Optimizing Trash Collection in Manhattan

Andy Yang (hy586) , Bopeng Zhang (bz292), Lixing Li (ll963)

November 11, 2024

Abstract

This paper presents a mathematical model to quantize the efficiency, equity, and robustness of a given schedule. Using this model, we implemented binary search and hill climbing algorithms to find an optimal efficient schedule under certain equity and robustness constraints. By iterating this optimization under an array of different constraints, we discovered a fundamental trade off between efficiency, equity, and robustness. Balancing efficiency with equity on the trade-off curve, the model was able to assign trash collection schedules and truck allocations for twelve sanitation districts. The model shows that, by enabling shared use of garbage trucks across districts, the fleet requirement can be reduced from 219 to 109 trucks. Additionally, we used principal component analysis and spatial segmentation to find the key factor for rat populations: residual waste and waste production variation. By reducing residual waste through optimized scheduling, the environmental carrying capacity for rats could be lowered.

Contents

1 Introduction

The densely populated borough of Manhattan generates millions of pounds of trash daily, posing a significant challenge for the New York City Department of Sanitation (DSNY). Trash bags left on sidewalks for prolonged periods that attract rats, increasing health risks and exacerbating public sanitation issues. Compounding the problem, Manhattan's narrow streets and heavy pedestrian traffic complicate trash collection logistics, requiring a well-coordinated strategy to ensure efficiency and minimize disruptions.

Currently, DSNY deploys a fleet of trucks to service twelve designated sanitation districts (MN1-MN12) across Manhattan. Each district has unique characteristics, including varying levels of waste production, population densities, and socio-economic conditions, which must be considered in designing an equitable and effective collection system. Additionally, unanticipated disruptions, such as vehicle breakdowns or spikes in waste volume during public events, can strain the system, underscoring the need for a resilient and adaptable strategy.

In this paper, we develop a mathematical model to optimize trash collection in Manhattan by addressing four core objectives: (1) design a plan to allocate garbage trucks to each district and a schedule for the trash pick-ups, (2) quantifying and balancing efficiency, equity, and model's robustness to potential disruptions, (3) determine a way to predict rat population and find out how the new schedule will effect rat population across the districts.

2 Mathematical Model

2.1 Model Overview

Our model takes a schedule S as input and outputs three quantitative indexes. These three indexes are: The Cost $\mathcal{C}(S)$; The Equity index $\mathcal{Q}(S)$; and the robustness index $Z(S)$. These indexes give a quantitative measurement on the efficiency of a schedule [1], equal right to sanitation, and robustness of schedule under unforeseen circumstances.

2.2 Term Definitions

- **District**: Manhattan's 12 designated sanitation districts. [1]
- **Truck**: The garbage truck that picks up trash in Manhattan area.
- **Cleaning Session**: A session refers to the 1-hour woring period that the garbage truck goes out and cleans up the garbage bags in a district. A cleaning session is called a morning cleaning session if it takes place from 7am to 8am. A cleaning session is called an evening cleaning session if it takes place from 6pm to 7pm. [1]
- **Cleaning day**: The day of a week cleaning sessions takes place. A cleaning day can contain a morning cleaning session or a even cleaning session or both. A week may have 2 or 3 cleaning days.

[1] See Section 2.3 assumption 1 for why cost is efficiency

- **Schedule**: A plan for garbage picking, which consists of number of cleaning days in each district and how much truck in each district.
- **Trash Pick Up Run**: One run refers to one truck picking up trash during one cleaning session.
- **Active Rat Signs**: The probability of failing rat inspection based on district.

2.3 Assumptions

1. **The efficiency of the model is determined only by the cost of implementing the schedule, which is determined only by the number of trash pick up runs.** We assume that efficiency of the model is maximized when the cost of the schedule is minimized. This means we are using less resources to collect the most amount of waste. The cost is proportional to the number of trash pick up runs in a schedule.

   Reason: The cost of the schedule completely results from paying the workers and gas for each trash pick up run, and the cost is constant. The costs for maintenance and emergencies is insignificant, so it can be ignored.

2. **The weekly schedule and variables are stable and time-independent** The schedule is a long term schedule that is constant for all weeks. For instance, if trucks operate on Tuesdays and Thursday for every week in a particular district, the number of trucks assigned, such as the distribution between morning and evening shifts, remains fixed each week. The data such as population and waste do not depend on time.

   Reason: In real world, we also don't want trash picking schedule to change too frequently. Change in data is also insignificant compare to the period of one week.

3. **Uncertainties can be modeled by Gaussian distributions.** This assumption accounts for various uncertainties that may affect waste production and cleaning. such as seasonal weather variation, illegal dumpings, and other environmental factors. As a result, both waste produced and cleaned in each district are assumed to follow a Gaussian distribution. Therefore, the difference between them, namely the remaining waste that haven't been collected will also be modeled as a normal distribution, enabling a more robust analysis of sanitation levels across districts.

   Reason: By central limit theorem, every distribution is approximately Gaussian.

4. **Details in each district is ignored.** We assume that population, waste production, trash pick up efficiency, etc. is uniform through out a district.

   Reason: This simplifies the model a lot and is detailed enough to determine a schedule for each district.

5. **Details in each specific trash pick run in the same district is ignored.** We Assume that there is no difference in trash picking capacity between each trash pick up run. We can use the same distribution to model the capacity for every trash pick up run, that is $W_{c,n}$, the distribution of waste collected for each trash pick up run, is

constant. The variation in trash collected results from the type of truck, the worker, weather and traffic condition or the type of trash is addressed by the same $\sigma_{c,n}$ the standard deviation of waste collected. The expected capacity of each trash pick run is determined only by the product of maximum trash picking capacity for each truck and a constant efficiency. We assume the maximum capacity is $R = 12$ tons and the efficiency is $\eta = 1$. [1]

Reason: The variation for trash pick up capacity is addressed by setting it as a distribution, so we can assume that every trash pick up run has the same distribution. The efficiency is difficult to determine without the details of specific trash pick run. For simplicity, we assume the efficiency is 1.

6. **Street sanitation level is determined only by the weight of left over trash.** In reality, the street cleanliness level maybe determined by the type of trash, pedestrians throwing trash, illegal dumping, etc. We ignore the difference between the type of trash and account for all sorts of trash throwing by a single distribution in waste production $W_{p,n}$.

   Reason: Trash produced by pedestrians and illegal dumping is addressed by the deviation in waste production. The "dirtiness" created by each type trash is relatively similar and insignificant compared to the absolute weight of the trash.

2.4 Variables and Functions

Notation	Definition	Units
n	An integer range 1 to 12 that represents the district number.	
$\mathcal{C}$	The cost of the schedule per week.	[Dollars]
$\mathcal{Q}$	The Equity Index, a number that measures equity across districts.	$\left[\frac{\text{Waste}}{\text{Person}\cdot\text{Area}}\right]$
Z_n	The Robustness Index of district n.	
$\mathcal{S}_n$	Sanitation Distribution of district n, a distribution that measures how clean a district is.	$\left[\frac{\text{Waste}}{\text{Area}}\right]$
$N_{m,n}$	The number of truck pick-up runs in the morning in district n.	[Truck]
$N_{e,n}$	The number of truck pick-up runs in the evening in district n.	[Truck]
N_{total}	The total number of available trucks.	[Truck]
k_n	The number of cleaning days per week in district n, it is either 2 or 3.	
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and standard deviation σ. It is defined as $\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{\sigma^2}}$.	
$W_{p,n}$	The distribution of waste produced by district n per week.	
$E[W_{p,n}]$	The expected waste produced by district n per week.	[Waste]
$\sigma_{p,n}$	The standard deviation of waste produced by district n per week.	[Waste]
$W_{C,n}$	The distribution of waste collected in district n per week.	
$E[W_{c,n}]$	The expected waste collected in district n per week.	[Waste]
$\sigma_{c,n}$	standard deviation of waste collected in district n per week.	[Waste]
R	The amount of garbage that a truck is able to collect in one cleaning session.	$\left[\frac{\text{Waste}}{\text{Truck}}\right]$
$\eta_{t,n}$	The efficiency of trash pick up at session t, which is either morning or evening, in district n.	
P_n	The population in district n .	[People]
A_n	The area of district n.	[Area]
S	The schedule.	
Z_{bound}	The lower bound of all Z_n .	
$\mathcal{Q}_{\text{bound}}$	The upper bound of $\mathcal{Q}$.	$\left[\frac{\text{Waste}}{\text{Person}\cdot\text{Area}}\right]$

Table 1: Notation and Definitions of Variables in the Model. Units left blank are dimensionless

2.5 Definitions

2.5.1 Cost (Efficiency)

- **Costs $\mathcal{C}$:** In our model, the efficiency is quantified by the cost of implementing the schedule per week, represented by $\mathcal{C}$. Maximizing efficiency is defined as minimizing the cost required to maintain the sanitation needs of each district. In particularly, it considers the total number of trucks assigned to morning and evening cleaning sessions across districts, multiplied by the number of cleaning sessions scheduled each week for each district. By assumption 1, the total cost is proportional to the total number of trash pick up runs. Thus,

$$\mathcal{C} = c \sum_{n=1}^{12} k_n \left(N_{m,n} + N_{e,n}\right) \tag{1}$$

  where:

  - c is cost per trash pick up run
  - k_n is the number of cleaning sessions per week for district n
  - $N_{m,n}$ is the number of the trucks used in morning cleaning sessions for district n
  - $N_{e,n}$ is the number of trucks used in the evening cleaning sessions for district n

2.5.2 Sanitation

- **Waste Produced ($W_{p,n}$):** The waste produced by each district every week is not neccessarily the same, thus can be modeled by a Gaussian distribution by assumption 3.

$$W_{p,n} \sim \mathcal{N}\left(E[W_{p,n}], \sigma_{p,n}^2\right) \tag{2}$$

  where:

  - $\mathcal{N}$ is the Normal Distribution.
  - $E[W_{p,n}]$ is the expected amount of trash produced each week based on past empirical data for district n.
  - $\sigma_{p,n}$ is the standard deviation of garbage produced based on past empirical data for district n.

- **Waste Collected ($W_{c,n}$):** The waste collected by pickup trucks are not necessarily the same for each week. Many factors such as weather conditions, traffic conditions, etc. may be essential in determining this value. It is modeled by Gaussian distribution by assumption 3. By assumption 5, the waste collection is the same for each trash pick up runs. Thus, it can be modeled by the same distribution.

$$W_{c,n} \sim \mathcal{N}\left(E[W_{c,n}], \sigma_{c,n}^2\right) \tag{3}$$

  where:

- $\mathcal{N}$ is the Normal Distribution,
- $E[W_{c,n}]$ is the expected amount of trash collected each week by trucks in district n. This depends on schedule parameters. Our model estimates this parameter to be

$$E[W_{c,n}] = Rk_n\left(\eta_{m,n}N_{m,n} + \eta_{e,n}N_{e,n}\right) \tag{4}$$

  in which η accounts for the efficiency of a pickup run. R is a constant of how much waste each truck can collect. By Assumption 5, the ability for the trash trucks to pick up trash is the same. So the expected value is a simple multiplication of trash pick rum number and capacity. And we assumed $\eta = 1$, $R = 12$ tons from the information given in problem description. [1]

- $\sigma_{c,n}$ is the standard deviation of garbage cleaned for district n. The value can be calculated using the standard deviation of a binomial distribution. This makes sense as a binomial distribution as the amount of waste cleaned stems from the sum of binomial choices of whether a truck can function over the number of trucks.

$$\sigma_{c,n} = R\sqrt{(1 - P_n)P_n} \tag{5}$$

  where P_n is the probability of a pickup truck being able to perform its required duties despite unforeseen circumstances. This probability is a function of many district dependent real life variables such as weather, truck break downs, civil unrest, etc. In its simplest form, we can approximate this probability of as the product of dominant factors.

$$P_n = P_{\text{Truck}} \cdot P_{\text{Weather}} \cdot P_{\text{no civil unrest}} \tag{6}$$

  where:

  * P_{Truck} is probability of truck not malfunctioning every week
  * P_{Weather} is probability of good weather, which we will approximate as the probability of not having floods per week. This is district dependent
  * $P_{\text{No Civil Unrest}}$ is probability of no major civil unrest such as protests and strikes that might stop garbage collection happening per week

- **Sanitation Distribution ($\mathcal{S}_n$)[2]**: The Sanitation Distribution $\mathcal{S}_n$ for district n quantifies the level of cleanliness by measuring the amount of uncollected waste per unit area. It reflects the effectiveness of waste collection services in a district. A higher Sanitation Distribution indicates more residual waste per area, implying lower sanitation levels, whereas a lower index suggests better sanitation due to efficient waste collection. By assumption 6, the level of sanitation is determined only by the amount of left over trash, so we can simply substract the weight of waste produced by the weight of waste collected. By assumption 4, the waste production is uniformly spread

[2] Notice that the higher the expectation value of the Sanitation Distribution the more "dirty" a district is. This is because $\mathcal{S}_n$ is being defined as the distribution of leftover trash.

across the district. So the sanitation level is also uniform. For each district $\mathcal{S}_n$ can by represent by a single distribution that does not depend on specific locations.

$$\begin{aligned} \mathcal{S}_n &= \frac{1}{A_n}\left(W_{p,n} - W_{c,n}\right) \\ &= \mathcal{N}(\frac{1}{A_n}\left(E[W_{p,n}] - E[W_{c,n}]\right), \frac{1}{A_n^2}\left(\sigma_{p,n}^2 + \sigma_{c,n}^2 k_n(N_m + N_e)\right)) \end{aligned} \tag{7}$$

where:

- A_n is the area of district n.

- **Expected Trash Residual (δ_n)**: The expected trash residual δ_n for district n is the expected amount of trash that will remain in a district due to either unexpected amount of trash being produced or unexpected small amounts of trash being collected in the week.

$$\delta_n = \int_0^\infty \mathcal{P}_n(t) t dt \tag{8}$$

where $\mathcal{P}_n(t)$ is the probability that t amount of trash in left on the street measured in tons. Here,

$$\mathcal{P}_n(t) = A_n \mathcal{S}_n = \mathcal{N}(E[W_{p,n}] - E[W_{c,n}], \sigma_{p,n}^2 + \sigma_{c,n}^2 k_n(N_m + N_e)) \tag{9}$$

Therefore,

$$\delta_n = \frac{1}{\sqrt{2\pi}} \int_0^\infty \frac{1}{\sigma} t e^{\frac{(t-\mu)^2}{2\sigma^2}} dt \tag{10}$$

which simplifies to:

$$\delta_n = \frac{1}{2}\left(\mu \ \operatorname{erf}\left(\frac{\mu}{\sqrt{2}\sigma}\right) + \mu + \sqrt{\frac{2}{\pi}} \sigma e^{-\frac{\mu^2}{2\sigma^2}}\right) \tag{11}$$

where:

- μ is the mean of $\mathcal{P}_n(t)$, is equal to $E[W_{p,n}] - R\eta k_n\left(N_{m,n} + N_{e,n}\right)$,
- σ is the standard deviation of $\mathcal{P}_n(t)$, and is equal to $\sqrt{\sigma_{p,n}^2 + \sigma_{c,n}^2 k_n(N_m + N_e)}$,
- $\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt$ is the error function.

2.5.3 Equity

Equity Index ($\mathcal{Q}$)[3] In this model, equity is represented by the equity index $\mathcal{Q}$, which measures fairness in trash collection across different districts. According to the United

[3] Notice that the higher the index is, the more "unequal" it is. This is because $\mathcal{Q}$ is a standard deviation. Careful readers may notice that most of our variables has name that is exact opposite of its physical meaning. (Equity Index measures unequal-ness, Sanitation Distribution measures dirtiness, Cost measures inefficiency) But the inconsistencies can make you stay focused.

States Environmental Protection Agency's Office of Resource Conservation and Recovery[2], equity is defined as providing equal exposure to waste management services for each people.

The quality of waste management services can be quantified by the amount of left over trash density, that is $\frac{\delta_n}{A_n}$. By assumption 6, the density is uniform in each district. The service provided on each person is then determined by dividing the service quality by the number of people, which gives $\frac{\delta_n}{A_n P_n}$.

Equality means the service provided for each person is relatively the same for each district. The standard deviation is a good way to measure how *unequal* a schedule is. Therefore, the equity index is calculated by measuring the standard deviation of sanitation service levels per capita within each district. The higher the equality index is, the more unequal it is.

$$\mathcal{Q} = \sqrt{\frac{1}{12}\sum_{n=1}^{12}\left[\frac{\delta_n}{P_n A_n} - \left(\overline{\frac{\delta_n}{P_n A_n}}\right)\right]^2} \tag{12}$$

where:

- δ_n is expected left over trash in district n,
- P_n is the population of district n,
- $\frac{\delta_n}{P_n A_n}$ is the left over trash per area per capita for district n.
- $\left(\overline{\frac{\delta_n}{P_n A_n}}\right) = \frac{1}{12}\sum_{n=1}^{12}\frac{\delta_n}{P_n A_n}$ is mean left over trash per area per capita for all districts n.

2.5.4 Robustness

Robustness Index (Z_n): In this model, Z_n is a measure of how likely the schedule is able to finish its cleaning quotas despite uncertainties in waste output and waste collection. It be quantified using the *z-score* (in statistics) of $t = 0$.

$$Z_n = \frac{0 - E[\mathcal{S}_n]}{\sigma_{\mathcal{S}_n}} \tag{13}$$

where:

- $E[\mathcal{S}_n]$ is the expected value of $\mathcal{S}_n$
- $\sigma_{\mathcal{S}_n}$ is the standard deviation of $\mathcal{S}_n$

2.6 Optimization and Constraints

The objective is to maximize efficiency by optimizing our cost function $\mathcal{C}(S)$, while keeping our constraints—the Sanitation Distribution $\mathcal{S}_n(S)$ and Equity Index $\mathcal{Q}_n(S)$ within specified thresholds.

To ensure the model prioritizes citizens' health and equitable treatment, we aim to minimize the cost function $\mathcal{C}(S)$ while enforcing strict thresholds for both the Sanitation Distribution $\mathcal{S}_n(S)$ and Equity Index $\mathcal{Q}(S)$. These indices represent essential standards for public health and social fairness in Manhattan, reflecting a commitment to equality and the right to health that must not be compromised. Consequently, rather than allowing any trade-off between these standards and cost, we frame equity and sanitation as non-negotiable constraints in our optimization model, ensuring cost reduction only within the bounds of these critical requirements.

Therefore, we choose the minimize cost while treating equity and Sanitation Distribution as constraints.

2.6.1 Independent Variables and Domain

Our independent variable is the schedule S for trash pick up runs for each district. A schedule for each district is determined by the number of cleaning days in a week k_n and the number of trash pick up runs in each session of the day $N_{m,n}$ and $N_{e,n}$. Therefore, the overall schedule is a tuple of $3 \times 12 = 36$ integers: k_1 to k_{12}, $N_{m,1}$ to $N_{m,12}$,and $N_{e,1}$ to $N_{e,12}$. So $S \in \mathbb{N}^{36}$,

$$S = (k_1, ..., k_{12}, N_{m,1}, ..., N_{m,12}, N_{e,1}, ..., N_{e,12}) \tag{14}$$

The domain of the input is determined in the following way. The cleaning frequency is either 2 or 3, so for each k_n, $k_n \in \{2, 3\}$ The number of truck pick up runs cannot be higher than the number of total trucks, so $N_{m,n}, N_{e,n} \in [0, N_{\text{total}}]$.Therefore, the domain of the schedule is

$$S \in \{2, 3\}^{12} \times [0, N_{\text{total}}]^{12+12} \tag{15}$$

2.6.2 Optimization function

We want to minimize the cost of the schedule, that is, we want

$$\mathcal{C}(S) = c \sum_{n=1}^{12} k_n \left(N_{m,n} + N_{e,n}\right) \tag{16}$$

to be minimized.

2.6.3 Schedule Constraints

We require the total number of trucks used in each time of the day is lower than the total number of trucks or else the trucks is not enough. That is

$$\sum_{n=1}^{12} N_{m,n} \leqslant N_{\text{total}} \text{ and } \sum_{n=1}^{12} N_{e,n} \leqslant N_{\text{total}} \tag{17}$$

However, we can loosen this bound by stagger the schedule. For example, for two districts with 3 cleaning days, one district can clean on Monday, Wednesday, and Friday. The other district can clean on Tuesday, Thursday, and Saturday. Now, a single truck can be used in 2 morning sessions. Similarly, if there are 3 districts with 2 cleaning days, a single truck can be used in 3 morning sessions. On average, a truck can be used in $\frac{6}{k_n}$ sessions of the same time. Therefore, the new bound is

$$\sum_{n=1}^{12} N_{m,n} \frac{k_n}{6} \leqslant N_{\text{total}} \text{ and } \sum_{n=1}^{12} N_{e,n} \frac{k_n}{6} \leqslant N_{\text{total}} \tag{18}$$

2.6.4 Sanitation Constraints

The sanitation level for each district is determined by $\mathcal{S}_n$, which is a distribution that indicates the probability of leftover wastes. The most basic requirement is that the expected leftover trash $E[\mathcal{S}_n]$ is at most 0. If the expected leftover trash is greater than 0, then there will be more and more trash accumulated on the street.

$$E[\mathcal{S}_n] = \frac{1}{A_n} \left(E[W_{p,n}] - E[W_{c,n}]\right) = \frac{1}{A_n} \left(E[W_{p,n}] - R\eta k_n \left(N_{m,n} + N_{e,n}\right)\right) \tag{19}$$

So

$$E[W_{p,n}] - R\eta k_n \left(N_{m,n} + N_{e,n}\right) \leqslant 0 \tag{20}$$

However, we would want our schedule to be more robust. If the expected leftover trash is exactly 0, then if the waste produced is more the the waste collected ability, which happens about 50% of the times, then the pick up runs can't clean up all the trash.

We want a more robust schedule such that even if the waste production is slightly higher or the waste collection is slightly lower due to truck break down, weather, or traffic, we still have a high chance that the trash can be all cleaned up. Therefore, our stronger constraint is a bound on the probability that some trash is leftover, which is determined by the area under the distribution to the right of 0.

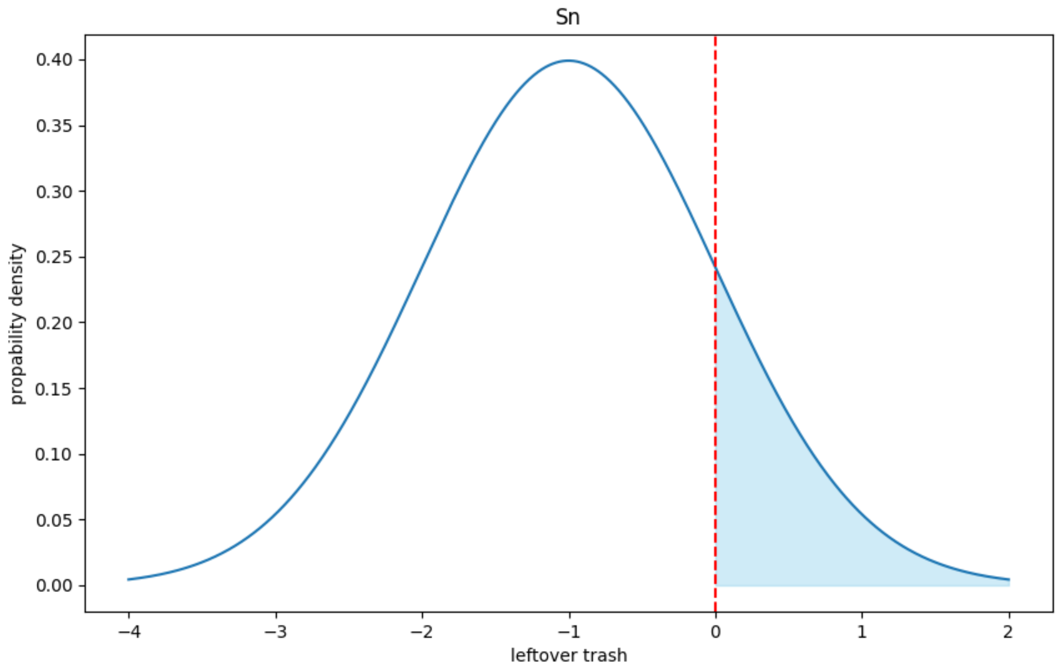

Figure 1: A visualization of the probability that some trash is left over (the shaded light-blue area).

Computing the actual probability is tedious. Instead, since the probability is directly related to the z-score of the distribution at 0, we can set a bound on the z-score. We define

$$\begin{aligned} Z_n &= \frac{0 - E[\mathcal{S}_n]}{\sigma_{\mathcal{S}_n}} \\ &= \frac{R\eta k_n \left(N_{m,n} + N_{e,n}\right) - E[W_{p,n}]}{\sqrt{\sigma_{p,n}^2 + \sigma_{c,n}^2 k_n (N_m + N_e)}} \end{aligned} \tag{21}$$

We want

$$Z_n \geqslant Z_{\text{bound}} \text{ where } Z_{\text{bound}} \geqslant 0 \tag{22}$$

Note that this constraint is stricter than the one we previously proposed because the previous constraint that $E[\mathcal{S}_n] \leqslant 0$ is equivalent to $Z_n \geqslant 0$.

2.6.5 Equity Constraints

Last but not least, we want the equality index of the schedule to be regulated by some constraint.

$$\mathcal{Q} \leqslant \mathcal{Q}_{\text{bound}} \tag{23}$$

2.7 Model Summary

In order to determine the most efficient schedule for trash cleaning while taking equity, sanitation level, and robustness in to account,

we minimize the function

$$\mathcal{C}(S) = c \sum_{n=1}^{12} k_n \left(N_{m,n} + N_{e,n}\right) \tag{F.1}$$

with constraints

$$\sum_{n=1}^{12} N_{m,n} \frac{k_n}{6} \leqslant N_{\text{total}} \tag{C.1}$$

$$\sum_{n=1}^{12} N_{e,n} \frac{k_n}{6} \frac{6}{k_n} \leqslant N_{\text{total}} \tag{C.2}$$

$$Z_n \geqslant Z_{\text{bound}} \tag{C.3}$$

$$\mathcal{Q} \leqslant \mathcal{Q}_{\text{bound}} \tag{C.4}$$

3 Optimization Algorithm

3.1 Overview

The domain of the function are integers, so the simplest algorithm is enumerating through all the possible schedules in the domain and check whether it satisfies the constraints and whether it minimizes cost. This approach is not feasible because the time complexity will be $\mathcal{O}(N^{24})$.

We will reduce the domain by grouping $k_n \left(N_{m,n} + N_{e,n}\right)$ into a single variable T_n, which represents the total number of trash pick up runs in district n per week. This reduces the time complexity down to $\mathcal{O}(N^{12})$. We will tightening the bounds for each T_n by calculating the lowest acceptable T_n from the z-score constraint and the upper bound to a z-score that is unnecessarily clean. This reduces N, the search space for each T_n, to about 100.

Next, we will use binary search to find the minimum cost $\mathcal{C}_{\min}$. This reduces time complexity to $\mathcal{O}(N^{11} \log C)$, where C is the search space for cost. More specifically, the binary search will "propose" a guess for the minimum value for $\mathcal{C}$ and check if there is any possible schedule that satisfies the constraints. Since the constraint on $\mathcal{Q}$ is an upper bound, we know that if for a given $\mathcal{C}$ the minimum $\mathcal{Q}$ is greater than $\mathcal{Q}_{\text{bound}}$, then there is no schedule that satisfy the constraint.

The problem has now become finding the minimum $\mathcal{Q}$ for a given $\mathcal{C}$. However, 100^{11} is still an unfeasible amount of calculation. We could use hill climbing algorithm to find a local minimum for $\mathcal{Q}$, which takes constant amount of time $\mathcal{O}(M)$, where M is the maximum allowed iterations. The total complexity is $\mathcal{O}(M \log C)$.

Any valid plan with cost $\mathcal{C}_{\min}$ works, but we still need a more specific schedule that determines $k_n, N_{m,n}, N_{e,n}$ that is reasonable. One measure is to minimize the total number of trucks needed, and we can simply enumerate all possibilities in $\mathcal{O}(2^{12})$.

3.2 Simplifications

In assumption 5, we assume that the efficiency of picking trash in morning session is the same as picking trash in evening session. So

$$Z_n = \frac{R\eta k_n \left(N_{m,n} + N_{e,n}\right) - E[W_{p,n}]}{\sqrt{\sigma_{p,n}^2 + \sigma_{c,n}^2 k_n (N_m + N_e)}}$$

The equity index $\mathcal{Q}$ is dependent only on the expected leftover trash for each district δ_n. δ_n is dependent only on the Sanitation Distribution $\mathcal{S}_n$.

$$\begin{aligned}\mathcal{S}_n &= \mathcal{N}(\frac{1}{A_n}\left(E[W_{p,n}] - E[W_{c,n}]\right), \frac{1}{A_n^2}\left(\sigma_{p,n}^2 + \sigma_{c,n}^2 k_n (N_m + N_e)\right)) \\ &= \mathcal{N}(\frac{1}{A_n}\left(E[W_{p,n}] - R\eta k_n \left(N_{m,n} + N_{e,n}\right)\right), \frac{1}{A_n^2}\left(\sigma_{p,n}^2 + \sigma_{c,n}^2 k_n (N_m + N_e)\right))\end{aligned}$$

Since $\mathcal{S}_n$ only depends on $k_n \left(N_{m,n} + N_{e,n}\right)$, $\mathcal{Q}$ only depends on $k_n \left(N_{m,n} + N_{e,n}\right)$.

$$\mathcal{C}(S) = c \sum_{n=1}^{12} k_n \left(N_{m,n} + N_{e,n}\right)$$

Since c is a constant, the function $\mathcal{C}$ is optimized when $\sum_{n=1}^{12} k_n \left(N_{m,n} + N_{e,n}\right)$ is optimized, we can drop the c when implementing the algorithm.

Since all functions depend only on the total number of trash pick up runs in each district we let $T_n = k_n \left(N_{m,n} + N_{e,n}\right)$ and choose to optimize with the variable T_n, which reduces the dimension from 36 to 12. Now, the simplified optimization model is

$$\text{minimize } \mathcal{C} = \sum_{n=1}^{12} T_n \tag{F.1}$$

$$\sum_{n=1}^{12} T_n \leqslant 12 N_{\text{total}} \tag{C.1}$$

$$Z_n = \frac{R T_n - E[W_{p,n}]}{\sqrt{\sigma_{p,n}^2 + \sigma_{c,n}^2 T_n}} \geqslant Z_{\text{bound}} \tag{C.2}$$

$$\mathcal{Q} \leqslant \mathcal{Q}_{\text{bound}} \tag{C.3}$$

Now, the possible values of each T_n ranges from 0 to $6N_{\text{total}}$ (when $k_n = 3, N_{m,n} = N_{e,n} = N_{\text{total}}$). But, we can find a tighter bound for T_n. Consider the constraint (C.2). We can solve for a lower bound for T_n.

$$T_n \geqslant \frac{2RE[W_{p,n}] + Z_{\text{bound}}^2 \sigma_{c,n}^2 + \sqrt{\left(2RE[W_{p,n}] + Z_{\text{bound}}^2 \sigma_{c,n}^2\right)^2 - 4R^2 \left(E[W_{p,n}]^2 - Z_{\text{bound}}^2 \sigma_{p,n}^2\right)}}{2R^2} \tag{24}$$

For branch cutting, we don't want a district to be overly cleaned because marginal benefit for putting more trucks to district significantly decreases. Therefore, we could set an upper bound for the z-score $Z_{\max}$. $Z_{\max} = 5$ is a good upper bound because that means the probability of having leftover trash is 0.00002%. Similarly, the upper bounded for T_n can be solved.

$$T_n \leqslant \frac{2RE[W_{p,n}] + Z_{\max}^2\sigma_{c,n}^2 + \sqrt{\left(2RE[W_{p,n}] + Z_{\max}^2\sigma_{c,n}^2\right)^2 - 4R^2\left(E[W_{p,n}]^2 - Z_{\max}^2\sigma_{p,n}^2\right)}}{2R^2} \tag{25}$$

We later will plug in data and find that this upper bound given by $Z_{\max} = 5$ is less than N_{total}. So we can use this tighter constraint in place of constraint (C.1).

Now, our model becomes

$$\text{minimize } \mathcal{C} = \sum_{n=1}^{12} T_n \tag{F.1}$$

$$T_{n,\text{lower}} \leqslant T_n \leqslant T_{n,\text{upper}} \tag{C.1}$$

$$\mathcal{Q} \leqslant \mathcal{Q}_{\text{bound}} \tag{C.2}$$

3.3 Binary Search

If we can clean up Manhattan with a lower cost, then we can certainly achieve the same thing with a higher cost. This provides the basis for our binary search algorithm.

The binary search finds the minimal $\mathcal{C}$ such that there exists a schedule within the constraint. The boundaries of the binary search can be determined by the boundaries of T_n

$$\mathcal{C}_{\text{lower}} = \sum_{n=1}^{12} T_{n,\text{lower}}$$

$$\mathcal{C}_{\text{upper}} = \sum_{n=1}^{12} T_{n,\text{upper}}$$

we first look at the middle cost $\mathcal{C}_{\text{mid}}$. If there is a valid schedule with this cost, then either there is a valid schedule with a lower cost or this is the optimized cost. For either case, we can look at the lower half of the interval. If there is a no valid schedule with this cost, then this cost must be too low. So we can hope to find a valid schedule in the upper half of the interval.

Here is the pseudo code for the binary search. The function isValid(mid) returns true if there exists a valid schedule with $\mathcal{C}$ = mid and return false otherwise. We will implement the function isValid($\mathcal{C}$) in the next session.

Algorithm 1 Binary Search

```
left = C_lower
right = C_upper
while left < right do
    mid = (low + high) / 2
    if isValid(mid) then
        right = mid
    else
        left = mid + 1
    end if
end while
return mid
```

3.4 Hill Climbing Algorithm

Now, our goal is to decide whether there is a schedule with cost $\mathcal{C}$ that is within the constraints. The most fundamental method is that we find the minimal equity index $\mathcal{Q}_{\text{min}}$. If $\mathcal{Q}_{\text{min}} > \mathcal{Q}_{\text{bound}}$, we can guaranteed that there is no valid schedule. This is an optimization problem on a discreet space, so we will use the hill climbing algorithm.

Our independent variable is now the total number of trash pick up runs for the first 11 districts, and $T_{12} = \mathcal{C} - \sum_{n=1}^{11} T_n$. Our search space is the bounds of the first 11 districts.

$$\prod_{n=1}^{11} [T_{n,\text{lower}}, T_{n,\text{upper}}]$$

This is a "rectangular" domain. And we propose that the minimum is near the diagonal of this "rectangle" because the number of runs is scheduled evenly based on the needs of each district, which results in a lower equity index. That is,

$$\frac{T_{n,\text{initial}} - T_{n,\text{lower}}}{T_{n,\text{upper}} - T_{n,\text{lower}}} = \frac{\mathcal{C} - \mathcal{C}_{\text{lower}}}{\mathcal{C}_{\text{upper}} - \mathcal{C}_{\text{lower}}} \tag{26}$$

Which simplifies to

$$T_{n,\text{initial}} = \frac{\mathcal{C} - \mathcal{C}_{\text{lower}}}{\mathcal{C}_{\text{upper}} - \mathcal{C}_{\text{lower}}} (T_{n,\text{upper}} - T_{n,\text{lower}}) + T_{n,\text{lower}} \tag{27}$$

We later checked that starting from the diagonal is much better than starting from a random position within the domain.

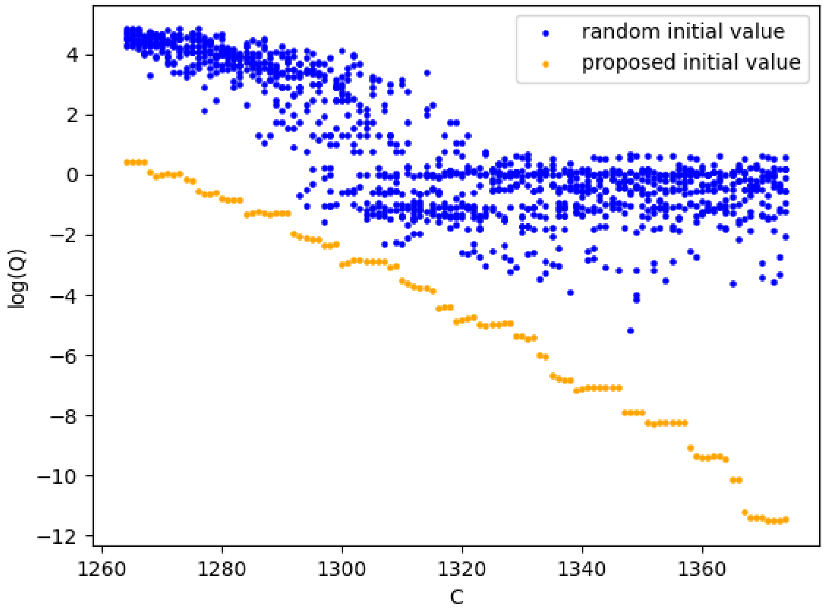

Figure 2: The minimum $\mathcal{Q}$ for different $\mathcal{C}$ returned by the algorithm starting from a random initial value verse starting on the diagonal. We can see that the minimum value found by starting from on the diagonal (orange) is always lower than starting from a random position (blue). This makes the diagonal a good initial value.

Next, we determine whether the equity index for the current schedule $T_1, ..., T_{12}$ is lower than the bound. If not, then we will try to change our schedule a little bit by either increase or decrease the total number of trash pick runs for district 1 to 11 by one. To keep the cost constant, we will compensate by decreasing or increasing T_{12} by 1. We computer the equity index for all possible directions and move to the schedule with the lowest $\mathcal{Q}$.

If $\mathcal{Q}$ cannot be lower in any direction, we are at the local minimum. If the our number of steps is higher than the maximum number of iterations, which we decided to be 500 steps, we will also return.

Algorithm 2 Hill Climbing Algorithm

```
T = T_initial
while iteration < max iterations do
  if Q(T)< Q_bound then
    return true
  end if
  for ΔT in each direction do
    compute Q(T+ΔT)
    record the T'=T+ΔT that results in a minimum Q(T')
  end for
  if no direction give a lower Q then
    return false
  end if
  T=T'
  iteration+=1
end while
return false
```

3.5 Schedule Determination

Using the above algorithms we will obtain a minimum cost $\mathcal{C}_{\text{min}}$ and at least one valid schedule that attains this minimum. It is possible that there are other valid schedule that attains the minimum, and they are equally efficient by assumption 1.

Now we have an array of number of total trash pick up runs for every district $T_1, ..., T_{12}$. Although any schedule that provides these number of runs is equally efficient and valid, we need to decide a more detailed schedule.

We want the schedule to be stable, that is, the number of trucks is the same for all morning or evening sessions through out the week for each district. This gives a more preferable schedule for the workers.

We determined that $k = 3$ for all districts is better than $k = 2$ because there will be fewer trash pick runs during the same cleaning session. This reduces traffic congestion.

The next criteria is the minimum required number of trucks. One schedule is better than the another if it needs fewer number of trucks. This gives more robustness to the schedule because fewer trucks operating means more trucks are resting, giving them a more flexible time for maintenance. The total number of required truck is the maximum number of trucks operating across districts in the same time of the same day.

Suppose some districts cleans trash on Monday(M), Wednesday(W), Friday(F) (plan A), while others cleans trash on Tuesday(Tu), Thursday(Th), Saturday(S) (plan B). Let A be the set of district indices that use plan A and B be the set of district indices that use plan B. A and B must be disjoint and any $n \in \{1, ..., 12\}$ is either in A or B.

Although by assumption 5, morning sessions and evening sessions are equally efficient, for other reasons, such as rat problem and traffic, some district might prefer morning sessions over evening sessions or vice versa. Suppose district n want m_n fraction of their trash pick up runs to tack place in the morning.

To sum up our new definitions and assumptions:

- $k = 3$ for all districts.
- We want to minimize the maximum number of trucks operating in the same time.
- Each district either use cleaning plan A or B. A and B is the set of district indices that use plan A or B, respectively.
- $N_{m,n} = \frac{1}{3}T_n m_n$ and $N_{e,n} = \frac{1}{3}T_n(1 - m_n)$

The total number of trucks operating on each session is:

- M W F mornings: $\sum_{n \in A} \frac{1}{3} T_n m_n$
- M W F evenings: $\sum_{n \in A} \frac{1}{3} T_n (1 - m_n)$
- Tu Th S mornings: $\sum_{n \in B} \frac{1}{3} T_n m_n$
- Tu Th S evenings: $\sum_{n \in B} \frac{1}{3} T_n (1 - m_n)$

Thus, the required trucks is

$$\max\left(\sum_{n \in A} \frac{1}{3} T_n m_n, \sum_{n \in A} \frac{1}{3} T_n (1 - m_n), \sum_{n \in B} \frac{1}{3} T_n m_n, \sum_{n \in B} \frac{1}{3} T_n (1 - m_n)\right) \tag{28}$$

We could enumerate all 2^{12} possibilities (whether district n is in A or not) and determine the one that minimizes the required number of trucks.

Algorithm 3 Schedule Determination

```
for every possible combination of planning do
    compute M = max(∑_{n∈A} (1/3)T_n m_n, ∑_{n∈A} (1/3)T_n(1 − m_n), ∑_{n∈B} (1/3)T_n m_n, ∑_{n∈B} (1/3)T_n(1 − m_n))
    record the combination that attains the minimum M
end for
return that combination
```

3.6 Summary

- By grouping variables and solving constraints, we simplified the model to the form

$$\text{minimize } \mathcal{C} = \sum_{n=1}^{12} T_n \tag{F.1}$$

$$T_{n,\text{lower}} \leqslant T_n \leqslant T_{n,\text{upper}} \tag{C.1}$$

$$\mathcal{Q} \leqslant \mathcal{Q}_{\text{bound}} \tag{C.2}$$

- We check whether there is a valid schedule with cost $\mathcal{C}$. The minimum cost that has a valid schedule is the $\mathcal{C}_{\min}$ we look for. We uses binary search to reduce the number of costs we need to check.

- We find valid schedule for a given $\mathcal{C}$ by hill climbing algorithm. We start from a good initial guess and minimize $\mathcal{Q}$ by walking towards the fastest descending direction.

- After we obtain the number of required trash pick up runs for each district T_n. We determine a detailed schedule by enumerating all possibilities and choose the optimal one based on the minimal total truck required.

4 Model Implementation and Result Analysis

4.1 City Data

District	Area (m^2)	Waste per Week ($tons$)	σ_p ($tons$)	σ_c ($tons$)	Population
MN01	3782141	692.5	14.8	0.870159	78,390
MN02	3504182	776.7	13.7	0.870159	92,445
MN03	4355469	1448.5	35.9	0.870159	163,141
MN04	4579317	1080.3	18.4	0.870159	131,351
MN05	4068815	560.5	19.7	0.870242	63,600
MN06	3594881	1155.5	16.6	0.870226	155,614
MN07	4938146	1855.7	34.8	0.870226	222,129
MN08	5113884	1985.4	33.6	0.870226	231,983
MN09	3891940	1026.6	22.1	0.870226	110,458
MN10	3630694	1199.45	42.84	0.870242	130,440
MN11	6145154	1254.9	14.9	0.870159	125,771
MN12	7243585	1703.6	69.1	0.870226	180,206

Table 2: District-wise Statistics on Area, Waste, and Population[3],[5]. σ_p is calculated from the standard deviation of four years of trash data. Calculations for σ_c are according to Equations 5 and 6 and can be found in Appendix A

4.2 Trade Offs

With population, area, and wastes data, we still need $\mathcal{Q}_{\text{bound}}$ and $\mathcal{Z}_{\text{bound}}$ to implement the algorithm. However, there is no clear standard or data to help determine the best value for these bounds. Therefore, we will pick the most reasonable $\mathcal{Q}_{\text{bound}}$ and $\mathcal{Z}_{\text{bound}}$ by looking at their trade off curves.

4.2.1 Equity and Efficiency

Recall in section 2.3, the total number of trash pick runs is an indicator of efficiency and in section 2.5.3 the standard deviation of $\frac{\delta_n}{P_n A_n}$ between districts is a measurement of equity. We expect a trade off between $\mathcal{Q}$ and $\mathcal{C}$ because increasing the number of trash pickup runs can guarantee the each district's expected trash residuals $\delta_n \to 0$, thus decreasing overall $\mathcal{Q}$. Using the Algorithm 2 detailed in section 3 that calculates the lowest possible Cost given an Equity upper bound, we can find a fundamental limit to where the maximum Equity a schedule can achieve given a total cost is. We can produce a trade-off curve by optimizing $\mathcal{C}$ for each highest constraint on $\mathcal{Q}$ at a fixed lowest constraint on Z.

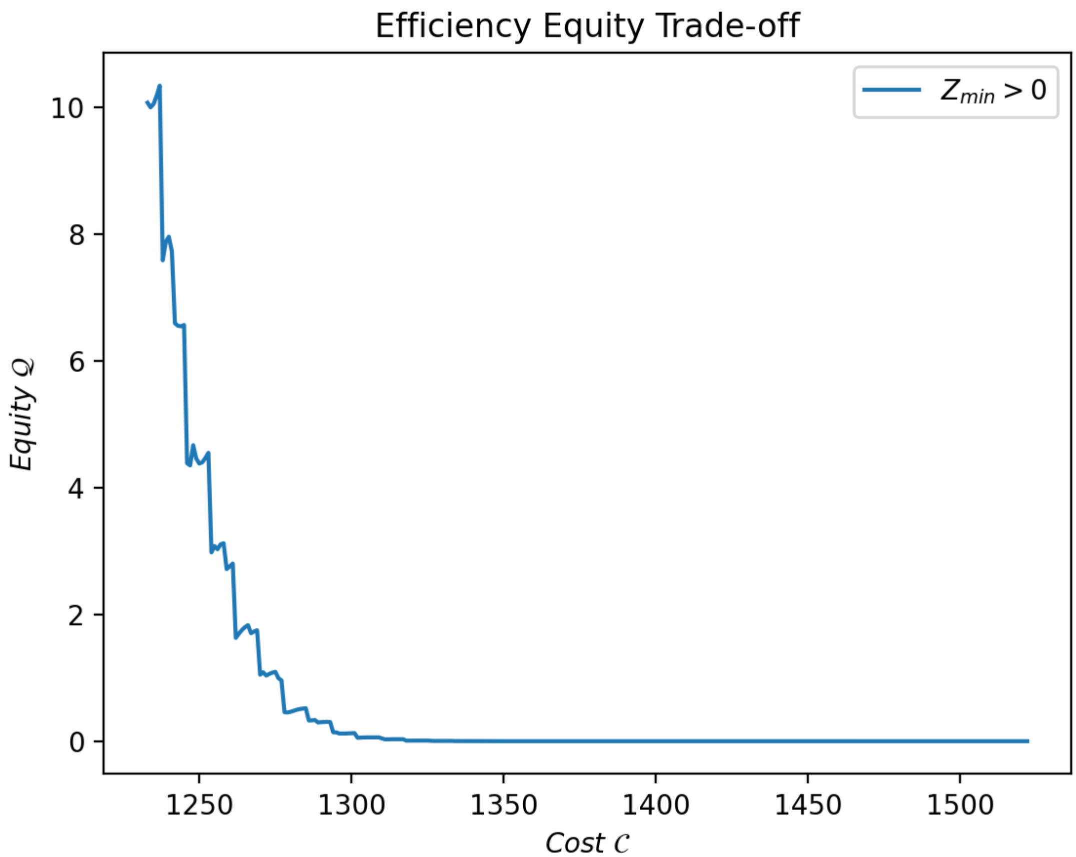

Figure 3: A trade off curve between $\mathcal{C}$ and $\mathcal{Q}$ with a constant $Z_{min} > 0$ constraint. Notice that the Z_{min} is the Z_{bound} from the previous section. Every point (C, Q) on the C-Q plane can represent a possible schedule with cost C and equity Q. The line draws a delimitation boundary for which the region below it is impossible to achieve given the restrictions we have put on the model. Here, Q has units of SI which are $\frac{\text{kg of waste}}{\text{km}^2\text{Person}}$ and C has units of truck pickup runs.

The figure clearly shows that the minimized $\mathcal{Q}$ of a schedule is inversely proportional to its Cost $\mathcal{C}$, as expected. Note that for this particular trade-off curve we have only graphed $\mathcal{C} \in [\mathcal{C}_{min}, \mathcal{C}_{max}] \approx [1240, \infty)$. While the upper bound has been arbitrary set because $\mathcal{Q} \to 0$ as $\mathcal{C} \to \infty$, the lower bound is set by constraint in section 2.6.4, which is related to Robustness. The following section will address the problem.

4.2.2 Efficiency and Robustness

Recall in section 2.5.4, Robustness is a measure of how likely a schedule is able to finish its cleaning quotas despite uncontrollable uncertainties. Similar to the relation between Efficiency and Equity, we expect a trade off relationship between Efficiency and Robustness. This is due to Robustness being proportional to the amount of waste cleaned per week, which is proportional to the Cost / Efficiency of a schedule. Thus, for a fixed C, there should be a fundamental limit to the highest Z it can achieve. Looking back at Figure 2, the lower bound of $\mathcal{C}$ is because there exist a certain $\mathcal{C}_{min}$ such that below it there will not enough waste collection capacity to satify the condition that $Z > 0$ for all districts

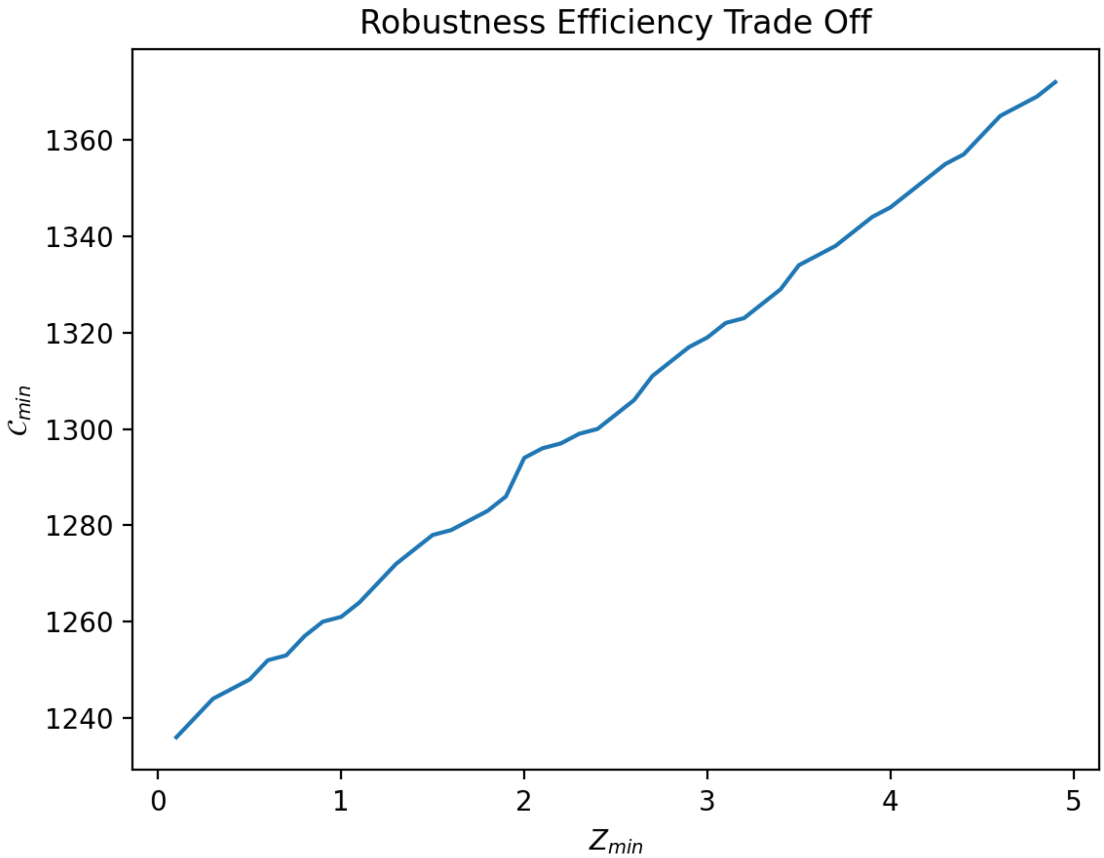

Figure 4: A trade off curve between $\mathcal{C}$ and Z_n with no constraint on Q. Here, C has units of truck pickup runs and Z is dimensionless. Any region above the curve is an allowed region in C-Z parameter space and any region below the curve is the fobidden region. We can see that a minimum of $\mathcal{C} \sim 1240$ is needed for $Z = 0$ which means that the $E[W_{c,n}] = E[W_{p,n}]$ where DSNY is collecting the bare minimum amount of waste to satisfy constraint 2.6.4.

4.3 Optimizing Efficiency and Equity

Given the trade off relation between Equity and Cost, along with Robustness and Cost, it is possible to find a point on both of these curves that optimizes both cost and equity. Equity-index and Cost-index should be the schedule's primary concern.

In C-Q Parameter space, $\left|\frac{\partial Q}{\partial C}\right|$ can serve as an ratio of how a schedule values C compared to Q. Looking at Figure 4, there are two extreme regions that prioritize one index far over the other. Schedules that reside in the region on the trade off curve in C-Q parameter space where $\left|\frac{\partial Q}{\partial C}\right| >> 1$ prioritize C over Q, as opposed to schedules that reside on the trade off curve in C-Q parameter space where $\left|\frac{\partial Q}{\partial C}\right| << 1$ prioritize Q over C. A region where both Q and C are equally weighted can be estimated using the condition where

$$\frac{\partial Q}{\partial C} \sim mean\left(\frac{\partial Q}{\partial C}\right) \tag{29}$$

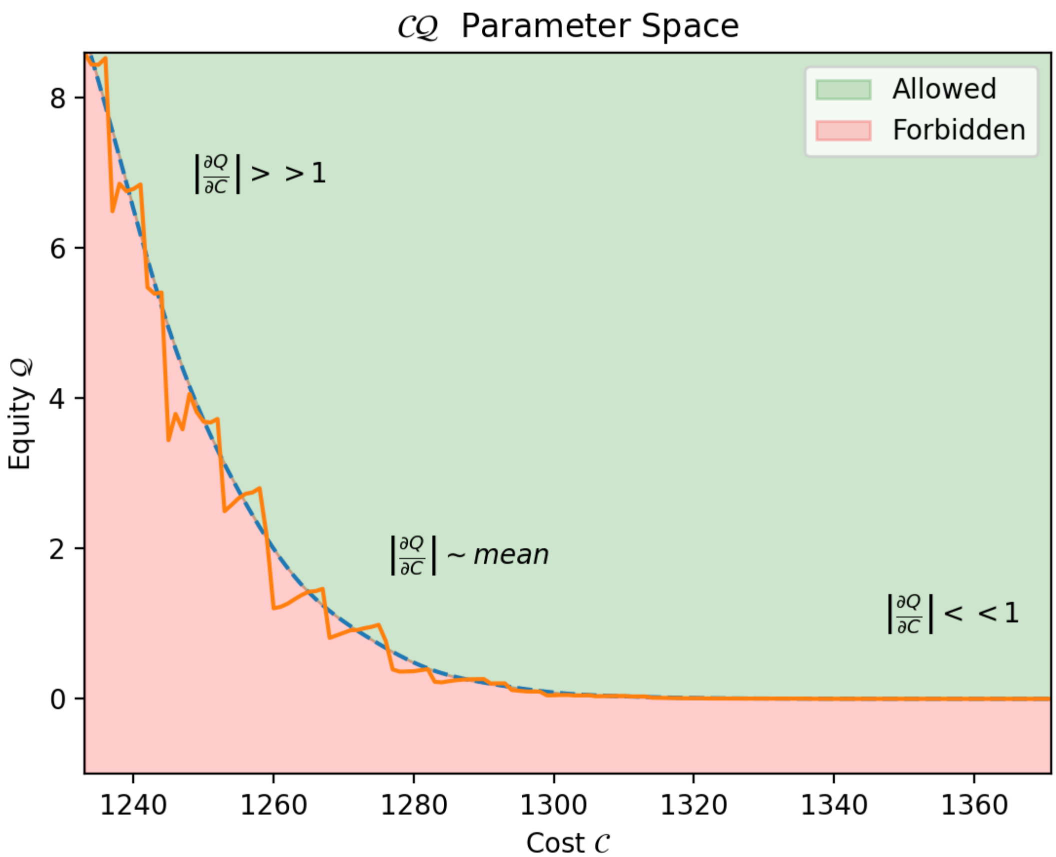

Figure 5: The parameter space of $\mathcal{C}$ and $\mathcal{Q}$ under constraint $Z_{min} > 0$. Dashed line is trade off curve smoothed out by convoluting with a Gaussian. Green is the allowed region of schedules according to our model and constraints and red is the forbidden region of schedules according to our model and constraints. Regions of interest are labeled according to previous paragraph. Here, Q has units of SI which are $\frac{\text{kg of waste}}{\text{km}^2\text{Person}}$ and C has units of truck pickup runs.

For current purposes, without an explicit preference for Equity or Efficiency, we shall only be considering schedules in the region satisfying Equation 29. Using computed results from Algorithm 1 and 2 we can numerically determine $\mathcal{C}$, $\mathcal{Q}$ of the schedule and T_n, Z_n for each district.

$$\mathcal{C} \sim 1276 \quad \mathcal{Q} \sim 0.771 \tag{30}$$

District	T_n	Z_n
MN01	60	1.713
MN02	67	1.805
MN03	126	1.715
MN04	93	1.791
MN05	49	1.339
MN06	99	1.768
MN07	159	1.443
MN08	170	1.554
MN09	89	1.775
MN10	106	1.662
MN11	106	1.002
MN12	152	1.725
Average	107	1.608

Table 3: T_n is the number of pickup runs for district n per week that the schedule must satisfy. Z_n is the Robustness index for district n based on this schedule. Here, the Z_{min} is capped by Z_{11} so for this model, $Z_{min} = 1.002$. Here, Z_{min} is the lowest Robustness index out of all districts. $\bar{Z}_n = 1.608$ which translates to 96.3% probability of collecting all waste.

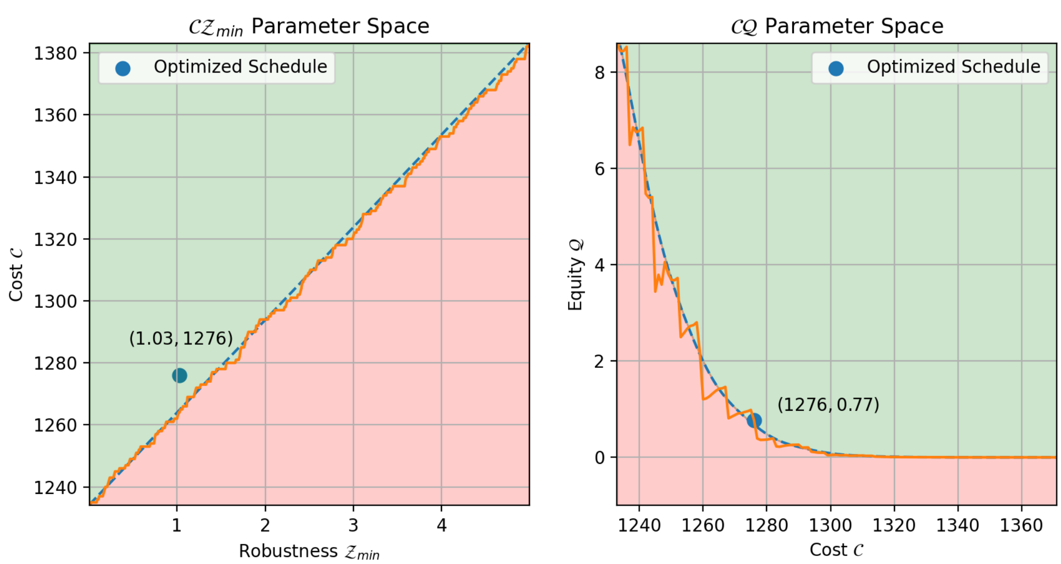

Figure 6: The optimal schedule plotted inside of $\mathcal{C}Z_{min}$ and $\mathcal{CQ}$. Units for the indexes are same as previous figures. Orange lines are trade off curves simulated by algorithm 1 and 2. The dashed line in CZ space is a linear fit to simulated data. The dashed line in CQ space in smoothed out trade off curve by convoluting a Gaussian. Note that while the schedule balances $\mathcal{C}$ and $\mathcal{Q}$ optimally, it is yet the most optimal in CZ space as it is not on the trade-off line

4.4 Final Optical Schedule

After using Algorithm 1 and 2 to derive the properties (See table above and Equation 30) of a schedule that balances Efficiency and Equity equally. We can then use these properties to come up with a specific schedule and the total amount of trucks required to staff this schedule. [4] The final schedule from derived is,

District	Schedule	Trucks per Cleaning Session / Trucks Allocated
MN01	M W F	10
MN02	M W F	12
MN03	T Th S	21
MN04	M W F	16
MN05	T Th S	9
MN06	T Th S	17
MN07	M W F	27
MN08	M W F	29
MN09	M W F	15
MN10	T Th S	18
MN11	T Th S	18
MN12	T Th S	26
Total (No Sharing)		**219**
Total (Sharing)		**109**

Table 4: This is the final optimal schedule calculated per Algorithm 3

The minimum amount of trucks required to implement the schedule is 219 if sharing trucks between districts are not allowed. The number can be drastically lowered to 109 if sharing between district are allowed. The total number of trucks allocated is the number of trucks per cleaning session if sharing trucks between districts is not allowed. This is because a district must be allocated at least the number of needed trucks per cleaning session to complete its duties. There is no column listing the number of trucks allocated per district if sharing is allowed because if trucks can freely move around then the act of allocating a truck to a single district is not well defined.

5 Rats Problem

5.1 Definitions of Variables

Before we start the Analysis of Correlations, we first state that under the section Rats Problem:

- R_n represents the population of rats in the district n,

[4]For specifics of how and what simplifications we made please see Section 3.5 and Algorithm 3. Schedule A in 3.5 is M W F here and Schedule B in 3.5 is T Th S here

- σ_n represents the standard deviation of produced waste in district n per week,
- δ_n represents the expected trash residual waste for district n,
- org_n represents the organic waste in district n,
- W_n represents waste produce in the district n.
- p_n represents resident population in the district n
- We will also introduce the superscript of j when discussing the Spatial Segmentation.

5.2 Assumption about Rat Population

1. **Closed Rat Population System.** The rat population in the sanitation districts is treated as a closed system, meaning there is no immigration of rats from outside areas into the 12 districts, nor emigration of rats from these districts to external areas. Additionally, this assumption implies that rats largely remain within their home ranges and do not frequently move across multiple districts.

   Reason: According to urban rodent ecology studies, the Norway rats, which are common in cities like Manhattan, have home ranges typically within 50 to 150 meters [4]. Also, Manhattan is surrounded by water, creating a natural barrier that limits movement in and out.

2. **The population of rats should directly related to the active rats signs.** Though we don't have specific data for rats population in each districts, there is the data about the "active rats signs" [6], and we could assume the population rats at each districts should be proportionate to this number.

   Reason: While there may be variations in reporting accuracy or detection rates within districts, the assumption is that these factors average out and do not significantly impact the overall correlation between active rat signs and actual rat populations at the district level.

3. **Rats Tend to Cluster Together.** Rats are more likely to cluster together within localized areas rather than dispersing widely across multiple districts. Therefore, we can group districts together, which we will implement in 5.3.2.

   Reason: As social animals, rats exhibit a natural tendency to stay close to one another, forming groups within their home ranges. For example, they live in large colonies with overlapping generations, establishing intricate burrow systems with shared channels and chambers. This social behavior reinforces the likelihood of rats clustering in specific areas. [7]

4. **The rats' food primarily comes from uncollected residual garbage.** The rats' food primarily comes from uncollected residual garbage. Although we recognize that rats may obtain food from other sources—such as sneaking into homes and scavenging food—our focus remains on the significant role of uncollected garbage on curbside as their main food source.

Reason: While rats may indeed obtain food from occasional sources like entry into buildings, these are generally unstable and limited. Additionally, the food from restaurant waste or litter on the curbside provides a more accessible and reliable food supply, making it a primary factor in sustaining rat populations.

5.3 Analysis of Correlations

Our goal is to identify the variable most strongly associated with the percentage of active rat signs. However, the calculated correlation values show that none of the examined factors showed a strong correlation with rat activity (Figure 7).

More interestingly, when we analyzed the correlation of each variable with the percentage of active rat signs across different years, we observed a consistent decline in these correlations over time. This trend suggests that rat population dynamics may be influenced by underlying patterns or factors not fully captured by the individual variables we studied. Initially, we considered using Factor Analysis to identify these potential latent factors. However, due to a lower-than-desired Kaiser–Meyer–Olkin (KMO) test score of 0.5s, we instead applied Principal Component Analysis (PCA).

To address these findings, we employed a two-step analytical approach combining Principal Component Analysis and Spatial Segmentation. This approach enabled us to capture the main sources of variance in the data and identify localized patterns across districts, providing a clearer understanding of the dominant trends in waste, population, and rat activity.

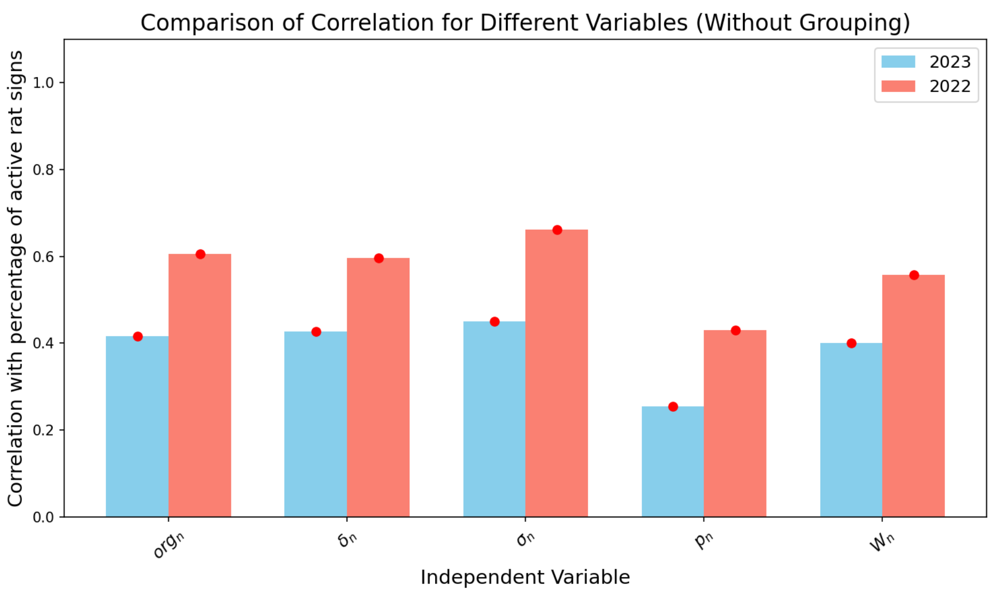

Figure 7: Comparison of yearly correlations between independent variables and failed rat inspections (indicating active rat presence) for 2022 (red) and 2023 (blue), without any district grouping.

5.3.1 Principal Component Analysis

As mentioned, given the limitations of individual correlations and the potential for underlying factors affecting rat populations, we applied Principal Component Analysis (PCA) to better understand the primary dimensions of variation among our independent variables, which allows us to condense the dataset into a set of principal components, each representing a combination of variables that capture the major patterns in waste levels, population density, and related factors across districts. By reducing the dimensionality of the dataset, PCA helps us identify the primary sources of variance without assuming strong correlations among specific variables.

Using the elbow method, we determined the optimal number of components by examining the explained variance ratio for each principal component. The Figure 8 reveals a clear *elbow* at the second component, suggesting that the first two components capture the majority of the variance in the data. **Therefore, retaining two components allows us to capture around 90% of the total variance, providing a simplified yet comprehensive view of the dominant trends without including unnecessary components.**

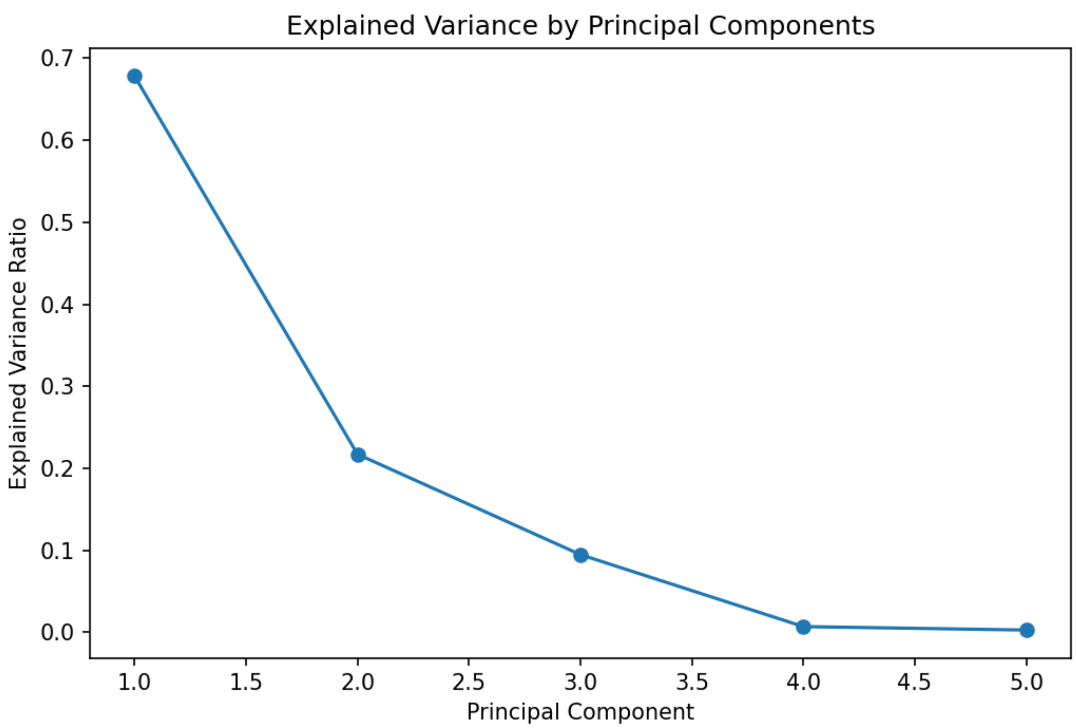

Figure 8: Explained Variance by Principal Components

The table 5 below displays the loadings of each variable on the first two principal components (PC1 and PC2) obtained from the PCA. These loadings indicate the contribution of each variable to the respective component:

- **PC1 (Overall Waste and Population Influence):** This component has high loadings for W_n (0.51), p_n (0.49), δ_n (0.46), and σ_n (0.46), suggesting it represents a general influence of waste levels and population density across districts. The high loadings on these variables indicate that PC1 captures patterns of W_n and the p_n factors, which are likely to impact rat activity at a broad level.

- **PC2 (Organic Waste and Waste Variability):** This component has a strong loading on org_n (0.69), while δ_n (-0.48) and σ_n (-0.43) have moderate negative loadings. PC2 appears to focus on org_n, suggesting that factors related to org_n might be strongly relevant in understanding rat activity.

Variable	**PC1**	**PC2**	**Interpretation (based on high loadings)**
org_n	0.29	0.69	PC2: Strong influence from organic waste
δ_n	0.46	-0.48	PC1/PC2: Waste variability across districts
σ_n	0.46	-0.43	PC1/PC2: Waste level variability
p_n	0.49	0.26	PC1: Population influence, with some influence on PC2
W_n	0.51	0.18	PC1: Overall waste levels

Table 5: Loadings for Each Variable on Principal Components and Interpretation

To explore these patterns further, we applied **segmentation analysis** based on spatial characteristics. By grouping districts according to these spatial factors, we aimed to uncover localized effects that may correlate with rat activity in ways that the broad PCA components might not fully capture.

5.3.2 Spatial Segmentation Analysis

The rationale for spatial segmentation is that rats do not recognize district boundaries and will move within neighboring districts. This movement aligns with our closed-system assumption, as there is still no any immigration or emigration across the entire set of districts. Nonetheless, rat populations in one district can influence neighboring districts, especially if a district has a high rat population, though this influence remains localized. For instance, while rats may move from District 12 to District 11, they are unlikely to travel from District 12 to District 5, thus preserving more localized interactions.

To account for this, We have re-grouped our districts as follows:

- Districts 1, 2, and 3 are grouped together
- Districts 4, 5, and 6 are grouped together
- Districts 7 and 8 are grouped together
- Districts 9, 10, 11, and 12 are grouped together

Afterwards, we introduce and use the superscript j to represent each areas. For example, δ^j will denote the $\sum_{n=1}^{3} \delta_n$.

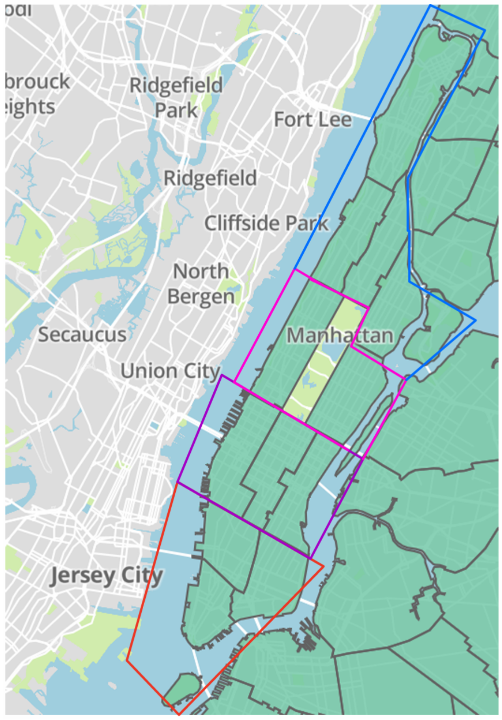

Figure 9: Illustration of district grouping, with each color representing a distinct group of districts.

From the new Comparison of Correlations for Different Variables Figure 10, we could find that the spatial segmentation approach has highlighted δ^j and σ^j as the most strongly correlated variables with active rat signs, suggesting they may be key factors influencing rat populations. The detailed interpretation are:

- **The Strongest Correlations for δ^j and σ^j:** The variables δ^j and σ^j show high and consistent correlations with active rat signs in both 2022 and 2023, indicating a strong association with rat activity across spatial groups. This suggests that variability in these factors may be significant drivers of rat activity.
- **Low Correlation for Organic Waste:** org^j has the lowest correlation with active rat signs among all the variables in both 2022 and 2023. This implies that org^j may not be a primary factor influencing rat populations within these spatial groups.
- **Consistency Over Time:** The similar correlation patterns between 2022 and 2023 suggest that the factors driving rat activity are relatively stable year over year. This consistency implies that the spatial segmentation may be capturing inherent regional characteristics that influence rat activity persistently over time.

The strong correlations of δ^j and σ^j with active rat signs are intuitive, as they directly relate to the availability and variability of food sources for rats. In urban environments, rats rely heavily on accessible waste for sustenance, and areas with higher levels of uncollected residue, δ^j, provide more food supplies. This encourages larger rat populations.

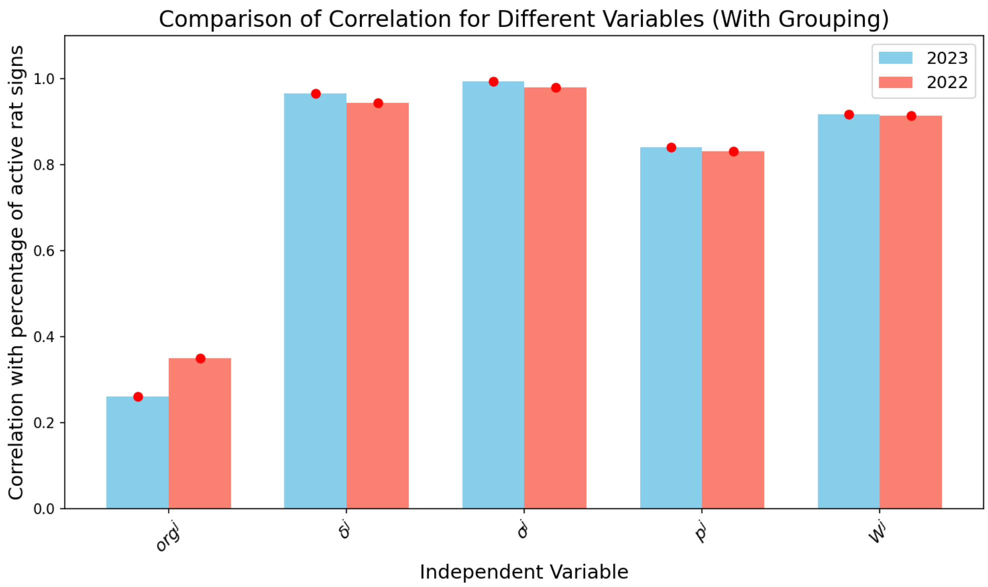

Figure 10: Comparison of yearly correlations between independent variables and failed rat inspections (indicating active rat presence) for 2022 (red) and 2023 (blue), after grouping adjacent/neighboring districts.

To summarize what we have done in the part Analysis of Correlations, the combined use of PCA and spatial segmentation has allowed us to gain both a broad and localized understanding of the factors influencing rat activity across districts. PCA identified general patterns in w_j and p_n dynamics, with p_n and W_n levels as primary contributors to the main variance in the dataset. However, spatial segmentation revealed that specific factors, namely δ^j and the σ^j, are more strongly correlated with active rat signs across grouped districts. This suggests that while p_n and W_n broadly shape the variance in the data, it is the δ^j and σ^j are likely to directly influence rat populations at a localized level. Together, these methods provide complementary insights, with PCA capturing the overarching variance structure and segmentation pinpointing critical localized drivers of rat activity.

5.3.3 Summary of Correlations Analysis

1. **Weak Individual Correlations**: We initially tried to find single factors strongly linked to rat activity but found no strong correlations, and these decreased over time.

2. **Possibility of Hidden Factors**: This suggested that unseen (latent) factors might be influencing rat populations.

3. **Attempted Factor Analysis**:

- Considered using **Factor Analysis** to identify underlying factors.
- Couldn't proceed due to low Kaiser–Meyer–Olkin (KMO) test score, indicating unsuitable data conditions.

4. **Switched to Principal Component Analysis (PCA)**:
   - Chose **PCA** as it simplifies data by highlighting main components accounting for most variation.
   - PCA doesn't require strict data conditions.

5. **Results from PCA**:
   - Found that two principal components captured most of the data's variance.
   - Simplified our analysis while retaining essential information.

6. **Spatial Segmentation**:
   - Grouped neighboring districts to account for rat movement across adjacent areas.
   - Recognized that rats move freely within local regions but not across distant districts.

7. **Findings from Spatial Segmentation**:
   - Identified stronger correlations between certain variables and rat activity within these groups.
   - Highlighted localized patterns that broad analysis might miss.

8. **Combined Approach for Better Understanding**:
   - Used PCA for overall trends.
   - Applied spatial segmentation for local patterns.
   - This combination provided clearer insights into key factors influencing rat populations.

5.4 Effect of New Strategy on Rats Population Modeling

We incorporate findings from our **Analysis of Correlations**. Our Spatial Segmentation analysis identified that both the expected trash residual, δ^j, and variability in waste collected, σ^j, are strongly correlated with active rat signs in the districts j.

5.4.1 Rat Population

The growth of the rat population in districts j over time t is modeled by the logistic equation, which reflects the population's natural growth while accounting for environmental limits.

$$\frac{dR^j}{dt} = r^j \cdot R^j \left(1 - \frac{R^j}{K^j}\right) \tag{31}$$

where:

- $R^j(t)$ is the rat population at time t in district n.
- r^j is the intrinsic growth rate in districts j.
- K^j is the carrying capacity in districts j, which depends on the δ_n and σ_n.

5.4.2 The Carrying Capacity and Intrinsic Growth Rate

Based on the conclusion from our analysis of correlations, we could reasonably assume that the the carrying capacity K^j for rats is positively influenced by δ^j and σ^j. To capture the diminishing returns effect—where increasing δ^j or σ^j has a reduced incremental impact—we use a logarithmic model:

$$K^j = \alpha \cdot \log(\delta^j + 1) \tag{32}$$

where:

- α is the sensitivity coefficients.

Recall our trash strategy mentioned above where we minimize the cost function $\mathcal{C}(S)$, while keeping our constraints—the Sanitation Distribution $\mathcal{S}_n(S)$ and Equity Index $\mathcal{Q}_n(S)$ within specified thresholds. Notice that our strategy will make the δ_n smaller than the original, which indicates the δ^j_{new} also smaller than δ^j_{old}. However, the σ_n, the standard deviation of waste produced by district n per week, is independent of the strategy, so $\sigma^j_{new} = \sigma^j_{old}$. Therefore, our strategy will only update carrying capacity, and can be represented as:

$$K^j_{new} = \alpha \cdot \log(\delta^j_{new} + 1) \tag{33}$$

Since we mentioned $\delta^j_{new} < \delta^j_{old}$, we conclude that $K_n^{\text{new}} < K^{\text{old}}$, reflecting a reduction in the environmental capacity for sustaining rat populations under improved conditions. Therefore, our strategy will solve the rats problem by reducing the upper bound of rats population in the long-term.

6 Strength, Weakness and Future Improvements

6.1 Strength

- **Valid:** According the waste production data, Manhattan produced about 17% of all waste in New York City. Suppose the number of trash truck assigned to Manhattan is also 17% of all trash trucks (2230 in total). Manhattan requires about 380 trash trucks, which is close to the number we obtained (109 to 219). Therefore, our prediction is within a reasonable bound. Moreover, our proposed model for measuring leftover trash strongly correlations with the population for rats after spatial segmentation. This also validates our model. [1][8]

- **Adjustable Robustness:** We have accounted for the uncertainties in the both waste production and waste collection, so that the schedule ensures that probability of fail to collect all the trash is lower than a given point. The user of the model can adjust the level of robustness of the model by setting Z_{bound} to a desired value. On the other hand, the user can easily take multiple source of uncertainty, such as weather, traffic, truck break down into account by computing the deviation for waste collection $\sigma_{c,n}$.

- **Simple Application:** The amount of waste productions and waste collection could be determined by multiple factors. However, we simplified all the uncertainties into two distributions, making the model easy to implement.

- **Fast Algorithm:** The schedule of the model is over 32 dimensional. Enumerating all the possible schedules is not feasible. However, we reduced our algorithm to a complexity of $\mathcal{O}(N \log N)$. A fast algorithm gives us more insights to the model because we can adjust the constraints to see the trade offs, relations, and dynamics between each indices. This enables us to set a more reasonable and carefully considered standard.

- **Spatial Segmentation for Localized Insights:** Grouping districts improves accuracy by recognizing rats' localized movement, offering targeted insights into rat activity patterns within segmented areas.

6.2 Weakness and Future Improvements

- **Unused Variables:** Although we have account some variables, we made them simple due to time constraints. These variables includes: η the efficiency of each trash pick up runs, m_n the preferred schedule for morning and evening, k_n the number of cleaning days in each district.

  Improvements: We can set non-trivial values for these variables.

- **Ignored Detail of Districts:** We did not consider the details of each district. We ignore the waste production, population, traffic difference between blocks.

  Improvements: We can factor the city into smaller blocks and the keep the same algorithm as for the districts. The new method should model reality in greater detail.

- **Ignored time duration for leftover trash** We only considered the amount of leftover waste but fail the consider the amount of times they are left over, which hugely impact public health and rat population. Additionally, summer poses a greater risk than winter due to higher temperatures, which elevate health hazards and potentially intensify rat activity.

  Improvements: We can consider the difference in contribution to sanitation level for morning sessions and evening sessions. Or we can compute the expected time of trash leftover before they are collected and use that as our new indicator for sanitation.

7 Conclusion

In conclusion, we developed a model to quantize the efficiency, equity, and robustness of a given schedule. Using this quantitative model to optimize efficiency under different constrains, we discovered a fundamental trade-off between efficiency, equity, and cost. Using trade-off curves we found reasonable amount of equity and robustness. Together with data from the city we found an efficient schedule for truck allocation in city cleaning sessions (Section 4.3). By sharing trucks across districts, we demonstrated a potential 50% reduction in the required fleet.

Our model incorporates a quantitative measure of efficiency and fairness, examining their trade-offs and enhancing robustness by treating waste production and collection as distributions, with adjustable robustness through the Z_{bound} parameter. (Section 2)

Additionally, we identified a strong correlation between rat populations and both leftover trash and waste production variability, showing that reducing leftover trash through our optimized schedule can effectively decrease rat populations. (Section 5)

8 Executive Summary

Dear Head of Waste Management Division,

We are honored to present an innovative solution for optimizing Manhattan's trash collection schedule. Our model is designed to address the critical aspects of robustness, equity, and cost-effectiveness in managing waste collection across the city. Our model computes that only 109 Trucks are needed to implement the schedule derived. This is a significant reduction compared to the currently an estimated number of 300 refusal collection trucks are active throughout Manhattan. [5]

Our model anticipates real-world uncertainties, such as fluctuations in waste volume, weather impacts, and potential truck breakdowns, ensuring reliable collection with minimal leftover trash on the curbs. Given New York City's scale and the challenges posed by increasingly unpredictable climate patterns, we believe this model's robust framework will be particularly valuable. Our model's recommended schedule has an estimated average probability of 94.52% success rate in collecting all trash bags in Manhattan.

The model also minimizes costs while promoting equity across districts. The model can balancing these priorities through flexible parameter settings. Even if our current plan does not fully align with all logistical requirements, we believe our model offers a valuable tool for refining and enhancing schedule decisions.

While our model offers robust solutions for optimizing collection schedules by accounting for key uncertainties, we want to acknowledge its current limitations. Our model does not yet fully incorporate variations between districts, differences in waste production, population density, or the nuances between morning and evening cleaning sessions. However, with minor adjustments, these factors can be readily integrated to enhance the model's accuracy.

We have also studied the relationship between trash and the rat population, identifying that variability in waste production and leftover trash are significant contributors to rat proliferation. By keeping leftover trash within a certain limit, our model helps curb conditions favorable to rats, supporting efforts to manage this issue in the city. We believe that, coupled with stronger trash regulation enforcement, our model can play a significant role in addressing both waste and rat-related challenges.

Thank you for considering this innovative approach to waste management in Manhattan. We look forward in contributing to a cleaner and more efficient city.

Sincerely,
Andy, Bopeng, Lixing

[5] This estimation is from the ratio of waste produced in Manhattan over NYC applied to DSNY's fleet size

References

[1] Problem Description: Of rats and garbage men sanitation workers URL: https://e.math.cornell.edu/sites/mcm/old_problems/CMCM_2024_rats_and_garbage.pdf

[2] Best Practices for Solid Waste Management: A Guide for Decision-Makers in Developing Countries - Equity in Solid Waste Management by the United States Environmental Protection Agency Office of Resource Conservation and Recovery. URL: https://www.epa.gov/system/files/documents/2023-07/SWM_Equity-Final.pdf

[3] New York City Department of City Planning. NYC Population FactFinder. Accessed November 10, 2024. URL: https://popfactfinder.planning.nyc.gov/#10.68/40.7479/-74

[4] Urban Norway rats (Rattus norvegicus) have limited 50–150 m home ranges, influenced by local factors, supporting closed-system modeling. URL: https://www.frontiersin.org/journals/ecology-and-evolution/articles/10.3389/fevo.2019.00013/full#B48

[5] Annual Report: New York City Curbside and Containerized Municipal Refuse and Recycling Statistics. URL: https://www.nyc.gov/assets/dsny/downloads/resources/statistics/annual-curbside-collection/dsny-collections-annual-fy2023.pdf

[6] Annual Report: Rat inspections in New York City Curbside. URL: https://a816-dohbesp.nyc.gov/IndicatorPublic/data-explorer/mice-and-rats/?id=2434#display=map

[7] The social life of Norway rats (Rattus norvegicus). URL: https://elifesciences.org/articles/54020

[8] and https://a860-gpp.nyc.gov/concern/nyc_government_publications/vh53wz49j?locale=en

[9] List of Civil Unrest Events https://en.wikipedia.org/wiki/List_of_incidents_of_civil_unrest_in_New_York_City

[10] NYC flooding probability per district https://dcp.maps.arcgis.com/apps/webappviewer/index.html?id=1c37d271fba14163bbb520517153d6d5

[11] 2019 DSNY Financial Report https://council.nyc.gov/budget/wp-content/uploads/sites/54/2024/03/827-DSNY.pdf#:~:text=The%20Fiscal%202025%20Preliminary%20Financial%20Plan%20%28Preliminary%20Plan%29,an%20annual%20average%20expenditure%20growth%20rate%20of%204.5%25.

A Calculating σ_c

District	P_{Truck}	$P_{Weather}$	$P_{NoCivilUnrest}$	σ_c
MN01	0.00025	0.99980769	0.005	0.87015967
MN02	0.00025	0.99980769	0.005	0.87015967
MN03	0.00025	0.99980769,	0.005	0.87015967
MN04	0.00025	1.	0.005	0.87015967
MN05	0.00025	0.99996154	0.005	0.87024291,
MN06	0.00025	0.99996154	0.005	0.87022626
MN07	0.00025	0.99996154	0.005	0.87022626
MN08	0.00025	0.99996154	0.005	0.87022626
MN09	0.00025	0.99996154	0.005	0.87022626
MN10	0.00025	1.	0.005	0.87024291,
MN11	0.00025	0.99980769	0.005	0.87015967,
MN12	0.00025	0.99996154	0.005	0.87022626

Table 6: Calculate σ_c for each district. The probabilities are calculated to be per week

P_{truck} is estimated using the number of vehicle maintenance contracts ordered by DSNY in 2019. Assuming each contract is for repairing a single truck, we divide number of contracts by total number of trucks DSNY has. $P_{Weather}$ is calculated by looking at annual flood probabilities. Only floods are factored because floods are the main weather hazard for those living in NYC. $P_{NoCivilUnrest}$ is estimated by counting diving the number of days that happen to have major civil unrest over the last 4 years by the total number of days.[9],[10], [11]

B Code

https://github.com/lllx125/CMCM