.10 — (Remarks on code availability): No, I focused on the substance of the paper.

Reply: Thank you again for your detailed comments and suggestions!

AR.4 Response to Reviewer 2

Point 2.1 — This paper addresses a well-known and longstanding challenge in using user-generated, crowdsourced data – e.g., calls to 311 hotlines to report hazardous incidents like downed trees – to inform the provision of municipal services, namely that the true incident rate is unobserved, and that different incidents are reported at different rates and with different (unobserved) initial delays. The paper provides a novel strategy based on Bayesian regression and Poisson rate estimation for estimating heterogeneous incident reporting rates and reporting delays, conditional on an incident having occurred. Rather than leverage external (approximate) ground truth data, as other methods do, the strategy outlined in this paper relies on using duplicate reports for the same incident. The paper also includes compelling empirical illustrations of the method to data from NYC's Department of Parks and Recreation, as well as two municipal departments in Chicago.

[7] `https://mc-stan.org/docs/2_28/stan-users-guide/parameterizing-centered-vectors.html`

To my knowledge, this approach is new, and a sensible strategy for tackling the underlying statistical challenges involved in measuring heterogeneous reporting rates when ground truth data are unknown/unavailable. The methodological choices (e.g., using a zero-inflated model, and a Bayesian approach, given the high dimensionality of incident characteristics) strike me as appropriate for the suggested applications. I really like the emphasis on real-world relevance for policymaking, and could see this approach being used by a much wider variety of municipal (and state and federal) agencies (e.g., police departments receiving 911 calls about publicly observable incidents, housing inspectors receiving complaints, etc) to inform and improve decision-making. At a purely technical level, my impression is that the novelty is limited; the contribution consists of combining fairly standard statistical methods (and a bit of theory) to convincingly address a real-world challenge in a new way.

The paper is fairly clearly written (see more on this below), and has a comprehensive set of appendices that include various robustness checks (the discussion of possible sources of bias in Appendix D.1 was appreciated). To the extent that they are included directly, the statistics presented seem appropriate and are given with uncertainty (e.g., the posterior summaries in Table 1a). The code and exact data used in most of the analysis (some data are private) are publicly available, easy to reproduce via CodeOcean, and it seems like the analysis would be straightforward to reproduce on one's own machine or to modify for a new problem.

Below, I list my main comments, questions, and suggestions first, followed by some minor corrections and typos and such. I hope these will be useful to the authors as they revise this paper.

Reply: Thank you for your detailed comments and feedback! We believe that resolving these concerns has made for a stronger paper.

Point 2.2 — At a high level, the writing is clear, but there are numerous typos throughout (that could have been caught by a spellcheck, e.g., "Chicago" is spelled "Chciago" on p. 16, "statistically" is spelled "statisitically" on p.8, "occurrence" is spelled "occurence" on p. 12, and many others). I'd encourage the authors to take careful pass through the entire manuscript to polish it and make it publication-quality.

Reply: Thank you for this suggestion. We have carefully gone through the manuscript, including checking grammatical and spelling errors, and ensured that the writing is up to the standard of a publication.

Point 2.3 — Similarly, the latter half of the paper is written in a noticeably less careful manner than the first half. The first part of section 4 stood out to me in this regard; it was quite hard to follow exactly what was going on without skipping back and forth to other parts of the paper (e.g., to section 6 and to various tables in the appendix) that weren't referenced. Again, I would encourage the authors to take a careful pass through the entire manuscript to make sure it contains enough information at the appropriate points (and is properly organized) for a reader to easily follow.

Reply: Thank you for this suggestion. We have made several changes to the structure of the paper that should make it more accessible to readers, and include more information in the relevant results sections to reduce the amount of back and forth necessary.

Point 2.4 — I would like to see a deeper discussion of *inequity* overall. The results given in the paper largely boil down to "estimated reporting rates are higher over here than over there, or for these

incident types compared to those incident types, and that's likely to be inequitable." However, I think the method presented here is quite powerful, and can be leveraged more than the simple descriptive statistics in sections 4 and 5 to provide some insight into inequity. Perhaps provide a definition or two (of some quantitative notions of inequity) and use that to contextualize the results?

Reply: We have modified our paper to more clearly center and discuss inequity.

First, in Section 3, we have changed the main Spatial map that we show in the main text, to illustrate reporting inequity as it relates to socioeconomic characteristics – what is the overall association of each census tract with reporting rates, as a function just of socioeconomic characteristics. The associated text now partially reads:

> We next study the relationship between socioeconomic characteristics and reporting rates in each census tract, by fitting our model with the incident-level characteristics and a set of socioeconomic characteristics jointly (log income per capita, fraction of white residents, fraction of renter, median age, and fraction of residents with college degree, but *not* population density); then, we use the socioeconomic coefficients to learn the cumulative association for each census tract. Figure 2 shows the resulting map—there are substantial spatial differences in reporting across census tracts as explained by demographics. For example, downtown and midtown Manhattan (the blue region in the middle) have substantially higher values than Harlem and the Bronx (the red region to the north of the blue region), and the latter is substantially socioeconomically disadvantaged compared to the former. In Appendix D.8.3, we further show that these census tract-level values further correlate with voter participation rates in each tract – i.e., disparities in 311 system usage further relate to other forms of civic participation and representation. We note that these associations cannot be explained via correlations with population density – as shown in Appendix Tables 8 and 9, further controlling for population density does not substantially affect the other coefficients. These reporting disparities suggest substantial downstream effects in how quickly incidents are addressed, even if the agency does not prioritize one demographic group over another after receiving reports.

This figure replaces (in the main text) the previous spatial plot, which showed spatial differences between census tracts, including *all* factors beyond incident-level characteristics (like incident risk or tree size) – e.g., including population density, or spatial coefficients unexplained by any census variable. We believe that doing so highlights socioeconomic inequity in the system.

Second, as recommended, we have added a notion of "resolution parity" and an accompanying figure in the main text Section 4, around how long end-to-end delays are in each Borough, for the comparable highest-risk incidents. The associated text now partially reads:

> Figure 3b converts the end-to-end delays to relative differences from the city-wide median for such incidents – we calculate incident-level overall days and then compare Borough-specific median delays to the citywide median.
>
> ...
>
> Substantial end-to-end resolution inequities. There are large differences in end-to-end median delays to address incidents across Boroughs. *Even though the analysis is restricted to incidents that all received the highest work order priority level,* there are differences

in how long incidents take to be reported, inspected, and worked on. The cumulative differences are meaningful and point to substantial inequity: highest-risk incidents are resolved within 2 days in Manhattan, and only within 14 days in Queens. If the city instead had *resolution parity* – incidents of similar risk levels being addressed at similar delays after incident occurrence – then incidents in Manhattan would be addressed approximately 2x slower, and in Queens approximately 2.5x more quickly than in the status quo.

Point 2.5 — I'd also like to see a more thorough discussion of the role of population density, which is the naive first-order explanation for apparent spatial inequities. The authors do include population density as a covariate in some of their analyses, but why not include that and census tract indicators (that are used in the spatial analysis)? Maybe there are more reports (and faster service) in some places just because there are more people (and hence more people who would benefit from the problem being fixed)? More ambitiously, is there any way to get even an approximate measure of foot traffic, rather than just a residential measure?

Reply: Thank you for bringing up this concern – we agree that it is important to understand whether population density is driving the spatial inequity results.

To alleviate the concern that population density is the sole driving force behind all spatial inequities, we conduct an analysis that jointly estimates coefficients for a selected set of socioeconomic variables that includes population density. This set of variables is condensed from the full set of variables we present in Table 2, by eliminating those with apparent collinearity: for example, fractions of residents identifying as white, Black, and Hispanic have high degrees of collinearity, and we omit the latter two in this analysis. The results of this analysis are presented in Appendix Table 8. We find that even when included alongside population density, most socioeconomic variables still significantly contribute to the reporting rate, with directions the same as when they are included alone in the regression. For example, the fraction of white residents, the fraction of renters, and log income per capita are still positively correlated with the reporting rate. This analysis supports the notion that the measured inequities appear beyond population density.

Second, to your point on measurements of foot traffic – we attempted to include such data but ultimately did not. One potential option was to use foot traffic data collected by cities. However, upon closer inspection, measures of foot traffic often suffer heavily from inconsistencies and selection biases in the methods they are collected. For example, one set of such data available on the NYC OpenData platform[8] looks at 114 junctions across New York City and measure the volume of pedestrians passing these junctions. However, these junctions vary in their sizes and accessibility, and are already a subset of all junctions in New York that could be considered 'busy' – and thus may not be representative of their census tract. Commercial data providers, e.g., Unacast or SafeGraph may provide more comprehensive data on foot traffic using cell phone trace data, but now charge large fees. We can proceed and attempt to purchase such data if the review team deems this analysis essential – at a high level, we are skeptical that this will substantially change the results, given that adding population density does not seem to.

Point 2.6 — I'd like to see some discussion, even if it's not extended, of how NYC DPR currently uses reporting data to inform their inspections and response more generally. Do they send inspectors out as

[8] https://data.cityofnewyork.us/Transportation/Bi-Annual-Pedestrian-Counts/2de2-6x2h

some function of the number of reports? Presumably their process also involves the incident type, the location of their inspectors, things like that? Can more detail be provided about how specifically DPR would change their practices in response to more detailed knowledge of reporting delays?

Reply: We now more explicitly discuss both the status quo decision-making and potential intervention in more detail – you are correct that incident type is paramount, though report times and amount and logistics issues (where inspectors are, their workload) also matter. At the beginning of Section 4, we now state:

> After an incident is reported, NYC DPR may schedule an inspector to travel to the report site to analyze the incident. This scheduling decision is made by the agency based on reports – more hazardous report types are prioritized, as are incidents with more reports and incidents that were reported earlier. Then, based on the inspector's risk determination, the incident may be scheduled for a maintenance crew to address the issue. In each case, operational concerns such as locations and schedules of individual employees and incidents also play a role.

Then, in Section 4 when discussing interventions, we now have an expanded discussion, stating:

> Agencies could incorporate reporting delay estimates into their inspection and maintenance scheduling policies: (a) prioritize *individual incidents* which are estimated to have had substantial reporting delays – for example, for two incidents of the same type, prioritizing the incident with the earlier estimated time of occurrence, instead of report time as in the status quo; (b) invest more resources overall in *neighborhoods* with higher overall estimated reporting delays to respond to reported incidents – doing so would shorten inspection and work order delays to render overall time from incident to resolution more equitable.

Point 2.7 — Figure 3 on p. 12 is very interesting. Can more insight be provided into the nature of the heterogeneity in delays? My naive guess would be that the distribution of incident types is very different in Manhattan vs. the other boroughs, with incidents that require immediate response much more prevalent in Manhattan. But is that actually the case?

Reply: It is true that counts and distributions of incidents vary substantially across Boroughs. To account for this in the analysis, to produce this plot, (and associated figures in the appendix), we grouped incidents by their *work order priority level* as provided to us by DPR – the agency converts risk ratings (on several dimensions) after inspections into a single priority grouping, that they use to prioritize work orders (within each Borough/community group, work orders are essentially prioritized in order of priority, alongside operational reasons). All incidents included in Figure 3 are of the highest ("A") work order prioritization level and thus considered "equally important" by the agency to address.

To understand the heterogeneity in delays across incident severity, Appendix D.10 contains the same figures for the "B" and "C" prioritization groups. Indeed, median delays (in each of the reporting, inspection, and work order stages) are longer for the less urgent incidents. For example, work order delays in the Bronx and Staten Island are 28 and 22 days, respectively, for level B, and 103 and 118 days for level C. We note that similar plots emerge if we instead grouped by category (e.g., most of group "A" has report type "Hazard") – these results are provided in Appendix D.10.

We have added to the writing to clarify these points.

Point 2.8 — On P. 13, "Estimating relative potential of different interventions" subsection. I think this could be fleshed out more. A couple ideas: (1) it would be interesting to see reporting/inspection/work order delay times for one specific kind of incident (say, downed trees) where the intervention type and speed is clear; (2) it seems like it would be pretty simple to provide some quick estimates of some type of overall utility gain DPR would get from following one of the ideas laid out (e.g., what if DPR were to prioritize responding faster in certain neighborhoods? What would that look like? Which neighborhoods would get more resources? If one were going to deploy targeted advertising encouraging reporting, where should one do so?)

Reply: Thank you for these comments – as discussed in the response to the previous point, we perform these breakdowns both by incident type and by work order priority levels (see Figures in Appendix D.10). These figures all tell a similar story, which we summarize in the main text as follows:

> Our work provides a principled approach to compare reporting and response delays, enabling such analysis. For example, reducing work order delays (and to a lesser extent, inspection delays) in Staten Island, The Bronx, and Queens would substantially mitigate Borough-level inequity – e.g., reducing inspection and work order delays down to one day each (as in Manhattan and Brooklyn) would make end-to-end delays in these Boroughs comparable to those in Brooklyn. However, to reach parity with Manhattan (without introducing new delays there), the city would need to increase the reporting rate in every other Borough, such as through advertising or proactive inspections. Similar insights hold for other risk prioritization levels in Appendix D.9.

We note that we are hesitant to go more detailed than this in this paper because budget decisions and scheduling for inspections and work orders are intricate processes: increasing resources by Borough affects all types of incidents in that Borough, but perhaps not evenly across types. Further, some of the inspection and work delays are unavoidable, due to storm events – even if Boroughs have adequate and equitable resources for daily events, a storm might create many incidents in a location, that cannot be handled by the "standard" resources. We note that, in ongoing work, we are developing approaches to audit and improve agency responses in the presence of such factors – we hope to be able to say more about *how* to improve inspection and work order scheduling in the future, but the current methods do not help us with the *how* – only the "where."

Point 2.9 — The authors could consider creating a package to help less-technical users (e.g., practitioners) use these methods on their own.

Reply: Thanks for this suggestion – it made us realize that the role of our previously provided code was to help others replicate our exact results and that it would be separately useful to have a simple code base that allows others to apply our results in their setting. We have now created such a simple Python file with example usage, here: `https://github.com/ZhiLiu724/reporting_rate_estimation`. We believe that this code base will be more readily usable by others.

Point 2.10 — (abstract) some inconsistencies in saying "white" (e.g., intro) vs. "White" (e.g., abstract).

Reply: Thank you for pointing out this inconsistency. We have standardized with 'white' for all such usage, consistent with the Associated Press guidelines: `https://blog.ap.org/announcements/why-we-will-lowercase-white`.

Point 2.11 — (Section 4) reference section 6 as providing more detail on the data, etc, at the beginning of this section.

Reply: Thank you for your suggestion, we have made this reference.

Point 2.12 — P. 8, beginning of section 4.1, "Base" covariates are mentioned but haven't been defined and I don't know where I can look those up (edit: they are defined in Section 6; please mention that "more detail can be found in section 6, including further details on our model specifications"). Are these the covariates in Table 1a, on p.9? What are the values of INRiskAssessment? What units is the Tree Diameter variable measured in?

Reply: Thank you for these suggestions, we have changed the way these covariates are referenced and mentioned, and believe the updated manuscript should be clearer on this matter. The risk assessment scores range from 0 to 12, with 12 given to the highest risk incidents, and we now mention this information in Appendix Table 5. The tree diameter at breast height variable is calculated (by the inspectors) by dividing the tree circumference at breast height (measured in inches at approximately 54 inches / 137cm above the ground) by 3.14159 and rounded to the nearest whole inch.

Point 2.13 — P. 9, Table 1a, give descriptive names for the covariates (not, e.g., "INSPcondition[T.Dead]"). Also, no need to give the 50th percentile of each posterior distribution, just say that they're all basically the same as the means.

Reply: Thank you for these suggestions, we have changed the way they are presented in Table 1a, and believe that this improves interpretability. We have also omitted the column for the 50^{th} percentiles, which are indeed very close to the mean in all cases.

Point 2.14 — P.9, Table 1b, how are these numbers calculated (say, for the Manhattan example incidents)? It might be nice to give one example calculation.

Reply: Thank you for pointing out that this part is not fully explained. Here we give the example calculation for the Hazard, tree in Poor condition, risk assessment score 12 incident in Manhattan. The mean delay is calculated as:

$$1/\exp\Big(\underbrace{-3.229}_{\text{Intercept, tree in Poor condition}} + \underbrace{1.418}_{\text{Hazard}} + \underbrace{0.438}_{\text{Manhattan}} + \underbrace{\frac{12-6.4915}{2.1788}\times 0.240}_{\text{Standardized risk assessment score}}\Big) \approx 2.2$$

There are a few points worth mentioning with this calculation. First, we take the exponential of the sum of these coefficients, in accordance with the specification of the Poisson regression we fit; we further take the reciprocal, since the mean of an Exponential random variable is the reciprocal of its rate. Second, the coefficient estimate for the dimension of 'tree in Poor condition' is integrated into the estimate for the intercept term; for a tree in other conditions, an additional appropriate coefficient estimate needs to be added to the exponent. Third, tree size and risk assessment scores

are standardized in the train set, thus in the calculation, we need to do the same standardization process as when we obtained the train set. Since we are concerned with trees of average size, the tree size variable is standardized to 0 here and omitted.

To help readers understand, we give a similar breakdown in Appendix Section D.5.

Point 2.15 — P.10, Figure 2, caption text "coefficients on spatial coefficients" doesn't make sense. Also, one shouldn't assume the reader knows, based just on the map in Figure 2(a), where "downtown Manhattan" is, Queens, etc. Perhaps annotate Fig. 2(a) accordingly?

Reply: Thank you for pointing out this mistake. We have revised the caption and the writing around this figure so that it should be more accessible to the readers.

Point 2.16 — P.12, very bottom, why were risk assessment scores discretized in this application but not in the model fit in Section 4 (i.e., the one presented in Table 1a)? Why not include the interaction between borough and risk category in that model as well?

Reply: Thank you for this comment. We believe the right modeling choice when calculating reporting rates is to treat risk assessment scores as continuous when trying to understand reporting delays, so that it captures fine-grained differences among incidents.

Discretizing it as DPR does in Section 4, however, helps us align our estimates of the reporting delays with the work order delays: in practice, DPR prioritizes work orders by discretized levels of the risk assessment scores. Following this practice allows us to be sure that the work order delay differences across Boroughs are not due to differences in true risk profiles (as you suggest in an above point) – the agency has determined that these incidents should be equally prioritized. Thus, for this set of analyses, we calculate the work order delays by discretized risk assessment scores in Figure 3 and Appendix D.7, and in order to be consistent, we trained a model on discretized risk assessment scores to estimate the reporting delay.

The results shown in Table 1a, on the other hand, are chosen with interpretability and brevity in mind. To analyze whether this choice matters, we reproduce Table 1a with risk assessment scores discretized, in Appendix Table 13. We feel that the current presentation is best suitable to the purpose of each section, but would be happy to present other results if the review team believes that it's necessary. We note that the two approaches yield qualitatively similar results.

Point 2.17 — (Section 6.1) what exactly are the incident-level covariates? More is said about this on p.15, but e.g., the values that the 'inspection results' variable can take on weren't provided, at least not that I saw. How is location recorded in the data?

Reply: Thank you for pointing these out. In the updated Section 7.2, we give a more detailed summary of the covariates selected. Namely, the incident-level covariates include a risk assessment score given by the inspector, the condition of the tree at the time of inspection, report *Category*, and tree size. Location is encoded in the data as latitude-longitude, which we convert to a census tract via a spatial join. We provide more information on, e.g., possible values for these covariates in Appendix Table 5.

Point 2.18 — (Appendix D) why is Table 5 so small? Please make it the same, standardized (and readable) size as other tables in the data.

Reply: We have fixed Table 5 and ensured all tables in the manuscript are readable.

Point 2.19 — (Remarks on code availability): The CodeOcean capsule ran without any issues (it takes a while, though!). The code (including the README file) and data are generally well organized and it seems like the analysis (with public data) would be pretty easy to reproduce on one's own machine.

Reply: Thank you for your careful reading, including of the CodeOcean capsule!

AR.5 Response to Reviewer 3

Point 3.1 — Summary: In this paper, the authors propose a Poisson estimation method to identify heterogeneous reporting delays using duplicate reports about the same incident. They also provide a theorem to justify their method. Then they apply their method to New York data and Chicago data. They find out that there are substantial spatial disparities in reporting rates after controlling for incident characteristics. Finally, they explain how their method can help people to come up with practical solutions and insights.

The paper is mainly divided into the following parts: introduction, model and research question, empirical method, heterogeneity in NYC and Chicago, discussion of the application of findings, and data processing.

Strength: The authors propose a reasonable model that takes the spatial disparities in reporting rates into account and estimates the reporting rates with duplicate reports. The theory is easy to understand and the authors apply the method to two real datasets to support their statements. The experiments are clear and they also explain how the findings of their methods could potentially help with real-world problems. Overall, the method is reasonable in solving the problem that the authors aim at.

Reply: Thank you for your careful reading, kind words, and, especially, for pushing us on more careful writing of the theoretical statement.

Point 3.2 — Critics: The proof of the main theorem is not carefully written. The author skips some steps and does not define every variable and function clearly. Some indexes also seem to be wrong, which makes the proof a bit confusing to read. Considering that this is the main and only theoretical result in this paper, I think it needs to be written more carefully and clearly. In the paper, the authors use many words to describe something, but sometimes they skip some proofs and it would be better if they write all the derivations clearly.

Reply: Thank you for your suggestions. We have carefully rewritten most of the theoretical results section, making it more clear and concrete. At a high level, we now provide a more accessible version of the theorem in the main text. An extended (but equivalent) version of the theorem with everything more carefully stated in mathematical terms that are commonly used for general Poisson processes is provided in the appendix, along with the proof.

The proof is now also more carefully written, with each step of the derivation explained in words. The proof technique remains the same: we find conditions under which we could separate the full likelihood of the data into conditional likelihoods, and the terms involving the reporting rate λ conform to a Poisson distribution likelihood.

Please see detailed replies to your questions and more details on our changes below and in the manuscript.

Point 3.3 — P6 2nd paragraph: why the stopping time is independent of the process parameter lambda?

Reply: In the current writing, we avoid using the term 'stopping time' to refer to the observation interval start and end. We realized that in defining stopping times, appropriate probability spaces associated with them need to be specified, which reduces the accessibility to the writing. Instead, we state the exact conditions we need for the start and end time to satisfy, in order for the theorem statement to hold.

We briefly discuss here why our choices used in the empirical section satisfy conditions (a) and (b) in the statement of Theorem 1. For the start of the observation period S, we use the time of the first report. Conditional on observing the first report and its time, denoted by $\tilde{t}_1$, we have $S = \tilde{t}_1$ with probability 1, and thus its distribution is independent of the reporting rate λ, satisfying condition (a). For the end of the observation period E, we use the minimum of inspection/work order time, and a fixed time period after the first report. In practice, though inspection and work order times may be dependent on the *number* of reports observed (e.g., agencies tend to respond faster when there are more reports about the same incident), conditional on the reports observed in the past, exactly when such inspections or work orders actually take place is not regulated by the reporting process; rather, it depends on various factors internal to the agency (e.g., how many inspectors are available), and thus independent of the reporting rate λ. Adding a minimum with a fixed time maintains its independence of λ.

Point 3.4 — P7 4th paragraph: why do you choose $\bar{T}$ in this way, and is there a criterion to choose this?

Reply: We require the end of the interval to be before the actual resolution time of the incident. The addition of $\bar{T}$ is to help ensure this condition, in cases where an inspection or work order is not logged or not representative of the incident resolution time. There is no gold standard in choosing $\bar{T}$: larger $\bar{T}$ lets us use more data, but risk underestimating the reporting rate if it extends beyond the true resolution time, while smaller $\bar{T}$ restricts the size of the data we can use in training. For this reason, we evaluate different choices of $\bar{T}$ and report the results of each choice in the Appendix.

Point 3.5 — P7 6th paragraph: why the reporting rate is defined as equation (4)? Can you explain the reason?

Reply: This definition has mainly two benefits. First, the parameter space for a Poisson process is $\lambda \in (0, \infty)$, this definition naturally satisfies this condition; second, by defining λ like this, the log-likelihood of the data as a function of α and β only involves evaluating $\alpha + \beta^T\theta$ and $\exp(\alpha + \beta^T\theta)$, which are easy to compute.

It is worth noting that such a definition for Poisson regressions is widely used in modeling count data (see, e.g., Section 3.2 in Regression Analysis of Count Data by Cameron and Trivedi) and built into some off-the-shelf estimation packages (e.g., Generalized Linear Models within Statsmodels in

Python[9].) Adopting this approach thus further allows us to follow best practices – and, in our new simplified code repository, it allows us to use Statsmodels to fit the model.

Point 3.6 — P8 3rd paragraph: the authors mention that there can be other specifications. Can you give some examples and explain the pros and cons?

Reply: Under our current theorem, other specifications are also possible.

For the interval start time, an alternative would be $\epsilon > 0$ days after the first report. Compared to directly using the first report time, this of course would mean that we are not using the data available to us as effectively.

For the interval end time, a trivial alternative in our setting would be $x > 0$ days after the first report. We discuss why this alternative is data-inefficient in footnote 5 – for this reason, we choose the minimum of such a constant and the data available to us (inspection and work order times).

Point 3.7 — P22 4th paragraph, the index is confusing, what is interval start, is it $\tilde{t}_i$? If t_i^0 is the time between the interval start and the first report, why $\tilde{T}_i$ is equal to the sum of t_i^m without $m = 0$?

Reply: Thank you for pointing out this concern. We have rewritten the proof and believe this concern is no longer present.

Point 3.8 — P22 1st equation 2nd line: why the sum in the first bracket is over subscript j? Are you summing over different incidents?

Reply: Thank you for pointing out this concern. We have rewritten the proof and this summation is no longer needed.

Point 3.9 — P22 How do you marginalize out t_i to obtain the final result?

Reply: Thank you for pointing out this concern. In our updated proof, we no longer need to marginalize out the observed reporting times. In fact, we find that the likelihood of the data does not depend on the exact realizations of these reporting times under a homogenous Poisson process model, but rather only depends on the observation interval length $e_i - s_i$. Thus, marginalizing out these reporting times is no longer needed.

Point 3.10 — Would it be possible to deal with the case where the report rate changes over time?

Reply: We conjecture that results similar to what we obtain in Theorem 1 should extend to non-homogenous Poisson process models, but the likelihood function would further depend on the exact realizations of the reporting times (jump times), and would be substantially more complicated, as we would now be learning a reporting rate function $\lambda(\tau)$ for when the incident is aged $\tau > 0$, as opposed to one number. In other words, we expect that the identification theoretical results would extend, but empirical estimation given the missing data challenge would be difficult.

As empirical evidence in Appendix Figure 7 suggests that reports come in at roughly the same rate, independent of how long it has been after the first report, we focus on the homogenous case in our manuscript, which is more accessible to present and benefits the discussion on its association

[9] `https://www.statsmodels.org/dev/glm.html`

with covariates of interest; we leave this conjecture for future work. In the interest of precision, we have removed any statements about non-homogeneous processes in the main text.

Point 3.11 — What would be a future direction, and what's the limitation?

Reply: Thank you for this question. We now have an extended discussion of these points in Section 5 (discussion and conclusion) of the paper, including in what other contexts others can (and cannot) apply our methods.

Point 3.12 — (Remarks on code availability): Properly documented code.

Reply: Thank you again for your careful reading, including of the code.

References

[1] Gabrielle Kruks-Wisner. Seeking the local state: gender, caste, and the pursuit of public services in post-tsunami india. *World Development*, 39(7):1143–1154, 2011.

[2] New York City Emergency Management. Nyc emergency management issues travel advisory tuesday morning through tuesday night, 2020. URL `https://www.nyc.gov/site/em/about/press-releases/20200803_pr_nycem-issues-travel-advisory.page`. Accessed: 2023-09-01.

[3] Daniel T. O'Brien. *The Urban Commons: How Data and Technology Can Rebuild Our Communities*. Harvard University Press, December 2018. ISBN 978-0-674-97529-3. Google-Books-ID: JttwDwAAQBAJ.

[4] Daniel Tumminelli O'Brien, Robert J. Sampson, and Christopher Winship. Ecometrics in the Age of Big Data: Measuring and Assessing "Broken Windows" Using Large-scale Administrative Records. *Sociological Methodology*, 45(1):101–147, August 2015. ISSN 0081-1750. doi: 10.1177/0081175015576601. URL `https://doi.org/10.1177/0081175015576601`. Publisher: SAGE Publications Inc.

[5] Daniel Tumminelli O'Brien, Dietmar Offenhuber, Jessica Baldwin-Philippi, Melissa Sands, and Eric Gordon. Uncharted Territoriality in Coproduction: The Motivations for 311 Reporting. *Journal of Public Administration Research and Theory*, 27(2):320–335, April 2017. ISSN 1053-1858. doi: 10.1093/jopart/muw046. URL `https://doi.org/10.1093/jopart/muw046`.

Decision Letter, first revision:

Date: 2nd November 23 10:58:47
Last Sent: 2nd November 23 10:58:47
Triggered By: Fernando Chirigati
From: fernando.chirigati@us.nature.com
To: ng343@cornell.edu
CC: computationalscience@nature.com
BCC: fernando.chirigati@us.nature.com
Subject: AIP Decision on Manuscript NATCOMPUTSCI-23-0304B
Message: Our ref: NATCOMPUTSCI-23-0304B

2nd November 2023

Dear Dr. Garg,

Thank you for submitting your revised manuscript "Quantifying Spatial Under-reporting Disparities in Resident Crowdsourcing" (NATCOMPUTSCI-23-0304B). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Computational Science, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please note that, once we send our requirements, we will likely request for a quick turnaround (approximately 5 days) for the revision to be submitted to us, as we would like to publish your paper before the end of the year. Please let us know if you have any questions or if you think this won't be possible.

TRANSPARENT PEER REVIEW
Nature Computational Science offers a transparent peer review option for original research manuscripts. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. **Please remember to choose, using the manuscript system, whether or not you want to participate in transparent peer review.**
Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer

names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

Thank you again for your interest in Nature Computational Science. Please do not hesitate to contact me if you have any questions.

Best,
Fernando

--
Fernando Chirigati, PhD
Chief Editor, Nature Computational Science
Nature Portfolio

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance: https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Reviewer #1 (Remarks to the Author):

I have now reviewed the revised manuscript, and I commend the authors on a thorough and effective revision, which I believe has resulted in an excellent paper and strong contribution to the science of administrative data. Congratulations! I was especially impressed by the use of the hurricane as essentially an instrumental variable or exogenous event to validate.
I only have two remaining comments:
--Social scientists and public administrators who come across this paper are going to be concerned about the validation question from the outset, so it would be good to state prominently (maybe even in more than one place if that makes sense) that this will be reported in Section 6.
--My point about the measurement strategy employed by O'Brien and colleagues was less about the semantics and more about the fact that they also used time elapsed as an indicator of tendency to report. In O'Brien et al.'s (2015) Sociological Methodology paper (and the re-tellings in The Urban Commons (Chapter 2; 2018) and Urban Informatics (Chapter 6; 2022; ui.danourban.com), they measure the time between the identification of a street light outage and the date on which it was reported by a member of the public as one indicator of "custodianship," which is described as a subcomponent of "civic response rate."

Again, congratulations on a great piece of work.

Reviewer #2 (Remarks to the Author):

The authors have done an excellent job of addressing all my comments and concerns with their initial submission. From reading their responses to the other reviewers, it seems they have done a thorough job in responding to those comments as well.

Reviewer #3 (Remarks to the Author):

After reading the revised manuscript, I believe the concerned raised has been addressed.

Author Rebuttal, first revision:

nature research

Round 2 Authors' response: "Quantifying Spatial Under-reporting Disparities in Resident Crowdsourcing"

Dear Editorial Team,

Thank you for the careful read of our revised paper. We are grateful to the editorial team and reviewers for their time and thoughtful comments. Based on the review team's feedback, this draft has two minor changes, as we detail in response to Reviewer 1.

AR.1 Response to Reviewer 1

Point 1.1 — Social scientists and public administrators who come across this paper are going to be concerned about the validation question from the outset, so it would be good to state prominently (maybe even in more than one place if that makes sense) that this will be reported in Section 6.

Reply: Thank you for your suggestion. We now point the readers to the validation section in a prominent place in the introduction section, with the following sentence:

> *The accuracy of these estimates are further validated in Section 5.2.3.*

It is worth noting that, based on suggestions by the journal editor, we now cut down the validation section in the main text to only include the part about validation using storms (which appears to be relatively more convincing based on reviewers' feedback), and delayed the rest of that section to the Supplementary Information.

Point 1.2 — My point about the measurement strategy employed by O'Brien and colleagues was less about the semantics and more about the fact that they also used time elapsed as an indicator of tendency to report. In O'Brien et al.'s (2015) Sociological Methodology paper (and the re-tellings in The Urban Commons (Chapter 2; 2018) and Urban Informatics (Chapter 6; 2022; ui.danourban.com), they measure the time between the identification of a street light outage and the date on which it was reported by a member of the public as one indicator of "custodianship," which is described as a subcomponent of "civic response rate."

Reply: Thank you for pointing us to these literature. We have revised our writing in three places, and believe the current writing better reflect the relationship of our work with the literature as a result. These changes are detailed below.

First, in the introduction section in the main text, in addition to citing these works, we now accompany them with the description

and even those who do submit reports do so with different delays,

as motivation of our work.

In Supplementary Information Section 1, as part of the detailed literature review on differential underreporting, we mention that

> *O'Brien et al. [29] and O'Brien [28] measure differences in the delay of reporting, as an indicator of the citizens' custodianship of their environment. They do so by measuring the delay from when researchers observe the incident during a street audit, to when the incidents are reported by the public.*

Perhaps the most relevant part of our work to the literature you pointed to is in Supplementary Information Section 4.8.4, where we compare reporting rates with measures of voter participation. We now open this section with the following paragraph:

> *The disparities in reporting behavior along socioeconomic variables highlight individual-level behavioral heterogeneity in resident crowdsourcing, and civic engagement at large. For example, O'Brien et al. [29] use the reporting delay as part of their measure for the 'civic response rate', that relates to other forms of activities as well. The level of civic engagement can be measured by many other means, chief among which is participation in political voting. In this section, we validate our coefficient estimates on socioeconomic variables using this idea.*

nature research

Final Decision Letter:

Date: 13th November 23 17:09:42
Last Sent: 13th November 23 17:09:42
Triggered By: Fernando Chirigati
From: fernando.chirigati@us.nature.com
To: ng343@cornell.edu
CC: computationalscience@nature.com
BCC: fernando.chirigati@us.nature.com
Subject: Decision on Nature Computational Science submission NATCOMPUTSCI-23-0304C
Message: 13th November 2023

Dear Dr. Garg,

I am delighted to tell you that your manuscript NATCOMPUTSCI-23-0304C has been accepted for publication in Nature Computational Science.

As discussed, due to the exceptional nature of your work, we will publish your paper on an accelerated schedule. **Please carefully review the details below and contact us immediately at computationalscience@nature.com if you have any travel plans or other conflicts that may make you unable to respond to us for the next 5-7 days.**

In approximately 2 business days you will receive a link to choose the appropriate publishing options for your paper and complete the appropriate grant of rights necessary to publish your work. As it is vital that this process not be delayed, we strongly encourage you to whitelist the email address do-not-reply@springernature.com to ensure that this message is received.

You will receive a link to your electronic proof via email with a request to make any necessary corrections as soon as possible. You will find that we have made minor changes to enhance the clarity of the text and to ensure that your paper conforms to the journal's style so we ask that you review these proofs carefully to ensure that we have not inadvertently introduced errors or altered the sense of your text in any way.

Please return your proof within 24 hours of receiving it. If you have any questions about your proofs or anticipate any delays please contact rjsproduction@springernature.com immediately.

Once a publication date is set for your paper, the Springer Nature press office will be in touch with the full embargo details. We request that you do not send out your own publicity or contact any journalists until you hear from us that the paper has a confirmed publication date.

If you would like to inform your Public Relations or Press Office about your paper, we suggest that you do so immediately to allow them as much time as possible to prepare an appropriate press release and organize publicity if they choose to do so. Please include your manuscript tracking number NATCOMPUTSCI-23-0304C and the name of the journal, which they will need if they contact our press office.

Please note that Nature Computational Science is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com.

If you have not already done so, we strongly recommend that you upload the step-by-step protocols used in this manuscript to the Protocol Exchange. Protocol Exchange is an open online resource that allows researchers to share their detailed experimental know-how. All uploaded protocols are made freely available, assigned DOIs for ease of citation and fully searchable through nature.com. Protocols can be linked to any publications in which they are used and will be linked to from your article. You can also establish a dedicated page to collect all your lab Protocols. By uploading your Protocols to Protocol Exchange, you are enabling researchers to more readily reproduce or adapt the methodology you use, as well as increasing the visibility of your protocols and papers. Upload your Protocols at www.nature.com/protocolexchange/. Further information can be found at www.nature.com/protocolexchange/about .

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

An online order form for reprints of your paper is available at https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

Best,
Fernando

--
Fernando Chirigati, PhD
Chief Editor, Nature Computational Science
Nature Portfolio

P.S. Click here if you would like to recommend Nature Computational Science to your librarian - this will link directly to the Recommend page.

http://www.nature.com/subscriptions/recommend.html#forms

** Visit the Springer Nature Editorial and Publishing website at www.springernature.com/editorial-and-publishing-jobs for more information about our career opportunities. If you have any questions please click here.**