[94867] Project Data and Model Plan

Members: Jamie Lim, Lisa Yeung, Dai Ling Wu, Shun Tomita

I. Data Sources

Dataset Name	Tentative use of the dataset	Sources
311 Calls Services	We would like to use this dataset as our major source of data. This dataset contains a row record of incidents received by the 311 hotline by the city of San Francisco. Within the dataset, we are interested in using the category of 'Street and Sidewalk Cleaning', which entails all the incidents regarding this category. We will be using a sub-category called 'Human and Animal Waste' to proxy the demand of public toilets. To proxy homeless populations, we would use the category of 'Encampments', which entails all of the incidents regarding complaints on places with one or more tents, vehicles, or structures. These incidents are a close proxy of homeless populations as they usually reside in these forms of shelters. As the dataset is provided by the city government, it is overall pretty clean and is ready to use. The key fields that we need such as longitude and latitude have no missing values.	https://data.sfgov.org/City-Infrastructure/311-Cases/vw6y-z8j6
Google Map API	We would like to use Google Map API to retrieve data on already available toilets to model the demand of additional toilets. In Google Map API, we can query with keywords and location to get a list of toilets in that location. For example, we can enter the keyword 'Starbucks' near 'Union Square' to fetch all Starbucks(which offer toilets for the public) near Union Square. By iterating through all the keyword and location pairs, we will have a list of toilets for the public and its coordinates. We will use this to approximate the degree of lack of available toilets.	https://developers.google.com/maps/documentation/places/web-service

II. Model Plan

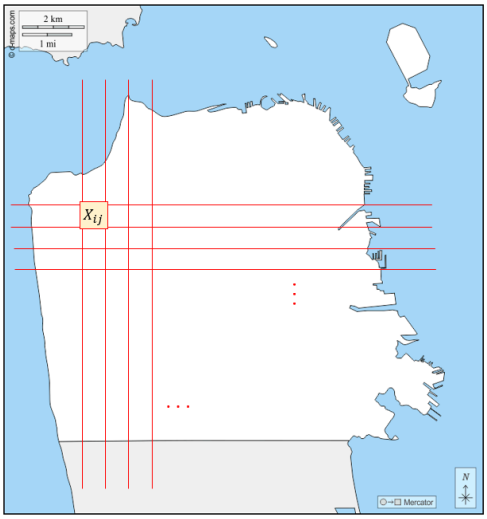
1. Model Selection

Our project aims to identify the optimal locations and the optimal number of public toilets to install in San Francisco. We chose to implement optimization techniques as we want to find a set of optimal values of decision variables under certain conditions.

Among many optimization techniques, we specifically selected the Integer Linear Programming (ILP) model because the number of toilets to be installed in an area has to be an integer (e.g. we cannot install 2.5 toilets) and the objective function and the constraints in our model are linear in nature which will be further illustrated in the following model formulation part.

2. Model Formulation

(A) Decision Variables



We are going to divide the San Francisco area into M by N matrix. Each cell would have a uniform distance across longitude and latitude and corresponds to an actual region in San Francisco.

The decision variables in our model would be Xij (i=1,2,3,..,M, j=1,2,3,...N) which refers to the number of public toilets to be installed in area ij. Therefore, our model would have M*N number of decision variables.

The number of decision variables may vary depending on how granular we want the model to be. In our preliminary model, we plan to optimize 900 (=30*30) decision variables.

(B) Parameters

The data sources we collected will be manipulated to generate three score matrices. First is *Uij* which refers to the score for uncleanness in area ij. This is measured by the number of human and animal waste related 311 calls in area ij. Second is *Lij* which refers to the score for access to public toilets in area ij. We will derive this score from the number of publicly accessible toilets in area ij and will subtract this component from the public utility function. Last is *Sij* which refers to the score for susceptibility to waste in area ij. This will be measured by the number of encampment related 311 calls which we think could be a proxy for the homeless population generating human waste. Overall, the higher the score, the higher the need for public toilets in that area.

(C) Objective Function

The ultimate objective of this project is
]to maximize public utility (denoted as *PU*) by supplying public toilets in areas where they are most needed and thereby promoting public hygiene.

$$\max_{X_{ij}} \boldsymbol{PU} = 0.5\sum_{i=1}^{M}\sum_{j=1}^{N} X_{ij}U_{ij} - 0.2\sum_{i=1}^{M}\sum_{j=1}^{N} X_{ij}L_{ij} + 0.3\sum_{i=1}^{M}\sum_{j=1}^{N} X_{ij}S_{ij}$$

Public utility function consists of three components using the above three score matrices. When decision variable Xij is multiplied by a high Uij, Lij, or Sij score, more utility is generated in area ij and thus, increases the total public utility.

The weights are assigned according to the importance of each component in determining how many toilets should be installed in each area. We think that uncleanness plays the most important role, followed by susceptibility and lack of accessible toilets so we assign 0.5, 0.3, and 0.2, respectively.

(D) Constraints

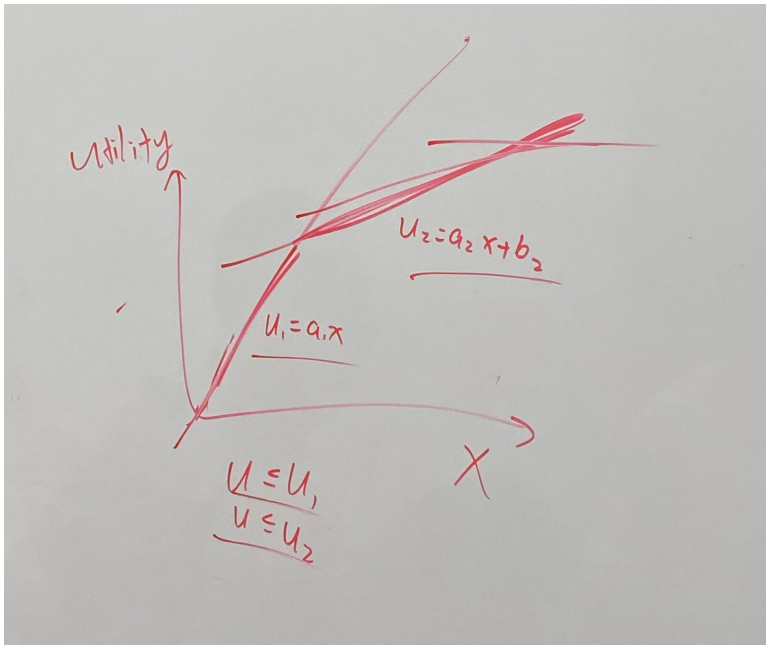
Budget	San Francisco has a budget for the Pit Stop Program of $8.6 Million with each toilet costing $200K a year to operate. $$\sum_{i,j} X_{ij} * \$200K \leq \$8.6M$$
Lower bound	We want to impose an upper limit (denoted by t) to how many toilets can be installed in area ij so that toilets are not heavily concentrated in a few areas. $$X_{ij} \leq t \quad for\ all \quad i,j$$
Upper bound	We also want to guarantee that at least some number of toilets (denoted by k) are installed in area ij when there is a certain level of need for public toilets (denoted by n) in that area. $$X_{ij} \geq k \ \ if \ \ 0.5 * X_{ij} * U_{ij} + 0.2 * X_{ij} * L_{ij} + 0.3 * X_{ij} * S_{ij} \geq n$$
Contiguity	When a public toilet is installed in area ij, some portion of the need for public toilets in nearby areas will be met by the toilet in area ij. We will group five contiguous areas together in determining optimal provision of public toilets and impose an upper limit (denoted by u) on the total number of public toilets that can be installed in that region. $$X_{ij} + X_{(i-1)j} + X_{(i+1)j} + X_{i(j-1)} + X_{i(j+1)} \leq u$$
Non-negativity	The number of public toilets to be installed in area ij should be greater than or equal to 0. $$X_{ij} \geq 0$$
Integer constraint	The number of public toilets to be installed in area ij should be integer.

	$X_{ij} \in \mathbb{Z}$

Questions:

1. Weights assigned to each matrix are arbitrary. Do we need to validate ? Or do we try different weights? Provided by clients.
    a. Clients wouldn't define it even in practice. You need to try multiple values and compare the results.

2. How do we try different sizes on each grid?
    a. 1 - 4 blocks
    b.
3. Linear relationship between number of toilets and public utility (using integer linear programming). Is it valid? Or can we use non-linear ? How ? How to define the relationship between the number of toilets and public utility?
    a. Piecewise linear constraints.



Suggestion from professor:

1. Incorporating maintenance issue would be great
2. Useful Reference: Gates Foundation Toilet