New York City's Bike System

Short description and reference. Simulation of the flow of bike trips between bike stations in NY City. The optimization problem is the allocation of a constrained number of bikes (600) to available docks within the city at the start of rush hour, so as to minimize, in simulation, the expected number of potential trips in which the rider could not find an available bike at their preferred origination station, or could not find an available dock at their preferred destination station. We call such trips "negatively affected trips." [1].

Problem description

- Deterministic inputs: We divided the bike stations in 4 groups using k-nearest neighbors, and the number of bikes in each group at 7:00 AM is the deterministic input.
- Randomness: We consider a directed graph between the bike stations, where each pair of bike stations has two directed edges, and we divided these edges in 4 groups. One random input is the number of bike trips in each of those groups. The other random inputs are the number of bike trips between each pair of bike stations, and the duration of the trips.
- Output: Number of negatively affected trips.

Information sources

- The expected number of negatively affected trips given the number of bike trips in each of four groups where each group was selected deterministically. So, we have one information source for each possible configuration of number of trips in the four groups.

Contact Saul Toscano-Palmerin, st684@cornell.edu, Cornell Peter Frazier group

Detailed Description We consider a queuing simulation based on New York City's Bike system, in which system users may remove an available bike from a station at one location within the city, and ride it to a station with an available dock in some other location within the city. The optimization problem that we consider is the allocation of a constrained number of bikes (600) to available docks within the city at the start of rush hour, so as to minimize, in simulation, the expected number of potential trips in which the rider could not find an available bike at their preferred origination station, or could not find an available dock at their preferred destination station. We call such trips "negatively affected trips."

We used 329 actual bike stations, locations, and numbers of docks from the Citi Bike system, and estimate demand for trips using publicly available data from Citi Bike's website [2].

We simulate the demand for trips between each pair of bike stations using an independent Poisson process, and trip times between pairs of stations follows an exponential distribution. If a potential trip's origination station has no available bikes, then that trip does not occur, and we increment our count of negatively affected trips. If a trip does occur, and its preferred destination station does not have an available dock, then we also increment our count of negatively affected trips, and the bike is returned to the closest bike station with available docks.

We divided the bike stations in 4 groups using k-nearest neighbors, and let x be the number of bikes in each group at 7:00 AM. We suppose that bikes are allocated uniformly among stations within a single group. Then we consider a directed graph between the bike stations, where each pair of bike stations has two directed edges, and we divided these edges in 4 groups. If the edge (i, j) is in a group, then (j,i) is also in that group. The randomness of the simulator is divided in two parts: w and z. The random vector w is the number of bike trips in each of those groups during the period of our simulation, and so w follows a multivariate Poisson distribution and its components are independent. The random vector z contains all other random quantities within our simulation.

We define $h(x, w, z)$ by the number of negatively affected trips occurred in the simulator, and the feasible set A is defined by $\left\{(x_1, \ldots, x_4) : \sum_{i=1}^{4} x_i = 600, x_i \in \mathbb{N}, 0 \leq x_i \leq 600 \text{ for } 1 \leq i \leq 4\right\}$. We enumerate the set $\mathbb{N}^4$, say $\mathbb{N}^4 = \{q_i\}_{i \geq 0}$. So, we are interested in solving the following problem

$$\begin{aligned}\min_{x\in A} g(x) = \min_{x\in A} \mathbb{E}_{w,z}\left[h(x,w,z)\right] &= \min_{x\in A} \mathbb{E}_w\left[\mathbb{E}_z\left[h(x,w,z) \mid w\right]\right] \\ &= \min_{x\in A} \sum_{i\geq 0} \mathbb{P}(w=q_i)\, \mathbb{E}_z\left[h(x,w,z) \mid w=q_i\right] \\ &= \min_{x\in A} \sum_{i\geq 0} p_i f_i(x)\end{aligned}$$

where $p_i = \mathbb{P}(w = q_i)$ and $f_i(x) = \mathbb{E}_z\left[f(x, w, z) \mid w = q_i\right]$ for $i \geq 0$. In other words, $\{f_i\}_{i\geq 0}$ are the information sources and our objective function g is the weighted sum of these sources.

In this problem, we can only get noisy observations of the information sources, specifically for given i and x we can only observe the empirical mean

$$y_i(x) = \frac{1}{M} \sum_{j=1}^{M} h(x, q_i, z_j) = f_i(x) + \epsilon_i$$

where M is a parameter fixed by the user, $z_j \sim p(z \mid w = q_i)$ where $p(z \mid w)$ is the conditional density of z given w, and by the Central Limit Theorem ϵ_i is normally distributed with mean zero. This means that we can only get noisy observations of g: $\sum_{i\geq 0} p_i y_i(x)$.

The following diagram illustrates the previous problem,

$$\begin{array}{cccccc} & f_0(x) & \rightarrow & y_0(x) = f_0(x) + \epsilon_1 \rightarrow & \\ & f_1(x) & \rightarrow & y_1(x) = f_1(x) + \epsilon_2 \rightarrow & \\ & \vdots & & \vdots & \\ x \rightarrow & f_n(x) & \rightarrow & y_n(x) = f_n(x) + \epsilon_n \rightarrow & z = \sum_{i\geq 0} p_i y_i(x) \\ & \vdots & & \vdots & \\ & \vdots & & \vdots & \end{array}$$

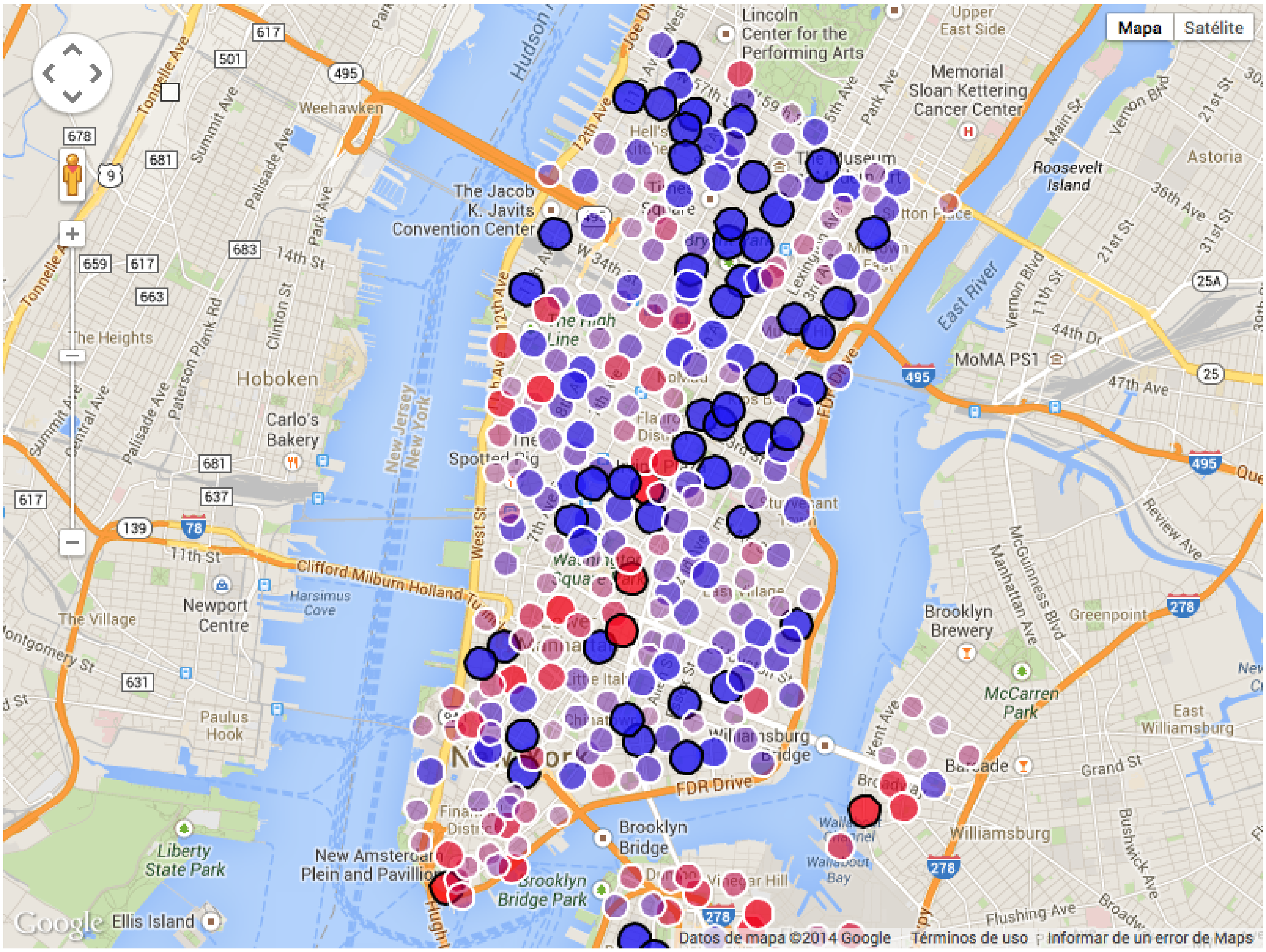

Figure 1: Location of bike stations (circles) in New York City, where size and color represent the ratio of available bikes to available docks.

References

[1] Saul Toscano-Palmerin and Peter Frazier. Stratified bayesian optimization. 2015. http://toscanosaul.github.io/saul/.

[2] Citi bike website. https://www.citibikenyc.com/, accessed May 2015.