End-to-end Auditing of Decision Pipelines

Benjamin Laufer [1] Emma Pierson [2] Nikhil Garg [3]

Abstract

Many high-stakes policies can be modeled as a sequence of decisions along a pipeline. We are interested in auditing such pipelines for both efficiency and equity. Using a dataset of over 100,000 crowdsourced resident requests for po- life-tentially hazardous tree maintenance in New York City, we observe a sequence of city government decisions about whether to inspect and work on a reported incident. At each decision in the pipeline, we define parity definitions and tests to identify inefficient, inequitable treatment. Disparities in resource allocation and scheduling across census tracts are reported as preliminary results.

1. Introduction

Major cities use resident crowdsourcing (or "co-production"), in which the public reports problems (such as downed trees and power-lines, or potholes) to the government (Clark et al., 2020; Liu & Garg, 2022). For example, New York City's 311 system received over 2.6 million requests in 2021. A resident report triggers a pipeline of bureaucratic actions: an inspection involving an agency member visiting the incident location, and then a work order to fix the issue if necessary. Each pipeline stage involves both an allocation decision (whether to inspect/work), and a scheduling one (when to do so). Figure 1 illustrates the specific pipeline we study.

We are interested in auditing such pipelines for both efficiency and equity. Sequential decision-making within pipelines (Arunachaleswaran et al., 2020) requiring predictive inference (Kleinberg et al., 2015) occurs in many critical domains with equity concerns – including education (Estrada et al., 2016) and criminal justice (Rehavi & Starr, 2014). However, such pipelines are often difficult to study empirically: (1) pipelines are long, often spanning many years and multiple decision-makers; and (2) there is substantial unobserved confounding between decision stages (Knox et al., 2020).

Our empirical focus is on policy decisions made by the New York City Department of Parks and Recreation (DPR) about its street trees. Street trees are important: they provide saving temperature reductions in urban spaces (Ziter et al., 2019), and when they fall they can cause significant infrastructure damage and death. DPR is responsible for managing about 700,000 such trees and fielding about 100,000 reports annually (ranging from hazards to pruning requests). Due to resource constraints, it only inspects about $\frac{2}{3}$ of reported incidents and conducts work after about $\frac{1}{2}$ of the inspections. This setting has numerous advantages as an avenue to more generally audit responsible sequential decision-making along pipelines: (1) the pipelines are centralized and short, with most incidents being resolved within a few months, providing sufficient outcome data to rigorously audit performance; (2) there is arguably little unobserved confounding,[1] allowing us to focus on general selective labels (Lakkaraju et al., 2017) and unfairness accumulation challenges; and (3) regular discussions with DPR officials provide us both vital contextual knowledge and a potential avenue to change operations as a result of our findings.[2] Liu & Garg (2022) recently studied heterogeneous reporting behavior by the public in the same empirical context; we focus on the decisions made by the agency in response to public reports.

2. Research Questions and Challenges

Our pipeline model is illustrated in Figure 1. After an $\mathtt{incident}$ (such as a dangling tree limb), a resident may submit a $\mathtt{report}$ after some delay (cf. Liu & Garg (2022)). The agency schedules an inspection (which we refer to as $\mathtt{insp}$) time $t_{\mathtt{report}\to\mathtt{insp}}$ after the report, for a subset of the reported incidents. One of the outcomes of the inspection is a risk rating r;[3] crucially, these risk ratings do not depend

[1]Department of Information Science, Cornell Tech, New York, USA [2]Department of Computer Science, Cornell Tech, New York, USA [3]Department of Operations Research and Information Engineering, Cornell Tech, New York, USA. Correspondence to: Benjamin Laufer <bdl56@cornell.edu>, Emma Pierson <ep432@cornell.edu>, Nikhil Garg <ng343@cornell.edu>.

ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA, 2022.

[1]Reports are viewed by foresters via a centralized electronic dashboard, all the features of which we have access, cf. (Hangartner et al., 2021). Discussions with agency officials also confirm that decisions are made primarily based on variables we observe.

[2]For example, while much of the reporting data is public, DPR has provided us with internal data on inspections and work orders.

[3]From the city's Tree Risk Management standards, this risk rating is a combination of three sub-ratings: 1) likelihood of tree fail-

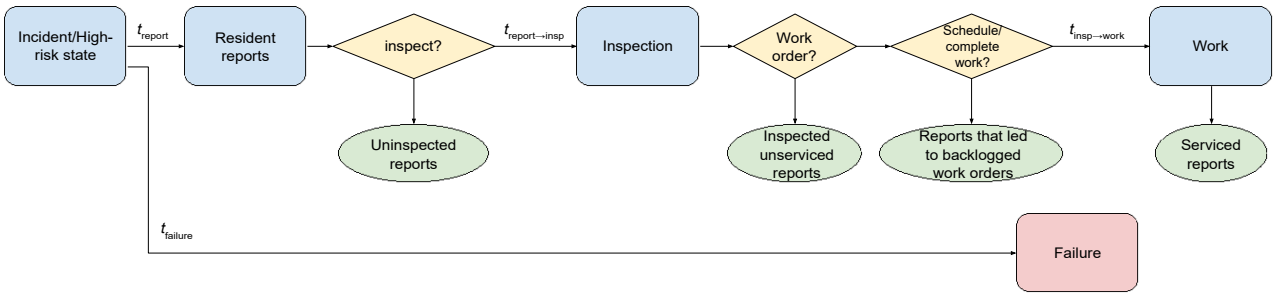

Figure 1. The pipeline of decisions and events that occurs when an incident occurs leaving a tree in potentially high-risk status. Blue and red squares correspond to events, yellow diamonds correspond to the city's operational decisions, green circles represent the states of various observations (or rows) in our dataset.

on operational constraints such as work order capacity. The inspector may also create a work order (`work order created`) through a dashboard, a subjective decision that primarily depends on r but may also depend on capacity. Ultimately, work is scheduled and `completed` by a maintenance crew after $t_{insp \to work}$ days, or else the order is backlogged and not `completed`.

We wish to audit the overall efficiency and equity of this pipeline: is the agency prioritizing the riskiest incidents, and, if not, is it unduly prioritizing some neighborhoods or populations over others? However, there are several challenges in answering those questions.

1. Ultimately, cumulative decisions across stages matter, and inefficiencies or inequities at each stage may either accumulate or cancel out in the aggregate (D'Amour et al., 2020; Rehavi & Starr, 2014). Analyzing one stage in isolation may give an inaccurate view. Furthermore, both the allocation (which incidents are inspected/worked on) and the scheduling (inspection/work delay) dimensions must be incorporated into any single measure.

2. There is a selective labels challenge: as in many high-stakes decision-making pipelines, our access to data is censored in a non-random fashion (Lakkaraju et al., 2017). We observe the assessed risk (and resulting work order status) only for reports that are inspected – and thus lack direct knowledge of whether uninspected incidents are high-risk and in need of work.

3. True distributions of tree risk may differ across neighborhoods, rendering naive measures of inequity invalid. In particular, if a city inspects trees above a risk threshold which remains equitable and consistent across neighborhoods, simply examining the mean risk of inspected trees may still differ across neighborhoods, implying inequity even though the thresholds remain consistent. (This is a specific example of infra-marginality, which complicates analyses of inequity in other contexts like criminal justice as well (Ayres, 2002; Simoiu et al., 2017; Pierson et al., 2018; 2020)).

These challenges are general to pipeline decision-making, and we develop an auditing framework that addresses them, drawing from related literatures with similar challenges. For space reasons, we include partial results here, focusing on economic equity considerations.

3. Methods and Metrics

In this section, we describe the methods used to conduct an end-to-end audit on DPR's decision-making pipeline, drawing on formal methods for assessing fairness proposed in recent years.

Sensitive attributes In discussing the relevant notions of fairness in auditing DPR decisions, we define a sensitive attribute g along which we would conclude a system is inequitable if neighborhoods with different values of g receive different levels of DPR resources for comparable incidents. Here, we present results when defining g as the median income in the tree's census tract, but our approach extends to other sensitive attributes as well.

Assessing inequity A first attempt at quantifying inequity is simply to examine how the rate of DPR decisions varies by income—e.g., how much likelier are incidents in higher-income neighborhoods to get inspected. This is a straightforward approach and we report these raw disparities. However, this approach suffers from a type of omitted variable bias—if higher-income neighborhoods receive more inspections, it could simply be because their incidents are truly riskier (e.g., their trees are ten years older on average, or have larger diameters). To adjust for this, we use two controls for differential riskiness. First, trees inspected by DPR receive a risk score r, which we can use as a control among the

ure 2) likelihood of impacting target 3) consequences of impacting target. This strategy is outlined on public NYC Parks documents.

inspected trees. To develop a riskiness measure that extends to uninspected trees, we train a machine learning model (Extreme Gradient Boosted Decision Tree (XGBoost) (Chen et al., 2015)) on historical data (using features related to the service request category, borough/community board, weekly capacity, month, and census tract demographics) to predict a tree's risk score r; we use $\hat{r}$ to denote the model prediction. $\hat{r}$ can be evaluated even on uninspected trees. (This approach is adapted from Jung et al. (2018)'s Risk-Adjusted Regression approach.) To assess inequity when adjusting for risk, we regress DPR's decision—e.g., whether or not to inspect a tree—on the risk measure r or $\hat{r}$ and the sensitive attribute g. This lets us assess whether neighborhoods of different incomes are treated differently when controlling for risk.

Evaluating fairness of a single pipeline step Because we are dealing with several stages in a decision pipeline where data are not observed for all reports, fairness considerations must be carefully dealt with at each particular step. To illustrate the process of evaluating fairness at one stage in the decision pipeline, here we detail one such step: whether a work order is conducted after an inspection, controlling for risk. Here we are able to observe the true risk r associated with a report, because an inspector visited the site and recorded their assessment. We limit this analysis to the set of trees which have received an inspection, which precedes a work order in almost all cases. In considering whether the DPR's decision is fair across neighborhoods of different income at this stage, we compare the probability $P(\texttt{work order} \mid r, \texttt{income})$ – i.e., whether the probability a tree receives a work order varies by income when controlling for risk. Specifically, we conduct a logistic regression where the dependent variable is `work order` and the regressors are `income` and r. The regression coefficient on `income` then captures how much DPR work order decisions vary by income when controlling for risk. We also conduct a similar process to inspect the decision to inspect given a report, for which we instead use predicted risk $\hat{r}$, using only covariates available before an inspection.

4. Fairness Definitions

4.1. Parity in Allocation Decisions

We define fairness definitions for allocating resources (including inspectors and workers) in NYC parks policies.

Definition 4.1. Inspection Parity: Given a `report` with sensitive group attribute g, the `insp` decision exhibits inspection parity between groups g_i and g_j if:

$$P(\texttt{insp} \mid g = g_i, \texttt{report}) = P(\texttt{insp} \mid g = g_j, \texttt{report})$$

This parity definition is defined on the set of all received reports, and that the conditioning is on the existence of a report, not its content.

Definition 4.2. Work Order Parity: Given a `report` with sensitive group attribute g that received an inspection, the `work order` decision exhibits work order parity if:

$$P(\texttt{work order} \mid g = g_i, \texttt{insp}) = P(\texttt{work order} \mid g = g_j, \texttt{insp})$$

Note that this parity definition is only defined on the set of reports that are inspected.

Definition 4.3. Work Completion Parity: Given a `report` with sensitive group attribute g that received an `insp` and `work order`, the `work completed` decision exhibits work completion parity if:

$$P(\texttt{work completed} \mid g = g_i, \texttt{work order}) = P(\texttt{work completed} \mid g = g_j, \texttt{work order})$$

This parity definition is only defined on the set of reports that receive a work order that either is or isn't completed.

4.2. Parity in Scheduling Decisions

Here we report the list of parity definitions for temporal (scheduling) efficiency in NYC parks policies.

Definition 4.4. Inspection Time Parity: Given an inspected `report` with sensitive group attribute g, the time from report to inspection $t_{report \to insp}$ is a decision made by NYC DPR. Such a decision exhibits inspection time parity if:

$$E[t_{report \to insp} \mid g = g_i] = E[t_{report \to insp} \mid g = g_j]$$

This parity definition is only defined on the set of reports that are inspected.

Definition 4.5. Work Time Parity: Given a `report` with sensitive group attribute g, the time from report to work completion $t_{insp \to work}$ is a scheduling decision made by NYC DPR. Such a decision exhibits work time parity if:

$$E[t_{insp \to work} \mid g = g_i] = E[t_{insp \to work} \mid g = g_j]$$

This parity definition is only defined on the set of reports that ultimately have completed work done.

4.3. Risk-adjusted Parity

Per Jung et al. (2018), risk-adjusted regression tests for parity include the (predicted or observed) risk as a regressor in order to directly compare parity among reports that are of the same risk level—whereas the above parity definitions could suffer from inframarginality issues.

When we report both non-risk-adjusted and risk-adjusted results, we intend to show a more naive measure of disparate impact (reported in the definitions above), and then see whether such disparities (if observed) are explained by differences in risk associated with the reports across groups.

Parity Defn.	Population	Coef. (naive)	Coef. (risk-adjusted)	Benefits...
Inspection Parity (4.1)	All reports	–	–	Lower-income
Work Order Parity (4.2)	Inspected reports	+	+	Higher-income
Work Completion Parity (4.3)	Reports w/ work order	+	+	Higher-income

Table 1. Equity tests for binary allocation decisions of whether to inspect, whether to order work, and whether to complete work. Parity is tested by performing a Logistic Regression of pipeline events on sensitive feature $\ln(\text{income})$, measured as the natural log of the median income for the census tract associated with a report, either as the sole regressor (naive) or controlling for predicted or observed risk as an additional regressor (risk-adjusted). Findings are preliminary and use a subset of available data and are therefore subject to change; however, all results are statistically significant at $p < 0.001$ except the naive inspection parity test.

Parity Defn.	Population	Coef. (naive)	Coef. (risk-adjusted)	Benefits...
Inspection Time Parity (4.4)	Inspected reports	–	–	Higher-income
Work Time Parity (4.5)	Reports w/ work completed	–	–	Higher-income

Table 2. Equity tests for temporal scheduling decisions of when to inspect and when to conduct work. Parity is tested by conducting an OLS regression of pipeline delays on sensitive feature $\ln(\text{income})$, measured as the natural log of the median income for the census tract associated with a report, either as the sole regressor (naive) or controlling for predicted or observed risk as an additional regressor (risk-adjusted). These preliminary directional findings use a subset of available data and are therefore subject to change; however, all results are statistically significant at $p < 0.001$.

5. Results

In this section, we present preliminary results that suggest that there exist lower-risk sites that receive systematically more attention and resources, faster, than higher-risk sites elsewhere in the city. Further, these sites are not uniformly geographically distributed; nor are patterns consistent across every stage in the decision-making pipeline.

Table 1 shows the allocative decisions at each step in the decision-pipeline: $\texttt{insp}$, $\texttt{work order}$, $\texttt{work completed}$. Notably, the inspection decisions made by DPR seem to devote disproportionate attention to lower-income neighborhoods, on average, even when adjusting for ML-predicted risk $\hat{r}$. This is perhaps due to more concerted audits being directed solely at the inspection stage.[4] However, looking at the other stages in the decision pipeline, a picture starts to emerge that is entirely different from what an inspection-auditor might conclude: In creating and completing work orders, the parks department disproportionately allocates resources to higher-income neighborhoods, even when controlling for observed report risk r.

Table 2 shows the temporal (or scheduling) equity considerations related to the DPR decision pipeline. Regression coefficients were produced using an OLS (rather than Logistic Regression) because the outcome various is a continuous time quantity rather than a binary decision. Coefficients suggest that at the same level of risk, a 10% increase in median neighborhood income is associated with about a $\frac{1}{2}$ day expedite in how long the neighborhood can expect DPR to conduct an inspection. For work orders, a 10% increase in median neighborhood income is associated with work being completed about 2 days faster, after an inspection.

Crucially, the hetergeneity of our results at different stages in the decision pipeline show that end-to-end analysis is necessary for auditing sequential decisions. It is only through systematically observing equity concerns at every component decision that a larger picture starts to emerge.

6. Conclusion

We develop a framework to audit sequences of decisions end-to-end for inefficiency and inequity. Using data from NYC Department of Parks and Recreation, we analyze sequential decisions in urban governance with arguably low unobserved confounding effects and relatively short time-frames between decisions. Using the dataset, we measure equity concerns at each decision along the pipeline using conditional-probability definitions. Preliminary results indicate that while inspection allocation decisions seem to over-allocate to lower-income neighborhoods, an end-to-end analysis reveals that for each subsequent decision in the pipeline, reports from lower-income-neighborhoods are less likely to receive work and it takes longer for work to occur, on average, on potentially hazardous trees.

[4] Inspection audits are mentioned in DPR documentation and publicly accessible reports.

References

Abner, K. A., Li, J., and Garner, C. Food inspections and bias in chicago. 2019.

Arunachaleswaran, E. R., Kannan, S., Roth, A., and Ziani, J. Pipeline interventions. arXiv preprint arXiv:2002.06592, 2020.

Ayres, I. Outcome tests of racial disparities in police practices. Justice research and Policy, 4(1-2):131–142, 2002.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., et al. Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4):1–4, 2015.

Clark, B. Y., Brudney, J. L., Jang, S.-G., and Davy, B. 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, 2020.

D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., and Halpern, Y. Fairness is not static: deeper understanding of long term fairness via simulation studies. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 525–534, 2020.

Estrada, M., Burnett, M., Campbell, A. G., Campbell, P. B., Denetclaw, W. F., Gutiérrez, C. G., Hurtado, S., John, G. H., Matsui, J., McGee, R., et al. Improving underrepresented minority student persistence in stem. CBE—Life Sciences Education, 15(3):es5, 2016.

Hangartner, D., Kopp, D., and Siegenthaler, M. Monitoring hiring discrimination through online recruitment platforms. Nature, 589(7843):572–576, 2021.

Harris, J. K., Mansour, R., Choucair, B., Olson, J., Nissen, C., and Bhatt, J. Health department use of social media to identify foodborne illness—chicago, illinois, 2013–2014. MMWR. Morbidity and mortality weekly report, 63(32): 681, 2014.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., and Mitchell, M. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 560–575, 2021.

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

Kannan, V., Shapiro, M. A., and Bilgic, M. Hindsight analysis of the chicago food inspection forecasting model. arXiv preprint arXiv:1910.04906, 2019.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. Prediction policy problems. American Economic Review, 105(5):491–95, 2015.

Knox, D., Lowe, W., and Mummolo, J. Administrative records mask racially biased policing. American Political Science Review, 114(3):619–637, 2020.

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

Lee, M. S. A. and Singh, J. Risk identification questionnaire for unintended bias in machine learning development lifecycle. Available at SSRN, 2021.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. Delayed impact of fair machine learning. In International Conference on Machine Learning, pp. 3150–3158. PMLR, 2018.

Liu, Z. and Garg, N. Equity in resident crowdsourcing: Measuring under-reporting without ground truth data. arXiv preprint arXiv:2204.08620, 2022.

McBride, K., Aavik, G., Kalvet, T., and Krimmer, R. Co-creating an open government data driven public service: The case of chicago's food inspection forecasting model. In Proceedings of the 51st Hawaii International Conference on System Sciences, 2018.

Pierson, E., Corbett-Davies, S., and Goel, S. Fast threshold tests for detecting discrimination. In International Conference on Artificial Intelligence and Statistics, pp. 96–105. PMLR, 2018.

Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., Ramachandran, V., Barghouty, P., Phillips, C., Shroff, R., et al. A large-scale analysis of racial disparities in police stops across the united states. Nature human behaviour, 4(7):736–745, 2020.

Rehavi, M. M. and Starr, S. B. Racial disparity in federal criminal sentences. Journal of Political Economy, 122 (6):1320–1354, 2014.

Sadilek, A., Caty, S., DiPrete, L., Mansour, R., Schenk, T., Bergtholdt, M., Jha, A., Ramaswami, P., and Gabrilovich, E. Machine-learned epidemiology: real-time detection of foodborne illness at scale. NPJ digital medicine, 1(1): 1–7, 2018.

Saltzer, J. H., Reed, D. P., and Clark, D. D. End-to-end arguments in system design. ACM Transactions on Computer Systems (TOCS), 2(4):277–288, 1984.

Shaikh, S., Vishwakarma, H., Mehta, S., Varshney, K. R., Ramamurthy, K. N., and Wei, D. An end-to-end machine learning pipeline that ensures fairness policies. arXiv preprint arXiv:1710.06876, 2017.

Simoiu, C., Corbett-Davies, S., and Goel, S. The problem of infra-marginality in outcome tests for discrimination. The Annals of Applied Statistics, 11(3):1193–1216, 2017.

Singh, S., Shah, B., and Kash, I. A. Fair decision-making for food inspections. arXiv preprint arXiv:2108.05523, 2021.

Suresh, H. and Guttag, J. A framework for understanding sources of harm throughout the machine learning life cycle. In Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 1–9. 2021.

Ziter, C. D., Pedersen, E. J., Kucharik, C. J., and Turner, M. G. Scale-dependent interactions between tree canopy cover and impervious surfaces reduce daytime urban heat during summer. Proceedings of the National Academy of Sciences, 116(15):7575–7580, 2019.

A. Further Related Works

The Parks Department does not only make one decision per report—as in many operational decisions, the NYC Parks policies involve sequences of dependent choices which may be modeled as a pipeline. Pipeline models have proven useful in identifying sources of harm in machine learning (Suresh & Guttag, 2021; Hutchinson et al., 2021; Liu et al., 2018). This way of identifying risks and interventions has emerged as a theme in theoretical work (Arunachaleswaran et al., 2020) as well as technical approaches to identifying biases and fairness concerns (Shaikh et al., 2017; Lee & Singh, 2021). Appropriately considering the sequential nature of DPR's decisions requires an auditing strategy that is end-to-end (Saltzer et al., 1984), meaning each component decision is considered to make conclusions about the system as a whole.

Chicago Food Inspections A number of scholars have directed attention to using data to improve food and safety inspections, most notably for Chicago (Singh et al., 2021; Kannan et al., 2019; Abner et al., 2019; McBride et al., 2018; Harris et al., 2014; Sadilek et al., 2018). We note two particular observations from these studies: First, inspections do not prompt subsequent resource allocation in the form of manual work that takes time and effort–the sequential nature of NYC's pipeline of decisions is a departure from Chicago's food inspections. Second, the allocational dimension of inspection problems is not relevant in the Chicago Food inspections case because all incidents are ultimately inspected and risk prediction models only impact the temporal or scheduling priority that inspections receive (Kannan et al., 2019) . As such, Singh et al. (2021)'s development of fairness notions for food inspections informs our below definitions but, appropriately for their setting, does not rely on as many conditional pipeline steps and probabilistic formalisms; and only considers the termporal dimension of fairness.