A Bayesian Spatial Model To Correct Under-Reporting in Urban Crowdsourcing

Gabriel Agostini, Emma Pierson†, Nikhil Garg†

Cornell Tech, New York, United States
gsagostini@infosci.cornell.edu, emma.pierson@cornell.edu, ngarg@cornell.edu

Abstract

Decision-makers often observe the occurrence of events through a reporting process. City governments, for example, rely on resident reports to find and then resolve urban infrastructural problems such as fallen street trees, flooded basements, or rat infestations. Without additional assumptions, there is no way to distinguish events that occur but are not reported from events that truly did not occur–a fundamental problem in settings with positive-unlabeled data. Because disparities in reporting rates correlate with resident demographics, addressing incidents only on the basis of reports leads to systematic neglect in neighborhoods that are less likely to report events. We show how to overcome this challenge by leveraging the fact that events are *spatially correlated*. Our framework uses a Bayesian spatial latent variable model to infer event occurrence probabilities and applies it to storm-induced flooding reports in New York City, further pooling results across multiple storms. We show that a model accounting for under-reporting and spatial correlation predicts future reports more accurately than other models, and further induces a more equitable set of inspections: its allocations better reflect the population and provide equitable service to non-white, less traditionally educated, and lower-income residents. This finding reflects heterogeneous reporting behavior learned by the model: reporting rates are higher in Census tracts with higher populations, proportions of white residents, and proportions of owner-occupied households. Our work lays the groundwork for more equitable proactive government services, even with disparate reporting behavior.

1 Introduction

Urban crowdsourcing is key to identifying and resolving problems such as fallen street trees and flooded basements, in both emergency and daily contexts. For example, New York City's 311 system received over 3 million service requests in 2021 (NYC Open Data 2023). However, reporting is *heterogeneous* – different neighborhoods, even when facing similar problems, report problems at different rates (Liu, Bhandaram, and Garg 2023; Kontokosta, Hong, and Korsberg 2017; Minkoff 2016; O'Brien et al. 2017), and under-reporting often correlates with socioeconomic factors such as race, ethnicity, and income. If agencies primarily address incidents that are reported, then under-reporting leads to downstream disparities. This mistargeting of resources, especially when it results in inequity, is a substantial concern and is a stated priority research area for the Federal Emergency Management Agency (FEMA) in the United States (Federal Emergency Management Agency 2023).

While previous work has quantified the magnitude of reporting disparities, they do not estimate the probability that an event truly occurred at each location, which is essential for resource allocation. A challenge in estimation is that, without further assumptions, *unreported* incidents cannot be distinguished from incidents which *did not occur*. This is analogous to *positive-unlabeled* (PU) machine learning (Liu et al. 2003; Shanmugam and Pierson 2021), where datapoints are either labeled positive or unlabeled, and the latter group consists of both true positives and negatives.

PU learning problems are unsolvable without further assumptions on the data generating process (Bekker and Davis 2020). How do we make progress? Our insight is that many urban phenomena are *spatially correlated*, and we can use this correlation to distinguish under-reporting from true lack of event occurrence. For example, in our empirical application, we use reporting data for *flooding* after a storm. If an area does not report flooding but all of its spatial neighbors do, that area is likely to have experienced flooding (but not reported it); conversely, if no neighbors report flooding, that area is not likely to have experienced flooding. Spatial correlation departs from the standard PU learning setup, where the data points are assumed to be independent.

To encode spatial correlations, we build on top of a spatial Bayesian model developed in the ecology literature (Spezia, Friel, and Gimona 2018). The model uses a latent indicator variable at each location to encode whether an event truly occurs there; latent variables at adjacent locations are spatially correlated according to an Ising model (Onsager 1944). If an event does occur, the probability it generates an (observed) report varies as a function of location demographics. In semi-synthetic simulations, we show that the model infers where events have truly occurred more accurately than baseline models. For example, the model significantly outperforms a Gaussian Process (GP) in terms of AUC with an improvement of 0.14.

Using this model, we develop a novel framework to iden-

*†Co-senior author

tify non-reported events in urban crowdsourcing and show it leads to more efficient and equitable resource allocation. We apply the framework to street flooding reports after storms in New York City obtained from public 311 data (NYC Open Data 2023). NYC uses such report data to allocate post-storm resources, such as addressing wood debris, clogged catch basins, or building water leaks (City of New York 2023). Storms cause severe damage; Hurricane Ida in 2021 was responsible for the largest rainfall hour in the city's history, over 7 billions of dollars in damage to infrastructure and the transportation system, and at least 13 deaths, heavily skewed along socioeconomic lines (Newman 2021).

Our framework—which accounts for both heterogeneous under-reporting and spatial correlation—outperforms baseline approaches in terms of *efficiency*: using report data immediately after the storm, it better predicts reports made days later. Such prediction can facilitate a more timely allocation of resources. Our framework also leads to more *equitable* allocation of inspections: the allocations are more in line with population proportions, as opposed to other models that are likely to rate minority and lower-income neighborhoods as lower priority for inspections, due to under-reporting. Finally, we show how to pool model estimates *across* storms to learn historical heterogeneous reporting patterns: we find that reporting rates are higher in Census tracts with higher median incomes, proportions of white residents, and proportions of owner-occupied housing.

Overall, our work (a) leverages spatial correlation in a Bayesian machine learning model to overcome the positive-unlabeled challenge in crowdsourcing; (b) develops a framework to validate and apply the model for more efficient and equitable emergency management response to a given storm event, and to pool reports across storms to identify under-reporting patterns; (c) applies the framework to real-world data, showing that it can substantially improve both the efficiency and equity of government responses to crowdsourcing. Our framework can improve responses in other contexts with spatial correlation, such as in public health and power outages. Together, we provide and validate a novel approach to resource *allocation* in the presence of under-reporting.

We further provide an open-source Python implementation of our approach for application in other contexts[1].

2 Related Work

Our work relates to multiple threads of prior literature from PU Learning, Bayesian methods in ecology, flood prediction, and urban crowdsourcing.

Our setting is one with **positive-unlabeled** data. Without further assumptions, even the proportion of true positive points—the *prevalence*—is unidentifiable because a positive-class, unlabeled data point is indistinguishable from a negative-class point. Hence, PU learning methods *must* make further assumptions (Bekker and Davis 2020); e.g., a common assumption is that each true positive point has the same uniform probability of being labeled positive (Elkan and Noto 2008). Even this strong assumption, which often does not hold in real-world settings where the labeling probability is non-uniform (e.g. when reporting is heterogeneous), is not sufficient. In contrast, our work overcomes the challenge by leveraging *spatial correlation*.

Methodologically, our work builds on approaches from the **ecology** literature, which seeks to count animal populations in the presence of detection errors (Heikkinen and Hogmander 1994; Sicacha-Parada et al. 2021; Della Rocca and Milanesi 2022; Xu et al. 2023). Santos-Fernandez et al. (2021) fit a Bayesian model to correct for misreporting errors in coral detection, leveraging spatial correlation and individuals that analyze coral in multiple locations. Most relevant is work by Spezia, Friel, and Gimona (2018). Their model assumes that the true probability of animal species presence is described by an Ising model, and observed presence is described by a reporting process. We build a framework to effectively use and validate this approach in urban crowdsourcing, showing that the model is predictive of future reports, can guide equitable resource allocation, and can be pooled across multiple events.

This work's specific empirical application – urban flood detection – is complementary to the substantial machine learning work on **flood prediction**, using precipitation and seasonal climate information and data from satellites or sensors (see Mosavi, Ozturk, and Chau (2018) for a comprehensive review). Mauerman et al. (2022), for example, use a Bayesian latent variable model to predict seasonal floods in Bangladesh through the reconstruction of historical satellite data. Agonafir et al. (2022) study infrastructure correlates of flooding using 311 reports in NYC. We believe that reporting data is a valuable complementary data source to sensors and satellites, and is especially temporally and spatially granular in urban environments; however, for both efficiency and equity, it is important to quantify and correct heterogeneous under-reporting. Future work could incorporate such outside sensor data into our model. Our approach is also applicable to reporting contexts beyond flooding.

There is a large literature quantifying disparities in **urban crowdsourcing**; a consistent challenge is disambiguating between low reporting rates and low ground truth rates. Liu, Bhandaram, and Garg (2023) show that time-stamped, duplicate reports about the same event can be used to identify the reporting process; O'Brien, Sampson, and Winship (2015) send researchers to neighborhoods to document ground-truth conditions. We contribute a method that leverages spatial correlation and, unlike other methods, *predicts* the probability an event has occurred in each location.

Finally, our work relates to a much broader literature on methods to quantify and compensate for the effects of missing and imperfect data in inequality-related contexts, including healthcare, policing, education, and government inspections (Coston, Rambachan, and Chouldechova 2021; Rambachan et al. 2021; Movva et al. 2023; Franchi et al. 2023; Laufer, Pierson, and Garg 2022; Guerdan et al. 2023; Zink, Obermeyer, and Pierson 2023; Cai et al. 2020; Pierson 2020; Liu, Rankin, and Garg 2024; Balachandar, Garg, and Pierson 2023; Obermeyer et al. 2019; Kleinberg et al. 2018; Zanger-Tishler, Nyarko, and Goel 2023; Jung et al. 2018; Garg, Li, and Monachou 2021; Lakkaraju et al. 2017; Arnold, Dobbie, and Hull 2022). This broader literature considers many

[1] Available at github.com/gsagostini/networks_underreporting/

types of missingness besides the PU-missingness we study here, and many types of identification approaches besides the spatial correlations leveraged here.

3 Model, Inference, and Framework

Our model captures three characteristics common to many urban crowdsourcing systems: (a) the city does not observe *ground truth* data (where incidents actually occurred), only *reports*; (b) there is *under-reporting*, i.e., not all incidents are reported, and under-reporting may be *heterogeneous* across demographic groups; (c) incidents are *spatially correlated*.

Formally, consider a network G with N nodes and adjacency matrix E. Each node i has two binary state variables. First, $A_i \in \{-1, +1\}$ denotes the latent, ground-truth state; second, $T_i \in \{0, 1\}$ denotes the observed, reported state. In the flood setting, $A_i = 1$ if a flood occurred in that node and -1 if not, while $T_i = 1$ denotes that there was a report for flooding at the node. We observe reports T_i and the network G, but not incidents A_i.

Our specific approach follows that of Spezia, Friel, and Gimona (2018). **Ground truth states** $A_1, \dots A_N$ are generated according to an Ising model with two real-valued parameters, θ_0 and θ_1, controlling the event *incidence rate* and spatial *correlation* respectively. The probability distribution of the vector $\vec{A} \in \{\pm 1\}^N$ is:

$$\Pr(\vec{A}) = \frac{\exp\left(\theta_0 \sum_i A_i + \theta_1 \sum_{i,j} A_i A_j \cdot E_{ij}\right)}{\mathcal{Z}(\theta_0, \theta_1)} \quad (1)$$

where $\mathcal{Z}(\theta_0, \theta_1)$ is an intractable *partition function* ensuring the distribution is normalized. As proven by Besag (1974), the conditional distribution for a single node A_i given all other nodes, is, with positive spatially correlation $\theta_1 > 0$:

$$\Pr(A_i = 1 \mid A_k \; \forall k \neq i) = \frac{1}{1 + \exp\left(-2\left(\theta_0 + \theta_1 \sum_j A_j \cdot E_{ij}\right)\right)} \quad (2)$$

A **report** at node i only depends on the incident state at i and a *reporting rate* ψ_i, i.e.,

$$\Pr(T_i = 1 \mid A_i = 1) = \psi_i. \quad (3)$$

As in PU learning, we assume that there are no false positive reports: $\Pr(T_i = 1 \mid A_i = -1) = 0$.

We fit and compare two models for reporting rates ψ_i:

- With **homogeneous reporting**, $\psi_i = \alpha$ is assumed constant across nodes.
- With **heterogeneous reporting**, report rates ψ_i are a function of demographic factors of node i. That is, given M node-specific features $X_{i1} \dots X_{iM}$,

$$\psi_i = \text{logit}^{-1}\left(\alpha_0 + \sum_{\ell=1}^{M} \alpha_\ell X_{i\ell}\right), \quad (4)$$

  where the coefficients $\alpha_0, \dots \alpha_M$ are learned latent parameters shared across nodes.

We discuss some of the modeling choices in Section 7.

3.1 Inference Procedure

Given reporting data $\{T_i\}$ and a spatial network with known edges $\{E_{ij}\}$, we use a Gibbs sampling MCMC procedure for posterior inference: namely, at each iteration, we draw each latent value from its conditional distribution given the current values of all the other variables. All variables are initialized at random. Model priors are in Appendix F.

We provide code with our submission – we note that we modify the procedure of Spezia, Friel, and Gimona (2018), to speed up inference, such as by jointly sampling some of the parameters within the outer Gibbs routine and implementing dynamic step size optimization. We believe that the public Python code release will enable other practitioners in crowdsourcing or ecology settings to apply such methods.

Sampling θ_0 and θ_1: The conditional distribution of θ_0 and θ_1 given all other variables depends only on $\vec{A}$. We cannot directly compute it due to the presence of the partition function $\mathcal{Z}$ in eq. (1) (Murray, Ghahramani, and MacKay 2006). This normalization constant is intractable, as it must be evaluated for 2^N values of the ground-truth vector $\vec{A}$.

We use the Single-Variable Exchange Algorithm (SVEA) to circumvent this difficulty (Møller et al. 2006). The SVEA is a Metropolis-Hasting type sampling algorithm that introduces an auxiliary variable $\vec{w}$ to cancel two terms with the partition function when computing the acceptance ratio. To do so, $\vec{w}$ must be sampled from the same distribution family as $\vec{A}$. We generate auxiliary variables from the Ising model distribution in eq. (1) using the Swendsen-Wang algorithm with 50 burn-in samples (Swendsen and Wang 1987; Wolff 1989). This is an efficient method to sample from an Ising Model with $\theta_1 > 0$ (Park et al. 2017; Cooper et al. 2000).

Sampling A_i: We sample each of the A_i through Gibbs sampling. The conditional distribution of A_i depends on θ_0, θ_1, T_i, and A_j for j such that $E_{ij} = 1$. If the corresponding $T_i = 1$, then the no false positives assumption leads to $\Pr(A_i = 1 \mid \cdot) = 1$. Otherwise, $\Pr(A_i \mid \cdot)$ is the conditional probability implied by eqs. (2) and (3).

Sampling ψ_i: In the *homogeneous reporting model*, we fit a single parameter α to describe the reporting rate. Given a beta prior, the conditional distribution of α is a beta distribution depending on the numbers of incidents that 1) occurred and are reported and 2) occurred and are not reported.

In the *heterogeneous reporting model*, the conditional distribution for the coefficients $\alpha_0, \dots \alpha_M$ can be found by fitting a Bayesian logistic regression of the reports T_i on the demographic features, restricted to the nodes for which the current latent ground truth parameters are positive ($A_i = 1$). We compute ψ_i following eq. (4).

Sampling Hyper-Parameters. We draw 3 chains with 40,000 samples each, after 20,000 burn-in samples. During burn-in, we optimize the auxiliary variable sampling step size. Further hyper-parameters are described on Appendix F.

Inferred Values. Our procedure gives posterior samples for each of $\theta_0, \theta_1, \alpha_\ell$ (and thus induced ψ_i), and A_i. We use these parameters to calculate $\Pr(A_i = 1|\cdot)$.

3.2 Framework Overview

We use the above model as a foundation and present a novel framework for resource allocation in the presence of under-reporting. Our framework for evaluation and application, implemented in the remainder of the paper, is as follows.

First, we evaluate the approach *semi*-synthetically using the real spatial map, showing model expressivity (that it can generate the full range of storm data), parameter recovery (that estimated parameter point estimates are correct and posteriors are calibrated), and predictive performance (that it correctly identifies unreported events A_i).

Second, we evaluate performance on real storm data, without needing external ground truth data. In particular, we show that, using data in the initial hours after the storm, the model predicts *future* reports in the following days.

Third, we demonstrate the approach's *application* to more efficient and equitable resource allocation. We show that, using the model predictions of ground truth unreported events $P(A_i)$, the resulting allocation better matches the population distribution and in particular does not deprioritize populations with lower reporting rates, unlike other approaches.

Finally, we *pool* parameters across storms (using a Bayes factor approach as described in Appendix A) to robustly characterize under-reporting behavior over time. Leveraging estimates of under-reporting behavior across storms would further aid in proactive resource allocation for future storms.

4 Data and Models

Data. We apply our model to flood reports in New York City. We primarily use Census tracts as nodes in our graph; two tracts are adjacent if they share a border. Minimum-distance edges are drawn whenever needed to make the graph connected (e.g. linking Staten Island to Brooklyn). For node-specific features X, we use demographic variables obtained from Census data, which include population, socioeconomic status, and racial composition measurements.

We use 311 resident reports of street floods. For our main results, we look at reports in the week of September 1st through September 8th, following Hurricane Ida. The 311 reports dataset is publicly available in the NYC Open Data portal, supporting replication of our results. We split the data into train and test reporting periods: we fit our models with reports created until 8% of the census tracts have received at least one report—this threshold is reached around 4 hours after the storm starting time. The remaining reports are used for evaluation. Out of the 2221 census tracts in our network, 177 tracts report at least one flooding incident during the training period, and 346 tracts report during the test period.

When evaluating historical under-reporting, we also consider reports of other floods in New York City. We look at data following the passage of Hurricane Henri (August 2021) and Tropical Storm Ophelia (September 2023). Further details on reports and data are presented in Appendix G.

Models Evaluated. We present results from four models: (a) Our model with *homogeneous reporting*; (b) Our model with *heterogeneous reporting*, with reporting probability varying as a function of demographic features; (c) a *spatial baseline* model in which we predict test-time reporting as the fraction of a node's neighbors that reported during training period; (d) a *Gaussian Process* (GP) baseline model, a standard approach to leverage geographic information. The baselines reflect approaches to incorporating spatial correlation, without explicitly modeling the reporting process as distinct from the incident occurrence process. Other modeling approaches (such as graph neural networks) may also be appropriate and perform well in terms of prediction, but it may be challenging to use such models to separately recover reporting from ground truth processes – we leave such approaches to future work.

5 Semi-Synthetic Simulation Experiments

We verify that our models correctly recover the true parameters and latent states in semi-synthetic data settings where the ground truth parameters and latent state are known. To generate semi-synthetic data, we begin with the real NYC spatial network E and demographic features X, and then for each of the two reporting models (homogeneous and heterogeneous), we generate latent states A and reports T through MCMC sampling assuming the corresponding data generating process. The observed data given to each model is T, E, and (for the heterogeneous reporting model) X. For all experiments, 500 trials were performed; for each trial, we re-sample new values of the latent parameters, re-generate A and T, and run two MCMC chains for inference. Here, we report results when data is drawn according to the heterogeneous model; details and other results, including validating model expressivity, are in Appendix D.

Calibration and Identifiability. We verify that our inference procedure correctly recovers the true data generating parameters, a standard identifiability check (Chang et al. 2021; Pierson et al. 2019). We find an overall high correlation between the recovered parameters and the true, latent parameter values. Correlation values all exceed 0.60. For the regression slope coefficients α_ℓ, which we more directly analyze, correlations of 0.91 and higher are observed. We also verify that our confidence intervals are *calibrated*, another standard check (Wilder, Mina, and Tambe 2021): at each significance level, whether posterior distribution confidence intervals cover the correct fraction of true values.

Predictive Performance. An advantage of semi-synthetic data (in contrast to real data) is that the ground truth latent states A_i are known. We can therefore compare the model's inferred event probabilities $\Pr(A_i)$ to the true latent states A_i. Table 1 reports model AUC (area under the ROC curve), comparing the heterogeneous reporting model to the baselines and the homogeneous model *when data is drawn following the heterogeneous model*. We find that correctly accounting for heterogeneous under-reporting has strong predictive performance, increasing increases AUC by 0.122 from the homogeneous reporting model, by 0.129 from the spatial baseline, and by 0.142 from the GP baseline—all at a significance level of 10^{-4} or less. These values are repeated alongside analogous results for RMSE in Appendix D.

Overall, our semi-synthetic experiments validate that our model correctly recovers the true latent parameters, includ-

Model	AUC	95% CI
Heterogeneous Reporting	0.642	(0.637, 0.647)
Homogeneous Reporting	0.520	(0.515, 0.525)
GP Baseline	0.501	(0.497, 0.505)
Spatial Baseline	0.513	(0.509, 0.518)

Table 1: In simulation, average AUC to predict latent ground-truth A_i according to each model. Nodes with observed training reports are excluded, as they are perfectly predicted by all models by definition. Confidence intervals were obtained through bootstrapping with 10,000 iterates.

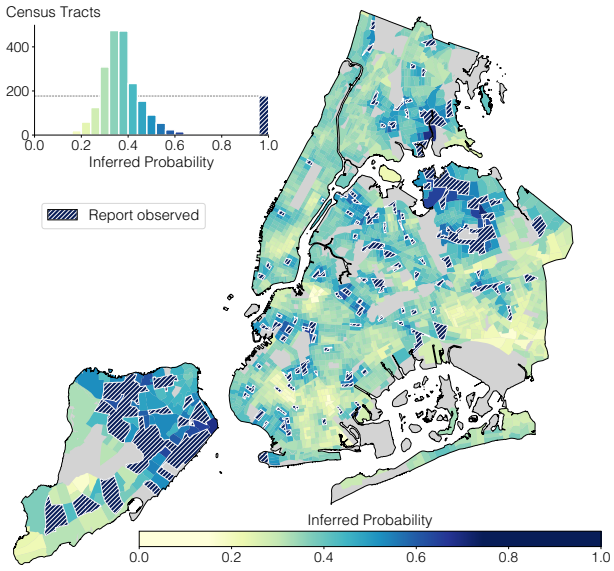

Figure 1: Model-inferred probabilities $\Pr(A_i)$ that each New York City Census tract is flooded after Hurricane Ida, from the heterogeneous reporting model. Hatched lines indicate tracts that reported during the training period.

ing how demographic covariates influence reporting rates. In this way, the use of spatial correlation overcomes the PU learning identifiability challenge. Further, our model outperforms baselines in the ability to infer the unobserved ground truth states A_i. Finally, the simulations demonstrate that, when true reporting processes are heterogeneous, assuming homogeneous under-reporting worsens estimation.

6 Empirical Results

We now apply our framework to NYC 311 data. First, we fit the models to training report data from Hurricane Ida and show that our models outperform the baseline in terms of *efficiency*, i.e., predicting future reports. Then, we show that accounting for heterogeneous under-reporting leads to *more equitable* allocation of resources. Finally, looking at results across multiple storms, we investigate the socioeconomic and demographic features that are mostly associated with heterogeneity in under-reporting.

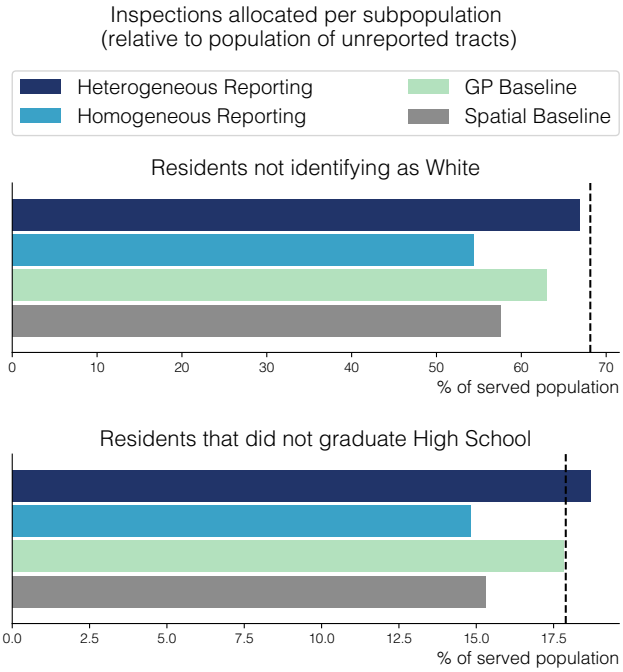

Figure 2: Demographic disparities when allocating resources to 100 census tracts (among those that do not report), using inferred flood probabilities from the four models. The horizontal axes shows the proportion of all residents served by the inspections (i.e. those who reside in the 100 inspected census tracts) who are non-white and do not have a high school degree, computed as a weighted average from the proportions on inspected tracts. Dashed lines represent the total proportion of residents in tracts without a report who are non-white and do not have a high school degree.

Prediction of Floods and Future Reports. Figure 1 shows, for the *heterogeneous reporting model*, the inferred probability of flood by census tract, $\Pr(A_i)$. This map indicates that there is a substantial spatial correlation in reporting, but also that there is likely substantial under-reporting: many tracts have several neighboring tracts with reports, but did not report themselves. The positive spatial correlation is captured by our positive estimate for the spatial correlation parameter θ_1 (0.15, 95% CI (0.08, 0.21)). Convergence diagnostics indicate that the inference procedure converged, with maximum $\hat{R} = 1.03$ for the latent parameters.

We evaluate the four models by how well they predict *future reports.* Unlike the simulation results discussed in Section 5, we cannot evaluate the models in terms of predicting ground truth A_i: we do not have access to *true* flooding events, just reports – lack of such ground truth data indeed motivates the city to use 311 reporting data. We fit the model using train time data and then compare the model estimates of $\Pr(T)$ to future reports T during the test period. This prediction task is decision-relevant because better prediction of future reports would allow the agency to proactively allocate resources (though, as we discuss below, the agency should still be aware of heterogeneous under-reporting).

Model	AUC Estimate	AUC 95% CI	RMSE Estimate	RMSE 95% CI
Heterogeneous Reporting	0.680	(0.646, 0.713)	0.355	(0.338, 0.371)
Homogeneous Reporting	0.682	(0.649, 0.714)	0.360	(0.343, 0.376)
GP Baseline	0.629	(0.595, 0.662)	0.417	(0.400, 0.434)
Spatial Baseline	0.647	(0.616, 0.678)	0.395	(0.377, 0.412)

(a) AUC and RMSE point estimates and confidence intervals for each of the four models.

Model A	Model B	Δ_{AUC}	Δ_{AUC} p-value	Δ_{RMSE}	Δ_{RMSE} p-value
Heterogeneous Reporting	Homogeneous Reporting	-0.002	0.831	-0.004	0.004
	GP Baseline	0.051	0.003	-0.062	$< 10^{-3}$
	Spatial Baseline	0.033	0.009	-0.040	$< 10^{-3}$
Homogeneous Reporting	GP Baseline	0.054	$< 10^{-3}$	-0.058	$< 10^{-3}$
	Spatial Baseline	0.035	$< 10^{-3}$	-0.035	$< 10^{-3}$

(b) AUC and RMSE changes. Two-sided p-values report whether model A and model B differ significantly in performance.

Table 2: Performance metrics for the four models in predicting future reports. Confidence intervals were obtained by bootstrapping the tracts with 10,000 iterates.

Table 2 reports AUC and RMSE for each model, along with bootstrapped 95% confidence intervals and relative improvements. The models accounting for under-reporting achieve better predictive performance than the baseline models that do not. The p-values shown test for positive *differences* between each pair of models. Performance improvements are statistically significant at the 0.05 level for each of our models over the baselines. In practice, an improvement in report prediction – when converted to estimates of ground truth A_i – would translate to more efficient resource allocation, as city governments can anticipate what areas will need attention after a disaster.

Equitable Inspection Allocation. Consider an agency that is allocating resources (such as emergency response, maintenance, or inspections) after storms, in response to reports. Table 2 suggests that our models could lead to more *efficient* allocations, as they are more predictive of future reports after the first day. Here, we analyze how *equitable* these allocations are, under each of the three models we study. We consider the task of allocating a fixed number of resources to Census tracts without reports ($T_i = 0$) – as most tracts receive no report, the agency can allocate some resources to such tracts. We suppose that the agency first infers flood probabilities $\Pr(A_i)$ for each tract in which no report was received, $T_i = 0$. It then allocates resources to the k tracts with the highest inferred probabilities $\Pr(A_i)$.

Figure 2 shows the fraction of resources allocated to non-white residents and residents without a high school degree, when $k = 100$, alongside the population fractions of the tracts without a report. The model accounting for heterogeneous reporting allocates resources more in line with the population distribution, especially in comparison to the homogeneous reporting model which accounts for under-reporting but not *differences* between populations. Note that the model does so while achieving similar predictive performance, as established above.

Additional results in Appendix E show these results are robust to other values of k and for other socioeconomic and demographic factors. The results establish that taking into account heterogeneous under-reporting leads to less deprioritization of non-white and socioeconomically disadvantaged populations (more in line with population distributions).

Socioeconomic Factors of Heterogeneous Reporting. Figure 2 suggests that the heterogenous reporting model identifies and corrects for demographic disparities in under-reporting. We now analyze such differences directly. To understand *persistent* reporting behavior across storms, we also run our model for Hurricanes Ophelia (September 2023) and Henri (August 2021) and pool together the feature coefficients. The pooling method and (qualitatively identical) results for individual storms are in Appendix A.

Using the pooled regression coefficients, we estimate report rates ψ_i for each census tract, with results shown in Figure 3. Demographic patterns emerge – e.g., the Upper West and Upper East Side neighborhoods in Manhattan, with higher median incomes, are estimated to have a higher reporting rate ψ_i than surrounding areas, even though few reports were received in those areas. In contrast, the Bronx (relatively lower incomes) is estimated to have a low reporting rate. Taking a weighted average across neighborhoods, the reporting rate for white populations is, on average: 24% higher than Black populations, 18% higher than Hispanic populations, and 12% higher than Asian populations.

Next, we ask: what demographic factors are associated with under-reporting? Figure 4 shows the pooled posteriors for each coefficient α_ℓ in equation 4: population, income, education, race/ethnicity, age, and household ownership.

We find that there are significant demographic disparities in reporting. As expected, a higher population is associated with higher reporting rates. However, other demographic factors are also associated with reporting rates. Higher proportions of white residents are positively correlated with reporting rate even when controlling for the other five demographic features considered, consistent with the different average report rates per subpopulation shown in Figure 3. Median age and fraction of households occupied by a renter are negatively correlated with higher reporting, suggesting that neighborhoods with older, home-owning pop-

ulations tend to receive reports at a higher rate. These estimates are consistent with prior work on demographic disparities in reporting in the NYC 311 system (Liu, Bhandaram, and Garg 2023). These estimates further explain the results in Figure 2: the heterogeneous reporting model can identify and correct for these reporting disparities when calculating $P(A_i|\cdot)$, the probability that an area is actually flooded (even when no one submitted a report).

Further discussion, as well as similar analyses for other features, can be found in Appendix C.

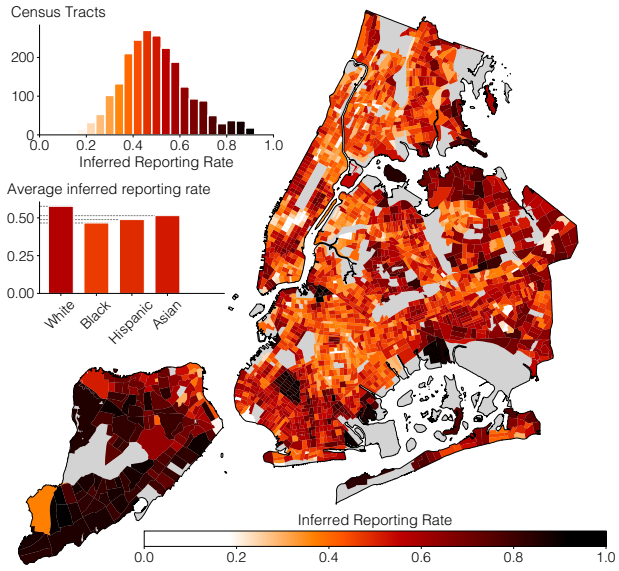

Figure 3: Model-inferred report rates ψ_i per census tract, from the heterogeneous reporting model. The report rates range from near 0.1 to 0.9. Weighted averages of report rates per racial composition are shown in a bar plot.

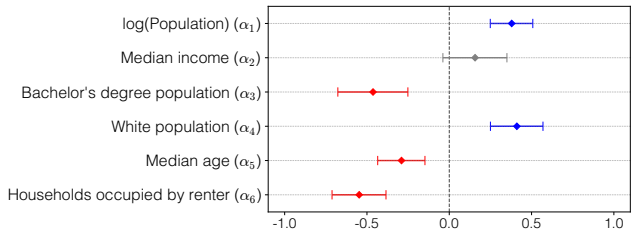

Figure 4: Estimated multivariate coefficients after pooling the three storms. Features were all standardized. Confidence intervals shown, and estimates with insignificant non-zero association colored in grey.

7 Discussion

This work shows the promise of leveraging the *spatial correlation* of incidents to quantify and correct for *under-reporting* in city crowdsourcing systems. Using a case study of flood reporting in New York City, we show that a Bayesian spatial model can accurately recover ground truth under-reporting patterns on semi-synthetic data by leveraging spatial correlation—a challenging task in missing data (PU learning) settings like the one we study. Using this model, we develop a framework for more accurate prediction of future flood reports compared to baselines, and allocation of resources that are more in line with population proportions (not neglecting neighborhoods with larger proportions of non-white and lower-income residents) – improving both efficiency and equity.

Future work might extend the model in several directions. (1) First, we fit the model on binarized reporting data: i.e., whether a location has *any reports*, as opposed to the *count* of reports. While, in our setting, a binary variable is sufficient to capture most of the variation (96.5% of locations have zero or one report) the modeling approach here could plausibly be extended to accommodate count data as well. (2) Second, we assume that flooding is spatially correlated according to an Ising model, and do not explicitly model shared infrastructure and weather patterns. We adopt this model given the well-studied nature of Ising models, which aids estimation. Future work – as driven by the model setting – may choose to modify these choices, though model identifiability may be a challenge. (3) Third, in crowdsourcing settings, we can sometimes incorporate external data, such as (potentially noisy or non-randomly distributed) sensors. That is: if we have access to data from a different source (e.g. sensors to detect floods that are placed sparsely through the city) which reveals that $A_i = 1$ even though $T_i = 0$ for some i, will estimates improve? This question is practically important: often inspections occur in clusters of regions, and it would be important for decision-makers to be able to incorporate the results of negative inspections in the model. (4) Fourth, 311 data is richer than data in other under-reporting settings (such as in ecology) due to spatiotemporal correlations in reporting rates across different incident types or different events of the same type. For example, report rates for flood events may be related to report rates for pest infestations or noise complaints, as both may be related to socioeconomic or other factors. We further note that we can use the pooled estimates of reporting rates across storms when new storms occur, leading to faster, more efficient and equitable allocation. All of these directions represent important opportunities for future work.

Acknowledgments

The authors thank Sidhika Balachandar, Serina Chang, Zhi Liu, Allison Koenecke, Matt Franchi, and Arkadiy Saakyan for feedback. This research was supported by a Meta research award, Google Research Scholar award, a Cornell Tech Urban Tech grant, NSF CAREER #2142419, a CIFAR Azrieli Global scholarship, a LinkedIn Research Award, and the Abby Joseph Cohen Faculty Fund.

References

Agonafir, C.; Pabon, A. R.; Lakhankar, T.; Khanbilvardi, R.; and Devineni, N. 2022. Understanding New York City street

flooding through 311 complaints. *Journal of Hydrology*, 605: 127300.

Arnold, D.; Dobbie, W.; and Hull, P. 2022. Measuring racial discrimination in bail decisions. *American Economic Review*, 112(9): 2992–3038.

Balachandar, S.; Garg, N.; and Pierson, E. 2023. Domain constraints improve risk prediction when outcome data is missing. *NeurIPS ML4H Symposium*.

Bekker, J.; and Davis, J. 2020. Learning from positive and unlabeled data: a survey. *Machine Learning*, 109(4): 719–760.

Besag, J. 1974. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2): 192–225.

Cai, W.; Gaebler, J.; Garg, N.; and Goel, S. 2020. Fair allocation through selective information acquisition. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 22–28.

Chang, S.; Pierson, E.; Koh, P. W.; Gerardin, J.; Redbird, B.; Grusky, D.; and Leskovec, J. 2021. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589(7840): 82–87.

City of New York. 2023. Severe Weather. https://portal.311.nyc.gov/article/?kanumber=KA-03457. Accessed: 2023-07-05.

Cooper, C.; Dyer, M. E.; Frieze, A. M.; and Rue, R. 2000. Mixing properties of the Swendsen–Wang process on the complete graph and narrow grids. *Journal of Mathematical Physics*, 41(3): 1499–1527.

Coston, A.; Rambachan, A.; and Chouldechova, A. 2021. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*, 2144–2155. PMLR.

Della Rocca, F.; and Milanesi, P. 2022. The New Dominator of the World: Modeling the Global Distribution of the Japanese Beetle under Land Use and Climate Change Scenarios. *Land*, 11(4): 567. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Elkan, C.; and Noto, K. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, 213–220. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-60558-193-4.

Federal Emergency Management Agency. 2023. Appendix 2: Learning Agenda. https://www.fema.gov/about/strategic-plan/appendices/learning-agenda. Accessed: 2023-07-05.

Franchi, M.; Zamfirescu-Pereira, J.; Ju, W.; and Pierson, E. 2023. Detecting disparities in police deployments using dashcam data. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 534–544.

Garg, N.; Li, H.; and Monachou, F. 2021. Standardized tests and affirmative action: The role of bias and variance. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 261–261.

Guerdan, L.; Coston, A.; Holstein, K.; and Wu, Z. S. 2023. Counterfactual Prediction Under Outcome Measurement Error. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1584–1598.

Heikkinen, J.; and Hogmander, H. 1994. Fully Bayesian Approach to Image Restoration with an Application in Biogeography. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(4): 569–582. Publisher: [Wiley, Royal Statistical Society].

Jung, J.; Corbett-Davies, S.; Shroff, R.; and Goel, S. 2018. Omitted and included variable bias in tests for disparate impact. *arXiv preprint arXiv:1809.05651*.

Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2018. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1): 237–293.

Kontokosta, C.; Hong, B.; and Korsberg, K. 2017. Equity in 311 Reporting: Understanding Socio-Spatial Differentials in the Propensity to Complain. ArXiv:1710.02452 [cs].

Lakkaraju, H.; Kleinberg, J.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2017. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275–284.

Laufer, B.; Pierson, E.; and Garg, N. 2022. End-to-end Auditing of Decision Pipelines. In *ICML Workshop on Responsible Decision-Making in Dynamic Environments. ACM, Baltimore, Maryland, USA*, 1–7.

Liu, B.; Dai, Y.; Li, X.; Lee, W.; and Yu, P. 2003. Building text classifiers using positive and unlabeled examples. In *Third IEEE International Conference on Data Mining*, 179–186.

Liu, Z.; Bhandaram, U.; and Garg, N. 2023. Quantifying spatial under-reporting disparities in resident crowdsourcing. *Nature Computational Science*.

Liu, Z.; Rankin, S.; and Garg, N. 2024. Identifying and Addressing Disparities in Public Libraries
with Bayesian Latent Variable Modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Mauerman, M.; Tellman, E.; Lall, U.; Tedesco, M.; Colosio, P.; Thomas, M.; Osgood, D.; and Bhuyan, A. 2022. High-Quality Historical Flood Data Reconstruction in Bangladesh Using Hidden Markov Models. In Tarekul Islam, G. M.; Shampa, S.; and Chowdhury, A. I. A., eds., *Water Management: A View from Multidisciplinary Perspectives*, 191–210. Cham: Springer International Publishing. ISBN 978-3-030-95722-3.

Minkoff, S. L. 2016. NYC 311: A Tract-Level Analysis of Citizen–Government Contacting in New York City. *Urban Affairs Review*, 52(2): 211–246. Publisher: SAGE Publications Inc.

Mosavi, A.; Ozturk, P.; and Chau, K.-w. 2018. Flood prediction using machine learning models: Literature review. *Water*, 10(11): 1536.

Movva, R.; Shanmugam, D.; Hou, K.; Pathak, P.; Guttag, J.; Garg, N.; and Pierson, E. 2023. Coarse race data conceals disparities in clinical risk score performance. *arXiv preprint arXiv:2304.09270*.

Murray, I.; Ghahramani, Z.; and MacKay, D. J. C. 2006. MCMC for Doubly-Intractable Distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06, 359–366. Arlington, Virginia, USA: AUAI Press. ISBN 0974903922.

Møller, J.; Pettitt, A. N.; Reeves, R.; and Berthelsen, K. K. 2006. An Efficient Markov Chain Monte Carlo Method for Distributions with Intractable Normalising Constants. *Biometrika*, 93(2): 451–458. Publisher: [Oxford University Press, Biometrika Trust].

Newman, A. 2021. 43 Die as Deadliest Storm Since Sandy Devastates the Northeast. *The New York Times*.

NYC Open Data. 2023. 311 Service Requests from 2010 to Present. https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9. Accessed: 2023-07-05.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Onsager, L. 1944. Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition. *Physical Review*, 65(3-4): 117–149. Publisher: American Physical Society.

O'Brien, D. T.; Offenhuber, D.; Baldwin-Philippi, J.; Sands, M.; and Gordon, E. 2017. Uncharted Territoriality in Coproduction: The Motivations for 311 Reporting. *Journal of Public Administration Research and Theory*, 27(2): 320–335.

O'Brien, D. T.; Sampson, R. J.; and Winship, C. 2015. Ecometrics in the Age of Big Data: Measuring and Assessing "Broken Windows" Using Large-scale Administrative Records. *Sociological Methodology*, 45(1): 101–147.

Park, S.; Jang, Y.; Galanis, A.; Shin, J.; Stefankovic, D.; and Vigoda, E. 2017. Rapid Mixing Swendsen-Wang Sampler for Stochastic Partitioned Attractive Models. ArXiv:1704.02232 [cs, stat].

Pierson, E. 2020. Assessing racial inequality in COVID-19 testing with Bayesian threshold tests. *NeurIPS ML4H Workshop*.

Pierson, E.; Koh, P. W.; Hashimoto, T.; Koller, D.; Leskovec, J.; Eriksson, N.; and Liang, P. 2019. Inferring multidimensional rates of aging from cross-sectional data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 97–107. PMLR.

Rambachan, A.; et al. 2021. Identifying prediction mistakes in observational data. *Harvard University*.

Santos-Fernandez, E.; Peterson, E. E.; Vercelloni, J.; Rushworth, E.; and Mengersen, K. 2021. Correcting Misclassification Errors in Crowdsourced Ecological Data: A Bayesian Perspective. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 70(1): 147–173.

Shanmugam, D.; and Pierson, E. 2021. Quantifying Inequality in Underreported Medical Conditions. *arXiv preprint arXiv:2110.04133*.

Sicacha-Parada, J.; Steinsland, I.; Cretois, B.; and Borgelt, J. 2021. Accounting for spatial varying sampling effort due to accessibility in Citizen Science data: A case study of moose in Norway. *Spatial Statistics*, 42: 100446.

Spezia, L.; Friel, N.; and Gimona, A. 2018. Spatial hidden Markov models and species distributions. *Journal of Applied Statistics*, 45(9): 1595–1615.

Swendsen, R. H.; and Wang, J.-S. 1987. Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, 58(2): 86–88. Publisher: American Physical Society.

Wilder, B.; Mina, M.; and Tambe, M. 2021. Tracking Disease Outbreaks from Sparse Data with Bayesian Inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6): 4883–4891.

Wolff, U. 1989. Collective Monte Carlo Updating for Spin Systems. *Physical Review Letters*, 62(4): 361–364. Publisher: American Physical Society.

Xu, L.; Rolf, E.; Beery, S.; Bennett, J. R.; Berger-Wolf, T.; Birch, T.; Bondi-Kelly, E.; Brashares, J.; Chapman, M.; Corso, A.; et al. 2023. Reflections from the Workshop on AI-Assisted Decision Making for Conservation. *arXiv preprint arXiv:2307.08774*.

Zanger-Tishler, M.; Nyarko, J.; and Goel, S. 2023. Risk scores, label bias, and everything but the kitchen sink. *arXiv preprint arXiv:2305.12638*.

Zink, A.; Obermeyer, Z.; and Pierson, E. 2023. Race Corrections in Clinical Models: Examining Family History and Cancer Risk. *medRxiv*, 2023–03.