The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1691–1713
https://doi.org/10.1214/24-AOAS2003

ESTIMATING REPORTING BIAS IN 311 COMPLAINT DATA

BY KATE S. BOXER[1,a], BOYEONG HONG[2,c], CONSTANTINE E. KONTOKOSTA[2,d] AND DANIEL B. NEILL[1,b]

[1]*Machine Learning for Good Laboratory, Center for Urban Science and Progress, New York University,* [a]*kb145@nyu.edu,* [b]*daniel.neill@nyu.edu*

[2]*Marron Institute of Urban Management, New York University,* [c]*bh1555@nyu.edu,* [d]*ckontokosta@nyu.edu*

Systems such as "311" enable residents of a community to report on their environments and to request nonemergency municipal services. While such systems provide an important link between community and government, resident-generated data suffer from reporting bias, with some subpopulations reporting at lower rates than others. Our research focuses on defining the underreporting of heating and hot water problems to New York City's 311 system and developing methods to estimate under-reporting. First, we estimate nonreporting by fitting a latent variable model, which estimates both the probability of an underlying heating problem conditional on building characteristics, and the probability of reporting a problem conditional on population characteristics. Second, we analyze "less-than-expected" reporting: buildings with fewer 311 calls than expected, as compared to similarly-sized buildings with similar estimated problem durations. Together, these analyses determine neighborhoods and neighborhood-level socioeconomic characteristics that are predictive of underreporting of heating and hot water problems. Our approaches can aid government agencies wishing to use resident-generated data to assist in constructing fair public policies.

1. Introduction. Resident-generated data[1] consists of data that are intentionally created by individuals for accessible use by policy-makers and the public for purposes including improved governance and more responsive public services (Meijer and Potjer (2018)). For example, many cities now have "311" systems through which residents can report on their environments, request information, and request nonemergency municipal services through phone calls, texts, web, or other communication modalities (City of New York (2021)). Increasingly, such resident-generated data are used to help make policy decisions and allocate resources, and can be used as data for machine learning models incorporated into automated tools for policy decision support (City of New Orleans (2019)).

Here we focus on New York City's 311 system (NYC 311) (City of New York (2021)), which receives more than 8,000,000 resident-generated reports annually (Kontokosta et al. (2017), Zha and Veloso (2014), Nadeau (2011)). Of these reports, approximately 20% can be considered a "complaint" about city services or conditions, which require additional follow-up by the relevant city agency (Nadeau (2011)). Our research focuses on residential heating and water issues in New York City. NYC's Department of Housing Preservation and Development (HPD), the agency responsible for monitoring housing conditions in NYC, inspects and subsequently issues violations for inadequate heating and hot water at building level. Residential building inspections primarily occur after a 311 complaint is filed through NYC's 311

Received June 2023; revised November 2024.

Key words and phrases. Resident-generated data, citizen-generated data, 311 data, city analytics, latent variable models, reporting bias, positive and unlabeled learning.

[1]Resident-generated data has traditionally been referred to as citizen-generated data. However, we prefer "resident-generated" because individuals who reside in a specific geographic area but do not meet the legal definition of citizenship generate these data as well.

system, stating that at least one residential unit in the building is experiencing a heating and hot water problem.

Despite recent work to increase the accessibility of NYC 311 (Minkoff (2016)), there is a concern that these data are biased due to systematic differences in residents' propensity to report a problem (Kontokosta and Hong (2021), McLafferty, Schneider and Abelt (2020)). Therefore, using resident-generated data at face value to fit decision support tools, without interrogating potential data biases, can result in misallocation of community services and resources and can further reinforce societal biases which may harm under-served populations. These biases are potentially exacerbated because most of New York City's building inspections are conducted reactively, in response to 311 resident complaints, and, therefore, a problem that is not reported to 311 is both less likely to be observed in the data, and less likely to be addressed in a timely manner by the city. Thus, identifying underreporting areas and subpopulations could assist policy-makers in reaching out to these groups to ensure that their housing issues are addressed.

Heating and hot water problems are serious quality-of-life issues which represent the largest category of NYC 311 complaints (35.9% of total complaints). As such, there is a critical need to understand geographic and demographic disparities, both in the frequency of these problems and the probability that they are reported to the city via 311. However, obtaining an unbiased ground truth of heating and hot water issues would require some form of resident surveys, ubiquitous sensing (e.g., Internet of Things), or inspections throughout all or a large random sample of buildings. These solutions are costly, and they often suffer from common survey sampling errors (Kelly and Swindell (2002)) or are nascent initiatives that are inconsistently deployed and unevenly distributed (Kontokosta (2016), Zheng et al. (2014)). Without reliable "gold standard" data, previous studies, such as Minkoff (2016) and White and Trump (2018), identify geographic and demographic variability in 311 complaints but are unable to distinguish whether this variation is due to differences in reporting, differences in the underlying rates of problems, or a combination of the two.

Here we address the challenge of estimating underreporting in 311 data using two distinct and novel methodological approaches. Our first approach estimates "nonreporting" in which a building has a heating or hot water problem during the legally mandated heating period, but no resident of the building places a 311 call about that problem. In this case we have no record in the 311 call log that such a problem has occurred, making it impossible to distinguish whether that particular building failed to report its problem or simply did not have a problem to report. Nevertheless, when aggregating data across many buildings in neighborhoods throughout a city, we can estimate the expected frequency of problems conditional on building characteristics and use this information to estimate the "nonreporting" rate for different neighborhoods and neighborhood-level demographic characteristics. To accomplish this, we design a latent variable model, which we fit with expectation-maximization to estimate the distribution of underlying unreported heating and hot water problems. We then use this model to identify which neighborhood socioeconomic characteristics are associated with nonreporting.

Our second methodological approach estimates "less-than-expected" reporting. When a building places at least one 311 call for a given problem in a given heating season, we can both observe the total number of calls placed as well as estimate the duration of that problem based on the timing of those calls. This allows us to ask a different question: conditional on a problem having occurred and at least one 311 call being placed and controlling for building size (number of residential units) and problem duration, which buildings are placing a lower-than-expected number of calls? To achieve this, we define groupings of buildings of similar size, experiencing a similar length of heating and hot water problems, and rank the normalized call volumes (311 calls per unit per day) within each grouping. We reaggregate these buildings by neighborhood and test whether each neighborhood has a significantly

higher than expected proportion of less-than-expected reporting buildings, adjusting for multiple testing and controlling the overall False Discovery Rate. Furthermore, we fit a linear regression model to determine which neighborhood-level socioeconomic characteristics are associated with less-than-expected reporting.

We apply these methods for estimating "nonreporting" and "less-than-expected reporting" in conjunction and observe that they provide highly consistent results regarding which neighborhoods and subpopulations are underreporting their heating and hot water problems to 311.

The contributions of this work include:

1. We define a taxonomy for underreporting and separate analysis approaches for nonreporting and less-than-expected reporting, which can be easily adapted to a variety of other real-world scenarios, such as crime reporting.
2. To our knowledge, we are the first to develop a latent variable model for reporting bias in 311 complaints. This model integrates both attributes which are assumed independent of reporting behaviors (structural building characteristics) and attributes which are associated with the propensity to report (socioeconomic characteristics) to estimate a distribution of the latent variable, that is, underlying heating and hot water problems.
3. We integrate domain knowledge about the 311 system in our modeling logic to estimate nonreporting and exploit the nuances of the 311 complaint data schema to estimate less-than-expected reporting.
4. Our analyses provide actionable information about NYC neighborhoods with a higher density of underreporting and neighborhood-level socioeconomic characteristics which are associated with underreporting, which can help government and advocacy agencies improve accessibility to 311 and reduce the impacts of unreported heating and hot water problems in New York City.

The remainder of the paper proceeds as follows: Section 2 describes the NYC 311 data used for analysis and our methods for estimating nonreporting and less-than-expected reporting. The results of applying these methods to the NYC 311 data are presented in Section 3, and Section 4 contains a concluding discussion.

2. Materials and methods.

2.1. *Data*. We integrate publicly available New York City and U.S. Census data. The primary data source is the NYC 311 records file of heating and hot water complaints during the months of October through May, which is when building owners are required by law to provide adequate heat throughout the building.[2] We considered all data records representing heating and hot water complaints in residential buildings with two or more units, consisting of 1,636,248 records between 1 October 2010 and 31 May 2018.

The New York City Department of Finance Primary Land Use and Tax Lot Output (PLUTO) dataset and the Mayor's Office of Sustainability Oil Boiler dataset, which together include property attributes such as building class, building size, construction year, boiler type, and boiler age, are used as structural building characteristics. We use the Census Block Group (CBG) as our areal unit of analysis for neighborhood boundaries since it provides the most granular geographical unit of demographic and socioeconomic data provided by the U.S. Census Bureau.

[2]Section 27-2029 of the NYC Administrative Code states that from October 1st to May 31st from 6 a.m. to 10 p.m., if the outdoor temperature falls below 55 degrees Fahrenheit, those living dwellings that are legally required to provide heating must be kept at 68 degrees Fahrenheit or above. Between 10 p.m. and 6 a.m., those living dwellings that are legally required to provide heating must be kept at 55 degrees Fahrenheit or above, when outdoor temperatures fall below 40 degrees Fahrenheit.

TABLE 1
Notation reference table for nonreporting and less-than-expected reporting analyses

Singular instance symbol	Description of a singular instance	All instances set symbol ($\forall j, i, y$)
b_i	Structural building characteristics and boiler data for building i.	b
d_i	Neighborhood-level demographic and socioeconomic characteristics of the Census Block Group that building i resides in.	d
u_i	Number of residential units in building i.	u
X_{iy}	The number of calls placed to 311 from building i during heating season y.	X
C_{iy}	Indicator variable for whether at least one call was placed to 311 from building i during heating season y, that is, $C_{iy} = \mathbf{1}\{X_{iy} > 0\}$.	C
I_{iy}	Indicator variable for whether building i experienced an underlying heating and hot water problem at some point during heating season y.	I
a_{iy}^{j}	Indicator variable for whether a singular unit j in building i called 311 about a heating and hot water problem during heating season y.	a
t_{iy}	Estimate of the number of days building i experienced heating/hot water problems for the heating season y.	t

*The set of indicator variables I contain only positive and unlabeled data, that is, $I_{iy} = 1$ if $C_{iy} = 1$, and I_{iy} is unobserved if $C_{iy} = 0$. The variables a_{iy}^{j} and t_{iy} are unobserved but estimated in our models; the remaining variables are fully observed.

We also include data on voter turnout, using voter registration data from the Board of Elections in the City of New York and voter participation based on the New York State Board of Elections data for the November 2017 election cycle, as a measurement of political and civic engagement.

Our analysis considers all buildings in New York City with two or more residential units, for the heating seasons of 2010–2017, along with information about the number of calls placed during each heating season, structural building characteristics, neighborhood socioeconomic information, and building size (number of residential units). Table 1 contains all the variables and their associated descriptions used in the remainder of this section and Section 3.

2.2. *Method for estimating nonreporting*. The latent variable model discovers the structural characteristics of buildings which are correlated with underlying heating and hot water problems, and the socioeconomic neighborhood characteristics which are associated with at least one resident in a building reporting a heating and hot water problem to 311, conditional on there being an underlying heating and hot water issue. We provide an overview of our motivating problem, with reference to the literature on learning with positive and unlabeled data, in Section 2.2.1. We explain the structure of our model in Section 2.2.2 and describe how we used expectation-maximization to fit the model in Section 2.2.3.

2.2.1. *Motivating problem*. The approach we use to estimate nonreporting is motivated by the paradigm and methods defined in the machine learning subfield of positive and unlabeled learning (PU learning). In traditional classification, both positive and negative examples are used to fit a model. In PU learning the goal is to fit a model which predicts positive and negative instances, using only positive and unlabeled instances of data.[3]

[3]This described paradigm is also known as presence-only data in other fields such as ecology (Ward et al. (2009)).

Estimating nonreporting consists of learning from positive and unlabeled data because the 311 call log creates a record of only positive instances of a heating and hot water problem. All the remaining buildings during a heating season which did not call 311 either did not have a heating or hot water issue, which implies that they are unlabeled negative instances, or had an unreported heating and hot water issue, which implies that they are unlabeled positive instances. More formally, let I_{iy} be the indicator variable for whether building i experienced an underlying heating and hot water problem during heating season y, and C_{iy} be the indicator variable for whether that building reported that problem to 311 (i.e., at least one call was placed) during that heating season. If $C_{iy} = 1$, there is assumed to be an underlying heating and hot water problem and $I_{iy} = 1$. Otherwise, $C_{iy} = 0$, and I_{iy} is unlabeled (and could be either 1 or 0). This logic, which is implicit in PU learning (Bekker and Davis (2020)), supports the assumption in equation (12) that $\Pr(I_{iy} = 1 \mid C_{iy} = 1) = 1$.

There are two assumptions pertaining to the labeling mechanism which are often used in PU learning. The most common assumption, which most recent research utilizes, is that there is a fixed probability of a positive instance being labeled, regardless of that instance's attributes. This is formally called "selected completely at random" (SCAR) (Elkan and Noto (2008)). An alternative assumption, which we make here, is that labeling bias affects the labeling mechanism, and, therefore, an instance's other observed attributes affect the probability of that instance being labeled conditional on that instance having a positive value. This assumption is referred to as selected at random (SAR) (Bekker, Robberechts and Davis (2019)). We adopt the SAR assumption because the probability that an individual unit contacts 311 (given an underlying heating and hot water problem affecting the building) is associated with neighborhood demographic and socioeconomic characteristics which correlate with the propensity of individuals to report these problems (Kontokosta and Hong (2021), Minkoff (2016)); moreover, buildings with different numbers of units would be expected to report problems at different rates. We account for both of these sources of variation in our model below.

Expectation-maximization has been previously employed in fitting models from positive and unlabeled data. Ward et al. (2009) use expectation-maximization to iterate between estimating the unknown labels and using these estimated labels to fit a logistic regression with a case-control adjustment to the intercept. This case-control adjustment requires known baseline frequencies for the labels, $\Pr(I_{iy} = 1)$, and uses the SCAR assumption, both of which make this approach unsuitable for estimating underlying heating and hot problems from resident-generated data. Bekker, Robberechts and Davis (2019) also utilize an expectation-maximization algorithm in a fashion similar to the one we propose. While we never explicitly assign a propensity score (probability of being labeled) to each instance, their propensity function is similar to the ELBO_d component of our lower bound on the log-likelihood (equation (10)). Two differences in our approach are that we use a probability logic which exploits the structure of the 311 reporting process in New York City and our method of calculating the M-step includes fitting another set of variables of building characteristics which relate to the probability of an underlying issue while the E-step uses Bayes' rule.

Another significant difference is that the methods we propose involve using the latent variable model to calculate aggregate neighborhood-level expectations of heating and hot water problems rather than providing building-specific predictions. Therefore, this eliminates the need for certain parameters, such as an observed or estimated label frequency, or the application of further calibration techniques. Kato, Teshima and Honda (2019) propose an approach to learning from PU data with the SAR assumption by defining a density ratio which can be used for partial identification purposes to obtain an unbiased model. Using the class baseline frequencies, they find a threshold for the density ratio to perform classification. This method assumes known label frequencies, which our approach for nonreporting does not.

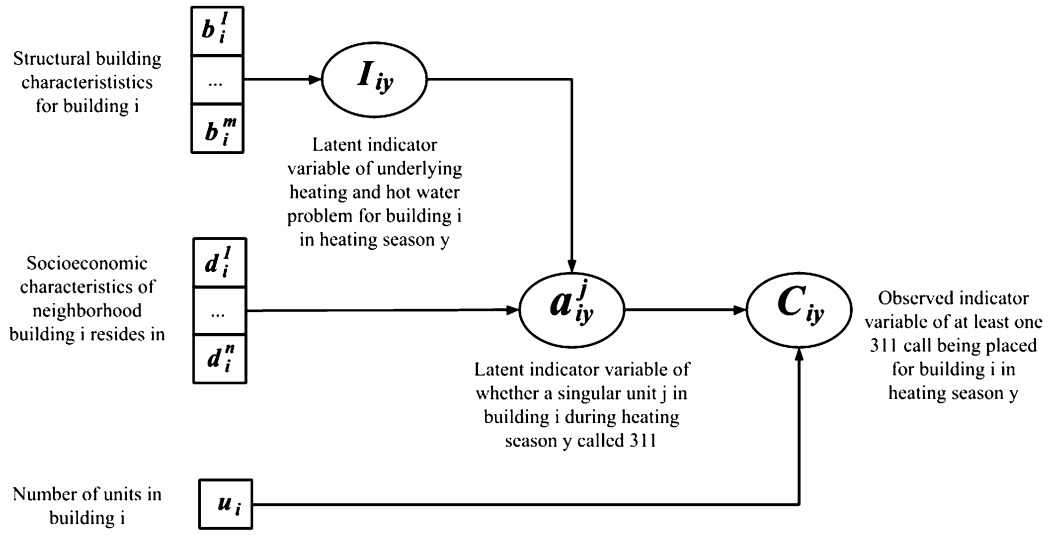

FIG. 1. *Latent variable model for predicting whether at least one* 311 *call was placed from building i during heating season y. The latent variable* I_{iy}*, which represents whether building i experienced an underlying heating and hot water problem at some point during the heating season y, serves as an indicator variable that mediates* $Pr(a_{iy}^j)$ *and* $Pr(C_{iy})$.

While they recommend using a specific mixture proportion estimation method proposed by Ramaswamy, Scott and Tewari (2016) to estimate the label frequencies, we have no reliable means of validating the estimated label frequencies, making this approach ill-suited for estimating nonreporting.

2.2.2. *Structure of the latent variable model.* As shown in Figure 1, our latent variable model defines three probabilities. First, it defines the probability that a building i has an underlying heating and hot water problem during a given heating season y, conditional on its structural building characteristics, $\Pr(I_{iy} = 1 \mid b_i)$. Second, it defines the probability that a singular unit j in building i will call 311 to report a problem during heating season y, conditional on the presence of an underlying heating and hot water problem and the socioeconomic characteristics of the neighborhood in which building i resides, $\Pr(a_{iy}^j = 1 \mid I_{iy}, d_i)$. Third, it defines the probability of at least one unit in building i calling 311 during heating season y conditional on the building having an underlying heating and hot water problem, neighborhood socioeconomic characteristics, and building size, $\Pr(C_{iy} = 1 \mid I_{iy}, d_i, u_i)$.

The probability of a building having an underlying heating and hot water problem for a given heating season takes the form of a logistic regression, as shown in equation (1), to produce a probability that is input into a univariate Bernoulli distribution defined in equation (2),

$$\Pr(I_{iy} = 1 \mid b_i, w_b) = \sigma(w_b \cdot b_i) = \frac{1}{1 + e^{-(w_b \cdot b_i)}}, \tag{1}$$

$$I_{iy} \sim \text{Bernoulli}\big(\Pr(I_{iy} = 1 \mid b_i, w_b)\big). \tag{2}$$

Note that equation (1) assumes the form of w_b and b_i as vectors and computes their inner product $(w_b \cdot b_i)$. We fit w_b, which are the linear model coefficients for the structural building characteristics, as part of the estimation procedure for the latent variable model described below.

Similarly, we define the probability of a singular unit j in building i placing a 311 call to report a heating and hot water issue during heating season y in equation (3). This probability

is also assumed to take the form of a logistic regression,

$$\Pr\big(a_{iy}^{j}=1 \mid I_{iy}, d_i, w_d\big) = I_{iy}\sigma(w_d \cdot d_i) = \frac{I_{iy}}{1+e^{-(w_d \cdot d_i)}}. \tag{3}$$

Note that equation (3) assumes the form of w_d and d_i as vectors and computes their inner product $(w_d \cdot d_i)$. We fit w_d, which are the linear model coefficients for the neighborhood-level socioeconomic characteristics, as part of the estimation procedure for the latent variable model described below. I_{iy} acts as an indicator variable that mediates $\Pr(a_{iy}^{j}=1 \mid I_{iy}, d_i, w_d)$: if there is no underlying heating or hot water problem, $I_{iy}=0$, then the probability $\Pr(a_{iy}^{j}=1 \mid I_{iy}, d_i, w_d)=0$. When there is an underlying heating and hot water problem, $I_{iy}=1$, then $\Pr(a_{iy}^{j}=1 \mid I_{iy}, d_i, w_d)$ functions like a logistic regression with input variables d_i and coefficients w_d.

Finally, we define the probability of at least one call being placed to 311 from building i during heating season y in equation (4), and the resulting indicator variable C_{iy} is drawn from a univariate Bernoulli distribution as defined in equation (5).

$$\Pr(C_{iy}=1 \mid I_{iy}, d_i, u_i, w_d) = 1-\big(1-\Pr\big(a_{iy}^{j}=1 \mid I_{iy}, d_i, w_d\big)\big)^{u_i}, \tag{4}$$

$$C_{iy} \sim \text{Bernoulli}\big(\Pr(C_{iy}=1 \mid I_{iy}, d_i, u_i, w_d)\big). \tag{5}$$

Equation (4) assumes that, conditional on the presence of an underlying heating and hot water problem, each of the u_i units in building i chooses whether or not to call 311 independently, with identical probabilities $\Pr(a_{iy}^{j}=1 \mid I_{iy}, d_i, w_d)$. Thus, the probability that at least one unit in building i calls 311 (and thus the probability that the problem is reported) is $1-(1-\Pr(a_{iy}^{j}=1 \mid I_{iy}, d_i, w_d))^{u_i}$. Given that we are not provided a 311 call log that includes consistent, granular, unit-level information about heating and hot water problems, the purpose of using the latent variables a_{iy}^{j}, while assuming that $\Pr(a_{iy}^{j}=1 \mid I_{iy}, d_i, w_d)$ is identical for all units in a given building during a given heating season, is to incorporate the number of units, u_i, in building i when fitting the coefficients of w_d, as shown in equation (4).

2.2.3. *Fitting the model with expectation-maximization.* For each building i, in each heating season y, the values of C_{iy}, b_i, d_i, and u_i are known from our data containing 311 call logs for that heating season, structural building characteristics, neighborhood socioeconomic characteristics, and building size information respectively. We wish to fit both the linear model coefficients w_b for the structural building characteristics, which are associated with an underlying heating and hot water problem, and the linear model coefficients w_d for the neighborhood socioeconomic characteristics, which are predictors of at least one unit placing a call to 311 if there is underlying heating or hot water problem. Our model also contains an unobserved latent variable I_{iy} representing whether each building has an underlying heating and hot water problem in each heating season. We refer to the variable sets for all buildings for all heating seasons as $C=\{C_{iy}\}$, $I=\{I_{iy}\}$, $b=\{b_i\}$, $d=\{d_i\}$ and $u=\{u_i\}$, as described in Table 1. For simplicity, for the remainder of this section we will default to using the variable set notation for the following equations, with the exception of equations where providing unit-level notation provides more technical clarity. We will estimate the latent variable distributions $\Pr(I_{iy} \mid C_{iy}, b_i, d_i, u_i)$ for all buildings across all heating seasons, which we refer to collectively as $q(I)$. Similarly, for $j \in \{0,1\}$, we use the notation $q(I=j)$ to collectively refer to the latent variable distributions $\Pr(I_{iy}=j \mid C_{iy}, b_i, d_i, u_i)$ for all buildings across all heating seasons.

To fit the coefficients and the latent variable distribution jointly, we use the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin (1977)). This iterative algorithm

alternates between the expectation (E) step, where the latent variable distribution $q(I)$ is estimated given fixed coefficients w_b and w_d, and the maximization (M) step, where the calculated latent variable distribution $q(I)$ from the most recent E-step is used to compute the coefficients w_b and w_d which maximize a lower bound on the complete log-likelihood.

Our joint probability model is $\Pr(C, I \mid b, d, u, w_b, w_d)$, and we use the marginal likelihood to calculate the probability $\Pr(C \mid b, d, u, w_b, w_d)$, where

$$\Pr(C_{iy} \mid b_i, d_i, u_i, w_b, w_d) = \sum_{j \in \{0,1\}} \Pr(C_{iy}, I_{iy} = j \mid b_i, d_i, u_i, w_b, w_d). \tag{6}$$

Therefore our objective is to fit the coefficients w_b and w_d to maximize the log-likelihood,

$$\hat{w}_b, \hat{w}_d = \underset{w_b, w_d}{\operatorname{argmax}} \log\bigl(\Pr(C \mid b, d, u, w_b, w_d)\bigr). \tag{7}$$

We must also calculate the conditional probability over I, $\Pr(I \mid C, b, d, u, w_b, w_d)$.

Following Blei, Kucukelbir and McAuliffe (2017), we define the evidence lower bound (ELBO) on the log-likelihood as

$$\text{ELBO} = \sum_{j \in \{0,1\}} q(I = j) \log\left(\frac{\Pr(C, I = j \mid b, d, u, w_b, w_d)}{q(I = j)}\right). \tag{8}$$

It can be seen easily (using Jensen's inequality) that

$$\begin{aligned}\text{ELBO} &\leq \log \sum_{j \in \{0,1\}} q(I = j) \frac{\Pr(C, I = j \mid b, d, u, w_b, w_d)}{q(I = j)} \\ &= \log \Pr(C \mid b, d, u, w_b, w_d).\end{aligned}$$

Moreover, it follows from the conditional independence assumptions of our model that $\Pr(C, I \mid b, d, u, w_b, w_d) = \Pr(I \mid b, w_b)\Pr(C \mid I, d, u, w_d)$. This allows us to decompose $\text{ELBO} = \text{ELBO}_b + \text{ELBO}_d + \text{ELBO}_q$, where

$$\text{ELBO}_b = \sum_{j \in \{0,1\}} q(I = j) \log \Pr(I = j \mid b, w_b), \tag{9}$$

$$\text{ELBO}_d = \sum_{j \in \{0,1\}} q(I = j) \log \Pr(C \mid I = j, d, u, w_d), \tag{10}$$

$$\text{ELBO}_q = \sum_{j \in \{0,1\}} -q(I = j) \log q(I = j). \tag{11}$$

On each iteration of EM, we first find a $q(I)$ that maximizes the ELBO, when w_b and w_d are fixed in the E-step, and then use that $q(I)$ to find the w_b and w_d, which maximize the ELBO in the M-Step. We alternate between these two steps until convergence:

E-Step

In the E-step of each iteration of the EM algorithm, we compute $\Pr(I \mid C, b, d, u, w_b, w_d)$ by Bayes' Theorem, and set $q(I) = \Pr(I \mid C, b, d, u, w_b, w_d)$.

To see that this step increases the ELBO, we can write

$$\begin{aligned}\text{ELBO} &= \sum_{j \in \{0,1\}} q(I = j) \log\left(\frac{\Pr(C \mid b, d, u, w_b, w_d)\Pr(I = j \mid C, b, d, u, w_b, w_d)}{q(I = j)}\right) \\ &= \sum_{j \in \{0,1\}} q(I = j) \log\bigl(\Pr(C \mid b, d, u, w_b, w_d)\bigr)\end{aligned}$$

$$+\sum_{j\in\{0,1\}} q(I=j)\log\left(\frac{\Pr(I=j\mid C,b,d,u,w_b,w_d)}{q(I=j)}\right)$$
$$=\log\big(\Pr(C\mid b,d,u,w_b,w_d)\big)-\mathrm{KL}\big(q(I),\Pr(I\mid C,b,d,u,w_b,w_d)\big),$$

where KL is Kullback–Leibler divergence and is thus minimized for $q(I)=\Pr(I\mid C,b,d,u,w_b,w_d)$.

To compute $\Pr(I\mid C,b,d,u,w_b,w_d)$ by Bayes' Theorem, we consider two scenarios, one in which at least one heating and hot water complaint is placed to 311 from building i during heating season y, $C_{iy}=1$, and one in which no heating and hot water complaints are placed to 311 from building i during heating season y, $C_{iy}=0$. For the first scenario, we have

$$\Pr(I_{iy}=1\mid C_{iy}=1,b_i,d_i,u_i,w_b,w_d)=1, \tag{12}$$

that is, we assume that if there was a 311 heating and hot water complaint placed from a resident of building i during heating season y, there was an underlying heating and hot water problem (in that building, for that heating season) which generated that call.

For the second scenario, we have

(13)
$$\begin{aligned}&\Pr(I_{iy}=1\mid C_{iy}=0,b_i,d_i,u_i,w_b,w_d)\\&=\frac{\Pr(I_{iy}=1\mid b_i,w_b)\Pr(C_{iy}=0\mid I_{iy}=1,d_i,u_i,w_d)}{\Pr(I_{iy}=1\mid b_i,w_b)\Pr(C_{iy}=0\mid I_{iy}=1,d_i,u_i,w_d)+\Pr(I_{iy}=0\mid b_i,w_b)\Pr(C_{iy}=0\mid I_{iy}=0,d_i,u_i,w_d)}\\&=\frac{\Pr(I_{iy}=1\mid b_i,w_b)\Pr(C_{iy}=0\mid I_{iy}=1,d_i,u_i,w_d)}{\Pr(I_{iy}=1\mid b_i,w_b)\Pr(C_{iy}=0\mid I_{iy}=1,d_i,u_i,w_d)+\Pr(I_{iy}=0\mid b_i,w_b)},\end{aligned}$$

where we have simplified the equation using our modeling assumption that $\Pr(C_{iy}=0\mid I_{iy}=0,d_i,u_i,w_d)=1$. The probabilities $\Pr(I_{iy}\mid b_i,w_b)$ are computed using equation (1) above, and the probability $\Pr(C_{iy}=0\mid I_{iy}=1,d_i,u_i,w_d)$ is computed using equations (3) and (4).

M-Step

In the M-step our goal is to compute the coefficients w_b and w_d which maximize the ELBO, as in equation (7). We note that only the ELBO_b and ELBO_d components of the ELBO are affected by w_b and w_d, respectively, and thus we can write

$$\hat{w_b}=\operatorname*{argmax}_{w_b}\sum_{j\in\{0,1\}} q(I=j)\log\Pr(I=j\mid b,w_b), \tag{14}$$

$$\hat{w_d}=\operatorname*{argmax}_{w_d}\sum_{j\in\{0,1\}} q(I=j)\log\Pr(C\mid I=j,d,u,w_d). \tag{15}$$

To fit the coefficients w_b that optimize ELBO_b, using the current estimate of $q(I)=\{q(I_{iy})\}$, we can use standard methods for fitting a weighted logistic regression $\Pr(I\mid b,w_b)$. Specifically, we create data for the weighted logistic regression by concatenating two versions of our original data. For each building i and each heating season y, we create a record with a positive label $(b_i,I_{iy}=1)$ and a record with a negative label $(b_i,I_{iy}=0)$ and assign them weights $q(I_{iy}=1)$ and $q(I_{iy}=0)$, respectively.

To fit the coefficients w_d that optimize ELBO_d, again using the current estimate of $q(I)$, we design a gradient ascent algorithm (Cauchy et al. (1847)). For a given building i and heating season y, the contribution of that instance to ELBO_d is

$$\begin{aligned}\mathrm{ELBO}_{diy}=&\,q(I_{iy}=1)\log\Pr(C_{iy}\mid I_{iy}=1,d_i,u_i,w_d)\\&+q(I_{iy}=0)\log\Pr(C_{iy}\mid I_{iy}=0,d_i,u_i,w_d).\end{aligned} \tag{16}$$

If $C_{iy} = 1$, then $q(I_{iy} = 1) = 1$, and $\text{ELBO}_{diy} = \log \Pr(C_{iy} = 1 \mid I_{iy} = 1, d_i, u_i, w_d)$. Plugging in the values from equations (3) and (4) above, the corresponding gradient is

$$\begin{aligned}\frac{\partial \text{ELBO}_{diy}}{\partial w_d} &= \frac{\partial}{\partial w_d} \log\left(1 - \left(1 - \frac{1}{1 + e^{-(w_d \cdot d_i)}}\right)^{u_i}\right) \\ &= \frac{d_i u_i e^{w_d \cdot d_i}}{((1 + e^{w_d \cdot d_i})^{u_i} - 1)(1 + e^{w_d \cdot d_i})}.\end{aligned} \tag{17}$$

If $C_{iy} = 0$, then $\log \Pr(C_{iy} \mid I_{iy} = 0, d_i, u_i, w_d) = 0$, and $\text{ELBO}_{diy} = q(I_{iy} = 1) \times \log \Pr(C_{iy} = 0 \mid I_{iy} = 1, d_i, u_i, w_d)$. Plugging in the values from equations (3) and (4) above, the corresponding gradient is

$$\begin{aligned}\frac{\partial \text{ELBO}_{diy}}{\partial w_d} &= \frac{\partial}{\partial w_d} q(I_{iy} = 1) \log\left(\left(1 - \frac{1}{1 + e^{-(w_d \cdot d_i)}}\right)^{u_i}\right) \\ &= -\frac{q(I_{iy} = 1) d_i u_i e^{w_d \cdot d_i}}{1 + e^{w_d \cdot d_i}}.\end{aligned} \tag{18}$$

During each iteration of the gradient ascent algorithm, w_d is adjusted by the partial derivative of ELBO_d with respect to w_d. Letting w_d^k denote the value of w_d on the kth iteration,

$$\begin{aligned}w_d^k &= w_d^{k-1} + \frac{\lambda}{n} \frac{\partial \text{ELBO}_d^{k-1}}{\partial w_d} \\ &= w_d^{k-1} + \frac{\lambda}{n} \sum_{i,y} \frac{\partial \text{ELBO}_{diy}^{k-1}}{\partial w_d},\end{aligned}$$

where λ is the constant learning rate, $\lambda = 0.1$, and n is the number of data instances. We continue updating w_d until $\text{ELBO}_d^k - \text{ELBO}_d^{k-1} \leq 10^{-6}$.

We alternate between the E-Step and M-Step. Each full iteration of performing the E-step followed by the M-step will be referred to as j, and the value of ELBO after each full iteration of EM will be referred to as ELBO^j. The algorithm terminates when $\text{ELBO}^j - \text{ELBO}^{j-1} \leq 10^{-4}$. The final w_b and w_d once the algorithm converges are the fitted coefficients $\hat{w}_b$ and $\hat{w}_d$ found by the EM algorithm for our latent variable model. To obtain confidence intervals for the coefficients $\hat{w}_b$ and $\hat{w}_d$, we fit the latent variable model multiple times on unique data randomly sampled from the original data with replacement, and then calculate the confidence intervals from these bootstrapped samples.

To validate that the proposed fitting method above recovers coefficients that model the relationship of the predictors with whether an underlying heating and hot water problem exists and whether a call is placed to 311 about an underlying heating and hot water problem, we include an analysis in Appendix A of the Supplementary Material (Boxer et al. (2025)). The validation method in Appendix A uses semisynthetic datasets that retain correlation relationships between predictors from our original data, to determine if the latent variable model can recover randomized preset coefficients used to generate the latent variable of an underlying heating and hot water problem and the observed indicator variable of whether a call was placed to 311.

2.3. *Method for measuring less-than-expected reporting.* The latent variable model described above estimates the probability that each building which did not report any heating and hot water complaints to 311 had an unreported heating and hot water problem, $\Pr(I_{iy} = 1 \mid C_{iy} = 0, b_i, d_i, u_i, \hat{w}_b, \hat{w}_d)$. However, buildings with a large number of residential units are likely to have at least one heating and hot water problem in a given heating season

($I_{iy} = 1$) and to have at least one unit report that problem ($C_{iy} = 1$). To gain additional insights into such buildings, one might ask instead about the *number* of 311 heating and hot water complaints X_{iy} made by building i during heating season y, subsetting the analysis to consider only those buildings that made at least one complaint during a given heating season ($X_{iy} \geq 1$) and are, therefore, present in the 311 call log data.

For this complementary analysis, we use the 311 call log data to estimate the total duration of heating and hot water issues for each such building i during a given heating season y and control for this estimate of the problem duration (as well as the number of residential units u_i) when calculating a quantile statistic of X_{iy}, which is used to measure the degree of less-than-expected call behavior. We then aggregate the building-level estimates across all heating seasons to the Census Block Group (CBG) level, to identify CBGs that have a significantly higher than expected proportion of less-than-expected reporting buildings. Finally, we perform a regression analysis to identify which socioeconomic characteristics are predictive of less-than-expected reporting at building level.

2.3.1. *Estimating the duration of a heating and hot water issue*. For a building i in which at least one unit contacted 311 about a heating and hot water problem during heating season y, our estimate of the total number of days that building i experienced heating and hot water problems during heating season y is referred to as t_{iy}.

To compute t_{iy}, we first observe that in the NYC 311 data, once a building's heating and hot water issue is resolved, all open 311 complaints about heating and hot water issues for that building are marked "closed" with the same resolution date. Therefore, we exploited this pattern to aggregate the 311 calls into a data set that contains the dates reported to 311 in which buildings were experiencing heating and hot water issues. Each row of this data set represents a building experiencing a heating and hot water problem, for a given period during a heating season, with its associated start date (as measured by the first 311 call with the given resolution date), resolution date, resolution description, and the number of 311 calls. While most of these problem periods for each building represent disjoint time periods, 0.025% of these problem periods overlap (a new problem was opened before a previous problem was closed). Such overlapping problem periods were merged into a single problem period.

The resolution description is used to determine if two adjacent temporal intervals should be merged into a single, longer interval. For example, if complaints are closed because the inspectors were unable to enter the building or a violation was identified, it is likely that the problem has not been resolved, and new calls placed within two days of the initial complaint's resolution date are assumed to correspond to the same underlying heating and hot water issue. On the other hand, if an inspection is performed and no issue is found, it is likely that calls placed after the initial complaint's resolution date correspond to a different problem, or the problem is intermittent, and the intervening days are excluded from the problem duration.

We further aggregate this dataset to calculate, for each building and each heating season, the number of 311 calls X_{iy} and the estimated number of days t_{iy} during which the building was experiencing a heating and hot water problem.

2.3.2. *Identifying less-than-expected reporting Census Block Groups*. For each heating season y, we expect that buildings with both a similar number of units u_i and a similar estimated number of days t_{iy} experiencing a heating and hot water issue would have comparable calling behavior. Therefore, we develop reference sets for each heating season by binning the u_i and t_{iy} values based on $\lfloor \log_2(u_i) \rfloor$ and $\lfloor \log_2(t_{iy}) \rfloor$. A reference set consists of a grouping of buildings for a given heating season with both a similar number of units and a similar number of days of experiencing a heating and hot water issue, for example, all reporting buildings for the 2013 heating season that have between four and seven units, with an estimated problem duration of between eight and 15 days during that season. Across all heating

seasons, 192 distinct reference sets were defined, where each combination b_{iy} of building i and heating season y, with $u_i \geq 2$ and $X_{iy} > 0$, is mapped to a reference set r_{iy}.

We first normalize the number of calls X_{iy} for each b_{iy}, dividing by the estimated duration in days of heating and hot water issues and the number of residential units in the building,

$$n_{iy} = \frac{X_{iy}}{t_{iy} u_i}. \tag{19}$$

Next, for each reference set r, we rank all of the normalized counts n_{iy} (number of 311 calls per unit per day of problem duration) associated with all $b_{iy} \in r$. We can then compute the quantile value Q_{iy} corresponding to each normalized count, where $|r_{iy}|$ represents the number of elements (buildings) in the reference set r_{iy},

$$Q_{iy} = \frac{\text{rank}(n_{iy})}{|r_{iy}|}. \tag{20}$$

In summary, the reference sets are used to control for factors that impact the number of calls placed by residents of a building: the number of residential units in the building, and the estimated number of days a building has a heating and hot water problem. These two factors influence calling behavior for a building but do not necessarily indicate whether reporting behaviors differ in relation to the neighborhood's socioeconomic characteristics or geography. The purpose of transforming the normalized count of 311 calls into quantile rankings is to map the normalized count of 311 calls, which we assume as being comparable for buildings in a reference set, into a measurement that can be comparable across reference sets, thus allowing us to better understand the role of geographic and socioeconomic characteristics.

The smallest reference set contains 121 buildings with 64+ residential units for the 2010 heating season with an estimated one to three days of heating and hot water problems. The largest reference set contains 3385 buildings with two to three residential units for the 2013 heating season with an estimated four to seven days of heating and hot water problems. The average number of buildings across all reference sets is 987.5. The average normalized call volume n_{iy} across all buildings and heating seasons is 0.055. The reference set of 823 buildings for 2010 with 64+ residential units and eight to 15 estimated days of heating and hot water problems had the lowest average n_{iy} of 0.002, and the reference set of 2199 buildings for 2017 with two to three residential units and one to three estimated days of heating and hot water problems had the highest average n_{iy} of 0.289.

Given these quantile values for each building and heating season, we aggregate the quantile values for each CBG across all buildings and heating seasons. To identify whether a CBG has a significantly higher rate of less-than-expected reporting buildings across all heating seasons, we perform a Kolmogorov–Smirnov test (KS test), which is a nonparametric test that determines whether two distributions are significantly different (Massey Jr (1951)). The quantile values range between 0 and 1. If reporting behavior (normalized 311 call volume n_{iy}) was independent of CBG, after stratifying by building size and estimated problem duration, this would result in the quantile values for each CBG being approximately uniformly distributed on [0, 1]. Thus, the null hypothesis for the KS test is that the quantile values for CBG j are drawn from the uniform distribution on [0, 1]. Since we are interested in identifying CBGs with less-than-expected reporting, we perform a one-sided KS test, where the alternative hypothesis is that the distribution of quantile values Q_{iy} for CBG j is drawn from a different underlying distribution, with cumulative distribution function (CDF) greater than the CDF of the uniform distribution at some value between 0 and 1. The KS test measures the maximum of the difference between the two CDFs and evaluates its statistical significance. For each CBG's quantile values, we perform the KS test and record the corresponding p-value.

To control for multiple testing when determining if a CBG has statistically significant less-than-expected reporting, we calculate each CBG's critical value, using the Benjamini–Hochberg procedure (Benjamini and Hochberg (1995)) using a false discovery rate of 0.05, and reject the null hypothesis for a CBG if its p-value is less than the Benjamini–Hochberg critical value. For CBGs with p-values less than the Benjamini–Hochberg critical value, we determine that there is a significantly higher proportion of buildings in those CBGs reporting heating and hot water issues to 311 less than expected, compared to the distribution we would expect to see if observed reporting behavior to 311 was uniformly distributed across CBGs.

2.3.3. *Modeling buildings' reporting behavior as a function of neighborhood-level socioeconomic predictor variables.* Using the neighborhood-level socioeconomic characteristics, d, as predictors, we fit a linear regression model to determine which socioeconomic characteristics are associated with lower normalized rates of reporting for a building, as compared to similarly sized buildings with similar estimated durations of heating and hot water problems. The labels we use for each reporting building in a given heating season are the quantile values, Q_{iy}, as described in Section 2.3.2. Recall that Q_{iy} compares a building's normalized calling behavior n_{iy} (number of 311 calls per unit per day of problem duration) to other buildings in its reference set r_{iy}. For a given heating season, buildings with lower Q_{iy} are considered to be reporting less-than-expected. We compute t-statistics and corresponding p-values with the null hypothesis that all coefficients w_d are equal to 0 and the alternative hypothesis that there is a (positive or negative) linear relationship between the independent variables and Q_{iy}. All independent variables (neighborhood characteristics) with p-values <0.05 are considered significant predictors of less-than-expected reporting.

3. Results.

3.1. *Latent variable model for nonreporting.* As described above, we fit the latent variable model, described in Section 2.2.2, to the NYC 311 data in order to identify neighborhoods (CBGs) and subpopulations with higher estimated probabilities of nonreporting. Table 2 presents the linear model coefficients $\hat{w}_b$ for the structural building characteristics (predictors of a building having an underlying heating and hot water problem) and the linear model coefficients $\hat{w}_d$ for the socioeconomic neighborhood characteristics (predictors of a unit placing a call to 311) fitted by the latent variable model. We obtained bootstrapped confidence intervals for the latent variable model by running it 200 times with random samples (with replacement) of the original data.

As shown in Table 2, various structural building characteristics are statistically significant predictors of an underlying heating and hot water problem. Most notably, older buildings without basements and with larger building depths are associated with more heating and hot water problems, while newer buildings with condominiums are associated with fewer problems. Several neighborhood demographic and socioeconomic characteristics are statistically significant predictors of whether a unit will contact 311 to report a heating and hot water problem. Variables associated with increased nonreporting in a CBG include a higher proportion of limited English speakers with Asian languages, a higher proportion of elderly individuals, a higher proportion of households with children under 18, a higher median rent, and a higher proportion of the population that votes. Variables associated with decreased nonreporting in a CBG include a higher proportion of limited English speakers with Spanish language and a higher proportion of Black individuals. Predictors with strong correlations within the socioeconomic characteristics set and the structural building characteristics set were removed to avoid issues of multicollinearity. The strongest remaining correlation ($r = 0.47$) within the socioeconomic characteristic predictors was between the proportion of population that

TABLE 2
Fitted linear model coefficients $\hat{w}_b$ and $\hat{w}_d$ for the latent variable model fit to NYC 311 data for $Pr(C = 1 \mid d, b, u, w_b, w_d)$

Predictors (b and d)	Coefficient (Std. error)	95% CI
Year built	−0.3482 (0.0034)***	(−0.3552, −0.3420)
Above grade full basement	−0.2692 (0.0042)***	(−0.2767, −0.2604)
Contains condominiums	−0.1260 (0.0032)***	(−0.1320, −0.1196)
Below grade full basement	−0.1042 (0.0043)***	(−0.1125, −0.0956)
Building has been altered twice	−0.0227 (0.0025)***	(−0.0272, −0.0176)
Below grade partial basement	−0.0179 (0.0039)***	(−0.0257, −0.0106)
Above grade partial basement	−0.0153 (0.0040)***	(−0.0229, −0.0074)
Falls within potential floodplain	−0.0111 (0.0019)***	(−0.0146, −0.0070)
Property area	−0.0088 (0.0041)*	(−0.0164, −0.0005)
Residential property area	0.0002 (0.0040)	(−0.0081, 0.0075)
Boiler age	0.0052 (0.0014)***	(0.0024, 0.0080)
Number of buildings	0.0080 (0.0028)**	(0.0026, 0.0134)
Boiler indicator variable	0.0197 (0.0018)***	(0.0163, 0.0232)
Number of floors	0.0419 (0.0028)***	(0.0374, 0.0484)
Building has been altered once	0.1008 (0.0026)***	(0.0964, 0.1067)
No basement	0.2080 (0.0032)***	(0.2025, 0.2148)
Building depth	0.5971 (0.0056)***	(0.5886, 0.6107)
Structural characteristics intercept	−1.5099 (0.0098)***	(−1.5363, −1.4981)
Proportion of population that votes	−0.1549 (0.0097)***	(−0.1767, −0.1386)
Proportion of households with children under 18	−0.1419 (0.0079)***	(−0.1548, −0.1237)
Proportion of elderly population (70+)	−0.0801 (0.0049)***	(−0.0939, −0.0747)
Proportion of limited English-speaking households with Asian languages	−0.0691 (0.0053)***	(−0.0811, −0.0603)
Median rent	−0.0608 (0.0079)***	(−0.0785, −0.0481)
Proportion of population with graduate education	−0.0367 (0.0098)***	(−0.0593, −0.0206)
Proportion of population with 40+ minute commute	−0.0080 (0.0041)*	(−0.0161, −0.0001)
Proportion of females	0.0094 (0.0035)*	(0.0017, 0.0154)
Proportion of population unemployed	0.0575 (0.0038)***	(0.0488, 0.0636)
Proportion of limited English-speaking households with Spanish language	0.1242 (0.0065)***	(0.1060, 0.1319)
Proportion of Black population	0.3499 (0.0101)***	(0.3319, 0.3715)
Socioeconomic intercept	−2.4882 (0.0093)***	(−2.5003, −2.4638)

All features are standardized to a mean of 0 and standard deviation of 1. Coefficients are produced from a singular run of the EM model using the full data. The standard errors and confidence intervals are calculated from 200 runs of the EM model with bootstrapped samples of the original data. *signifies $p < 0.05$, **signifies $p < 0.01$, and ***signifies $p < 0.001$. The coefficients of the structural building characteristic predictors are used to produce the probability of whether a building is experiencing a heating or hot water problem during a given heating season. Positive coefficients contribute to a higher probability of a building experiencing a heat or hot water problem. The coefficients of the socioeconomic characteristic predictors are used to produce the probability of a singular unit within a building during a given heating season contacting 311 about a heating or hot water problem. Therefore, positive coefficients contribute to a higher probability of a singular unit calling 311 about an issue whereas negative coefficients contribute to a higher probability of nonreporting for a singular unit during a given heating season.

votes and the proportion of population with graduate education. Within the building characteristic predictors, the strongest correlation ($r = 0.69$) was between the number of buildings and property area. For a discussion and simulations addressing practical identifiability of the latent variable model, please reference Appendix B of the Supplementary Material (Boxer et al. (2025)).

Using the coefficients $\hat{w}_b$ and $\hat{w}_d$ estimated from the latent variable model, we estimated the probability that each building i for each heating season y had an underlying heating and hot water problem, $\Pr(I_{iy} = 1 \mid C_{iy}, b_i, d_i, u_i, \hat{w}_b, \hat{w}_d)$. (Recall that, under our modeling assumptions, this probability equals 1 for all buildings with $C_{iy} = 1$.)

We then compute three quantities for each *CBG* j, which we use to create the plots below:

$$U_j = \sum_{i,y:\, i \in CBG_j} \mathbf{1}\{C_{iy} = 0\} \Pr(I_{iy} = 1 \mid C_{iy}, b_i, d_i, u_i, \hat{w}_b, \hat{w}_d), \tag{21}$$

$$R_j = \sum_{i,y:\, i \in CBG_j} \mathbf{1}\{C_{iy} = 1\}, \tag{22}$$

$$N_j = \sum_{i,y:\, i \in CBG_j} 1, \tag{23}$$

where U_j is the estimated number of buildings with unreported heating and hot water problems, R_j is the number of buildings with heating and hot-water problems that were reported to 311, and N_j is the total number of buildings, summed across all heating seasons.

First, we compute the estimated frequency of heating and hot water problems $\frac{R_j+U_j}{N_j}$, including both reported and unreported problems, for each CBG, as shown in Figure 2. This plot shows that there is a very high frequency of underlying heating and hot water problems in the Bronx and Upper Manhattan, a moderately high frequency in lower Manhattan, and various patches of increased problem frequency in Queens and Brooklyn.

Next, we compute the estimated probability that buildings in each CBG j will not report a heating and hot water issue, $\frac{U_j}{R_j+U_j}$. This estimated probability of nonreporting for each CBG is shown in Figure 3. Importantly, since nonreporting is computed as a fraction of the estimated number of underlying heating and hot water problems ($R_j + U_j$), the shaded

FIG. 2. *Estimated proportion of buildings experiencing a heating and hot water problem across all heating seasons* $\frac{R_j+U_j}{N_j}$ *for each Census Block Group in New York City, including both reported and unreported problems. For a color version of this figure, reference Figure S6 in Appendix D of the Supplementary Material (Boxer et al. (2025)).*

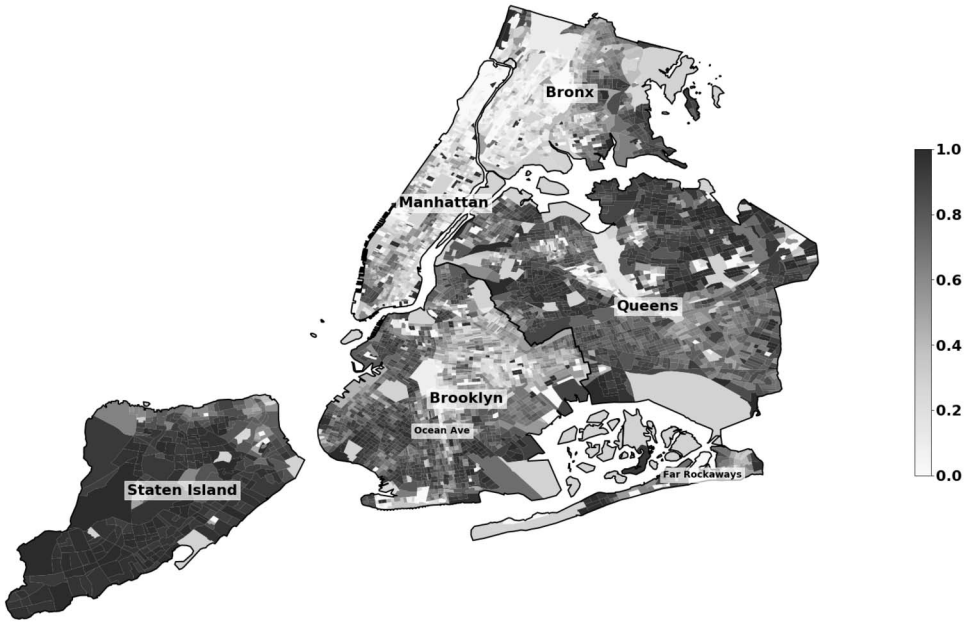

FIG. 3. *Estimated probability of nonreporting* $\frac{U_j}{R_j+U_j}$ *for each Census Block Group in New York City. For a color version of this figure, reference Figure S7 in Appendix D of the Supplementary Material* (Boxer et al. (2025)).

regions in Figure 3 represent nonreporting rates solely and not estimated underlying heating and hot water problems. The plot shows that, in many neighborhoods throughout New York City, there is a high proportion of buildings that do not report to 311, as compared to the estimated number of buildings experiencing a problem in a given heating season. In areas with a high proportion of smaller buildings, as shown in Figure 4, there is a higher probability of

FIG. 4. *Proportion of buildings with more than* 10 *units for each Census Block Group in New York City. For a color version of this figure, reference Figure S8 in Appendix D of the Supplementary Material* (Boxer et al. (2025)).

FIG. 5. *Estimated frequency of unreported heating and hot water problems among buildings that did not report any problems to* 311 *in a given heating season* $\frac{U_j}{N_j - R_j}$ *for each Census Block Group in New York City. For a color version of this figure, reference Figure S9 in Appendix D of the Supplementary Material* (Boxer *et al.* (2025)).

nonreporting. This is expected, given that smaller buildings, where presumably less residents reside, have a lower probability of placing at least one call to 311 per heating season about an underlying heating or hot water problem, as compared to buildings with more residential units. In areas with larger buildings (more than 10 units), such as Upper Manhattan and the West Bronx, the probability of nonreporting is smaller. We note that these areas have both a high density of underlying heating and hot water problems and a high probability that at least one unit in the building will call to report such problems.

Finally, Figure 5 estimates the probability that a building which did not report a heating and hot water problem to 311 in a given heating season had an underlying heating and hot water problem, $\frac{U_j}{N_j - R_j}$. Most notably, there is a high frequency of unreported underlying issues in Lower Manhattan and to a lesser extent in South Brooklyn and Queens.

It is important to note that this analysis of nonreporting only considers whether a building experiences a heating and hot water problem during a given heating season and does not take into account the number, severity, or duration of problems. Therefore, some buildings might experience less serious heating and hot water issues or might have receptive management which resolves these issues quickly, whereas other buildings with unresponsive management might lack heat and hot water for most of the heating season. Thus, we explicitly estimate and control for problem duration in our analysis of less-than-expected reporting described below.

3.2. *Less-than-expected reporting*. Using our method of estimating problem duration, calculating quantile values for each building's normalized 311 call volume for each heating season while controlling for building size and estimated problem duration, aggregating across buildings and heating seasons by CBG and testing each CBG for statistically significant less-than-expected reporting, as described in Section 2.3, we generated the map displayed in Figure 6. This plot shows a high density of less-than-expected reporting CBGs in Queens and the Far Rockaways. Additionally, it shows various smaller clusters of less-than-expected reporting in Manhattan, the Bronx, Brooklyn, and Staten Island.

FIG. 6. *Highlighted map of New York City showing Census Block Groups that were classified as having statistically significant less-than-expected reporting. Note that each Census Block Group is highlighted in one of three colors*: *dark blue*, *representing neighborhoods with statistically significant less-than-expected reporting*; *white*, *representing neighborhoods with as-expected reporting*; *or gray*, *representing neighborhoods with no available data. For a color version of this figure*, *reference Figure S*10 *in Appendix D of the Supplementary Material* (*Boxer et al.* (2025)).

Table 3 presents the results of a linear regression fitted to predict the quantile value Q_{iy} of a building's normalized reporting behavior within its reference set. Lower Q_{iy} values correspond to less-than-expected reporting. We observe that less-than-expected reporting is most significantly associated with: (i) a higher proportion of the population with limited English-speaking with both Asian languages and Spanish, (ii) a higher proportion of the population being elderly, (iii) a higher proportion of households with children under 18, (iv) higher median rent, (v) a lower proportion of Black individuals, and (vi) a lower proportion of unemployed individuals.

4. Discussion. Based on our results from the nonreporting analysis described in Section 3.1, there are a substantial number of buildings throughout New York City which experience a heating and hot water problem at some point during a heating season but do not contact 311. Some of these buildings might not contact 311 because they have other means of addressing heating and hot water issues outside of the NYC 311 system or because the issue was resolved quickly. It is important to note that the modeled probability of reporting, conditional on neighborhood demographic and socioeconomic characteristics and structural building characteristics, is higher for larger buildings since there are more residential units which might report a problem. By comparing Figures 3 and 4, we observe a higher estimated probability of nonreporting for CBGs with smaller buildings. While areas with larger buildings, such as Upper Manhattan, the lower Bronx, and certain areas of Brooklyn, like Ocean

TABLE 3

Linear regression as a function of neighborhood-level demographic and socioeconomic characteristics to predict each reporting building's quantile-value Q_{iy} in a given heating season. Q_{iy} is computed by ranking the building's normalized calling behavior n_{iy} (number of 311 calls per unit per day of problem duration) within its respective reference set r_{iy} for that heating season, as described in Section 2.3.2. Lower Q_{iy} corresponds to less-than-expected reporting

Predictors (d)	Mean (Std. dev.) of all reporting buildings per heating season	Mean (Std. dev.) of all reporting buildings in less-than-expected reporting neighborhoods	Coefficient	95% CI
Proportion of elderly population (70+)	0.074 (0.059)	0.082 (0.071)	−0.008***	(−0.009, −0.006)
Median rent	1187.139 (399.084)	1272.423 (409.180)	−0.006***	(−0.008, −0.005)
Proportion of households with children under 18	0.554 (0.295)	0.651 (0.262)	−0.005***	(−0.007, −0.004)
Proportion of limited English-speaking households with Spanish language	0.098 (0.117)	0.107 (0.122)	−0.005***	(−0.006, −0.003)
Proportion of limited English-speaking households with Asian languages	0.026 (0.075)	0.045 (0.100)	−0.004***	(−0.005, −0.003)
Proportion of females	0.522 (0.068)	0.516 (0.067)	−0.002**	(−0.003, −0.001)
Proportion of population that votes	0.227 (0.063)	0.230 (0.060)	0.000	(−0.001, 0.002)
Proportion of population with graduate education	0.097 (0.094)	0.109 (0.094)	0.000	(−0.001, 0.002)
Proportion of population with 40+ minute commute	0.498 (0.170)	0.484 (0.156)	0.002**	(0.001, 0.003)
Proportion of population unemployed	0.108 (0.075)	0.087 (0.062)	0.003***	(0.001, 0.004)
Proportion of Black population	0.313 (0.312)	0.179 (0.249)	0.015***	(0.013, 0.016)
Socioeconomic intercept			0.501***	(0.499, 0.502)

All features are standardized to a mean of 0 and standard deviation of 1 for the logistic regression.
*signifies $p < 0.05$, **signifies $p < 0.01$, and ***signifies $p < 0.001$.
The mean and std. dev. measurements contained in the second and third column refer to the mean and standard deviation of a predictor, before standardization, for all buildings per heating-season and only buildings per heating-season that reside in CBGs classified as having statistically significant less-than-expected reporting behavior to 311, respectively.

Avenue, have a lower probability of nonreporting, given the large number of underlying heating and hot water issues it is possible that these buildings have multiple problems during a single heating season, some of which may go unreported.

In Figures 2 and 5, we observe several CBGs with sharp variations in their rate of estimated underlying heating and hot water problems, as compared to their surrounding CBGs. As a case study, we examined the CBG in Staten Island that shows a notable increase in estimated unreported heating and hot water problems compared to its surrounding area in Figure 5. This CBG contains only three residential buildings with 80+ units each (265 residential units in total) with structural building characteristics that are associated with a higher probability of the buildings having underlying heating and hot water problems (no condominiums, higher building depths, no basements, etc). Only 12 calls in total were placed from residents of these buildings to 311 about heating and hot water issues for the heating seasons ranging from 2010 to 2017. Most importantly, these buildings are restricted to housing only the elderly population, with two of these buildings exclusively serving low-income elderly individuals. A high proportion of elderly individuals residing in a CBG is associated with higher rates of nonreporting. These buildings' structural characteristics, the lack of calls placed to 311 despite the large number of residential units per building, in conjunction with the increased

rate of elderly individuals residing in that CBG resulted in a higher estimated probability of unreported heating and hot water problems, as compared to the surrounding CBGs.

Based on our analyses for both nonreporting (Section 3.1) and less-than-expected reporting (Section 3.2), the borough of Queens has both a high probability of nonreporting for heating and hot water problems and a high density of buildings that report these problems to 311 significantly less than expected (given building size and estimated problem duration). While the former result could be explained by a high proportion of buildings with a smaller number of units, we note that the borough of Staten Island, which also has small buildings and a high nonreporting rate, does not have nearly as much less-than-expected reporting.

Our two analyses provide a high degree of consistency in identifying neighborhood-level demographic and socioeconomic characteristics which are predictive of both types of underreporting, including a high proportion of elderly individuals, limited English speakers with Asian languages, and households with children under 18 years old, as shown in Tables 2 and 3. Moreover, we observe one interesting difference between the two results in that a high proportion of limited English speakers with Spanish language was associated with a higher probability of less-than-expected reporting but a lower probability of nonreporting. The latter result is most likely because of a strong positive correlation between the proportion of limited English speakers with Spanish language and the proportion of buildings with more than 10 residential units, since such large buildings are very likely to have at least one unit report a problem to 311. Thus, we hypothesize, based on our analyses, that limited English speakers and elderly individuals are less likely to report a problem to 311; however, confirmation of these hypotheses would require more detailed unit-level data including both demographics and 311 call records.

Some of the socioeconomic predictors we found as positively associated with nonreporting, such as higher median rent and a higher proportion of the population with graduate education, we believe are indicators of buildings that have the structural characteristics of buildings that are likely to experience a heating and hot water problem during the heating seasons but have responsive management or have access to other methods of addressing heating and hot water problems. As mentioned above, our analysis of nonreporting does not explicitly control for the number of residential units; therefore, some of the socioeconomic characteristics we found as neighborhood-level negative predictors of nonreporting, such as a higher proportion of Black individuals and a higher proportion of limited English speakers with Spanish language correspond to subpopulations who tend to live in NYC neighborhoods with a higher proportion of 10+ unit buildings as well as lower median rents. Refer to Figure S5 in Appendix C of the Supplementary Material for explanatory plots (Boxer et al. (2025)). Therefore, these cannot conclusively be determined as negative predictors of nonreporting but do correspond to populations experiencing more frequent and potentially more serious heating and hot water problems. Given the lack of data about unit-specific calling behaviors, it is challenging to make claims about the reporting probability of these populations who often live in larger-unit buildings.

Understanding reporting behavior provides a necessary starting point for city agencies to better measure, evaluate, and address underlying problematic conditions. City agencies can encourage community engagement through directed outreach and education in neighborhoods identified in our analysis based on demographic and socioeconomic characteristics. In terms of constructing policies targeted at increasing access to 311 systems based on neighborhood-level socioeconomic characteristics, we believe our analysis of less-than-expected reporting, which controls for the duration of a heating and hot water problem and building size, has more useful results, suggesting that neighborhoods with a higher density of limited English speakers (including both Spanish and Asian languages), a higher density of households with children under 18, and a higher density of elderly individuals might benefit from increased outreach to reduce underreporting. Neighborhoods with a high density of

Black individuals and limited English speakers with Spanish language, such as Upper Manhattan and the lower Bronx, might benefit from increased attention to the structural issues, such as aging, poor-quality housing infrastructure and unresponsive landlords which lead to frequent heating and hot water problems and impact these communities' quality of life.

Our taxonomy of underreporting, including both nonreporting and less-than-expected reporting as well as the methods described here for estimating these two modes of underreporting, may be useful in analyzing reporting behavior across multiple resident-generated data sources in a variety of locations and could be adapted to other quantitative problems using positive and unlabeled data. Nevertheless, the specific conclusions we draw about structural building characteristics and neighborhood-level demographic and socioeconomic characteristics associated with underreporting are specific to heating and hot water problems within NYC. Research in a variety of social fields indicates that socioeconomic patterns can vary greatly across geographic locations (Darity Jr. et al. (2018)), and we, therefore, warn policymakers against extrapolating these results to draw conclusions pertaining to other cities or to underreporting of urban problems other than heating and hot water complaints.

Our analysis is limited to identifying neighborhood-level socioeconomic characteristics that are associated with a higher proportion of underreporting because the most granular socioeconomic data available are at the Census Block Group level. For example, our research suggests that areas with a higher density of limited English speakers with Asian languages tend to underreport, but this does not necessarily indicate that any individual building (or unit) with limited English-speaking residents is less likely to use 311. This limitation is a result of the lack of socioeconomic data available at the building and unit levels.

Similarly, a lack of unit-level information about 311 calls prevents our analyses from discerning between building-level and unit-level heating and hot water problems. This could invalidate our results if certain building characteristics and/or socioeconomic neighborhood characteristics correlate with unit-level issues but not larger building-level issues, or vice-versa. While we do not have the data to check for these issues, we note the high degree of consistency between the results of the nonreporting and less-than-expected reporting analyses, with the former using neighborhood-level socioeconomic characteristics as building-level predictors and the latter using socioeconomic characteristics as predictors for only reporting buildings and not using the building structural characteristics. Thus, we believe that the lack of granular data as to whether a heating and hot water issue is unit-specific or affects an entire building does not invalidate our results pertaining to neighborhoods and socioeconomic groups with higher levels of underreporting.

An additional limitation of the model is our assumption that each 311 heating and hot water complaint results from an underlying heating and hot water problem. However, it is possible that some 311 calls either did not correspond to actual problems ("false calls") or corresponded to problems that were transient or quickly resolved by the landlord ("unnecessary calls"). If there were a large number of false and unnecessary calls and if the distribution of such calls was demographically and geographically heterogeneous, this could bias our estimation of the underlying heating and hot water problems (Figure 2). Nevertheless, our primary focus is on understanding geographic and demographic variation in the propensity to report, and since underreporting is likely to result from either lack of information or lack of access to the 311 system, we believe that areas and subpopulations with higher rates of false and unnecessary calls are also less likely to have significantly high rates of underreporting.

Considering these limitations, we developed a framework for estimating underreporting and two analyses which consider both nonreporting and less-than-expected reporting within this framework. Our latent variable model for estimating nonreporting is a nontrivial extension of previous models and incorporates domain-specific knowledge about the building-level 311 reporting of heating and hot water issues. This method is specifically useful in that

it estimates a distribution of heating and hot water problems from resident-generated data, containing reporting biases, which differ across neighborhoods and subpopulations, without requiring a randomized survey or some proxy for ground truth. Obtaining such ground truth data is often prohibitively difficult and expensive, as discussed above, yet collection of such data even on a smaller scale would help to validate the modeling assumptions and analysis results described here. Nevertheless, we hope that these methods and results will assist city agencies and advocacy groups in improving access to 311 and increasing services to communities whose quality of life are impacted by heating and hot water problems.

Acknowledgments. We would like to thank the New York City Department of Information Technology and Telecommunications and NYC311 for providing data, insight, and feedback on preliminary versions of this work.

Constantine E. Kontokosta and Daniel B. Neill are co-senior authors of this work.

Funding. This work was partially supported by National Science Foundation grant IIS-1926470 and the NSF Program on Fairness in Artificial Intelligence in Collaboration with Amazon, grant IIS-2040898.

SUPPLEMENTARY MATERIAL

Appendices for estimating reporting bias in 311 complaint data (DOI: 10.1214/24-AOAS2003SUPP; .pdf). Appendices containing validation studies, discussions about identifiability, and additional figures.

REFERENCES

BEKKER, J. and DAVIS, J. (2020). Learning from positive and unlabeled data: A survey. *Mach. Learn.* **109** 719–760. MR4094532 https://doi.org/10.1007/s10994-020-05877-5

BEKKER, J., ROBBERECHTS, P. and DAVIS, J. (2019). Beyond the selected completely at random assumption for learning from positive and unlabeled data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 71–85. Springer, Berlin.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776 https://doi.org/10.1080/01621459.2017.1285773

BOXER, K. S., HONG, B., KONTOKOSTA, C. E. and NEILL, D. B. (2025). Supplement to "Estimating reporting bias in 311 complaint data." https://doi.org/10.1214/24-AOAS2003SUPP

CAUCHY, A. et al. (1847). Méthode générale pour la résolution des systemes d'équations simultanées. *C. R. Sci. Paris* **25** 536–538.

CITY OF NEW ORLEANS (2019). Catch basins. Available at https://nola.gov/dpw/catch-basins/. Accessed 2021-10-05.

CITY OF NEW YORK (2021). About NYC311. Available at https://portal.311.nyc.gov/about-nyc-311/.

DARITY JR., W., HAMILTON, D., PAUL, M., AJA, A., PRICE, A., MOORE, A. and CHIOPRIS, C. (2018). What we get wrong about closing the racial wealth gap. *Samuel DuBois Cook Center on Social Equity and Insight Center for Community Economic Development* **1** 1–67.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

ELKAN, C. and NOTO, K. (2008). Learning classifiers from only positive and unlabeled data. In *Proceedings of the* 14*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 213–220.

KATO, M., TESHIMA, T. and HONDA, J. (2019). Learning from positive and unlabeled data with a selection bias. In *International Conference on Learning Representations*.

KELLY, J. M. and SWINDELL, D. (2002). Service quality variation across urban space: First steps toward a model of citizen satisfaction. *J. Urban Aff.* **24** 271–288.

KONTOKOSTA, C. E. (2016). The quantified community and neighborhood labs: A framework for computational urban science and civic technology innovation. *J. Urban Technol.* **23** 67–84.

KONTOKOSTA, C. E. and HONG, B. (2021). Bias in smart city governance: How socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions. *Sustain. Cities Soc.* **64** 102503.

KONTOKOSTA, C. E., WEISS, M., SNIVELY, C. and GULICK, S. (2017). NYC311. Harvard Business School Case 818-056.

MASSEY JR, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *J. Amer. Statist. Assoc.* **46** 68–78.

MCLAFFERTY, S., SCHNEIDER, D. and ABELT, K. (2020). Placing volunteered geographic health information: Socio-spatial bias in 311 bed bug report data for New York City. *Health Place* **62** 102282. https://doi.org/10.1016/j.healthplace.2019.102282

MEIJER, A. and POTJER, S. (2018). Citizen-generated open data: An explorative analysis of 25 cases. *Gov. Inf. Q.* **35** 613–621.

MINKOFF, S. L. (2016). NYC 311: A tract-level analysis of citizen–government contacting in New York City. *Urban Aff. Rev.* **52** 211–246.

NADEAU, L. K. (2011). New York City unveils real-time 311 request map. Available at http://www.govtech.com/e-government/New-York-City-Real-Time-311-Map-021711.html. Accessed 2017-07-08.

RAMASWAMY, H., SCOTT, C. and TEWARI, A. (2016). Mixture proportion estimation via kernel embeddings of distributions. In *International Conference on Machine Learning* 2052–2060. PMLR.

WARD, G., HASTIE, T., BARRY, S., ELITH, J. and LEATHWICK, J. R. (2009). Presence-only data and the EM algorithm. *Biometrics* **65** 554–563. MR2751480 https://doi.org/10.1111/j.1541-0420.2008.01116.x

WHITE, A. and TRUMP, K.-S. (2018). The promises and pitfalls of 311 data. *Urban Aff. Rev.* **54** 794–823.

ZHA, Y. and VELOSO, M. (2014). Profiling and prediction of non-emergency calls in New York City. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

ZHENG, Y., LIU, T., WANG, Y., ZHU, Y., LIU, Y. and CHANG, E. (2014). Diagnosing New York city's noises with ubiquitous data. In *Proceedings of the* 2014 *ACM International Joint Conference on Pervasive and Ubiquitous Computing* 715–725. ACM, New York.