Urban Incident Prediction with Graph Neural Networks: Integrating Government Ratings and Crowdsourced Reports

Sidhika Balachandar[1], Shuvom Sadhuka[2], Bonnie Berger[2], Emma Pierson[1], Nikhil Garg[3]

[1] Department of Electrical Engineering & Computer Sciences, UC Berkeley
[2] Department of Electrical Engineering & Computer Science, MIT
[3] Department of Operations Research & Information Engineering, Cornell Tech
sidhikab@berkeley.edu, ssadhuka@mit.edu, bab@mit.edu, emmapierson@berkeley.edu, ngarg@cornell.edu

Abstract

Graph neural networks (GNNs) are widely used in urban spatiotemporal forecasting, e.g., predicting infrastructure problems. In this setting, government officials aim to identify in which neighborhoods incidents like potholes or rodents occur. The true state of incidents is observed via government inspection *ratings*. However, these ratings are only conducted for a sparse set of neighborhoods and incident types. We also observe the state of incidents via crowdsourced *reports*, which are more densely observed but may be biased due to heterogeneous reporting. First, we propose a multi-view, multi-output GNN-based model that uses both unbiased rating data and biased reporting data to predict the true latent state of incidents. Second, we investigate a case study of New York City urban incidents and collect a dataset of 9,615,863 crowdsourced reports and 1,041,415 government inspection ratings over 3 years and across 139 types of incidents. We show on both real and semi-synthetic data that our model can better predict the latent state compared to models that use only reporting data or only rating data. Finally, we quantify demographic biases in crowdsourced reporting, e.g., higher-income neighborhoods report problems at higher rates. Our analysis showcases a widely applicable approach for latent state prediction using heterogeneous, sparse, and biased data.

Model — https://github.com/sidhikabalachandar/nyc_urban_incident_model

Dataset — https://github.com/sidhikabalachandar/nyc_urban_incident_data

1 Introduction

Graph neural networks (GNNs) have emerged as expressive models for making predictions on graph-structured data, including for urban applications such as air quality monitoring, forecasting traffic flows, etc. (Xie et al. 2019; Roy et al. 2021; Brimos et al. 2023; Yu et al. 2023; Zhan and Datta 2024). In urban planning – our empirical setting – government officials often wish to know where incidents like rodent infestations or floods occur so they can make downstream resource allocation decisions; however, ground truth is often unobserved. GNNs are a powerful prediction tool, as they can encode spatial correlations of the ground-truth states across nodes in a graph (e.g., neighborhoods in a city) as well as across hundreds of types of incidents. For example, if a flood has occurred in one neighborhood, the adjacent neighborhoods are also likely to be flooded and might also experience related incidents like street damage.

Estimating latent ground truth for the hundreds of types of incidents in a city is challenging. Two sources of information are often used, each with its limitations. First, we observe the ground-truth state via *government inspections*, which generate *ratings* for neighborhoods. Importantly, these inspections are only conducted for some incident types and neighborhoods and thus ratings are sparsely observed. For example, New York City conducts street inspections for every street and rates them from 1-10, but each street is only rated about once every year. Second, we observe biased proxies of the latent state via crowdsourced *reports* of incidents. Compared to ratings, indicators of whether a report is observed are available for every incident type and neighborhood.

A challenge is that learning only from reporting data leads to biased predictions. Previous work has established that underreporting is pervasive and heterogeneous (O'Brien, Sampson, and Winship 2015; O'Brien et al. 2017; Clark et al. 2020; Kontokosta and Hong 2021; Agostini, Pierson, and Garg 2024; Liu, Bhandaram, and Garg 2024); in different neighborhoods that face similar incidents, residents often *report* those incidents at different rates. This bias presents an identifiability issue; if one neighborhood logs more reports than another, it is unclear whether the former has a worse ground truth, or if given the same ground truth, the latter is less likely to report. As a result, for example, Casey, Wilson, and Yokum (2018) found that in Washington, D.C., crowdsourced reports on rodents did not accurately predict the outcome of inspections. Moreover, differences in reporting often correlate with demographics, so learning only from reporting data risks introducing decision making disparities against underserved populations (Agostini, Pierson, and Garg 2024; Liu, Bhandaram, and Garg 2024).

In summary, we make the following contributions:

- We propose a multi-view, multi-output GNN-based model for our task named URBAN (Unbiased Ratings and Biased reports Aggregation Network). As summarized in Figure 1, we use both sparsely observed ground-truth rating data and frequently observed biased reporting data to predict ratings and to infer how the likelihood of reporting varies by demographics, conditional on ground-truth. URBAN's contribution lies in adapting GNNs for biased data settings by connecting multi-view datasets through a multi-output loss. Our model is available here: https://github.com/sidhikabalachandar/nyc_urban_incident_model.

- We investigate a case study of New York City incidents. We create a dataset of reports sourced from NYC 311-complaints (crowdsourced reports) and ratings sourced from government inspections. Our dataset presents two advantages. First, our dataset is heterogeneous, and through various train/test splits our data can be used to evaluate model performance on unseen time periods, neighborhoods, and incident types. Second, new data for a large set of types and locations is available daily, so our data provides an ideal setting to study distribution shifts over time with an automatic source of uncontaminated test sets. We share our data here: https://github.com/sidhikabalachandar/nyc_urban_incident_data.
- We show that our approach can predict ground-truth inspection ratings and quantify reporting biases. In both semi-synthetic and real data, we find that jointly using the sparsely observed ground-truth rating data and the frequently observed biased reporting data outperforms using reporting data alone (predicted ratings are $5.3\times$ more correlated) or using rating data alone (predicted ratings are up to $2.6\times$ more correlated). Finally, we quantify the biases in the reporting data, finding for example that conditional on ground-truth, neighborhoods with higher median income report incidents at higher rates.

Although our primary application is urban crowdsourcing, our approach is applicable to other tasks where both sparsely observed ground-truth data and frequently observed biased data are available. This includes other urban challenges (e.g., estimating air quality using resident reports and sparse sensor measurements (Sarto et al. 2016; Samal, Babu, and Das 2022)) and spatiotemporal processes (e.g., epidemic forecasting using internet search data and sparse official health reports (Kapoor et al. 2020; Hacker et al. 2022; Tomy et al. 2022; Wang et al. 2022a; Yu et al. 2023)).

2 Related Work

Our work is related to literature in three areas. We provide below a brief exposition of these areas, and provide more details in the appendix related works section.

Predicting urban incidents: There is a large literature on predicting urban incidents (Budde et al. 2014; O'Brien, Sampson, and Winship 2015; Lum and Isaac 2016; Sarto et al. 2016; Wang et al. 2017; Hu et al. 2018; Mosavi, Ozturk, and Chau 2018; Yu, Yin, and Zhu 2018; Gómez et al. 2019; Pan et al. 2019; Xie et al. 2019; Clark et al. 2020; Jiménez-Jiménez et al. 2020; Kontokosta and Hong 2021; Agonafir et al. 2022; Hacker et al. 2022; Mauerman et al. 2022; Farahmand, Xu, and Mostafavi 2023; Zaouche and Bode 2023; Agostini, Pierson, and Garg 2024; Gao et al. 2024), and prior work has shown that crowdsourced reporting data may not represent ground-truth states due to underreporting (Wang et al. 2017; Clark et al. 2020; Kontokosta and Hong 2021; Agostini, Pierson, and Garg 2024; Gao et al. 2024; Liu, Bhandaram, and Garg 2024; Liu and Garg 2024). Especially notable is the work of Casey, Wilson, and Yokum (2018), who show in Washington, D.C., that 311-reports alone are poor predictors of ground-truth ratings.

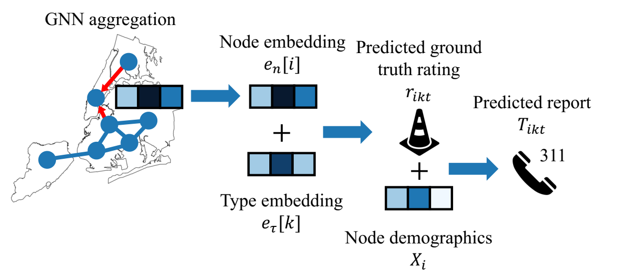


Figure 1: *Model.* We introduce URBAN (Unbiased Ratings and Biased reports Aggregation Network), a GNN-based model to estimate inspection ratings and reports of incidents. We model inspection ratings r_{ikt} using node i's learned embedding $e_n[i]$ and type k's learned embedding $e_\tau[k]$. We model reports T_{ikt} as a function of the rating r_{ikt} and demographics X_i.

Existing works in this area typically consider only one or a few related incident types; thus a novel contribution of our model is its ability to learn and make predictions for more than a hundred 311-complaint types. Additionally, many related works use statistical models that make strict assumptions (Agostini, Pierson, and Garg 2024), whereas our GNN-based model presents a flexible approach to modeling spatial correlation. Finally, we create a heterogeneous spatial dataset that we hope can be used by other researchers to study data bias in an urban application.

Using GNNs for biased data: A common approach to addressing selection bias in GNNs is *graph-based semi-supervised learning* (Song et al. 2023), which typically uses graph regularization (Zhou et al. 2003; Zhu, Ghahramani, and Lafferty 2003; Belkin, Niyogi, and Sindhwani 2006; Dai, Aggarwal, and Wang 2021; Verma et al. 2021; Li, Yin, and Chen 2023; Qian et al. 2023; Zhou et al. 2024) or graph embedding methods (Perozzi, Al-Rfou, and Skiena 2014; Tang et al. 2015; Grover and Leskovec 2016; Li, Xiong, and Hoi 2021; Sharma and Jones 2023) to infer labels for unlabeled datapoints.

These related works discuss how to effectively train GNNs on sparsely labeled data (e.g., inspection ratings) by generating proxy labels for the unlabeled datapoints. We extend this literature by identifying another approach to obtaining proxy labels: using data from a frequently observed biased proxy for the target of interest (e.g., whether a report is made is a biased proxy of the true rating).

Using GNNs for spatiotemporal modeling: GNNs are a natural fit for high-dimensional spatiotemporal modeling (Kapoor et al. 2020; Roy et al. 2021; Wang et al. 2022a,b; He et al. 2023; Pineda et al. 2023; Wu et al. 2024). Our study relates to literature on *multi-view GNNs*, which integrate information from multiple datasets or views (Shao et al. 2023; Cheng et al. 2018), and *multi-output GNNs*, which predict multiple outputs from a single graph by leveraging shared graph structure and correlations between the outputs (Peng et al. 2022; Samal, Babu, and Das 2022).

Related works that use GNNs for spatiotemporal modeling have identified that learning only from proxy data (reports) results in biased predictions. We extend this literature by identifying an approach to improve predictions: additionally learning from sparse ground-truth data (ratings). We show that our model, which uses both reporting and rating data, makes better predictions than a model trained only on reporting data or only on rating data. Additionally, while multi-view and multi-output GNNs are used for spatiotemporal tasks, prior work does not explicitly model biases.

3 Setting and Data

We now describe our setting. The goal is to estimate the ground-truth state of various types of incidents (e.g., rodents, floods, etc.). There are two sources of data for this task: frequently observed biased crowdsourced *reporting* data and sparsely observed unbiased inspection *rating* data. To our knowledge, prior urban prediction work has not considered the challenges and benefits of jointly learning from these two data sources—one reason is the lack of processed data combining reports and ratings across types. We formalize this setting and provide large-scale preprocessed data composed of 9,615,863 crowdsourced reports across 139 incident types and 1,041,415 government inspection ratings across 5 types. Additional details are provided in the appendix data section.

Dataset overview: We create a multi-view dataset of ratings and reports from New York City. We analyze all Census tracts[1] with valid demographic information (2292 tracts). We use all complaint types that receive at least one report for greater than 0.1% of all tract/week pairs. We study all data in the three year period from 2021-2023 discretized into 157 weeks. The main contribution is the data's heterogeneity. Our data can be used to evaluate models on unseen time periods, tracts, and types. Many related works only analyze *one or a few* types or do not model relationships between types (Agostini, Pierson, and Garg 2024; Liu, Bhandaram, and Garg 2024; Liu and Garg 2024; Xu et al. 2017; Clark et al. 2020; Kontokosta, Hong, and Korsberg 2017; Agonafir et al. 2022; Casey, Wilson, and Yokum 2018). In contrast, our data represents the majority of available types and our model can learn relationships between types.

Reports: We observe reports of incidents from New York City's resident reporting system, NYC311 (NYC Open Data 2024a). We collect data from 9,615,863 reports. We transform reports into *indicators* of whether a report of a type was observed in a Census tract during a particular week.

Ratings: For some Census tract, type, and week tuples, we observe inspection ratings. We source ratings from government inspections (e.g., street ratings, park ratings, etc.). A lower rating indicates a worse state; e.g., a street with a lower rating has more damage. We normalize (z-score) the ratings for each type across time and Census tracts.

We collect ratings from 1,041,415 inspections across five types: (i) street conditions (NYC Open Data 2023), (ii) park maintenance or facility conditions (NYC Open Data 2024d), (iii) rodents (NYC Open Data 2024e), (iv) food establishment/mobile food vendor/food poisoning (NYC Open Data 2024c), and (v) DCWP consumer complaints (NYC Open Data 2024b). We *do not* observe ratings for every tract/week pair. A summary is in the appendix data section.

Our model treats government inspections as a source of unbiased data, so we ensure that the inspections are not susceptible to bias. Selection bias can occur if an inspection is conducted in response to a crowdsourced report. We process the inspection data to only use inspections *not* conducted in response to crowdsourced reports, as verified by the data dictionaries or by filtering out responsive inspections. For example, rodent inspections are conducted both block-by-block in a scheduled manner and in response to 311-reports; the data does not distinguish between these reasons. As shown in the appendix data section, we identify the block-by-block inspections as those for which many other buildings in the same Census tract are inspected in the same week. We provide more details on how we filter out responsive inspections and validate our filtration process in the appendix data section.

Matching ratings and reports to geographic units: Next, we match government inspection ratings and crowdsourced reports to geographic regions and to each other. Each rating is for a fine-grained unit within a Census tract, e.g., street ratings are for street segments and rodent ratings are for buildings. For Census tract/type/week tuples with an observed rating, we match the rating to its corresponding fine-grained report. Thus, we may have multiple ratings and reports for each Census tract/type/week tuple. For Census tract/type/week tuples without observed ratings, we aggregate the reporting data at the Census tract level. Thus, we have one rating and report for each Census tract/type/week tuple. See the appendix data section for details.

4 Model

Approach overview: Our model is summarized in Figure 1. The goals are (i) to estimate the true state of an incident at a location – what is the true street condition in a neighborhood?; and (ii) to quantify biases in the reporting data – which neighborhoods systematically underreport incidents and how does reporting correlate with demographics? In many urban settings, models are fit using only the frequently observed reporting data, resulting in biased predictions (Xu et al. 2017; Casey, Wilson, and Yokum 2018; Li et al. 2020; Hacker et al. 2022). In contrast, our approach estimates the true state using both the frequently observed biased reporting data and the sparsely observed unbiased rating data.

Notation: Consider a network G with n nodes and adjacency matrix E. Nodes are indexed by i and represent neighborhoods, and edges connect adjacent neighborhoods. Each node i has features $X_i \in \mathbb{R}^D$, where D is the number of features. These features include demographic factors that may

[1] A Census tract is a geographic region defined by the U.S. Census Bureau. On average, each Census tract has thousands of inhabitants. There are 2326 total Census tracts in NYC.

influence reporting rates. There are τ total incident types indexed by k (e.g., rodents, floods, etc.). We index time by week t. The model uses two types of data: sparsely observed unbiased true state data (e.g., *inspection ratings*) and frequently observed biased data (e.g., *crowdsourced reports*).

Observed data: We observe inspection ratings $r_{ikt} \in \mathbb{R}$ which are unbiased observations of the true latent state, but are only available for a subset of nodes, types, and weeks. We also observe indicators of reports of incidents $T_{ikt} \in \{0, 1\}$, where $T_{ikt} = 1$ indicates that an incident of type k was reported for node i at week t. As discussed in §3, for some types, both inspection ratings and reports are available at the *sub-node* level. We provide the sub-node level notation of our URBAN model in the appendix model section.

Model: The *true* inspection rating r_{ikt} captures the condition of incident type k in node i during week t. The true ratings are drawn from some distribution f_r as a function of the node, type, and week, as follows:

$$\text{True inspection rating: } r_{ikt} \sim f_r(\cdot|i, k, t). \quad (1)$$

We now discuss how we *model* ratings. To model ratings, we learn a *node embedding* and a *type embedding*. Node i's embedding $e_n[i] \in \mathbb{R}^{E_{emb}}$, where E_{emb} is the embedding dimension, is a low-dimensional representation of a node and captures the node's attributes and position. The node embeddings are learned using a graph convolutional network (GCN) (Kipf and Welling 2017), which is a model that leverages graph data by iteratively aggregating and transforming features from neighboring nodes. Thus our node embeddings are *spatially correlated*, capturing the correlation of true incident occurrence across neighborhoods. We also learn type k's embedding $e_\tau[k] \in \mathbb{R}^{E_{emb}}$. The type embedding is a low-dimensional representation of a type and captures the type's features, similarity to other types, and relationship to nodes in the graph. Thus, our type embeddings capture correlations across types. We model ratings as:

$$\text{Pred. rating: } \hat{r}_{ikt} = e_n[i]^\top e_\tau[k] \quad (2)$$

In words, the predicted rating $\hat{r}_{ikt}$ is estimated as the dot product of node i's embedding and type k's embedding.

We model the probability of observing a report as a type-specific function of demographics and the inspection rating. We learn type-specific reporting parameters to account for the fact that different incident types have different reporting characteristics, a claim we confirm on our real rating data. We model the predicted probability of observing a report $\hat{P}(T_{ikt})$ as the following logistic (σ) function:

$$\text{Pred. prob. of report: } \hat{P}(T_{ikt}) = \sigma(\alpha^* r^*_{ikt} + \theta^{*\top} X_i), \quad (3)$$

where α^* is the coefficient on true rating (how relatively likely is an incident with rating r to be reported) and θ^* is the coefficient on node-specific demographic features (is a node with features X_i more likely to report, given the rating). α^* and θ^* are interpretable reporting coefficients that allow us to understand how each demographic covariate impacts reporting probabilities.

As we discuss further below, the values used for $r^*_{ikt}, \alpha^*, \theta^*$ depend on whether a rating r_{ikt} is observed in node/type/week tuple (i, k, t) and whether ratings for other nodes i' or weeks t' for type k are observed ($r_{i'kt'}$). We must consider these cases separately because we cannot simultaneously learn the rating r_{ikt} and the type-specific reporting coefficients $[\alpha_k, \theta_k]$—there is a fundamental unidentifiability between whether conditions are bad (low rating r_{ikt}) but the area does not make a report (small reporting coefficients $[\alpha_k, \theta_k]$) or conditions are good (high rating r_{ikt}) and no report is needed. We now describe the three different cases:

Case 1 – A rating r_{ikt} is observed: Here, the probability of observing a report is a function of demographics X_i and the true observed rating, $r^*_{ikt} = r_{ikt}$. We estimate type specific reporting coefficients, $[\alpha^*_k, \theta^*_k] = [\alpha_k, \theta_k]$.

Case 2 – No rating r_{ikt} is observed at node i but ratings $r_{i'kt'}$ for type k are observed at other nodes i' or times t': Here, we do not have access to node i's true rating, so we model the probability of observing a report as a function of demographics X_i and the *predicted ratings*, $r^*_{ikt} = \hat{r}_{ikt}$. We use type specific reporting coefficients, $[\alpha^*_k, \theta^*_k] = [\alpha_k, \theta_k]$, which are learned via nodes i' or times t' for which ratings $r_{i'kt'}$ are observed.

Case 3 – No ratings r_{ikt} for type k are observed at any node or time period: We do not observe the true rating, so we model the probability of observing a report as a function of demographics X_i and the *predicted ratings*, $r^*_{ikt} = \hat{r}_{ikt}$. We cannot simultaneously learn the rating r_{ikt} and type-specific reporting coefficients $[\alpha_k, \theta_k]$. Thus, we borrow information from types with observed ratings and use the mean reporting coefficients across these types, $[\alpha^*_k, \theta^*_k] = [\overline{\alpha}, \overline{\theta}]$.

For observed types k, we can recover type-specific reporting coefficients $[\alpha_k, \theta_k]$ to account for different types' likelihoods of being reported when incidents occur. For instance, residents may be more likely to report rodents than a noise complaint. We assume that the mean coefficients $[\overline{\alpha}, \overline{\theta}]$ are reasonable for types with unobserved ratings, i.e., the reporting coefficients transfer across types to some extent.

URBAN's modelling contribution lies in adapting GNNs for biased data, and our approach extends to other parameterizations of r_{ikt} and T_{ikt}. While our described model predicts constant ratings $\hat{r}_{ikt}$ and reporting probabilities $\hat{P}(T_{ikt})$ over time, our method generalizes to other GNN architectures, other reporting models, and spatio-temporal methods. We expect all these models *with multiple types of data* to perform well. Regarding time, as new data comes in, our model can be modified to throw away old data (this is how we train models on real data across different time splits), upweight new data, or use a spatiotemporal GNN. Full details are provided in the appendix model section.

Loss function: Our loss function is a weighted sum of four components that evaluate our URBAN model's performance in predicting reports and ratings:

(i) *Report loss for data points with* ***unobserved*** *ratings:* Binary cross entropy (BCE) loss between the true report status T_{ikt} and the predicted probability of a report $\hat{P}(T_{ikt})$ for data points with unobserved ratings.

$$\mathcal{L}_{\text{unobs}} = \sum_{ikt} \mathbb{1}\left(r_{ikt} \text{ unobs}\right) \cdot \text{BCE}(\hat{P}(T_{ikt}), T_{ikt}) \quad (4)$$

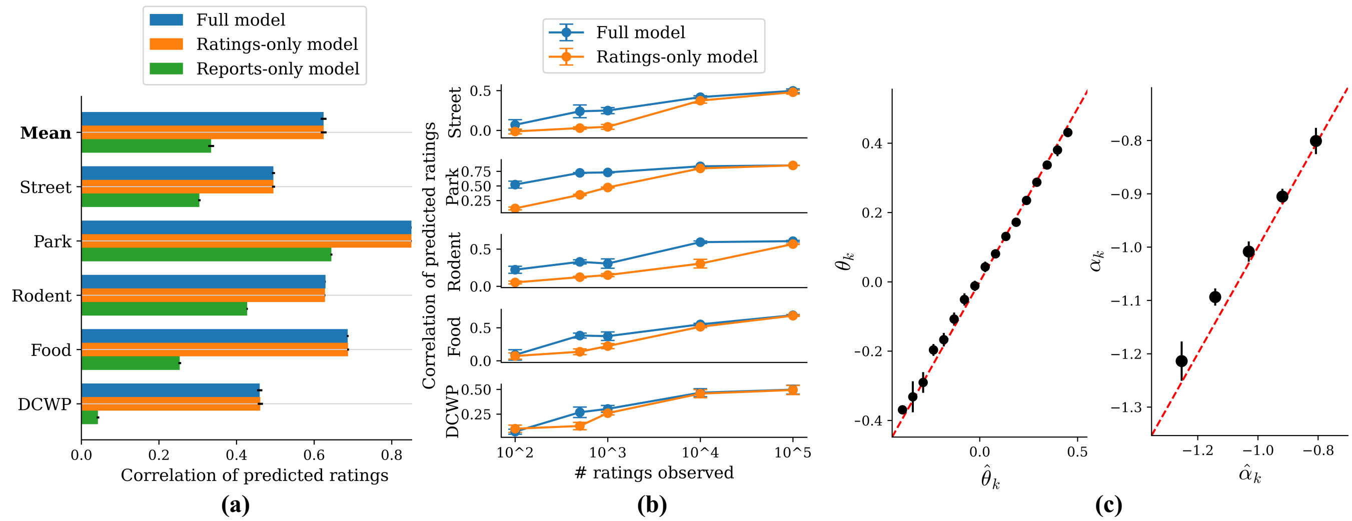


Figure 2: *Semi-synthetic results.* We calculate the correlation between the average predicted and true rating for each node/type pair. We find that (a) our full URBAN model predicts ratings that are more correlated with ground truth than a model that uses only reporting data, and (b) as ratings are more sparsely observed, for types where reports are predictive of ratings, our full model predicts ratings that are more correlated with ground truth than a model that uses only rating data. (c) Our model's predicted coefficients $[\hat{\theta_k}, \hat{\alpha_k}]$ match the true coefficients $[\theta_k, \alpha_k]$ for all types with observed ratings. Panels (a) and (c) show results averaged across 20 semi-synthetic datasets. Panel (b) shows results averaged across 5 datasets.

(ii) *Report loss for data points with **observed** ratings:* BCE between the true report status and the predicted probability of a report for fine-grained data points with observed ratings.

$$\mathcal{L}_{\text{obs}} = \sum_{ikt} \mathbb{1}\left(r_{ikt} \text{ obs}\right) \cdot \text{BCE}(\hat{P}(T_{ikt}), T_{ikt}) \tag{5}$$

(iii) *Rating loss:* Mean squared error (MSE) between the true rating r_{ikt} and the predicted rating $\hat{r}_{ikt}$.

$$\mathcal{L}_{\text{rating}} = \sum_{ikt} \mathbb{1}\left(r_{ikt} \text{ obs}\right) \cdot \text{MSE}(\hat{r}_{ikt}, r_{ikt}) \tag{6}$$

(iv) *Regularization loss:* L^2 norm of the predicted ratings for data points with unobserved ratings. We include this component to prevent our predicted ratings from growing excessively large.

$$\mathcal{L}_{\text{reg}} = \sum_{ikt} \mathbb{1}\left(r_{ikt} \text{ unobs}\right) \cdot \hat{r}_{ikt}^2 \tag{7}$$

The overall loss is:

$$\mathcal{L} = \mathcal{L}_{\text{unobs}} + \gamma_1 \cdot \mathcal{L}_{\text{obs}} + \gamma_2 \cdot \mathcal{L}_{\text{rating}} + \gamma_3 \cdot \mathcal{L}_{\text{reg}} \tag{8}$$

We use weights $\gamma_1, \gamma_2, \gamma_3$ and fix the weight on $\mathcal{L}_{\text{unobs}}$ to 1. We select weights via a hyperparameter search minimizing the RMSE of predicted reports and ratings on a validation split. To aid estimation in real data settings we use two additional losses. Full details are in the appendix model section.

5 Semi-synthetic Experiments

We now validate our approach on semi-synthetic data, using real reporting data but synthetic ratings. In this setting, our URBAN model is *well-specified* and captures the true relationship between ratings and reports. An advantage of our synthetic data is that we can know and vary the parameters. We verify that in these conditions our model can (i) better predict true ratings compared to using either the ratings or reports alone and (ii) recover the reporting parameters.

Experimental Setup

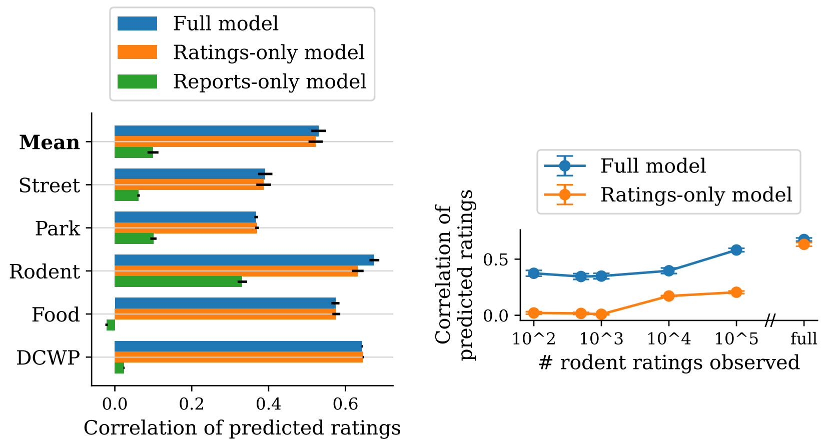
For our semi-synthetic experiments, we use real reporting data T_{ikt} as described in the §3, and we use demographic data X_i from the U.S Census Bureau (United States Census Bureau 2024a,b,c,d,e,f). We generate synthetic inspection ratings r_{ikt} by inverting eq. (3):

$$r_{ikt} = \frac{1}{\alpha_k}\left(\text{logit}(\mathbb{E}_t(T_{ikt})) - \theta_k^\top X_i\right). \tag{9}$$

$\mathbb{E}_t(T_{ikt})$ is the empirical frequency of T_{ikt} over all weeks in the dataset and $[\alpha_k, \theta_k]$ are the reporting coefficients. Our synthetically generated ratings r_{ikt} aim to distributionally match our real ratings. To match the sparsity pattern, we generate a synthetic version of each rating observed in the real data. To test the limits of our model with sparse data, we also run experiments where we further subsample the observed ratings. To reflect real-world reporting parameters, for each type k, we draw reporting parameter θ_k from a Gaussian with a standard deviation of 0.1 and a mean equal to the average reporting coefficient predicted by a logistic regression model run on the real rating data.[2] Full details are available in the appendix semi-synthetic experiments section.

We report results averaged across 20 trials. For each trial, we draw a set of coefficients $[\alpha_k, \theta_k]$; generate a new set of ratings; refit the model; and evaluate the predicted ratings,

[2]We set the intercept and α_k such that our generated r_{ikt} are zero mean and have a standard deviation of 1.



(a) **(b)**

Figure 3: *Real data results.* We calculate the correlation between the average predicted and true rating for each node/type pair. We find that (a) our full URBAN model predicts ratings that are more correlated with ground truth than a model that uses only reporting data, and (b) as ratings are more sparsely observed, for the rodent type where reports are predictive of ratings, our full model predicts ratings that are more correlated with ground truth than a model that uses only rating data. We plot the mean correlation and 95% CIs over contiguous two-year periods between 2021 and 2023.

reports, and coefficients. We use a time-based split. We train on data from January 2022 to June 2023 and test on data from July 2023 to December 2023. We assess the effect of using both reporting and rating data by comparing predictions from models using (i) both reporting and rating data (*full model*), (ii) only reporting data (*reports-only model*), and (iii) only rating data (*ratings-only model*).

To evaluate, we compare predictions to true reports and ratings. Our primary metric is the correlation between predicted ratings/probabilities of a report and the average true ratings/report statuses. This measure evaluates how well our model ranks incidents across nodes for each type. the The appendix semi-synthetic experiments section also reports RMSE which behaves similarly.

Semi-synthetic Data Results

Our URBAN model's predicted ratings and reports correlate with ground truth. As shown in the appendix semi-synthetic experiments section, across all types, the average correlation between our full model's predicted probability of a report and the true probability is $r = 0.49$ ($p < 0.001$ for 81% of types). Across all types with observed ratings, the average correlation between our full model's predicted rating and the true rating is $r = 0.62$ ($p < 0.001$ for all types). Our full model recovers the empirical average ratings for each node/type pair in the training data ($r = 1.00$) and the remaining error is due to the distribution shift between the train and test time periods.

The full model outperforms a model that uses only reporting data: Compared to the reports-only model, the full model predicts ratings that better correlate with ground truth ($r = 0.62$ for the full model vs. 0.33 for the reports-only model). Figure 2 (a) shows the improvement for each type with observed ratings. The reports-only model's ratings are poorly correlated with the ground-truth ratings because it *cannot* recover the true reporting coefficients.

The full model outperforms a model that only uses rating data when rating data is sparse: Next we compare our full model to the ratings-only model. We show that the full model is valuable when ratings are sparsely observed. We run experiments with artificial sparsity by subsampling ratings for each type with observed ratings. We plot the results in Figure 2 (b). For types where reports are predictive of ratings (reports-only model achieves a high correlation, e.g. park and rodent), as we decrease the number of observed ratings, the full model relies on reporting data and the predicted ratings remain correlated with the true ratings, while the ratings-only model approaches random prediction ($r = 0$). More details on the subsampled experiments are in the appendix model section.

The model recovers the true reporting coefficients: As shown in Figure 2 (c), our model recovers the true reporting coefficients $[\alpha_k, \theta_k]$ for incident types with observed ratings. This can only be done by using both types of data together. With only reporting data it is impossible to distinguish between a bad inspection rating that is never reported and a truly good inspection rating. With only rating data we cannot estimate reporting coefficients. Thus, to identify reporting coefficients, one must either use *both* rating and reporting data or make strong parametric assumptions. Additionally, our model only estimates type-specific reporting coefficients for incident types with observed ratings and we accordingly evaluate its performance on these types. For incident types without observed ratings, we use the average reporting coefficients learned across incident types with observed ratings.

The model can learn across types: In the appendix semi-synthetic experiments section we show that, compared to the reports-only model, our full model predicts ratings that are more correlated with ground truth even for types with fully unobserved ratings. Thus rating data for *observed* types helps in predicting ratings for *unobserved* types, and our model can generalize *across* types.

Overall, our results show that our approach helps if our URBAN model is well-specified. In the next section, we assess on real data.

6 Real-world Case Study: New York City Crowdsourced Reporting

We now evaluate our approach using real data, both in terms of predicting ratings and in recovering reporting biases.

Experimental Setup

We use real reporting data T_{ikt} and rating data r_{ikt}, as described in §3. As in the semi-synthetic experiments, we use demographic data X_i from the U.S Census Bureau (United States Census Bureau 2024a,b,c,d,e,f). When a government inspection occurs, we assume that the resulting rating remains constant over time until the next inspection occurs. We split our data into a train and test set using a time-based split, as is standard in urban prediction tasks (Yu, Yin, and Zhu 2018; Farahmand, Xu, and Mostafavi 2023; Huang et al. 2023; Agostini, Pierson, and Garg 2024). We create splits over every contiguous two-year period between January 2021 and December 2023 (13 total periods). For each two year time period, we train on the first 18 months of data and test on the last 6 months. We report results averaged across all 13 time splits.[3]

Results

Prediction on real data is more challenging than semi-synthetic data due to model misspecification. We model the probability of a report as a logistic function of demographics and ratings, which allows us to quantify how reporting rates vary by demographics. But in reality, it is likely that reports are generated by a more complex function with more complex inputs. Nevertheless, our model's predicted ratings and reports still correlate with ground truth. Across all types, the average correlation between our full model's predicted probability of a report and the true probability of a report is $r = 0.24$ ($p < 0.001$ for 83% of types). As shown in the appendix real data experiments section, across all types with observed ratings, the average correlation between our full model's predicted rating and the true rating is $r = 0.53$ ($p < 0.001$ for all types). Additionally, in the appendix real data experiments section we report RMSE results; we evaluate rating predictions using top-k coverage, expected calibration error (ECE), correlation and ECE gaps for demographic subgroups, and spatial equity under budget constraints as measured by the average demographics of Census tracts which receive resources compared to the overall demographics; and we discuss performance across Census tracts.

[3]We report CIs over the 13 overlapping two-year splits between Jan 2021 and Dec 2023.

Covariate	Mean coefficient
log(Population density)	0.250 ± 0.058
log(Median income)	0.173 ± 0.020
Bachelors degree population	0.159 ± 0.018
Households occupied by renter	0.115 ± 0.029
Median age	0.104 ± 0.016
White population	0.093 ± 0.012
True inspection rating	-0.197 ± 0.009

Table 1: *Demographic coefficients.* We report the average predicted univariate coefficients across types with observed ratings. The estimated coefficients capture known demographic factors: e.g., tracts that have a higher income are more likely to report incidents. We also report the average coefficient on the true inspection rating across all models. Tracts that have worse ratings are more likely to have reports. We report the mean coefficients and 95% CIs over all contiguous two year periods between 2021 and 2023.

The full model outperforms a model that uses only reporting data: The correlation between the full model's predicted ratings and the ground-truth ratings is $r = 0.53$ vs. 0.10 for the reports-only model. Figure 3 (a) shows the full model's improvement in predicting ratings for each type with observed ratings. We also compare to Gaussian process (GP) models trained only on the reporting data. We separately fit models for each type with observed ratings. The GP ($r = 0.03$) underperforms both the full model and the reports-only model showing the value of a flexible model that learns across multiple incident types and demographics.

Our key finding is that any model trained on only one source of data is limited. Even test split reports are a poor predictor of test split ratings ($r = 0.0180$ vs. 0.5303 for our full model). Prior work has also found that reports alone cannot accurately predict ground truth (Casey, Wilson, and Yokum 2018). This confirms that there is a low ceiling to the performance of any ML method trained only on reports.

The full model outperforms a model that only uses rating data when rating data is sparse: Next we compare our full model to the ratings-only model. We show that the full model is valuable when ratings are sparsely observed and reports are predictive of ratings. We run experiments with artificial sparsity by subsampling observed ratings. As shown in Figure 3 (b), for the rodent type, where reports are predictive of ratings, as we decrease the number of observed ratings, our full model's predicted ratings remain correlated with the true ratings, while the ratings-only model approaches random prediction ($r = 0$). For the other types with observed ratings, reports are not predictive of ratings (reports-only model achieves a low correlation, $r \leq 0.1$), the model's predictive power comes from rating data, and reporting data minimally improves rating predictions.

The model learns demographic predictors of under-reporting: θ_k measures the contribution of each demographic feature in X_i to the reporting rate. We estimate θ_k by fitting univariate variants of our full model. Each univariate model is trained on both rating and reporting data but only uses one demographic feature. We run a separate univariate model for each demographic feature in X_i. Table 1 shows that the inferred coefficients capture known demographic predictors of underreporting. We find that conditional on ground-truth, neighborhoods with higher population density, higher median income, higher proportion of the population with a bachelor's degree, older residents, higher proportion of white residents, and worse ground-truth all report incidents at higher rates; these estimates are consistent with those of other work (Kontokosta and Hong 2021; Agostini, Pierson, and Garg 2024; Liu, Bhandaram, and Garg 2024; Franchi et al. 2025) and add further evidence of demographic bias in crowdsourced reporting. Coefficients estimated by a multivariate model are in the appendix real data experiments section.

Our full model's predicted ratings are spatially correlated and capture correlations across incident types: We evaluate the interpretability of our model's predicted ratings by analyzing how well they correlate across nodes and types. We first analyze the predicted ratings' spatial correlation. For each node i, we create a vector $\mathbf{r}_i = \{r_{ikt}\}_{k=1}^{\tau}$ of predicted ratings over all types k. We use each node's $\mathbf{r}_i$ vector to cluster the nodes into 4 groups. As shown in the appendix real data experiments section, we find that our clusters are spatially correlated demographically different. We also show that our full model learns more spatially correlated ratings than the ratings-only model. The higher spatial correlations from our full model are consistent with prior work (Agostini, Pierson, and Garg 2024) and potentially rendes our full model to be more interpretable.

Next, we analyze our predicted ratings across complaint types. For each type k, we create a vector $\mathbf{r}_k = \{r_{ikt}\}_{i=1}^{n}$ of predicted ratings over all nodes and cluster the types into 8 groups. Each group contains a coherent cluster of types which we describe in the appendix real data experiments section. Additionally, we show that the dimension of highest variability of $\mathbf{r}_k$ captures type frequency (i.e., $\mathbb{E}_{it}[T_{ikt}]$). These results show that both our predicted and true ratings are spatially correlated and capture similarities across types.

7 Discussion

We address the challenging problem of estimating graph neural networks (GNNs) in settings where we observe biased outcome data. In these settings, nodes have a true latent state that is sparsely observed (e.g., via government inspection ratings). We also frequently observe biased proxies of the latent states (e.g., via crowdsourced reports). We propose a GNN-based model, URBAN, that jointly uses frequently observed biased reporting data and sparsely observed unbiased rating data. We apply our model to New York City 311-data and show that (i) our model makes better predictions of the ground-truth latent state compared to both a baseline model trained only on reporting data and a baseline model trained only on rating data; (ii) our model's inferred reporting coefficients capture known demographic factors associated with underreporting; and (iii) our model's learned ratings are interpretable and capture correlation between nodes and 311-complaint types.

Our findings have important implications for urban resource allocation decisions. NYC, like many municipalities, uses 311 data to help schedule inspections (NYC Open Data 2024e; Tussey and Yan 2025). Prior work has shown that learning *only* from reporting data can lead to biased predictions and downstream inequities in resource allocation. For example, Casey, Wilson, and Yokum (2018) found that in D.C., reports on rodents did not accurately predict ground truth inspection outcomes. Our results reinforce that decisions based *exclusively* on reporting data risk perpetuating demographic biases and misallocating resources, and that jointly modeling reporting data with ground truth outcomes can aid prediction.

There are several avenues for future work. Our model is an example of how to estimate reporting propensity, but future work could investigate whether other methods of incorporating reporting data could produce more accurate rating predictions. Another direction is to apply our model to other settings where demographic reporting disparities exist, such as, crime reporting, public health incident tracking, and across domains where equitable resource allocation depends on accurate understanding of underlying conditions.

Acknowledgments

The authors thank Vince Bartle, Serina Chang, Erica Chiang, Matt Franchi, Sophie Greenwood, Raj Movva, and Kenny Peng for helpful conversations. SB's work is supported by NSF GRFP Grant DGE #2139899. SS's work is supported by NSF GRFP Grant DGE #2141064 and the Hertz Fellowship. EP's work is supported by a Google Research Scholar award, an AI2050 Early Career Fellowship, NSF CAREER #2142419, a CIFAR Azrieli Global scholarship, a gift to the LinkedIn-Cornell Bowers CIS Strategic Partnership, the Survival and Flourishing Fund, Open Philanthropy, and the Zhang Family Endowed professorship. NG's work is supported by NSF CAREER IIS-2339427, NASA, the Sloan Foundation, and Cornell Tech Urban Tech Hub, Google, Meta, and Amazon research awards. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

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.

Agostini, G.; Pierson, E.; and Garg, N. 2024. A Bayesian spatial model to correct under-reporting in urban crowdsourcing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 21888–21896.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(11).

Brimos, P.; Karamanou, A.; Kalampokis, E.; Mamalis, M. E.; and Tarabanis, K. 2023. Explainable graph neural networks on linked statistical data for predicting Scottish house prices. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, 36–41.

Budde, M.; De Melo Borges, J.; Tomov, S.; Riedel, T.; and Beigl, M. 2014. Leveraging spatio-temporal clustering for participatory urban infrastructure monitoring. In *Proceedings of the First International Conference on IoT in Urban Space*, 32–37.

Casey, P. C.; Wilson, K. H.; and Yokum, D. 2018. A cautionary tail: A framework and case study for testing predictive model validity. *arXiv:1807.03860*.

Cheng, S.; Lu, F.; Peng, P.; and Wu, S. 2018. A spatiotemporal multi-view-based learning method for short-term traffic forecasting. *ISPRS International Journal of Geo-Information*, 7(6).

Clark, B. Y.; Brudney, J. L.; Jang, S.-G.; and Davy, B. 2020. Do advanced information technologies produce equitable government responses in coproduction: An examination of 311 systems in 15 US cities. *The American Review of Public Administration*, 50(3): 315–327.

Dai, E.; Aggarwal, C.; and Wang, S. 2021. NRGNN: Learning a label noise resistant graph neural network on sparsely and noisily labeled graphs. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 227–236. Association for Computing Machinery.

Farahmand, H.; Xu, Y.; and Mostafavi, A. 2023. A spatial–temporal graph deep learning model for urban flood nowcasting leveraging heterogeneous community features. *Scientific Reports*, 13(1): 6768.

Franchi, M.; Garg, N.; Ju, W.; and Pierson, E. 2025. Bayesian modeling of zero-shot classifications for urban flood detection.

Gao, Q.; Huang, Y.; Zerhouni, E. G.; and Zheng, Y. 2024. How to improve citizen engagement on public service platforms? The impact of government responsiveness.

Gómez, J. A.; Patiño, J. E.; Duque, J. C.; and Passos, S. 2019. Spatiotemporal modeling of urban growth using machine learning. *Remote Sensing*, 12(1): 109.

Grover, A.; and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864. Association for Computing Machinery.

Hacker, K. P.; Greenlee, A. J.; Hill, A. L.; Schneider, D.; and Levy, M. Z. 2022. Spatiotemporal trends in bed bug metrics: New York City. *PloS ONE*, 17(5): 1–14.

He, S.; Luo, Q.; Du, R.; Zhao, L.; He, G.; Fu, H.; and Li, H. 2023. STGC-GNNs: A GNN-based traffic prediction framework with a spatial–temporal Granger causality graph. *Physica A: Statistical Mechanics and its Applications*, 623: 128913.

Hu, T.; Zhu, X.; Duan, L.; and Guo, W. 2018. Urban crime prediction based on spatio-temporal Bayesian model. *PloS ONE*, 13(10): e0206215.

Huang, S.; Poursafaei, F.; Danovitch, J.; Fey, M.; Hu, W.; Rossi, E.; Leskovec, J.; Bronstein, M.; Rabusseau, G.; and Rabbany, R. 2023. Temporal graph benchmark for machine learning on temporal graphs. In *Advances in Neural Information Processing Systems*, volume 36, 2056–2073.

Jiménez-Jiménez, S. I.; Ojeda-Bustamante, W.; Ontiveros-Capurata, R. E.; and Marcial-Pablo, M. d. J. 2020. Rapid urban flood damage assessment using high resolution remote sensing data and an object-based approach. *Geomatics, Natural Hazards and Risk*, 11(1): 906–927.

Kapoor, A.; Ben, X.; Liu, L.; Perozzi, B.; Barnes, M.; Blais, M.; and O'Banion, S. 2020. Examining COVID-19 forecasting using spatio-temporal graph neural networks. *MLG workshop @ KDD'2020, epiDAMIK workshop @ KDD'2020*.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*.

Kontokosta, C.; Hong, B.; and Korsberg, K. 2017. Equity in 311 reporting: Understanding socio-spatial differentials in the propensity to complain. *arXiv:1710.02452*.

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. *Sustainable Cities and Society*, 64: 102503.

Li, J.; Xiong, C.; and Hoi, S. C. 2021. CoMatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9475–9484.

Li, Y.; Hyder, A.; Southerland, L. T.; Hammond, G.; Porr, A.; and Miller, H. J. 2020. 311 service requests as indicators of neighborhood distress and opioid use disorder. *Scientific Reports*, 10(1): 19579.

Li, Y.; Yin, J.; and Chen, L. 2023. Informative pseudo-labeling for graph neural networks with few labels. *Data Mining and Knowledge Discovery*, 37(1): 228–254.

Liu, Z.; Bhandaram, U.; and Garg, N. 2024. Quantifying spatial under-reporting disparities in resident crowdsourcing. *Nature Computational Science*, 4(1): 57–65.

Liu, Z.; and Garg, N. 2024. Redesigning Service Level Agreements: Equity and Efficiency in City Government Operations. In *Proceedings of the 25th ACM Conference on Economics and Computation*, 309–309.

Lum, K.; and Isaac, W. 2016. To predict and serve? *Significance*, 13(5): 14–19.

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*, 191–210. ISBN 978-3-030-95721-6.

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

NYC Open Data. 2023. Street Rating. https://data.cityofnewyork.us/Transportation/Street-Rating/mxi3-5xz5. Accessed: 2024-03-09.

NYC Open Data. 2024a. 311 Service Requests from 2010 to Present. https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9. Accessed: 2024-03-22.

NYC Open Data. 2024b. Department of Consumer and Worker Protection (DCWP) Inspections. https://data.cityofnewyork.us/Business/Department-of-Consumer-and-Worker-Protection-DCWP-/jzhd-m6uv. Accessed: 2024-07-10.

NYC Open Data. 2024c. DOHMH New York City Restaurant Inspection Results. https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j. Accessed: 2024-07-09.

NYC Open Data. 2024d. Parks Inspection Program – Inspections. https://data.cityofnewyork.us/dataset/Parks-Inspection-Program-Inspections/yg3y-7juh. Accessed: 2024-03-16.

NYC Open Data. 2024e. Rodent Inspection. https://data.cityofnewyork.us/Health/Rodent-Inspection/p937-wjvj. Accessed: 2024-04-04.

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.

Pan, Z.; Wang, Z.; Wang, W.; Yu, Y.; Zhang, J.; and Zheng, Y. 2019. Matrix factorization for spatio-temporal neural networks with applications to urban flow prediction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, 2683–2691. Association for Computing Machinery.

Peng, F.; Lu, W.; Tan, W.; Qi, K.; Zhang, X.; and Zhu, Q. 2022. Multi-output network combining GNN and CNN for remote sensing scene classification. *Remote Sensing*, 14(6).

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. DeepWalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 701–710. Association for Computing Machinery.

Pineda, J.; Midtvedt, B.; Bachimanchi, H.; Noé, S.; Midtvedt, D.; Volpe, G.; and Manzo, C. 2023. Geometric deep learning reveals the spatiotemporal features of microscopic motion. *Nature Machine Intelligence*, 5(1): 71–82.

Qian, S.; Ying, H.; Hu, R.; Zhou, J.; Chen, J.; Chen, D. Z.; and Wu, J. 2023. Robust training of graph neural networks via noise governance. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 607–615.

Roy, A.; Roy, K. K.; Ahsan Ali, A.; Amin, M. A.; and Rahman, A. M. 2021. SST-GNN: Simplified spatio-temporal traffic forecasting model using graph neural network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 90–102. Springer.

Samal, K. K. R.; Babu, K. S.; and Das, S. K. 2022. Multi-output spatio-temporal air pollution forecasting using neural network approach. *Applied Soft Computing*, 126: 109316.

Sarto, S. D.; Ranalli, M. G.; Bakar, K. S.; Cappelletti, D.; Moroni, B.; Crocchianti, S.; Castellini, S.; Spataro, F.; Esposito, G.; Ianniello, A.; et al. 2016. Bayesian spatiotemporal modeling of urban air pollution dynamics. In *Topics on Methodological and Applied Statistical Inference*, 95–103.

Shao, Z.; Xu, Y.; Wei, W.; Wang, F.; Zhang, Z.; and Zhu, F. 2023. Heterogeneous graph neural network with multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(11): 11476–11488.

Sharma, D.; and Jones, M. 2023. Efficiently learning the graph for semi-supervised learning. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, 1900–1910.

Song, Z.; Yang, X.; Xu, Z.; and King, I. 2023. Graph-based semi-supervised learning: A comprehensive review. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11): 8174–8194.

Tang, J.; Qu, M.; Wang, M.; Zhang, M.; Yan, J.; and Mei, Q. 2015. LINE: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, 1067–1077.

Tomy, A.; Razzanelli, M.; Di Lauro, F.; Rus, D.; and Della Santina, C. 2022. Estimating the state of epidemics spreading with graph neural networks. *Nonlinear Dynamics*, 109(1): 249–263.

Tussey, D.; and Yan, J. 2025. Principles for Open Data Curation: A Case Study with the New York City 311 Service Request Data. *arXiv:2502.08649*.

United States Census Bureau. 2024a. Educational attainment for the population 25 years and over. https://data.census.gov/table/ACSDT1Y2023.B15003?q=B15003. Accessed: 2023-12-18.

United States Census Bureau. 2024b. Hispanic or latino, and not hispanic or latino by race. https://data.census.gov/table/DECENNIALCD1182020.P9?q=P9. Accessed: 2023-12-18.

United States Census Bureau. 2024c. Income in the past 12 months (in 2023 inflation-adjusted dollars). https://data.census.gov/table/ACSST1Y2023.S1901?q=S1901. Accessed: 2023-12-18.

United States Census Bureau. 2024d. Median age by sex. https://data.census.gov/table/DECENNIALCD1182020.P13?q=P13. Accessed: 2023-12-18.

United States Census Bureau. 2024e. Tenure. https://data.census.gov/table/ACSDT1Y2023.B25003?q=B25003. Accessed: 2023-12-18.

United States Census Bureau. 2024f. Total population. https://data.census.gov/table/DECENNIALCD1182020.P1?q=p1. Accessed: 2023-12-18.

Verma, V.; Qu, M.; Kawaguchi, K.; Lamb, A.; Bengio, Y.; Kannala, J.; and Tang, J. 2021. GraphMix: Improved training of GNNs for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10024–10032.

Wang, L.; Adiga, A.; Chen, J.; Sadilek, A.; Venkatramanan, S.; and Marathe, M. 2022a. CausalGNN: Causal-based graph neural networks for spatio-temporal epidemic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12191–12199.

Wang, L.; Qian, C.; Kats, P.; Kontokosta, C.; and Sobolevsky, S. 2017. Structure of 311 service requests as a signature of urban location. *PloS ONE*, 12(10): e0186314.

Wang, Y.; Zheng, J.; Du, Y.; Huang, C.; and Li, P. 2022b. Traffic-GGNN: Predicting traffic flow via attentional spatial-temporal gated graph neural networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(10): 18423–18432.

Wu, L.; Hou, Z.; Yuan, J.; Rong, Y.; and Huang, W. 2024. Equivariant spatio-temporal attentive graph networks to simulate physical dynamics. In *Advances in Neural Information Processing Systems*, volume 36.

Xie, Z.; Lv, W.; Huang, S.; Lu, Z.; Du, B.; and Huang, R. 2019. Sequential graph neural network for urban road traffic speed prediction. *IEEE Access*, 8: 63349–63358.

Xu, L.; Kwan, M.-P.; McLafferty, S.; and Wang, S. 2017. Predicting demand for 311 non-emergency municipal services: An adaptive space-time kernel approach. *Applied Geography*, 89: 133–141.

Yu, B.; Yin, H.; and Zhu, Z. 2018. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, 3634–3640. AAAI Press.

Yu, S.; Xia, F.; Li, S.; Hou, M.; and Sheng, Q. Z. 2023. Spatio-temporal graph learning for epidemic prediction. *ACM Transactions on Intelligent Systems and Technology*, 14(2): 1–25.

Zaouche, M.; and Bode, N. W. 2023. Bayesian spatio-temporal models for mapping urban pedestrian traffic. *Journal of Transport Geography*, 111: 103647.

Zhan, W.; and Datta, A. 2024. Neural networks for geospatial data. *Journal of the American Statistical Association*, 0: 1–21.

Zhou, D.; Bousquet, O.; Lal, T.; Weston, J.; and Schölkopf, B. 2003. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, volume 16.

Zhou, X.; Liu, X.; Yu, H.; Wang, J.; Xie, Z.; Jiang, J.; and Ji, X. 2024. Variance-enlarged Poisson learning for graph-based semi-supervised learning with extremely sparse labeled data. In *International Conference on Learning Representations (ICLR)*.

Zhu, X.; Ghahramani, Z.; and Lafferty, J. D. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*, 912–919.