Redesigning Service Level Agreements: Equity and Efficiency in City Government Operations

ZHI LIU, Cornell Tech, USA
NIKHIL GARG, Cornell Tech, USA

We consider government service allocation – how the government allocates resources (e.g., maintenance of public infrastructure) over time. It is important to make these decisions efficiently and equitably – though these desiderata may conflict. In particular, we consider the design of Service Level Agreements (SLA) in city government operations: promises that incidents such as potholes and fallen trees will be responded to within a certain time. We model the problem of designing a set of SLAs as an optimization problem with different equity and efficiency objectives under a queuing network framework; the city has two decision levers: how to allocate response budgets to different neighborhoods, and how to schedule responses to individual incidents. We: (1) Theoretically analyze a stylized model and find that the "price of equity" is small in realistic settings; (2) Develop a simulation-optimization framework to optimize policies in practice; (3) Apply our framework empirically using data from NYC, finding that: (a) status quo inspections are highly inefficient and inequitable compared to optimal ones, and (b) in practice, the equity-efficiency tradeoff is not substantial: generally, inefficient policies are inequitable, and vice versa.

Manuscript submitted for review to the 25th ACM Conference on Economics & Computation (EC'24).

1 INTRODUCTION

Government, especially municipal government, makes allocation decisions over time: *when* and *where* to build infrastructure (such as roadways, parks, public transportation stations) or provide services (maintenance of resources, garbage collection, restaurant inspections). It must do so *efficiently* (invest resources toward the most urgent tasks) and *equitably* (do not unduly prioritize one neighborhood over another). However, these desiderata may conflict: "efficient" prioritization may mean that one area receives fewer services. We consider two aspects of this challenge: (a) **Analysis:** What does this efficiency-equity trade-off look like, i.e., when do we expect the price of equity to be large? (b) **Engineering:** How do we design efficiency and equitable policies in practice, in a data-driven manner?

We study these questions in the context of *policies* to respond to time-sensitive *incidents* – for example, scheduling inspections and maintenance crews in response to downed trees, flooding, or power outages. We consider the optimization of two government policy levers: **response budgets** in each neighborhood (i.e., number of workers who can respond to incidents) and **allocation guidelines** for how workers prioritize incidents of different types. Why these levers, instead of directly optimizing online, incident-level decisions? Spatial budget levels describe the status quo, for both administrative and logistical reasons: worker home offices are distributed throughout the city as determined by the budgets, and it is more efficient for a worker to respond to spatially nearby incidents. Thus, these levers *complement* daily incident-level decision optimization (which specific open incidents should a worker address that day), determining the feasibility of a specific daily decision (for example, the agency cannot easily inspect more incidents in a neighborhood in a day than their worker capacity and spatial distribution allows).

Furthermore, as we show, these levers alongside information on incident arrivals induce **Service Level Agreements** (SLA): promises by the government that incidents of type k will typically be addressed within z_k days – in our model, optimizing budgets and prioritization policies are equivalent to directly optimizing SLAs. Service Level Agreements have the following desirable properties: (a) They are commonly used to characterize and communicate system performance in cloud computing [Patel et al., 2009], various web services [Jin et al., 2002], and in city government in particular. For example, New York City has published SLAs for responses to service requests by residents;[1] e.g., the *Department of Parks and Recreation (NYC DPR)* will respond to a report of *Illegal Tree Damage* within *8 days*. In other words, they are *transparent* and externally auditable. (b) if met, they can translate to equity and efficiency desiderata; for example, more urgent types of incidents should have shorter response timelines, and overall (importance-weighted) delays in each neighborhood should not be disparate.

However, allocating relative budgets and designing SLAs is challenging: their (joint) feasibility depends on the available budget and incident arrival rates, which may change over time. Communication with NYC DPR indicates that current SLAs, though in theory promised to the public, are too inaccurate to meet or guide operations. As an illustration, Figure 1 shows publicly listed SLAs for two types of incidents in New York City, alongside how quickly these incidents were responded to in two Boroughs.[2] (Note: in this paper, we consider the allocation of *inspections* in response to requests, and so interchangeably use "response" and "inspect.")

We tackle this design question: (a) formulate and analyze a stylized queuing model, which induces a tractable optimization problem to determine budgets and incident type prioritization, giving

[1] https://data.cityofnewyork.us/City-Government/311-Service-Level-Agreements/cs9t-e3x8

[2] In NYC, the five main sub-city administrative units are called Boroughs. Agency sub-units for each Borough operate with some autonomy, and budgets are often divided into Borough-specific budgets. NYC DPR is actively planning to further centralize operations, a policy we analyze here.

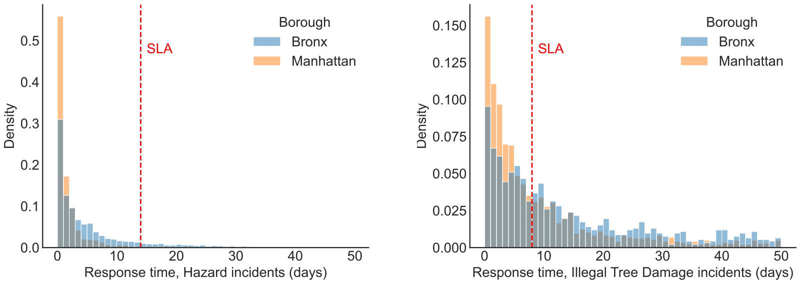


Fig. 1. The Department of Parks and Recreation of NYC responds to service requests on “Hazard" and “Illegal Tree Damage" events related to street trees, among others. We find that even conditional on the same category of incidents, the distribution of the response time (in this case defined by the time from first service request to the completion of an inspection) varies by Borough.[3] Moreover, substantial amounts of service requests are not responded to within the publicly available SLAs, with some others not inspected at all. The empirical distribution of the response times also does not correspond the priority defined in the SLAs: though hazard incidents have a looser SLA compared to illegal tree damage incidents (14 days versus 8 days), they are generally inspected sooner, reflecting their higher average risk rating.

insights on the price of equity (and efficiency) in such allocation settings; (b) extend the model to incorporate real-world complexities, which can be optimized to produce implementable policies, using a simulation-optimization framework; (c) empirically optimize policies using actual incident and inspection data from the New York City Department of Parks and Recreation.

A theoretical model to analyze SLAs. We model SLA design as an optimization problem under the framework of a queuing network, with decision levers: allocating inspection budgets (inspectors) to different neighborhoods and scheduling priorities for different types of incidents. We transform the problem to a convex formulation and simultaneously solve for an optimal budget allocation plan and inspection scheduling prioritization policy, to find the best, feasible SLA under some objective function. We show that our model is general enough to allow a large class of objective functions, encoding both equity and efficiency. Our model can be generalized to accommodate other administrative policies, such as when the city maintains a centralized inspection team. We theoretically study this model under a specific class of objectives, where the efficiency loss corresponds to risk-weighted SLAs across the city, and the equity loss measures the worst efficiency loss incurred by any neighborhood. Conceptually, we find that the “price of equity” in this setting – the efficiency loss from implementing the most equitable policy – is small, especially in realistic settings when risk distributions are similar across Boroughs, even as incident *numbers* may not be.

Simulation optimization framework to designing SLAs. The stylized model – though tractable for qualitative insights and a parametric starting point – excludes several real-life components (including non-Poisson incident arrival rates that are higher than the budget allows addressing) that prevent its use to optimize an actual city agency’s budget. We thus develop a simulation-optimization framework to optimize decisions in practice: the simulation inputs historical incident arrivals and worker capacities over time, simulates daily decisions according to a given policy (including per-area budget fractions and incident priorities), and calculates efficiency and equity metrics. Then, we can optimize over a class of policies in an outer loop, such as through a Bayesian

[3] These response times are calculated based on public data on inspections from 2015 to present, available at https://data.cityofnewyork.us/Environment/Forestry-Inspections/4pt5-3vv4

optimization framework. Optimal policies can be validated out-of-sample, such as by simulating their performance using data from a future time. The framework is thus flexible to incorporate high-fidelity simulations given a policy.

Empirical characterization of SLAs under different objectives. We empirically apply our simulation-optimization framework to design service level agreements for the Department of Parks and Recreation in New York City, for responses to service requests made by residents. We find: (a) optimal Borough (sub-divisions of NYC) budget allocations differ from status quo allocations, which are both highly inefficient and inequitable; (b) motivated by an ongoing debate for whether the agency should *centralize* response operations, we find that doing so would only provide a modest benefit over optimal Borough-specific budget allocations; (c) Optimal policies calculated using 2019 data are highly effective (outperforming actual decisions) in future years.

We further find that the *empirical* price of equity is indeed small: inefficient policies are also inequitable, and vice-versa. More precisely, adding explicit equity terms in the objective has a small impact, compared with designing better (even purely for efficiency) budget allocations or centralizing operations. We explain this finding as follows: inequity often implies inefficiency – substantial delays in one neighborhood also affect the overall city welfare, and so the space of Pareto optimal (in terms of efficiency vs equity) allocations is small. This finding suggests a “win-win” when compared to the status quo: we can improve both city-level outcomes and equity.

The rest of this paper is organized as follows. In the remainder of this section, we introduce related work. We introduce our model and theoretical results in Section 2, our simulation optimization framework in Section 3, and our empirical case study in Section 4.

1.1 Related work

Efficiency and equity in government operations. Our application – responses to service requests made by residents – relates to previous works on “co-production” [Brabham, 2015, Yuan, 2019] systems such as the 311 system in NYC, where city residents report incidents, triggering the need for responses. Recently, the upstream reporting behavior of city residents has been well-studied, and various aspects of inefficiency and inequitable usage have been identified [Agostini et al., 2024, Hacker et al., 2022, Kontokosta and Hong, 2021, Liu et al., 2023].

More closely related are the works that look into the cities’ responses to various reports, through the lens of efficiency and equity in public resource allocation. Previous theoretical works on this topic outline tradeoffs among equity criteria and between efficiency and equity [Freeman et al., 2020, Mashiat et al., 2022], in areas such as allocation of healthcare resources [Mhasawade et al., 2021]. Empirical works on this topic mainly focus on agency decision-making, and ask the question, “are status-quo response decisions equitable and efficient?". Most notably, Singh et al. [2022] study food inspection operations in Chicago to identify violations of fairness criteria due to idiosyncratic behavior of inspectors, suggesting algorithmic remedies; Laufer et al. [2022] study the response to forestry service requests in NYC under capacity constraints, and identify inequity in the decisions to inspect among neighborhoods, which is further associated with socio-economic factors; Rahmattalabi et al. [2022] and Jo et al. [2023] consider resource allocation under fairness objectives in homeless services, pointing out incompatibility between fairness objectives and developing a fair matching algorithm.

In contrast, considering time-sensitive incidents, our work mainly asks the question “how do we *make* the responses more equitable and efficient by adjusting the agency’s resource allocation policy?” Our work can incorporate insights from the above work: for example, it is possible to incorporate the incident type and location-specific reporting delay estimates by Liu et al. [2023] into our framework – if the city's objective is to equalize *occurrence to resolution end-to-end* response times, as opposed to *report to resolution* response times.

Design and planning of queuing systems in operations. Our problem broadly falls into the "capacity and flow assignment" category under the four types of optimization problems tied with the design and planning of queuing systems outlined by Kleinrock [1975]. Other works that fall into this category include a large literature on dynamic capacity allocation (e.g., [Andradóttir et al., 2003]), and in particular, designing hospital operations (e.g., [Bekker and de Bruin, 2010, Cochran and Roche, 2009, Green, 2006]). More closely aligned with our objectives are the works of Nowak et al. [2004] and Remesh Babu and Samuel [2019]; both consider service level agreement aware dynamic capacity allocation in the network service context. However, our work differs from these settings in that the capacity allocation decisions cannot be frequently adjusted in our setting – once the city determines on a set of budgets for each Borough, it might take years to revise such decisions, both due to administrative capacity and the challenge of moving workers to different offices.

Our model builds off of the work of Liu et al. [2001], who consider maximizing the profit from operations under SLA constraints; they consider the levers of *prioritizing* and *routing* different jobs to different servers, under the Generalized Processor Share [Parekh and Gallager, 1993] method. Our work introduces a lever of capacity allocation to this setting, but does not consider routing decisions (in context, it may be both administratively and logistically hard for a worker in one Borough to respond to incidents in another); these changes, besides being motivated by our application domain questions, render the optimization problem more tractable, allowing us to analytically characterize the price of equity under such a model. We further embed this model within an empirical simulation-optimization framework, finding approximately optimal policies in practice.

2 A STYLIZED MODEL TO DESIGN OPTIMAL SERVICE LEVEL AGREEMENTS

Our model has four aspects under a policy maker's control: how individual workers are allocated to incidents, how many workers there are in each neighborhood (Borough), the SLAs promised for each category (potentially in neighborhood-dependent manner), and an objective function formalizing their efficiency and equity goals. We further assume that the policy maker knows (can historically measure) characteristics of incidents, such as their arrival rates and average "riskiness."

These aspects are related as follows. Together, the worker allocation policy, budgets, and incident arrival rates determine the distribution of response times for each incident, and thus the Service Level Agreements. We assume that the policy maker's objective is a function just of the SLAs and the system constants (such as incident arrival rates and risk distributions) – together, these aspects can encompass standard metrics such average and tail response times, and their spatial distribution.

Our goal is two-fold: First, for a given policy objective, formulate and solve a corresponding (tractable) optimization problem to find an optimal SLA (and hence a worker allocation and budget policy). Second, characterize how solutions change as the objective function changes, and thus any potential trade-off between efficiency and equity. We note that our theoretical model is purposely stylized, to enable tractable insights. In practice, policy design must incorporate elements omitted in the theoretical analysis; we consider these components in Section 3 within a simulation optimization framework.

2.1 A queuing model for the inspection scheduling process

We start with a queuing-based model for how incidents are addressed: incidents occur over time, and join a corresponding queue. Workers service each queue according to an allocation policy over time, inducing response time distributions for each type of incident.

The queuing model. The inspection problem is modeled by a queuing network with multi-class single-server queues. There are two levers of policy: the *city* allocates worker budgets to Boroughs, and each Borough manages the allocation of workers to individual incidents.

The queuing network within each Borough operates as follows. We have a set of incident categories $k \in \mathcal{S}$ (e.g., Hazards versus less urgent incidents) for which we wish to define SLAs. Each category arrives according to a Poisson process with rate λ_k into their own queues, where each of these Poisson processes are mutually independent. Inspecting each incident takes up a random amount of time, distributed according to an Exponential random variable with unit mean regardless of their category.[4]

The city has a budget of C, in terms of the total capacity of the servers that can be allocated. The city first decides the capacity C_b that it wishes to allocate to each Borough $b \in \mathcal{B}$, where $\sum_{b \in \mathcal{B}} C_b \leq C$. Each Borough b maintains its own server with capacity C_b, serving the queues for incidents that occur in that Borough.

We assume that within each Borough, the server is managed with the Generalized Processor Share (GPS) scheme [Parekh and Gallager, 1993]. Under the GPS scheme, each SLA category is assigned a weight $\phi_{k,b}$ in each Borough b, such that $\sum_{k \in \mathcal{S}} \phi_{k,b} = 1$. At any given time t, on a server with capacity C_b, the (potentially fractional)[5] capacity devoted to category $k \in \mathcal{S}$ is $C_b \phi_{k,b} / \sum_{k' \in \mathcal{S} \cap \mathcal{K}(t)} \phi_{k',b}$, where $\mathcal{K}(t)$ denotes the set categories for which there is at least one pending incident (there is a "backlog" of incidents).[6] Within each queue, we assume a first-come-first-serve discipline of service. At a high level, $\phi_{k,b}$ determines the relative priority of different incident categories within a Borough – for example, a *Hazard* category may be higher risk on average, and so prioritized on average.

Why limit to this class of policies? (a) GPS policies are flexible and robust: on a server with capacity C_b, the minimum capacity devoted to any SLA category k would be $C_b \phi_{k,b} / \sum_{k' \in \mathcal{S}} \phi_{k',b} = C_b \phi_{k,b}$ while there are incidents in this category in backlog, so the backlog of other SLA categories will not affect our response to category k; on the other hand, the maximum possible capacity devoted to any SLA categories is the entire capacity C_b, whenever they are the only SLA category in backlog. (b) The policy reflects what is practiced by city agencies: a certain portion of workers would respond to certain categories of incidents, with adjustments from day to day based on backlog. (c) As we will show next, GPS policies naturally lead to well-defined notions of service level agreements.

Under this model, the decision variables for the city are to set the Borough level budgets C_b; for each Borough, it is to set their own incident allocation weights $\phi_{k,b}$.

Response times and SLAs. Reflecting practice, we consider SLAs in the following form:

"In Borough b, fraction 1-$\alpha_{k,b}$ of category k incidents are responded to within time $z_{k,b}$,"

where $\alpha_{k,b} \in (0, 1)$. It is thus important to quantify the tail distributions of the response time to each SLA categories. Mathematically, let $T_{k,b}$ be the generic random variable for the response time of category $k \in \mathcal{S}$ incidents, an SLA of the above form corresponds to

$$\mathbb{P}[T_{k,b} \geq z_{k,b}] \leq \alpha_{k,b}. \tag{1}$$

Assuming that these incidents are being processed on a server with capacity C_b, and the GPS weight of this category satisfies $C_b \phi_{k,b} > \lambda_{k,b}$ (i.e. there is guaranteed to be enough capacity to inspect all incidents of this category), following our above assumptions and classical results in

[4] Our theoretical framework can be easily extended to different processing times by category.

[5] Fractional capacity can be interpreted as randomized allocation.

[6] To finish formally describing the system, as standard in such queuing systems, we assume that there is an additional category of work that is always backlogged for which the city provides no guarantees – the worker services this category when there is no backlog in categories $k \in \mathcal{S}$.

queuing theory (e.g., see [Kleinrock, 1975] and [Liu et al., 2001]), the tail probability of the response time distribution is bounded by

$$\mathbb{P}[T_{k,b} \geq z_{k,b}] \leq \exp\left(-(\phi_{k,b}C_b - \lambda_{k,b})z_{k,b}\right). \tag{2}$$

Combining Equation (1) and Equation (2), to satisfy an SLA, a sufficient condition is to ensure

$$-(\phi_{k,b}C_b - \lambda_{k,b})z_{k,b} \leq \log(\alpha_{k,b}), \tag{3}$$

which we will henceforth refer to as the “SLA constraint". Intuitively, this constraint states that: when C_b budget is allocated to Borough b, and category k is assigned a GPS parameter $\phi_{k,b}$, then at least $1 - \alpha_{k,b}$ fraction of category k incidents in Borough b should be inspected within $z_{k,b}$ days, i.e., the SLA would be $z_{k,b}$ days. To simplify notation, we will consider $\log(\alpha_{k,b}) = \alpha$ for all SLA categories k. Note that, since $\alpha_{k,b} \in (0,1)$, $\alpha < 0$.

Policy maker objective. From the perspective of the city, the objective broadly contains two parts. First, SLAs across the whole city should reflect the principle of efficiency, in that more urgent incidents should be addressed sooner: if a more hazardous incident is left unattended for one day, the amount of risk it poses to surrounding residents is larger than a less hazardous one. However, since responding to service requests is a *public* service, the city must treat different people equitably: residents from different areas should not receive dramatically different levels of service.

We define two functions to capture these notions. Denote $g(\mathbf{z})$ and $f(\mathbf{z})$ as the efficiency loss function and the equity loss function, respectively, meaning a higher value of these functions represents less efficient and less equitable SLAs. We assume that both these functions are non-decreasing with respect to each element of $\mathbf{z}$, which are the SLAs assigned, and are convex in $\mathbf{z}$. Monotonicity is natural: increasing the SLA for one category of incidents while others remain the same represents an absolute deterioration in the level of service, and should not lead to Pareto improvement in either efficiency or equity. The intuition behind convexity is that the marginal cost of worsening SLAs is increasing. Within our stylized model, these assumptions lead to tractable optimization and, as we show, encompasses a large class of metrics.

2.2 An optimization problem for optimal SLAs

Putting this together, our optimization task is as follows.

$$\begin{aligned}
\min_{\mathbf{z},\phi,C_b} \quad & L(\mathbf{z}) = g(\mathbf{z}) + f(\mathbf{z}) && && \text{(4a)}\\
\text{s.t.} \quad & -(C_b\phi_{k,b} - \lambda_{k,b})z_{k,b} - \alpha \leq 0, && \forall k \in \mathcal{S}, b \in \mathcal{B}, && \text{(4b)}\\
& \sum_{k\in\mathcal{S}} \phi_{k,b} \leq 1, && \forall b \in \mathcal{B} && \text{(4c)}\\
& \sum_{b\in\mathcal{B}} C_b \leq C, && && \text{(4d)}\\
& \phi_{k,b} \geq 0, z_{k,b} \geq 0, C_b \geq 0. && && \text{(4e)}
\end{aligned}$$

The objective (4a) reflects both efficiency and equity losses, as functions of the SLAs $\mathbf{z}$; constraint (4b) ensures that the solution meets the set of SLAs encoded in $\mathbf{z}$, where the α comes from our definition of SLAs in equation (3); constraint (4c) comes from the GPS scheduling scheme; constraint (4d) enforces the overall budget constraint across Boroughs. As defined, problem (4) is non-convex as the Hessian of (4b) is not positive semi-definite. However, this problem can be reformulated to a convex program that has equivalent optimal solutions.

Proposition 2.1. *Let $\boldsymbol{x}^{-1}$ be the element-wise reciprocal of $\boldsymbol{x}$. Consider the reformulated problem.*

$$\begin{aligned}
\min_{\boldsymbol{x}, C_b} \quad & \tilde{L}(\boldsymbol{x}) = g(-\alpha \boldsymbol{x}^{-1}) + f(-\alpha \boldsymbol{x}^{-1}) && && \text{(5a)}\\
s.t. \quad & \sum_{k \in \mathcal{S}} x_{k,b} \leq C_b - \sum_{k \in \mathcal{S}} \lambda_{k,b}, && \forall b \in \mathcal{B} && \text{(5b)}\\
& \sum_b C_b \leq C, && && \text{(5c)}\\
& x_{k,b} > 0, C_b \geq 0. && && \text{(5d)}
\end{aligned}$$

The reformulated Problem (5) *is convex. Let $\{\boldsymbol{z}^*, \phi^*, C_b^*\}$ and $\{\boldsymbol{x}^*, C_b^*\}$ be the set of optimal solutions to Problems 4 and 5, respectively. Then we have*

$$z^*_{k,b} = -\frac{\alpha}{x^*_{k,b}}, \text{ and } L(\boldsymbol{z}^*) = \tilde{L}(\boldsymbol{x}^*).$$

The proof relies on the convexity and monotonicity of the objective function and is deferred to the Appendix. Given these properties, we will rely on solving Problem (5) in our subsequent analysis, and backtrack to find the optimal solutions to Problem (4). Note that the α term in the objective comes from our definition of the SLAs, and for general g and f cannot be omitted.[7]

Model discussion. What does this model capture? Namely, the four aspects under a policymaker's control. It captures how individual workers are assigned to each incident, through the GPS scheme and the decision variables ϕ; it captures how many workers are allocated to each Borough, through the capacity of the Borough servers using decision variables C_b; it captures the SLAs promised for each category and Borough through decision variables $z_{b,k}$; and finally, the objective function is open to configuration for policymakers to reflect their efficiency and equity goals.

We note that this model is also stylized in several crucial respects, that render it inappropriate to use to design *actual* government policies; we discuss some of these aspects in detail in Section 3.1, when detailing our simulation-optimization framework. In this section, we use the stylized model to draw insights on the structure of optimal policies. Furthermore, here we consider the administrative policy that the city first allocates budgets to Boroughs, and then Boroughs manage their responses. Crucial for the empirical analysis, our model can be generalized to other administrative policies, such as when the city manages a centralized server, or when budgets cannot be allocated and must stick to status quo levels. We will introduce these administrative policies in Section 3.2.

2.3 Analysis of the optimization model under specific objectives

Proposition 2.1 satisfies our first goal of formulating a tractable optimization problem to determine SLAs, budgets, and allocation policies. Here, we characterize the solutions of this optimization problem, for specific instantiations of the efficiency and equity objectives and relative weightings, to understand efficiency and equity tradeoffs in our allocation problem over time.

Risk-rating-based objective functions. An important aspect of an incident is the *risk* it poses if unaddressed (for example, the danger posed by a tree falling on a person or power line).[8] We assume that we can measure the average risk rating of $r_{k,b}$ for incidents of category k in Borough b.

[7] When a function $l(\cdot)$ is not homogeneous, the ordering may not be preserved when each argument is multiplied by a scaler: $l(\mathbf{x}) > l(\mathbf{y}) \not\equiv l(-\alpha\mathbf{x}) > l(-\alpha\mathbf{y})$, for some $\alpha < 0$.

[8] For example, in our empirical application motivation, we consider allocation decisions for inspections and work orders for incidents by the NYC DPR; a primary outcome for such inspections is a risk assessment, leading to risk ratings. Such risk ratings are then indeed used to determine work order scheduling priorities.

Motivated by the importance of heterogeneity in risk ratings, we define the following risk-based cost function for each Borough:

$$\mathrm{Cost}_b(\mathbf{z}) = \sum_{k\in\mathcal{S}} \lambda_{k,b} r_{k,b} z_{k,b}, \forall b \in \mathcal{B}, \tag{6}$$

which represents the sum of risk-rating-weighted SLAs in one Borough, and subsequently define these efficiency and equity loss functions:

$$g(\mathbf{z}) = \sum_{b\in\mathcal{B}} \mathrm{Cost}_b(\mathbf{z}), \tag{7}$$

$$f(\mathbf{z}) = \max_{b\in\mathcal{B}} \mathrm{Cost}_b(\mathbf{z}), \tag{8}$$

$$L_\gamma(\mathbf{z}) = \gamma g(\mathbf{z}) + (1-\gamma) f(\mathbf{z}), \tag{9}$$

where we assume risk rating r's are given as data, and $\gamma \in [0, 1]$ is a hyperparameter for the relative importance of the objectives: the larger γ is, the more weight is put on efficiency. In words, $g(\mathbf{z})$ represents the sum of costs across all Boroughs as the loss of efficiency, and $f(\mathbf{z})$ measures the largest cost of any Borough as the loss of equity. These formalizations also induce interpretable characterizations of the optimal solutions, corresponding to common notions of efficiency and equity. Given these functions, we now analyze the *efficiency-equity* tradeoff in allocation.

Proposition 2.2. *[Extreme efficiency prioritization] When $\gamma = 1$, the optimal solution to Problem* (5) *is such that $x_{k,b} \propto \sqrt{\lambda_{k,b} r_{k,b}}$. Consequently, the optimal solution to Problem* (4) *is such that*

$$z_{k,b} \propto \frac{1}{\sqrt{\lambda_{k,b} r_{k,b}}}, \ \forall k, b.$$

In other words, the optimal solution when we only care about efficiency is such that each type of incident will be assigned an SLA that is inversely proportional to the square root of its risk level: more urgent incidents should be inspected sooner. Interestingly, note that *budgets* may differ substantially across Boroughs, but per-incident service guarantees remain a function of their risk.

Proposition 2.3. *[Extreme equity prioritization] When $\gamma = 0$, the optimal solution to Problem* (5) *is such that $\sum_{k\in\mathcal{S}} r_{k,b}/(\lambda_{k,b} x_{k,b}) = M$ for some M, for all Boroughs $b \in \mathcal{B}$. Consequently, the optimal solution to Problem* (4) *is such that*

$$Cost_b(z) = \sum_{k\in\mathcal{S}} \lambda_{k,b} r_{k,b} z_{k,b} = M, \ \text{for some } M, \forall b \in \mathcal{B}.$$

In other words, the optimal solution when we only care about equity is such that all Boroughs would experience the same cost. Note that, within a Borough, more risky categories would still have shorter SLAs in this solution; in fact, they could have even *more* short SLAs than in the extreme equity case: for example, when there is only one category per Borough, $z_{k,b} \propto \frac{1}{\lambda_{k,b} r_{k,b}}$.

What are the practical implications of these two extremes? In the extreme efficiency case, all SLA categories are assigned SLAs that are only related to their risk level: the higher the risk rating, the sooner the incidents are promised to be inspected, without any consideration for where these incidents may be. However, in practice, each Borough has different geographical characteristics and thus different kinds of potential incidents – i.e., average risk ratings and incident arrival rates differ by area. Suppose all incidents in one Borough are uniformly riskier than those in another Borough; then, only optimizing for efficiency may result in a large divide in the level of service that the two Boroughs receive: one Borough will receive worse (risk-weighted) service. An equity term balances the differences: in the extreme equity case, there is a strict parity among the costs in all Boroughs. We formalize this discussion next, characterizing the "efficiency cost" of pursuing equity.

When is the efficiency and equity trade-off substantial? We consider the tradeoff between these objectives. Akin to the algorithmic fairness literature, we define *price of equity* as the difference in efficiency loss between most equitable and the the most efficient solutions.

Consider a simple case, where we are only concerned with the SLA of a single category of incidents ($|\mathcal{S}| = 1$, and hence we omit the subscript indicating category) in three Boroughs ($\mathcal{B} = \{1, 2, 3\}$). Then following Proposition 2.2 and Proposition 2.3, denoting the solution under extreme equity and efficiency prioritization as $\mathbf{z}^{eq}$ and $\mathbf{z}^{ef}$, respectively, we arrive at the following result:

Proposition 2.4. *[Price of equity] The* ***price of equity*** *can be measured by:*

$$g(\boldsymbol{z}^{eq}) - g(\boldsymbol{z}^{ef}) = -\alpha \frac{1}{C - \lambda_1 - \lambda_2} \left[\sqrt{\lambda_1 r_1} - \sqrt{\lambda_2 r_2}\right]^2 \geq 0.$$

In words, the *extra cost* of efficiency incurred by pursuing the most equitable solution compared with the most efficient solution is always non-negative and depends on two factors: (1) the relative difference between severity (measured in both risk and quantity) of incidents in the Boroughs $\lambda_1 r_1$ and $\lambda_2 r_2$, and (2) the amount of extra budget slack the agency has $C - \lambda_1 - \lambda_2$ (total server budget C minus the arrival rates).

When is the trade-off negligible? (1) When $\lambda_1 r_1 = \lambda_2 r_2$, i.e., both Boroughs face similarly severe incidents, we see that in this case, the price of equity $g(\mathbf{z}^{eq}) - g(\mathbf{z}^{ef}) = 0$. Otherwise, note that the Borough with the riskier incidents faces a *higher* cost at the most efficient solution. The intuition behind this is that, when the two Boroughs have the same risk profile, they are essentially indistinguishable by simply observing the risk of incidents that arrive. Thus, the problem becomes one of budget allocation between two homogenous groups. Similar to results from other works in the algorithmic fairness literature (e.g., [Cohen et al., 2022]), there is no trade-off between equity and efficiency – these two objectives are perfectly aligned. (2) When $C - \lambda_1 - \lambda_2$ is large, i.e., the city has enough budget to easily address all incidents quickly, and so that there is little tradeoff.

In contrast, when is the trade-off substantial? We see that two conditions are necessary: (1) when the two Boroughs face sufficiently different situations (measured by $\left|\sqrt{\lambda_1 r_1} - \sqrt{\lambda_2 r_2}\right|$), and (2) when the excess budget $C - \lambda_1 - \lambda_2$ is sufficiently small. This is also analogous to the equity-efficiency trade-off in other algorithmic decision-making setting. For example, in college admissions (e.g., [Garg et al., 2020]), if the admission capacity of a school is larger than the applicant pool, then admitting everyone would be both efficient and equitable; it is only when the capacity is smaller than the applicant pool and different groups of applicants are disparate when we observe a significant equity-efficiency trade-off. In our setting, when the excess budget $C - \lambda_1 - \lambda_2$ is large, no matter which budget allocation we choose, incidents in both Boroughs can be inspected within a reasonable time (as can be observed from the formula for z^{eq} and z^{ef}); it is only when the excess budget is small and incidents in the two Boroughs have disparate risks would we expect to see a large trade-off when prioritizing one Borough over another.

Similarly, we can also derive the following result, on the price of efficiency: the equity loss associated with adopting the most efficient solution.

Proposition 2.5. *[Price of efficiency] Without loss of generality, assume $r_1 \geq r_2$. The* ***price of efficiency*** *can be measured by*

$$f(\boldsymbol{z}^{ef}) - f(\boldsymbol{z}^{eq}) = -\alpha \frac{1}{C - \lambda_1 - \lambda_2} \sqrt{\lambda_2 r_2} \left(\sqrt{\lambda_1 r_1} - \sqrt{\lambda_2 r_2}\right) \geq 0.$$

As we'll see in the remainder of this paper, these insights extend to less stylized settings of incident arrival rates and agency policies.

3 SIMULATION OPTIMIZATION FRAMEWORK

Above, we formulate a theoretically tractable policy optimization problem, whose solutions yield insights regarding the tradeoffs between efficiency and equity in our temporal allocation setting. However, the *optimal policies* are not generally *deployable* – such theoretical tractability requires omitting modeling components that are important in practice.

Rather, we advocate a simulation-optimization approach for such policy optimization: (a) specify input data (here: incident arrivals and agency city-wide capacities), either calibrated to historical data or using it directly; (b) determine a class of policies over which one will optimize (e.g., GPS policies with Borough budgets, incident prioritization, etc); (c) simulate a given policy over time and evaluate its performance; (d) optimize over policies ; (e) Finally, evaluate the chosen "optimal policy" out-of-sample (such as using input data from another historical time period) to understand their robustness and to approximate future performance.

This procedure is of course more computationally expensive than solving the convex optimization Problem (5), which takes less than a second to solve (as opposed to the 2.5 days of computation required in our empirical application). However, policies are not updated often (e.g., agencies often plan budgets on a yearly cadence), and so the cost is not prohibitive. Such an approach can further trade off optimization tractability with modeling fidelity, by adjusting the complexity of the simulator and class of policies.

In this section, we provide methodological details for this procedure as applied to optimize NYC agency service allocation policies. The next section provides application data details and results.

3.1 Why a simulation optimization framework?

Our theoretical optimization framework makes at least two simplifications that prevent its use in practice in our application: (a) Historically, not all incidents are inspected: the number of service requests received far exceeds the capacity of the city agency (in our application, historical data suggests only around 60% of all requests are responded to); (b) as seen in Figure 4, incidents do not arrive according to a homogeneous Poisson arrival rate: there are daily variations not necessarily explainable by Poisson variation, seasonal effects, and emergency storm periods. As we explain next, these components prohibit tractable (convex) policy optimization.

Over-capacity queues and dropping incidents. Not all incidents, even in daily operating periods (outside of emergency periods), are inspected by the agency, let alone fixed through a work order. As expected, the fraction of incidents dropped correlates with incident importance (most Hazard reports are inspected, but few Root/Sewer/Sidewalk reports are); however, incident drops can also be an undesirable implication of historical policies: e.g., if a Borough is relatively understaffed, then it may be forced to drop more incidents. Thus, we need to make the fraction of incidents inspected a variable, either directly optimized as part of the policy or as depending on the capacities and prioritization. However, it turns out that doing so while maintaining theoretical tractability is challenging. Let $p_{k,b}$ for each category and Borough pair denote the fraction of service requests inspected.

First, suppose we fix constants $p_{k,b}$ and the total budget C as given and estimated from the historical data, i.e., substitute $\lambda_{k,b}$ in Problem (4) with $p_{k,b}\lambda_{k,b}$. Then, we would find that $\sum_{b\in\mathcal{B}}\sum_{k\in\mathcal{S}} p_{k,b}\lambda_{k,b} = C$, meaning that the total budget is the same as the total arrival rate of incidents. In this case, the only feasible solution would be that for all (k, b) we have $C_b\phi_{k,b} - p_{k,b}\lambda_{k,b} = 0$ (i.e., the arrival rate and the service rate of the queue are the same), and thus the SLAs $z_{k,b}$ could not be well defined.

Second, instead suppose $p_{k,b}$ is a decision variable, and we add a term h in the objective to penalize dropping incidents. The new problem, though still having a simple form, becomes either intractable or trivial, depending on the objective function:

$$\begin{aligned}
\min_{\mathbf{z},\phi,C_b,\mathbf{p}} \quad & L(\mathbf{z},\mathbf{p}) = g(\mathbf{z}) + f(\mathbf{z}) + h(\mathbf{p}) && && \text{(10a)}\\
\text{s.t.} \quad & -(C_b\phi_{k,b} - p_{k,b}\lambda_{k,b})z_{k,b} - \alpha \le 0, && \forall k \in \mathcal{S}, b \in \mathcal{B}, && \text{(10b)}\\
& \sum_{k\in\mathcal{S}} \phi_{k,b} \le 1, && \forall b \in \mathcal{B} && \text{(10c)}\\
& \sum_{b\in\mathcal{B}} C_b \le C, && && \text{(10d)}\\
& \phi_{k,b} \ge 0, z_{k,b} \ge 0, C_b \ge 0, p_{k,b} \in [0,1]. && && \text{(10e)}
\end{aligned}$$

If the objective function $h(\mathbf{p}) \equiv 0$ (i.e., deciding not to inspect incidents does not incur penalties), then trivially we should set $p_{k,b} = 0, \forall k, b$, so that all the SLAs are 0 days – not inspecting anything would result in the shortest response times of the things that are inspected. However, this would result in the solution being irrelevant in practice. On the other hand, if the objective function $h(\mathbf{p})$ is some non-trivial function of $\mathbf{p}$, then the non-convex nature of Constraint (10b) renders the problem intractable, even for simple forms of $h(\mathbf{p})$.

Arrival rates and service times. The theoretical analysis further depended crucially on incidents arriving according to a homogeneous Poisson process, and service times (how long it takes a worker to address an incident) being Exponentially distributed; these assumptions make it simple to derive the tail probabilities for how long it takes for an incident to be serviced, which are part of the SLA. While such assumptions are common in the queuing theory literature for tractability, they are not suboptimal for our setting for several reasons.

First, incident arrival rates do not necessarily follow homogeneous Poisson processes, displaying spatial and temporal (intra-week and seasonal) correlation, both during emergency periods and in routine operations. Second, the time it takes to address an incident is not necessarily Exponentially distributed: for example, in our *inspection* allocation application, an inspector travels to an incident and conducts an inspection, which may take an approximately constant amount of time; some incidents may require followup inspections at a later date, which would increase the total time it takes to address that incident but does not preoccupy the inspector's time in the meanwhile.

Suppose one wanted to incorporate these elements into Problem (4), using alternate functional forms of arrival and service times, for example incorporating spatial and temporal correlation of arrivals. Theoretically, this would require replacing constraint (4b) with another constraint that better reflects the mapping from budgets and prioritizations to the tail inspection probabilities that are induced; closed-form solutions may be intractable, depending on the functional forms chosen. Empirically, one would need to *calibrate* the functions using real-world data, and high-fidelity calibration may be challenging even if the theoretical challenge could be overcome.

In light of these obstacles, we leave theoretical advances to overcome them for future work and employ a simulation-based approach to finding near-optimal inspection policies in practice, while retaining the core of our principled approach – a parametric GPS scheme.

3.2 Empirical simulation-optimization framework

We now describe our simulation-optimization framework, informed by the above discussion. The agency inspects incidents every day, over a 10-year simulated period. As overviewed above, we need to specify the input data, policy class, and evaluation.

3.2.1 Input data: historical inspection arrivals and agency capacity. We directly use historical incident arrivals (time stamps, locations, and category) over a full year (in our application, 2019), with incidents arriving each day according to their true arrivals. As inspection budgets may vary throughout the year and week (e.g., fewer inspections on weekends), for each day we input the number of incidents from that year (in our application, 2019) the agency inspected throughout the entire city that day. (Of course, the policy will determine how this overall daily budget is distributed across Boroughs and incident categories). We repeat the year of data 3 times in each simulation, to further minimize boundary effects. These choices enable simulation realism without needing to calibrate functions for incident arrivals and overall capacity.

3.2.2 Policy class. We optimize over two classes of policies: Borough budget GPS policies as in our theoretical model, and one with city-wide budgets that allow more flexible cross-Borough allocation in response to real-time queue lengths.

Borough budget GPS policies. Each policy consists of two levers, identical to our GPS policy scheme described above: a set of budget allocations $\{C_b\}$ to each Borough that describes the *fraction* of the daily capacity available to that Borough (without loss of generality, we standardize them so that $\sum_b C_b = 1$); and two sets of parameters $\{\phi_{k,b}\}$ and $\{p_{k,b}\}$, which indicates how each Borough should manage their inspections under a GPS scheduling scheme, and what fraction of each type of incidents the agency intend to physically inspect.

City budget policies. In addition to evaluating Borough budget policies (where the city agency has two levers: budget allocation to Borough subdivisions and managing queues within each Borough), we consider another class of policies that assumes the city agency only manages *one centralized server*, i.e., it can flexibly allocate its workers across Boroughs in reaction to current queues, without needing to worry about e.g., travel time. We refer to this type of policy as **city budget** policies. City budget policies allow for more flexibility compared to each Borough managing its own server, as when one Borough is experiencing a surge of inspection requests, the capacity that would have been dedicated to other Boroughs can be concentrated on inspecting these. This added flexibility may induce higher administrative and logistic expenses: inspectors are usually stationed within each Borough, and there is a cost to transition them between Boroughs. Under this class of policies, the only lever is how the inspections should be managed, which is specified by two sets of parameters $\{\phi_{k,b}\}$ and $\{p_{k,b}\}$.

We note that these policies are not exhaustive or are likely to describe exactly how the agency behaves. However, it accurately reflects the agency data and our conversations with agency practitioners. Higher fidelity policies and simulators can be designed (e.g., those that incorporate city budgets but with travel costs for inspectors, spatial grouping of inspected incidents, or more deterministic prioritization of the highest priority incidents), at the cost of an increase in parameters.

3.2.3 Simulation given a policy and input data. We now describe our simulation. For precision, we detail the Borough budget GPS policy simulation; we detail the simulation process and input data for the city-wide budget policies in Appendix A.2 (in summary, the historical data used as input is the same, and we only make minor changes to the input policy parameters and the way counterfactual inspections are made using these policy parameters). We make use of two hyperparameters, review period length D and first-come-first-serve (FCFS) violation ρ to calibrate the simulator to historical inspections. For a period indexed by $t \in [T] := \{1, 2, \ldots, T\}$, our simulation process takes in historical arrivals of inspection requests $\mathcal{N}_{k,b}^{t}$, $t \in [T], \forall k, b$, and historical city-wide inspections performed $I^t, t \in [T]$ on these incidents on each day, and simulates outcomes (whether and when each incident is inspected) under counterfactual policies.

On each day t, $\mathcal{N}_{k,b}^t$ incidents of category k in Borough b arrive, and are immediately placed in their respective queues, waiting to be inspected. To decide the counterfactual number of inspections in each Borough on day t, we distribute the historical total inspections count I_t among Boroughs according to a Multinomial distribution governed by the budget distribution policy $\{C_b\}$:

$$\{I_b^t\} \sim \text{Multinomial}(I^t, \{C_b\}),$$

where I_b^t denotes the number of inspections allocated to Borough b on day t. This choice ensures that when the number of days in the simulation is large, the number of inspections in Borough b is close to a fraction C_b of the total number of inspections while allowing for day-to-day fluctuations.

Once I_b^t is determined, we further determine the number of inspections for each category of incident. Assuming the categories currently in backlog are $\mathcal{K}_b(t)$, and $\sum_{k \in \mathcal{K}_b(t)} \phi_{k,b} = \Phi_b(t)$, the number of inspections on category k, denoted by $I_{k,b}^t$ is generated through

$$\{I_{k,b}^t\} \sim \text{Multinomial}(I_b^t, \{\phi_{k,b}/\Phi_b(t)\}),$$

following the setup of the GPS scheme.[9] In most cases, $I_{k,b}^t$ is smaller than the actual number of k, b incidents in the backlog, denoted by $B_{k,b}^t$, and we inspect them randomly, by sampling $I_{k,b}^t$ incidents from the earliest $\rho(B_{k,b}^t - I_{k,b}^t) + I_{k,b}^t$ incidents in the backlog.[10]

As discussed above, historical incident arrival rates are larger than historical budgets can handle, so we specify inspection fractions $\{p_{k,b}\}$ as decision variables. After each period of D days, for each inspection request in the backlog, with probability $1 - p_{k,b}$ for appropriate k and b, we decide that it will not be inspected ("dropped"). This mimics such a process in practice: when faced with a list of inspection requests, inspectors with capacity constraints would only inspect those with perceived urgency above a certain threshold, and such decisions are reviewed on a weekly or monthly basis. If the incident is not dropped, it stays in its backlog, awaiting future inspections.

The following pseudo-code provides an overview of this simulation process.

3.2.4 Evaluation metrics. After completing each simulation run, we take the inspected incidents and calculate the empirical SLAs $\hat{z}_{k,b}$ by evaluating the actual 75 percentile of inspection delays for category k in Borough b. We further calculate the empirical inspection fraction $\hat{p}_{k,b}$ as the fraction of inspected incidents over the total number of arrivals. Note that both dropped incidents and incidents still in the backlog at the end of the simulation are considered uninspected (and thus each reduces $\hat{p}_{k,b}$). For each Borough, we calculate the cost incurred as

$$\text{Cost}_b(\hat{\mathbf{z}}, \hat{\mathbf{p}}) = \sum_{k \in \mathcal{S}} r_{k,b}\hat{z}_{k,b} + 100 r_{k,b}(1 - \hat{p}_{k,b}). \tag{11}$$

Where $r_{k,b}$ is the historical average risk level of an incident of that category and Borough. The penalty on the uninspected fractions of incidents is designed so that when no incidents are inspected, the cost incurred is equivalent to having a 100-day SLA.

[9] To see this, note that under GPS, the numbers of inspections for all categories are independent Poisson random variables; conditional on them having a sum of I_b^t, the distribution of these variables becomes multinomial.

[10] This mimics the fact that not all inspections are FCFS: more urgent incidents might receive faster inspection, even if they are reported later, as we see in historical data. Note that when $\rho = 1$, it is equivalent to randomly inspecting incidents in the backlog, and when $\rho = 0$, it is strictly FCFS.

Simulation process: adapted to historical data under Borough budget policy framework.

```
Data: number of days $T$; historical arrivals $\{\mathcal{N}^t_{k,b}\}$; historical inspections performed $\{I^t\}$.
Variables: Borough budgets $\{C_b\}$; GPS scheduling policy $\{\phi_{k,b}\}$; target inspection fraction $\{p_{k,b}\}$.
Hyperparameters: review period length $D$; FCFS violation $\rho$.
Result: inspected incidents with inspection delay, incidents still in backlog, and incidents dropped.
t ← 1, backlog ← $\emptyset$, inspected ← $\emptyset$, dropped ← $\emptyset$.
while $t \leq T$ do
    backlog ← backlog $\cup$ $\mathcal{N}^t_{k,b}$
    inspections in each Borough $\{I^t_b\}$ ← Multinomial($I^t$, $\{C_b\}$)
    for Borough $b \in \mathcal{B}$ do
        number of inspections for each category $\{I^t_{k,b}\}$ ← Multinomial($I^t_b$, $\{\phi_{b,k}/\Phi_b(t)\}$);
        inspections for each category ← backlog with indices Unif$\left(I^t_{k,b}, \rho(B^t_{k,b} - I^t_{k,b}) + I^t_{k,b}\right)$;
        inspected ← inspections for each category;
        backlog ← backlog \ inspections for each category;
    end
    if $t$ mod $D = 0$ then
        for incident $\in$ backlog do
            with probability $(1 - p_{k,b})$, backlog ← backlog \ incident, dropped ← dropped $\cup$ incident
        end
    end
    current day ← current day + 1;
end
```

Analogous to Section 2.3, we define the empirical efficiency and equity loss as

$$g(\hat{\mathbf{z}}, \hat{\mathbf{p}}) = \sum_{b \in \mathcal{B}} \text{Cost}_b(\hat{\mathbf{z}}, \hat{\mathbf{p}}), \tag{12}$$

$$f(\hat{\mathbf{z}}, \hat{\mathbf{p}}) = \max_{b \in \mathcal{B}} \text{Cost}_b(\hat{\mathbf{z}}, \hat{\mathbf{p}}), \tag{13}$$

and define the objective function as $L_\gamma(\hat{\mathbf{z}}, \hat{\mathbf{p}}) = \gamma g(\hat{\mathbf{z}}, \hat{\mathbf{p}}) + (1 - \gamma) f(\hat{\mathbf{z}}, \hat{\mathbf{p}})$. For a set of values of γ, we identify the policy with the smallest objective function value.

3.2.5 Policy optimization. For policy parameters, we perform a random search over the feasible domain. For example, for Borough budget policies, we first draw a set of budget allocations $\{C_b\}$ such that $\sum_{b \in \mathcal{B}} C_b = 1$, and then draw a set of GPS parameters $\{\phi_{k,b}\}$ such that for each $b \in \mathcal{B}$, we have $\sum_{k \in \mathcal{S}} \phi_{k,b} = 1$. We generate 20,000 sets of policy parameters in total for each policy class. We note that we conduct a random search out of simplicity; we could also proceed with a more sophisticated approach to choosing the next policy to simulate, such as through a Bayesian optimization framework. One challenge with doing so is that we want to simultaneously optimize for multiple objective functions (parameterized by γ), and so would need to choose policies in a way that does not prioritize a single choice of γ.

4 EMPIRICAL APPLICATION

We now apply our framework to data from the New York City Department of Parks and Recreation.

Data Source (Calendar Year)	Objective Function	Historical	Most efficient	Most equitable
2019	Efficiency loss	51,801	**17,201**	19,950
	Equity loss	16,595	5,091	**4,981**
2021	Efficiency loss	77,975	**11,843**	15,235
	Equity loss	21,246	**4,245**	4,872
2022	Efficiency loss	89,394	**17,238**	18,152
	Equity loss	21,671	6,908	**4,069**

Table 1. Performance of the most efficient and most equitable Borough budget GPS policies, when evaluated on data from 2019 ("training set"), 2021, and 2022 ("test set"), compared with performance evaluated from historical inspection data. A lower value indicates better performance, and the best performer of each row is indicated in bold. We omit 2020 due to a large storm dominating the number of incidents; during such emergency periods, the city activates cross-agency resources and does not follow its regular operations.

4.1 Data and Methods Description

We use publicly accessible data shared by the NYC DPR on the NYC Open Data Portal, which includes both inspection requests submitted to the forestry unit of DPR,[11] and inspections performed.[12]

We perform our main policy evaluations and selection using historical data from the calendar year 2019 and then out-of-sample evaluations on their performance using historical data from 2021 and 2022. During 2019, a total of 93,570 inspection requests were submitted, and a total of 51,610 inspections were performed. We exclude data from 2020 because, as can be observed from Figure 4, requests and inspections for that year are dominated by a tropical storm in August 2020; operational policies during such emergency periods differ substantially from those during regular periods – there are large cross-agency and cross-Borough resources allocated to the affected areas, and so optimal policies take a different structure and must be calculated separately.

A single policy simulation takes on average 13 minutes on a modern machine using 1 CPU core and 4GB of RAM (runtime varies across policies but does not exceed 20 minutes for any). As a result, running 20,000 policy evaluations for each of the two classes of policies using 200 CPU cores takes around 65 hours. In Figure 5, we present the objective function values of best policy evaluated versus number of policies evaluated, for both classes of policies and under various efficiency weights γ. We find that 20,000 policy evaluations are a reasonable amount for the marginal improvement to be small, indicating that the best policies evaluated in each class are close to the true optimal policies.[13]

4.2 Results

We now analyze optimal policies within each class. Table 1 contains, for each of 2019, 2021, and 2022, the efficiency and equity loss of three policies: the actual policy for that year, and the most efficient and equitable Borough budget policy, respectively, as calculated using 2019 data. Table 2 details how these policies allocate budgets across Boroughs (where for the historical policy, we only include 2019). Appendix Table 4 through Table 8 shows how these policies prioritize different categories and the resulting empirical SLAs in the five Boroughs in 2019. Figure 2 contains the cost

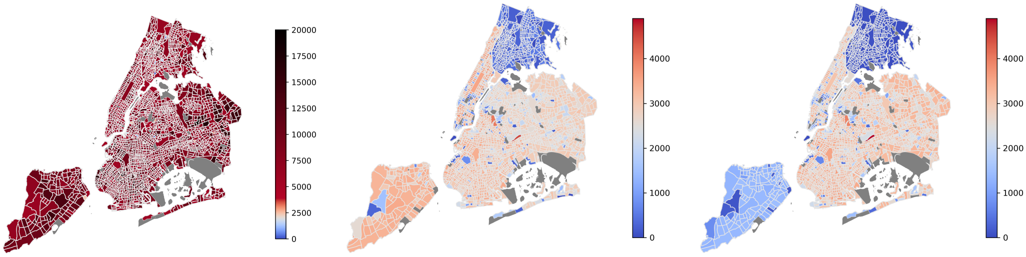
[11] https://data.cityofnewyork.us/Environment/Forestry-Service-Requests/mu46-p9is/

[12] https://data.cityofnewyork.us/Environment/Forestry-Inspections/4pt5-3vv4/

[13] We note that, given the nature of simulation optimization methods, obtaining proof of convergence to global optimum is inherently challenging, whereas our approach already provides reasonably good solutions within acceptable runtime.

Borough	Fraction of Inspection Requests	Budget Allocation		
		Historical	Most Efficient	Most Equitable
The Bronx	0.09	0.11	0.18	0.19
Brooklyn	0.32	0.28	0.35	0.28
Manhattan	0.10	0.09	0.06	0.06
Queens	0.37	0.39	0.35	0.32
Staten Island	0.12	0.13	0.07	0.14

Table 2. Comparison of budget allocation in historical inspections, and the most efficient and equitable Borough budget policies with the fraction of inspection requests received in each Borough.



(a) Historical inspections

(b) Most efficient

(c) Most equitable

Fig. 2. Cost borne by each census tract under three different inspection policies: (a) historical inspections, (b) the most efficient Borough budget policy, and (c) the most equitable Borough budget policy, evaluated on data from 2019. Instead of evaluating $\hat{p}_{k,b}$ (fraction of inspected incidents) and $\hat{z}_{k,b}$ (95% percentile of inspection delays) for each Borough b, we evaluate them for each census tract in NYC, which are finer-grained sub-divisions of Boroughs. We then calculate the cost for each census tract analogous to Equation (11). Note that the three plots are colored using the same scale. Census tracts with no incidents in 2019 are colored in gray. Qualitatively, the most efficient and equitable policies are far more similar to each other than they are to the status quo.

borne by each neighborhood (census tract) in New York City in 2019 according to each of these policies. Finally, Figure 3 contains the Pareto curve for each policy class, in terms of efficiency and equity loss. We now overview insights from this analysis.

Optimal Borough budget policies perform well and are robust. Consider Table 1, showing, for the class of Borough budget GPS policies, the most efficient and equitable policies (those with the lowest objective value with $\gamma = 1$ and $\gamma = 0$, respectively), and compare their performance with observed historical inspection plans. We find that the optimal policies are robust to out-of-sample evaluations: the most efficient policy remains the most efficient policy in future years (among the three in the table), and the most equitable policy remains the most equitable on data from 2022. Crucially, both optimal policies substantially outperform historical plans. As we discuss below, especially out of sample, the equity loss of the most efficient and equitable policies are similar. Further, note that, as shown in Figure 2, the cost in every Borough can be reduced substantially through such policies.

Comparing optimal budget allocations in Table 2 with historical inspection requests received, we find that the optimal allocations differ substantially from historical budgets in some cases. In

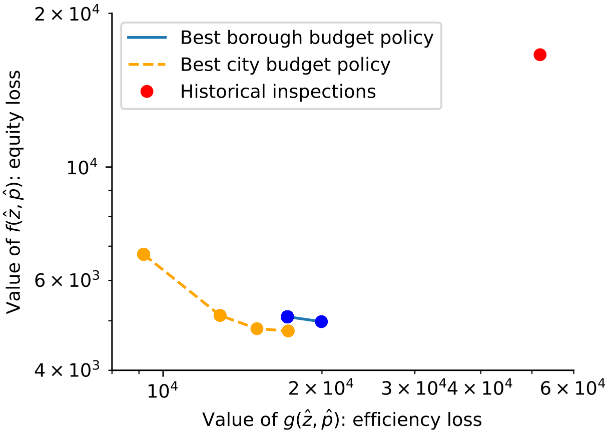


Fig. 3. Pareto frontier of equity-efficiency trade-off, under different classes of policies and historical inspections. Lower values on each axis indicate better performance: more efficient policies are more to the left, and more equitable policies are more to the bottom.

particular, the Bronx receives far more budget than historically given and Manhattan receives far less. Interestingly, while both the Bronx and Manhattan receive a similar fraction of incidents and similar historical budgets, the optimal (for either efficiency or equity) substantially prioritizes the Bronx. This may be because in general, incidents from the Bronx are more risky than those from Manhattan (see Appendix Table 3). Another portion of the improvement may be attributed to managing the queues effectively through the GPS scheme. For example, historically we see that even within the same category in the same Borough, inspections may not be first-come-first-served, thus contributing to the response time distribution having a heavier tail.[14]

The Equity-efficiency trade-off is small. In Figure 3, we map the Pareto frontier of equity-efficiency trade-off under different classes of policies and historical inspections. Restricted to the class of Borough budget policies, we see that the trade-off is very small. Indeed, Table 1 indicates that the most efficient policy is only 5% less equitable than the most equitable policy, whereas the most equitable policy is around 16% less efficient than the most efficient policy.

Drawing insights from Proposition 2.4 and Proposition 2.5, this is because when restricted to the same category, the difference in average risks of incidents from different Boroughs is relatively small. In Appendix Figure 6, we show the distribution of risk ratings for Hazards. Though different Boroughs receive substantially different *amount* of Hazard incident requests, out of the ones that were given risk ratings, the distributions are very similar, as are the average risk ratings.

Optimal Borough budget policies are close to optimal city budget policies. We next compare optimal Borough budget policies with city budget policies – to evaluate an upcoming change that will partially centralize operations. It is worth pointing out, that the most efficient city budget policy (representing the left-most tip of the dashed orange line) is a pathological case, where almost no resources are dedicated to Brooklyn (which receives a large number of inspection requests),

[14] In practice, FCFS is intended: currently each inspector faces a dashboard that sorts incidents by category and then by age when making inspection decisions. However, of the 34,688 incidents that were both reported and inspected in 2019, 20,103 of them were inspected later than at least one same-category, same-Borough incident that was reported later, indicating that FCFS is not perfectly implemented in practice. Future work is required to understand *why* there may be deviations.

and thus improves the overall efficiency at the cost of sacrificing one Borough and significantly worsening the equity loss. In practice, such a policy would rarely be implemented.

Excluding such a pathological policy, we see that the optimal Borough budget policies are very close to the optimal city budget policies – though there are some efficiency gains, the equity gain is almost negligible. Intuitively, such a finding suggests that though we can indeed make the inspections more efficient by pooling resources and centralizing response, this would not improve the equity by much – simply pooling resources does not give policymakers more levers in optimizing for equity.

More importantly, this finding is particularly relevant to practitioners, as it suggests that by reasonably allocating decentralized budgets and managing the queues well, we can come very close to the effect of implementing centralized policies, which may be more logistically complicated.

5 CONCLUSIONS

In this paper, we analyze and engineer equitable, efficient government resource allocation policies, with two government levers: per-area budgets and incident category prioritization. We provide a theoretical model, formulating a service level agreement design problem. Such a stylized model provides insights into the price of equity and serves as a foundation for our empirical simulation optimization framework, which can further capture a large class of policies. We apply our simulation framework to the design of SLAs in New York City and find that empirically, the trade-off between efficiency and equity is indeed small, and the optimal policies are robust to out-of-sample data. We further observe that by allocating resources well, decentralized agencies that are easier to run practically can perform almost as well as their centralized counterparts, which incur much higher logistical costs. This provides a valuable argument for an ongoing debate on whether and how agencies should centralize their response operations.

Several lines of future directions remain. First, our model and empirical analyses capture "non-emergency" periods, where the fluctuations in incident arrival and inspections are not dramatic. The ability to capture "emergency" periods, where a surge of incidents arrives, is a valuable extension. Second, our learned optimal policies outperform status-quo by a large margin. The ability to understand the contributions to this improvement by efficient budget allocation and effective management of queues could help us better understand the immediate next steps for agencies to implement our suggested policy changes. Lastly, our simulation optimization framework allows for more efficient zeroth order optimization methods, such as Bayesian optimization methods. Incorporating such methods may enable policymakers to conduct analyses similar to ours more frequently, thus being able to better adjust policy parameters.

REFERENCES

Gabriel Agostini, Emma Pierson, and Nikhil Garg. 2024. A Bayesian Spatial Model to Correct Under-Reporting in Urban Crowdsourcing. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

Sigrún Andradóttir, Hayriye Ayhan, and Douglas G Down. 2003. Dynamic server allocation for queueing networks with flexible servers. *Operations Research* 51, 6 (2003), 952–968.

René Bekker and Arnoud M de Bruin. 2010. Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research* 178 (2010), 45–65.

Stephen P Boyd and Lieven Vandenberghe. 2004. *Convex optimization.* Cambridge university press.

Daren C Brabham. 2015. *Crowdsourcing in the public sector.* Georgetown University Press.

Jeffery K Cochran and Kevin T Roche. 2009. A multi-class queuing network analysis methodology for improving hospital emergency department performance. *Computers & Operations Research* 36, 5 (2009), 1497–1512.

Maxime C Cohen, Adam N Elmachtoub, and Xiao Lei. 2022. Price discrimination with fairness constraints. *Management Science* 68, 12 (2022), 8536–8552.

Rupert Freeman, Nisarg Shah, and Rohit Vaish. 2020. Best of both worlds: Ex-ante and ex-post fairness in resource allocation. In *Proceedings of the 21st ACM Conference on Economics and Computation.* 21–22.

Nikhil Garg, Hannah Li, and Faidra Monachou. 2020. Dropping Standardized Testing for Admissions Trades Off Information and Access. *arXiv preprint arXiv:2010.04396* (2020).

Linda Green. 2006. Queueing analysis in healthcare. *Patient flow: reducing delay in healthcare delivery* (2006), 281–307.

Kathryn P Hacker, Andrew J Greenlee, Alison L Hill, Daniel Schneider, and Michael Z Levy. 2022. Spatiotemporal trends in bed bug metrics: New York City. *PloS one* 17, 5 (2022), e0268798.

Li-jie Jin, Vijay Machiraju, and Akhil Sahai. 2002. Analysis on service level agreement of web services. *HP June* 19 (2002), 1–13.

Nathanael Jo, Bill Tang, Kathryn Dullerud, Sina Aghaei, Eric Rice, and Phebe Vayanos. 2023. Fairness in contextual resource allocation systems: Metrics and incompatibility results. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 11837–11846.

Leonard Kleinrock. 1975. Queue Systems, Volume I: Theory. *Jonh Wiley & Sons* (1975).

Constantine E Kontokosta and Boyeong Hong. 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 (2021), 102503.

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

Zhi Liu, Uma Bhandaram, and Nikhil Garg. 2023. Quantifying spatial under-reporting disparities in resident crowdsourcing. *Nature Computational Science* (2023), 1–9.

Zhen Liu, Mark S Squillante, and Joel L Wolf. 2001. On maximizing service-level-agreement profits. In *Proceedings of the 3rd ACM conference on Electronic Commerce*. 213–223.

Tasfia Mashiat, Xavier Gitiaux, Huzefa Rangwala, Patrick Fowler, and Sanmay Das. 2022. Trade-offs between group fairness metrics in societal resource allocation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1095–1105.

Vishwali Mhasawade, Yuan Zhao, and Rumi Chunara. 2021. Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence* 3, 8 (2021), 659–666.

Dawid Nowak, Philip Perry, and John Murphy. 2004. Bandwidth allocation for service level agreement aware Ethernet passive optical networks. In *IEEE Global Telecommunications Conference, 2004. GLOBECOM'04.*, Vol. 3. IEEE, 1953–1957.

Abhay K Parekh and Robert G Gallager. 1993. A generalized processor sharing approach to flow control in integrated services networks: the single-node case. *IEEE/ACM transactions on networking* 1, 3 (1993), 344–357.

Pankesh Patel, Ajith H Ranabahu, and Amit P Sheth. 2009. Service level agreement in cloud computing. (2009).

Aida Rahmattalabi, Phebe Vayanos, Kathryn Dullerud, and Eric Rice. 2022. Learning resource allocation policies from observational data with an application to homeless services delivery. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1240–1256.

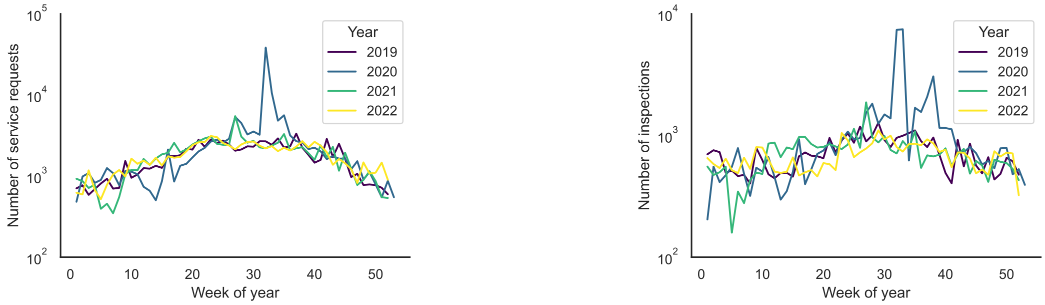
Kaippilly Raman Remesh Babu and Philip Samuel. 2019. Service-level agreement–aware scheduling and load balancing of tasks in cloud. *Software: Practice and Experience* 49, 6 (2019), 995–1012.

Shubham Singh, Bhuvni Shah, Chris Kanich, and Ian A Kash. 2022. Fair decision-making for food inspections. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–11.

Qianli Yuan. 2019. Co-production of Public Service and Information Technology: A Literature Review. *Proceedings of the 20th Annual International Conference on Digital Government Research* (2019).

A SUPPLEMENTARY MATERIALS

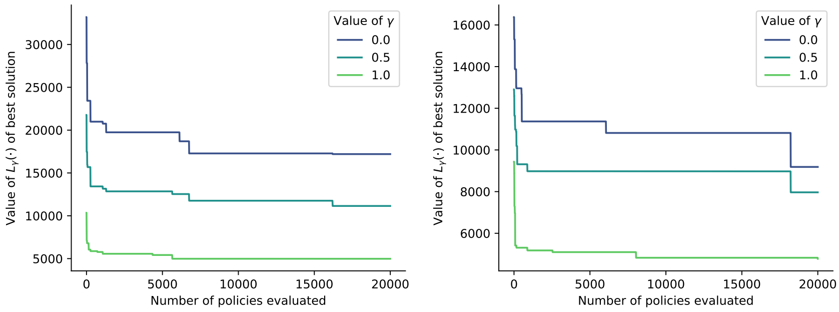
A.1 Supplementary information on empirical results



(a) Service(inspection) requests

(b) Inspections

Fig. 4. Number of service(inspection) requests and inspections by week from 2019 to 2022.



(a) Borough budget policy

(b) City budget policy

Fig. 5. Objective function value of best policy evaluated versus number of policies evaluated, under different values of γ (efficiency weight in the objective function). After 20,000 evaluations, the marginal improvement is small for both classes of policies, indicating the best policies evaluated are close to the true optimal policies.

	Bronx	Brooklyn	Manhattan	Queens	Staten Island
Hazard	6.96	6.82	7.01	7.36	7.44
Illegal Tree Damage	5.91	5.18	5.46	6.69	6.47
Other	6.86	6.19	6.18	7.67	7.75
Plant Tree	5.29	4.65	4.84	4.28	3.67
Prune	5.42	6.62	6.41	7.16	6.67
Remove Tree	6.69	6.73	5.85	7.21	7.14
Root/Sewer/Sidewalk	5.26	4.54	5.14	4.24	4.45

Table 3. Average risk rating of each category of incident in each borough.

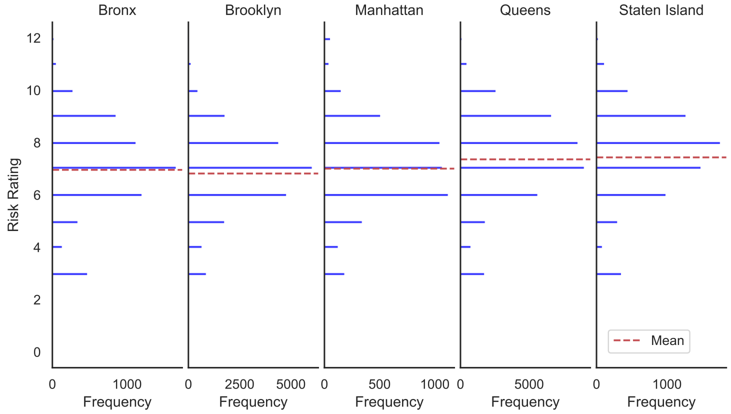


Fig. 6. Distribution of risk ratings for Hazard incidents across boroughs, with means of risk ratings within each borough marked out.

Category	Historical		Most Efficient		Most Equitable	
	SLAs (days)	Inspection Fraction	SLAs (days)	Inspection Fraction	SLAs (days)	Inspection Fraction
Hazard	36	0.75	1	1.00	1	1.00
Illegal Tree Damage	156	0.68	2	1.00	3	1.00
Other	149	0.76	1	1.00	5	1.00
Plant Tree	997	0.20	23	1.00	65	1.00
Prune	140	0.78	7	1.00	50	0.99
Remove Tree	197	0.84	1	1.00	2	1.00
Root/Sewer/Sidewalk	390	0.62	5	1.00	25	1.00

Table 4. SLAs and inspection fractions in Bronx, under the optimal Borough budget policies, compared with outcomes of historical inspections.

Category	Historical		Most Efficient		Most Equitable	
	SLAs (days)	Inspection Fraction	SLAs (days)	Inspection Fraction	SLAs (days)	Inspection Fraction
Hazard	65	0.82	367	0.68	436	0.86
Illegal Tree Damage	151	0.74	12	0.06	7	0.25
Other	176	0.53	1	1.00	1	1.00
Plant Tree	1393	0.26	15	0.03	24	0.33
Prune	586	0.35	46	0.55	50	0.69
Remove Tree	116	0.85	23	1.00	4	1.00
Root/Sewer/Sidewalk	688	0.72	14	0.02	6	0.01

Table 5. SLAs and inspection fractions in Brooklyn, under the optimal Borough budget policies, compared with outcomes of historical inspections.

Category	Historical		Most Efficient		Most Equitable	
	SLAs (days)	Inspection Fraction	SLAs (days)	Inspection Fraction	SLAs (days)	Inspection Fraction
Hazard	15	0.69	4	1.00	2	1.00
Illegal Tree Damage	305	0.51	17	0.11	10	0.00
Other	182	0.73	3	1.00	11	1.00
Plant Tree	1387	0.15	13	0.00	0	0.00
Prune	273	0.74	410	0.87	485	0.85
Remove Tree	196	0.61	16	0.08	12	0.06
Root/Sewer/Sidewalk	392	0.24	16	0.03	7	0.00

Table 6. SLAs and inspection fractions in Manhattan, under the optimal Borough budget policies, compared with outcomes of historical inspections.

Category	Historical		Most Efficient		Most Equitable	
	SLAs (days)	Inspection Fraction	SLAs (days)	Inspection Fraction	SLAs (days)	Inspection Fraction
Hazard	49	0.82	265	0.92	321	0.90
Illegal Tree Damage	477	0.69	1	0.15	2	1.00
Other	159	0.55	0	1.00	1	1.00
Plant Tree	855	0.24	1	0.03	2	0.01
Prune	479	0.33	1	0.06	2	0.42
Remove Tree	172	0.86	2	0.32	4	0.55
Root/Sewer/Sidewalk	595	0.66	2	0.01	1	0.00

Table 7. SLAs and inspection fractions in Queens, under the optimal Borough budget policies, compared with outcomes of historical inspections.

Category	Historical		Most Efficient		Most Equitable	
	SLAs (days)	Inspection Fraction	SLAs (days)	Inspection Fraction	SLAs (days)	Inspection Fraction
Hazard	21	0.58	2	1.00	273	0.86
Illegal Tree Damage	64	0.43	3	1.00	4	0.17
Other	86	0.63	2	1.00	2	1.00
Plant Tree	915	0.23	70	1.00	7	0.01
Prune	55	0.37	70	0.99	2	0.21
Remove Tree	53	0.74	56	1.00	3	0.21
Root/Sewer/Sidewalk	553	0.68	170	0.95	4	0.00

Table 8. SLAs and inspection fractions in Staten Island, under the optimal Borough budget policies, compared with outcomes of historical inspections.

A.2 Supplementary information on simulation optimization methods

In this section, we provide more details about the alternative policies that we evaluate within our empirical analyses.

City budget policy. Under this administrative policy, the city maintains a centralized server (group of inspectors), and the queues are defined for each {borough, category} pair: we maintain a different queue for incidents of each specification of {borough, category}).

To generate input to the simulation, we first draw a set of GPS parameters $\{\phi_{k,b}\}$ such that $\sum_{k\in\mathcal{S},b\in\mathcal{B}} \phi_{k,b} = 1$. We then set the target inspection fractions $\{p_{k,b}\}$ by $p_{k,b} = \frac{\phi_{k,b}\sum_t I^t}{\sum_t |N^t_{k,b}|}$, which is the ratio between the effective inspection capacity allocated and the total number of inspection requests. In cases where this ratio exceeds 1, we set $p_{k,b} = 1$ and adjust other parameters to be higher. We generate 20,000 sets of policy parameters in total. The simulation process for this type of policy is summarized in the following pseudo-code. Letting $\mathcal{K}_b(t)$ be the set of categories in backlog on day t in borough b, we denote $\Phi(t) = \sum_{k\in\mathcal{K}(t),b\in\mathcal{B}} \phi_{k,b}$.

Simulation process: under city budget with borough queues policy.

```
Data: number of days to simulate $T$; historical arrivals set each day $\mathcal{N}^t_{k,b}$, $t \in [T], \forall k, b$; historical total
      number of inspections performed each day $I^t, t \in [T]$; GPS scheduling policy defined by $\{\phi_{k,b}\}$;
      target inspection fraction defined by $\{p_{k,b}\}$.
Result: the set of inspected incidents with their response time, the set of incidents still in backlog, and
        the set of incidents dropped.
t ← 1, backlog incidents ← {}, inspected incidents ← {}, dropped incidents ← {}.
while $t \leq T$ do
    for incident $\in \mathcal{N}^t_{k,b}, \forall k, b$ do
        with probability $p_{k,b}$ append incident to backlog, otherwise append to dropped incidents
    end
    inspections on day $t$ in borough $b$ ← Multinomial($I^t$, $\{\phi_{k,b}/\Phi_b(t)\}$);
    inspected ← inspected ∪ inspections on day $t$ in borough $b$;
    backlog ← backlog \ inspections on day $t$ in borough $b$;
    current day ← current day + 1;
end
```

Evaluation metrics for alternative policies The evaluation metrics for these alternative policies are the same for the borough-budget policy, where after each simulation run, we evaluate the empirical SLAs $\hat{z}_{k,b}$ and the empirical inspection fractions $\hat{p}_{k,b}$, and calculate the cost incurred by each borough according to Equation (11).

B OMITTED PROOFS

Proposition 2.1. *Let $\boldsymbol{x}^{-1}$ be the element-wise reciprocal of $\boldsymbol{x}$. Consider the reformulated problem.*

$$\min_{\boldsymbol{x},C_b} \quad \tilde{L}(\boldsymbol{x}) = g(-\alpha\boldsymbol{x}^{-1}) + f(-\alpha\boldsymbol{x}^{-1}) \tag{5a}$$

$$s.t. \quad \sum_{k\in\mathcal{S}} x_{k,b} \leq C_b - \sum_{k\in\mathcal{S}} \lambda_{k,b}, \qquad \forall b \in \mathcal{B} \tag{5b}$$

$$\sum_{b} C_b \leq C, \tag{5c}$$

$$x_{k,b} > 0, C_b \geq 0. \tag{5d}$$

The reformulated Problem (5) *is convex. Let* $\{\boldsymbol{z}^*, \phi^*, C_b^*\}$ *and* $\{\boldsymbol{x}^*, C_b^*\}$ *be the set of optimal solutions to Problems 4 and 5, respectively. Then we have*

$$z_{k,b}^* = -\frac{\alpha}{x_{k,b}^*}, \ \textit{and}\ L(\boldsymbol{z}^*) = \tilde{L}(\boldsymbol{x}^*).$$

Proof of Proposition 2.1. First we show the convexity of problem 5. Since all constraints are linear in the decision variables, what remains is to show the objective function $\tilde{L}$ is convex in $\mathbf{x}$. Let $h(\mathbf{x}) = -\alpha\mathbf{x}^{-1}$, then $h(\mathbf{x})$ is convex in $\mathbf{x}$ since $\nabla^2 h = -2\alpha\mathbf{x}^{-3}\mathbf{I} \succ 0$ for positive x_k. Note that:

$$\tilde{L}(\mathbf{x}) = L \circ h(\mathbf{x}),$$

and by assumption, L is convex and non-decreasing, following classical results from convex analysis (e.g., see [Boyd and Vandenberghe, 2004]), $\tilde{L}$ is convex and non-decreasing.

Next we show the optimal objectives of the two problems coincide. By assumption, for constraint 4b to be feasible, we need $C_b\phi_{k,b} - \lambda_{k,b} > 0$ and $z_{k,b} > 0$. Thus this constraint is equivalent to:

$$z_{k,b} \geq \frac{-\alpha}{C_b\phi_{k,b} - \lambda_{k,b}}, \quad \forall k \in \mathcal{S}, b \in \mathcal{B} \tag{14a}$$

$$C_b\phi_{k,b} - \lambda_{k,b} > 0, \quad \forall k \in \mathcal{S}, b \in \mathcal{B} \tag{14b}$$

However, since we assume L to be non-decreasing, and there are no other constraints on $\mathbf{z}$, a necessary condition for optimality is that the optimal solution $\mathbf{z}^*$ must satisfy inequality 14a as an equality. This means that at optimality, the set of constraints $\{\mathbf{z}^*, \phi^*, C_b^*\}$ must satisfy are:

$$z_{k,b} = \frac{-\alpha}{C_b\phi_{k,b} - \lambda_{k,b}}, \quad \forall k \in \mathcal{S}, b \in \mathcal{B} \tag{15a}$$

$$C_b\phi_{k,b} - \lambda_{k,b} > 0, \quad \forall k \in \mathcal{S}, b \in \mathcal{B} \tag{15b}$$

$$\sum_{k\in\mathcal{S}} \phi_{k,b} \leq 1, \quad \forall b \in \mathcal{B} \tag{15c}$$

$$\sum_{b\in\mathcal{B}} C_b \leq C, \tag{15d}$$

$$\phi_{k,b} \geq 0, z_{k,b} \geq 0, C_b \geq 0. \tag{15e}$$

Now we make the substitution of $x_{k,b} = C_b\phi_{k,b} - \lambda_{k,b}$, and this set of constraints equivalently become

$$z_{k,b} = \frac{-\alpha}{x_{k,b}}, \quad \forall k \in \mathcal{S}, b \in \mathcal{B} \tag{16a}$$

$$x_{k,b} > 0, \quad \forall k \tag{16b}$$

$$\sum_{k\in\mathcal{S}} x_{k,b} - C_b \leq -\sum_{k\in\mathcal{S}} \lambda_{k,b}, \quad \forall b \in \mathcal{B} \tag{16c}$$

$$\sum_{b\in\mathcal{B}} C_b \leq C, \tag{16d}$$

$$C_b \geq 0, z_{k,b} \geq 0, \tag{16e}$$

where constraint 15a corresponds to 16a, constraint 15b corresponds to the positivity constraint, constraint 15c corresponds to 16c, and constraint 15d remains the same. Constraint 16a does not restrict the feasible region of $\mathbf{x}$ and can be further omitted. Furthermore, we find that $L(\mathbf{x}) = \tilde{L}(\mathbf{x})$ after this substitution. To conclude, at optimality, problem 4 and 5 have equivalent constraints and equivalent objective functions, thus the following must hold:

$$z_{k,b}^* = -\frac{\alpha}{x_{k,b}^*}, \ \text{and}\ L(\mathbf{z}^*) = \tilde{L}(\mathbf{x}^*).$$

We do note that, however, the two problems are not always equivalent: since z_k can violate the equality and still be feasible, a feasible solution to 4 may not always correspond to a feasible solution to 5 after such substitution.

□

Proposition 2.2. *[Extreme efficiency prioritization] When $\gamma = 1$, the optimal solution to Problem* (5) *is such that $x_{k,b} \propto \sqrt{\lambda_{k,b} r_{k,b}}$. Consequently, the optimal solution to Problem* (4) *is such that*

$$z_{k,b} \propto \frac{1}{\sqrt{\lambda_{k,b} r_{k,b}}}, \ \forall k, b.$$

Proof of Proposition 2.2. Since now both the efficiency and fairness loss functions are linear or piecewise linear in $\mathbf{z}$, we shall drop the constant α in the objective of problem 5, and focus on analyzing the following problem for this and the next proof, which should suffice given Proposition 2.1:

$$\begin{aligned} \min_{\mathbf{x}, C_b} \quad & \tilde{L}_\gamma(\mathbf{x}) = g_\gamma(\mathbf{x}^{-1}) + f_\gamma(\mathbf{x}^{-1}) && && \text{(17a)} \\ \text{s.t.} \quad & \sum_{k \in \mathcal{S}} x_{k,b} \le C_b - \sum_{k \in \mathcal{S}} \lambda_{k,b}, && \forall b \in \mathcal{B} && \text{(17b)} \\ & \sum_b C_b \le C, && \forall b \in \mathcal{B} && \text{(17c)} \\ & x_{k,b} > 0, \quad \forall k. && && \text{(17d)} \end{aligned}$$

Under the extreme efficiency case, the objective function becomes

$$\tilde{L}_1(\mathbf{x}) = \sum_{k,b} \lambda_{k,b} r_{k,b} / x_{k,b}.$$

Note that

$$\left(\sum_{k,b} \frac{\lambda_{k,b} r_{k,b}}{x_{k,b}} \right) \times \sum_{k,b} x_{k,b} = \sum_{k,b} \lambda_{k,b} r_{k,b} + \sum_{k,k',b,b'} \left(\lambda_{k,b} r_{k,b} \frac{x_{k',b'}}{x_{k,b}} + \lambda_{k',b'} r_{k',b'} \frac{x_{k,b}}{x_{k',b'}} \right) \tag{18}$$

$$\Rightarrow \sum_{k,b} \frac{\lambda_{k,b} r_{k,b}}{x_{k,b}} = \left(1 / \sum_{k,b} x_{k,b} \right) \left(\sum_{k,b} \lambda_{k,b} r_{k,b} + \sum_{k,k',b,b'} \left(\lambda_{k,b} r_{k,b} \frac{x_{k',b'}}{x_{k,b}} + \lambda_{k',b'} r_{k',b'} \frac{x_{k,b}}{x_{k',b'}} \right) \right) \tag{19}$$

$$\ge \left(1 / \left(C - \sum_{k,b} \lambda_{k,b} \right) \right) \left(\sum_{k,b} \lambda_{k,b} r_{k,b} + \sum_{k,k',b,b'} \left(\lambda_{k,b} r_{k,b} \frac{x_{k',b'}}{x_{k,b}} + \lambda_{k',b'} r_{k',b'} \frac{x_{k,b}}{x_{k',b'}} \right) \right) \tag{20}$$

$$\ge \left(1 / \left(C - \sum_{k,b} \lambda_{k,b} \right) \right) \left(\sum_{k,b} \lambda_{k,b} r_{k,b} + \sum_{k,k',b,b'} \left(\sqrt{\frac{\lambda_{k',b'} r_{k',b'}}{\lambda_{k,b} r_{k,b}}} + \sqrt{\frac{\lambda_{k,b} r_{k,b}}{\lambda_{k',b'} r_{k',b'}}} \right) \right) \tag{21}$$

where the first inequality is by combining constraints 17b and 17c and observing that $\sum_{k,b} x_{k,b} \le C - \sum_{k,b} \lambda_{k,b}$, and the second inequality is by the first-order condition of the function $h(x) = x + \frac{1}{x}, x > 0$. Note that the right-hand side of Equation (21) is a constant, that only depends on the problem parameters λ, r and C.

The conditions for these inequalities to become tight are 1) $\sum_{k,b} x_{k,b} = C - \sum_{k,b} \lambda_{k,b}$ and 2) $x_{k,b}/x_{k',b'} = \sqrt{\lambda_{k,b} r_{k,b}/\lambda_{k',b'} r_{k',b'}}, \forall k, k', b, b' \Leftrightarrow x_{k,b} \propto \sqrt{r_{k,b}}, \forall k, b$, which both hold under

$$x_{k,b} = \frac{\sqrt{\lambda_{k,b} r_{k,b}}}{\sum_{k',b'} \sqrt{\lambda_{k',b'} r_{k',b'}}} \frac{1}{C - \sum_{k',b'} \lambda_{k',b'}}, \ \forall k, b$$

and this constitutes a set of feasible solutions to problem 17, we conclude that this is in fact the set of optimal solutions. Consequently, the optimal solution to problem 4 satisfy

$$z_{k,b} = -\alpha \frac{\sum_{k',b'} \sqrt{\lambda_{k',b'} r_{k',b'}}}{(C - \sum_{k',b'} \lambda_{k',b'})\sqrt{\lambda_{k,b} r_{k,b}}} \propto \frac{1}{\sqrt{\lambda_{k,b} r_{k,b}}}.$$

In general, we note that the arguments used in this proof extends to the case where

$$\tilde{L}_1(\mathbf{x}) = \sum_{k,b} l_{k,b}/x_{k,b},$$

and $l_{k,b} > 0$ is some constant that does not depend on specific values of $\mathbf{x}$. The optimal solution under this objective would generally be

$$x_{k,b} \propto \sqrt{l_{k,b}},$$

and consequently

$$z_{k,b} \propto \frac{1}{\sqrt{l_{k,b}}}.$$

□

Proposition 2.3. *[Extreme equity prioritization] When $\gamma = 0$, the optimal solution to Problem* (5) *is such that $\sum_{k \in \mathcal{S}} r_{k,b}/(\lambda_{k,b} x_{k,b}) = M$ for some M, for all Boroughs $b \in \mathcal{B}$. Consequently, the optimal solution to Problem* (4) *is such that*

$$Cost_b(\boldsymbol{z}) = \sum_{k \in \mathcal{S}} \lambda_{k,b} r_{k,b} z_{k,b} = M, \ \textit{for some } M, \forall b \in \mathcal{B}.$$

Proof of Proposition 2.3. Under this case, the objective function in problem 17 becomes

$$\tilde{L}_0(\mathbf{x}) = \max_{b \in \mathcal{B}} \sum_{k \in \mathcal{S}} \lambda_{k,b} r_{k,b}/x_{k,b}.$$

We will show that any feasible solution $\mathbf{x}$ such that there exists at least two $b_1, b_2 \in \mathcal{B}$ where $\sum_{k \in \mathcal{S}} r_{k,b_1}/x_{k,b_1} \neq \sum_{k \in \mathcal{S}} r_{k,b_2}/x_{k,b_2}$ cannot be the optimal solution.

Without loss of generality, we assume that b_1 is the unique solution to

$$b = \arg\max_{b' \in \mathcal{B}} \sum_{k \in \mathcal{S}} \lambda_{k',b'} r_{k,b'}/x_{k,b'},$$

and $\sum_{k \in \mathcal{S}} r_{k,b_1}/x_{k,b_1} \geq \sum_{k \in \mathcal{S}} r_{k,b'}/x_{k,b'} + 2\epsilon$ for all other $b' \in \mathcal{B}$, for some $\epsilon > 0$. We note that cases with multiple maximizers can also be analyzed in this manner with an iterative approach.

Next, we pick an arbitrary k^*, and define

$$\sigma = \min\left\{ \frac{\epsilon x_{k^*,b_1}^2}{\lambda_{k^*,b_1} r_{k^*,b_1} + \epsilon x_{k^*,b_1}}, \frac{\epsilon x_{k^*,b_2}^2}{\lambda_{k^*,b_2} r_{k^*,b_2} + \epsilon x_{k^*,b_2}} \right\} > 0.$$

Define a new solution $\mathbf{x}^+$, where $x_{k^*,b_1}^+ = x_{k^*,b_1} + \sigma$ and $x_{k^*,b_2}^+ = x_{k^*,b_2} - \sigma$, and all other $x_{k,b}^+ = x_{k,b}$. Since $\mathbf{x}$ is feasible, and $\sum_{k,b} x_{k,b}^+ = \sum_{k,b} x_{k,b}$, by letting $C_{b_1}^+ = C_{b_1} + \sigma$ and $C_{b_2}^+ = C_{b_2} - \sigma$, we find $\mathbf{x}^+$

must also be feasible. However, we show that the objective function value will decrease. First note that, under $\mathbf{x}^+$,

$$\sum_{k\in\mathcal{S}} r_{k,b_1}/x_{k,b_1} > \sum_{k\in\mathcal{S}} r_{k,b_1}/x^+_{k,b_1} = r_{k*,b_1}/x^+_{k*,b_1} + \sum_{k\in\mathcal{S},k\neq k*} r_{k,b_1}/x_{k,b_1} > \sum_{k\in\mathcal{S}} r_{k,b_1}/x_{k,b_1} - \epsilon,$$

and

$$\sum_{k\in\mathcal{S}} r_{k,b_2}/x_{k,b_2} < \sum_{k\in\mathcal{S}} r_{k,b_2}/x^+_{k,b_2} = r_{k*,b_2}/x^+_{k*,b_2} + \sum_{k\in\mathcal{S},k\neq k*} r_{k,b_2}/x_{k,b_2} < \sum_{k\in\mathcal{S}} r_{k,b_2}/x_{k,b_2} + \epsilon.$$

In words, under x^+, b_1 remains the unique solution to

$$b = \arg\max_{b'\in\mathcal{B}} \sum_{k\in\mathcal{S}} r_{k,b'}/x_{k,b'},$$

so

$$\tilde{L}(\mathbf{x}^+) = \max_{b'\in\mathcal{B}} \sum_{k\in\mathcal{S}} r_{k,b'}/x^+_{k,b'} = \sum_{k\in\mathcal{S}} r_{k,b_1}/x^+_{k,b_1} < \sum_{k\in\mathcal{S}} r_{k,b_1}/x_{k,b_1} = \max_{b'\in\mathcal{B}} \sum_{k\in K_{\mathcal{S}}} r_{k,b'}/x_{k,b'} = \tilde{L}(\mathbf{x}).$$

To conclude, for any feasible solution $\mathbf{x}$ such that there exists at least two $b_1, b_2 \in \mathcal{B}$ where $\sum_{k\in\mathcal{S}} \lambda_{k,b_1} r_{k,b_1}/x_{k,b_1} \neq \sum_{k\in\mathcal{S}} \lambda_{k,b_2} r_{k,b_2}/x_{k,b_2}$ cannot be the optimal solution, therefore the optimal solution to problem 17 must be that there exists some value M such that $\sum_{k\in\mathcal{S}} \lambda_{k,b} r_{k,b}/x_{k,b} = M$, for all boroughs $b \in \mathcal{B}$. Consequently, invoking Proposition 2.1, the optimal solution to 4 is such that

$$\sum_{k\in\mathcal{S}} \lambda_{k,b} r_{k,b} z_{k,b} = M, \text{ for some } M, \forall b \in \mathcal{B}.$$

In general, we note that the arguments used in this proof extends to the case where

$$\tilde{L}_0(\mathbf{x}) = \max_{b\in\mathcal{B}} l_{k,b}/x_{k,b},$$

and $l_{k,b} > 0$ is some constant that does not depend on specific values of $\mathbf{x}$. The optimal solution under this objective would generally be

$$\text{Cost}_b(\mathbf{z}) = M, \text{for some } M, \forall b \in \mathcal{B}.$$

□

Since Proposition 2.5 is a direct corollary of Proposition 2.4, we provide the proof together.

Proof of Proposition 2.4 and Proposition 2.5. In the extreme efficiency prioritization case, we can directly derive the results from the conclusion in the proof of Proposition 2.2.

For the extreme equity prioritization case, setting

$$\lambda_1 r_1/x_1 = \lambda_2 r_2/x_2 = M,$$

we get the following constraints:

$$\begin{aligned}
&\frac{\lambda_1 r_1}{M} \leq C_1 - \lambda_1,\\
&\frac{\lambda_2 r_2}{M} \leq C_2 - \lambda_1,\\
&C_1 + C_2 \leq C.
\end{aligned}$$

By summing them together we get $\frac{\lambda_1 r_1+\lambda_2 r_2}{M} \leq C - \lambda_1 - \lambda_2$. Now note that the objective function is non-decreasing in x, thus non-increasing in M, which means at the optimal solution,

$$M = \frac{\lambda_1 r_1 + \lambda_2 r_2}{C - \lambda_1 - \lambda_2},$$

which yields x^{eq} and consequently the desired z^{eq}.

Substituting z^{eq} and z^{ef} into definitions of the efficiency loss $g(\cdot)$ and equity loss $f(\cdot)$ yields the price of equity and price of efficiency results.

□