[CS166] Final Project Report

April 22, 2020

Optimizing Ambulance Response Time Using Monte Carlo Simulation

A case study of London

Taha Bouhoun

Contents

1 Introduction

The efficiency of Emergency Medical Services is an essential indicator of the well-functioning of health systems. In this report, various strategies are compared in managing ambulance fleets to minimize their response time. Based on real-life data, I analyze then compare the findings of the simulation against the benchmark of EMS response time in London. Furthermore, I test the effect of varying the average speed of ambulances as well as the impact of the closure of the London Tower Bridge in the last quarter of 2016 on the average response time.

2 Methodology

2.1 Gathering data:

Compiling the road grid of London into a network is computationally expensive, so the approach was to coarse-grain the system to have regions (with a mean area of 1.6 km^2) as the building blocks. Using the open-source Uber Movement dataset[1], the city was constructed with roughly 1000 regions of Greater London made up with polygon shapes. Next, creating edges between two given regions relies on the number of coordinates that their polygons share (i.e., if a pair of regions shares at least one coordinates, then they're adjacent, and thus they're linked with an edge).

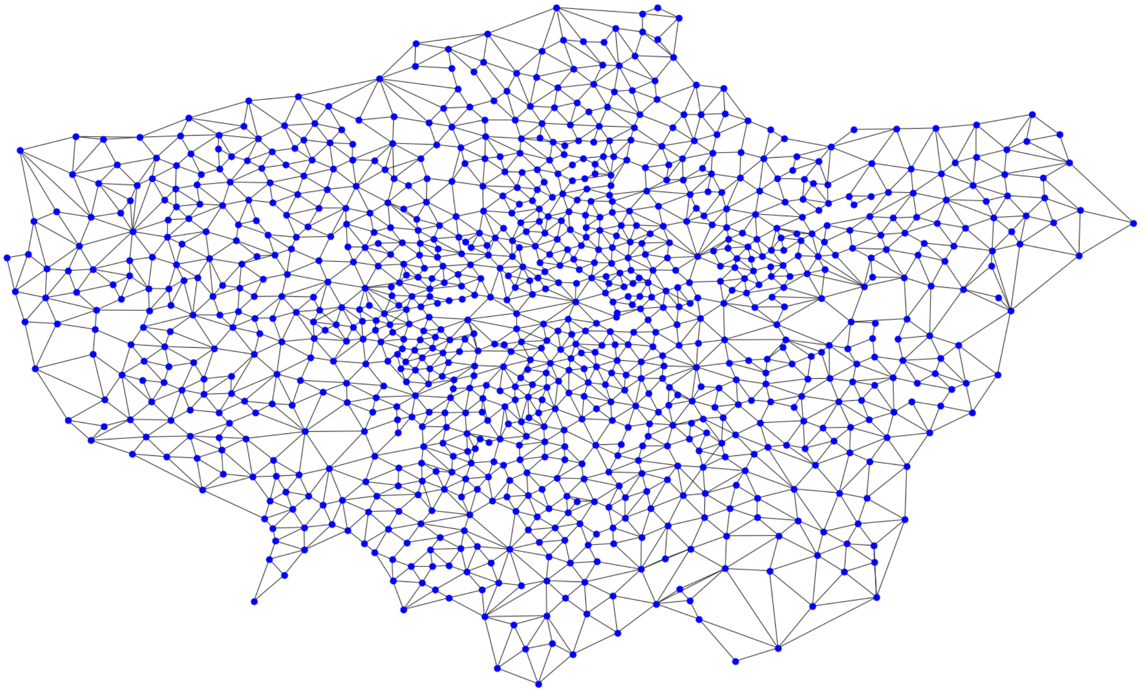


Fig 1. A network of Greater London generated
from the open source dataset of Uber Movement

[1] Uber Movement is an initiative to provide data and tools for cities to more deeply understand and address urban transportation challenges.

For a more focused analysis, a subsection of London's regions was selected containing 71 areas around the epicenter of the city (*see figure 2*). Besides, the dataset of Uber Movement populates the average traveling time between all pairs of regions which then served as edge-weights for the subsection of London network [2]

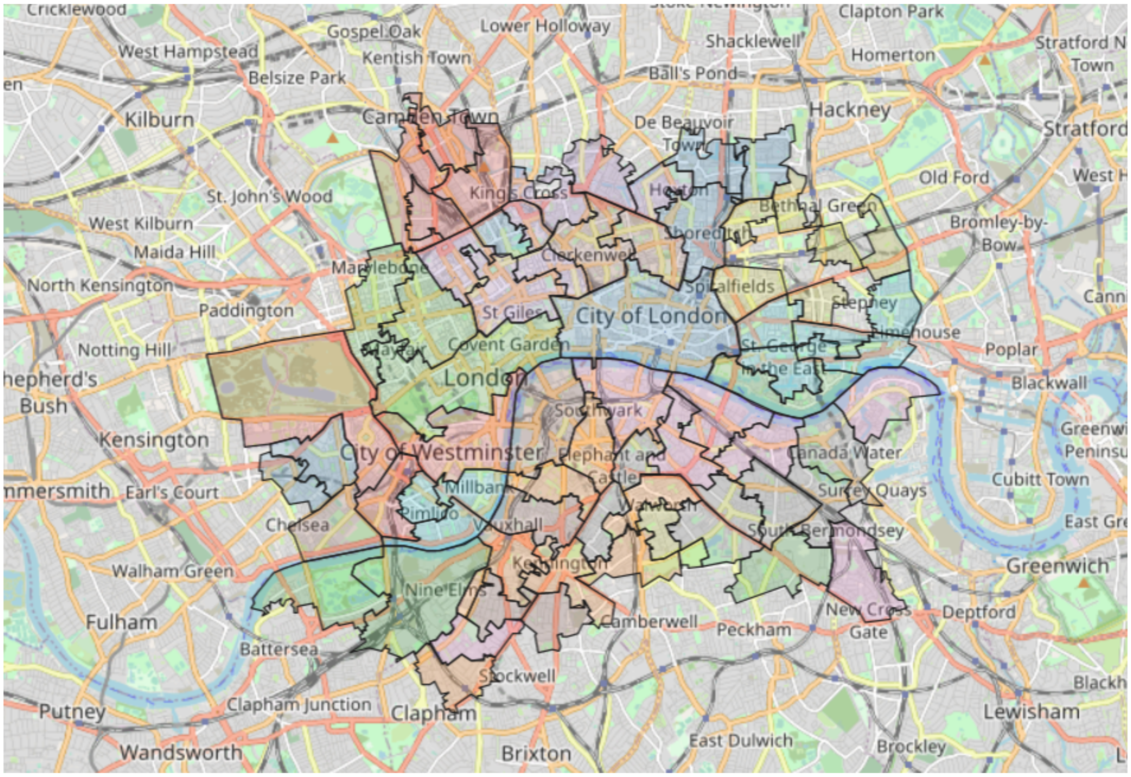


Fig 2. The regions of the selected subsection of London

The probability of requesting an ambulance differs across regions. Using the criminality rate published by the Metropolitan Police of London, each region was attributed to a crime rate metric indicated by the node color in the network (red for high crime rate and blue for low rates). The size of the nodes reflects the area of the region is km^2. Finally, hospitals in London were located using Google Maps API then associated with their respective region in London's subsection network. (i.e., if medical center coordinates fall within the polygon of a region then it's included into the region's node). [3] (*see figure 3*)

2.2 Assumptions

Although the network was constructed to approximate real settings, the following assumptions serve as a prior for interpreting the findings of the simulation:

- Ambulances are all coordinated with the same provider (i.e., all requests reach a central planner, then ambulances are dispatched given the proximity of the emergency request).

[2] The average traveling time between adjacent pairs of areas are based on data in the third quarter of 2016.

[3] Hospitals are colored in green on their real coordinate to show their location

- Crime rate is a holistic metric that pictures crimes that don't involve hospitalization. In other words, one can make an argument that touristic attractions inflate the crime rate because of pick-pocketing, which doesn't necessarily imply a high number of ambulance requests.
- The average travel time is based on Uber rides measurements, but ambulances can cut through traffic by skipping red lights or when cars move aside to make way for them. The simulation pictures this advantage by introducing `ambulance velocity` parameter.
- Hospitals are simulated to have an unconstrained number of ambulances.
- Emergencies are categorized by priority in real-life, but -for the sake of simplicity- all emergencies generated in the simulation have the same priority.

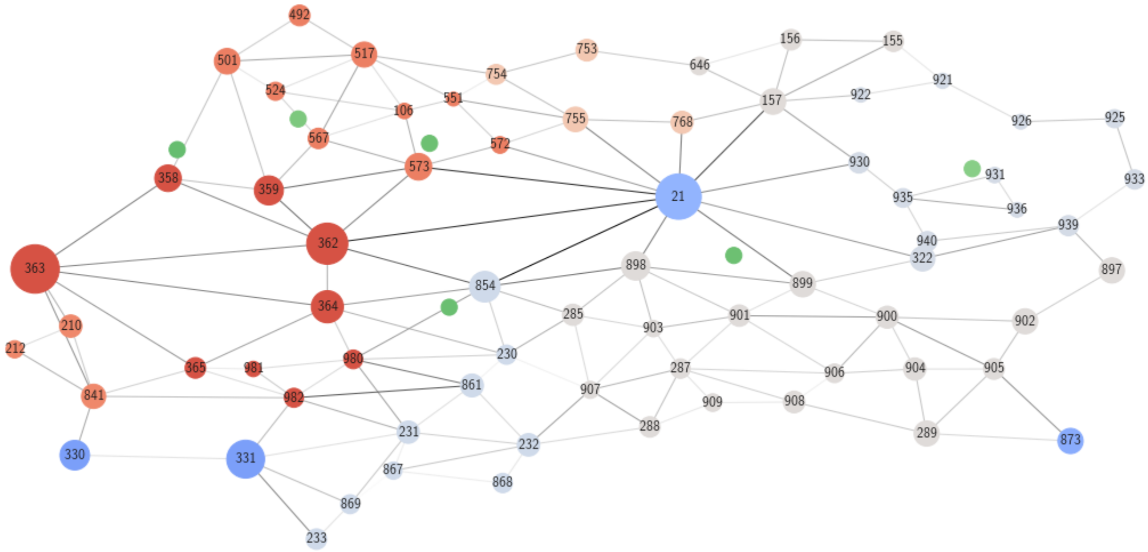


Fig 3. The network of the subsection of London

```
Type:  Graph
Number of nodes:  71
Number of edges:  156
Average degree:  4.3944
```

3 Modeling and Simulation

Given the underlying data structure, I designed an object-oriented program that takes the network as an input as well as other parameters (average speed of ambulances, hospital nodes, and the number of emergency requests). The approach is that all requests are generated at once, then the simulation loops through them to compute the average response time.

- `generate_requests`: The function generates emergency calls in different regions of the city using a random sampler weighted by the normalized criminality rate. (i.e., a region with relatively high criminality rate would have more frequent ambulance calls).

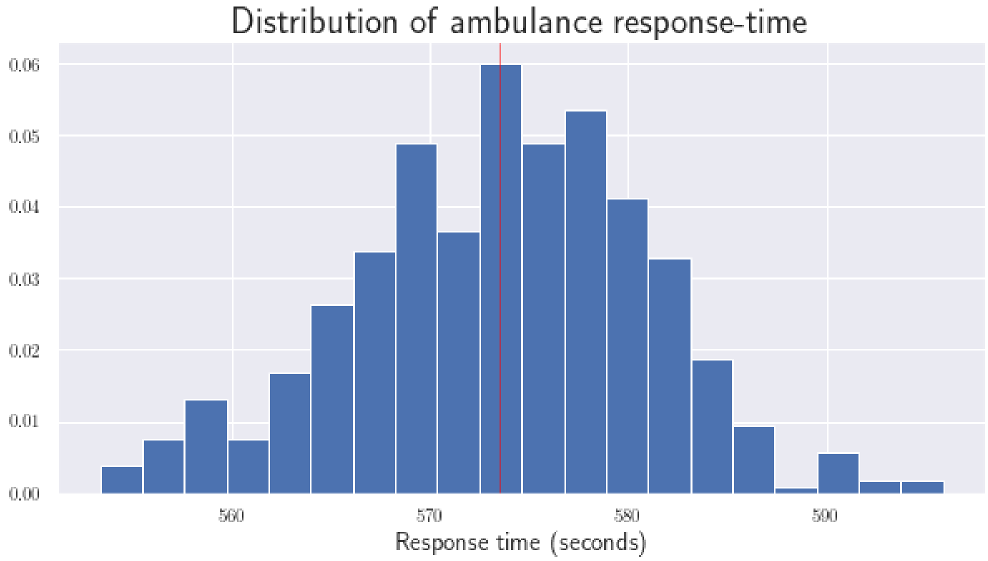
- `estimated_travel_time`: The function estimates the average travel time using the shortest path sequence between the hospital and the region from which the emergency call was generated. For each edge on the shortest path, a sample is taken from a normal distribution with mean set on the average travel time on the edge, and standard deviation of travel time on the edge. Finally, it returns the sum of edge samples.(*Note: both estimations of mean and standard deviation are extracted from Uber Movement data*)

- `shortest_path`: The function computes the shortest path between hospital nodes and the node from which the emergency call was generated following this process:

  - Edge: Using the Dijkstra method, the mean travel time of edges are used as weights to identify the fastest paths rather than the shorted paths (in terms of degrees of separation) to the emergency node. The function evaluates all the hospitals to determine the best option.
  - Dispatching: based on the radius of the region from which the ambulance is dispatched the estimated time of dispatching is calculated based on the average ambulance speed.
mean = time for the ambulance to travel the radius of the region
standard deviation = 5 seconds
  - Pick_up[4]: based on the region's radius from which the emergency call was generated
mean = time for the ambulance to travel the radius of the region
standard deviation = 10 seconds

4 Results and Analysis

The iterative nature of the simulation relies on two variables, the number of requests generated and the number of runs for each simulation. Using the travel time data from the third quarter of 2016, I controlled for the speed of ambulances at 20 Km/h then ran the simulation 500 times, generating 2000 requests in each iteration, which yielded the following histogram for the distribution of response time.



[4] Standard deviation is higher because the ambulance is less certain about the emergency call source.)

Node	Name	Count
845	St Thomas' Hospital	532
899	London Bridge Hospital	468
931	The Royal London Hospital	330
567	University College Hospital	282
358	King Edward VII's Hospital	263
573	Royal London Hospital for Integrated Medicine	125

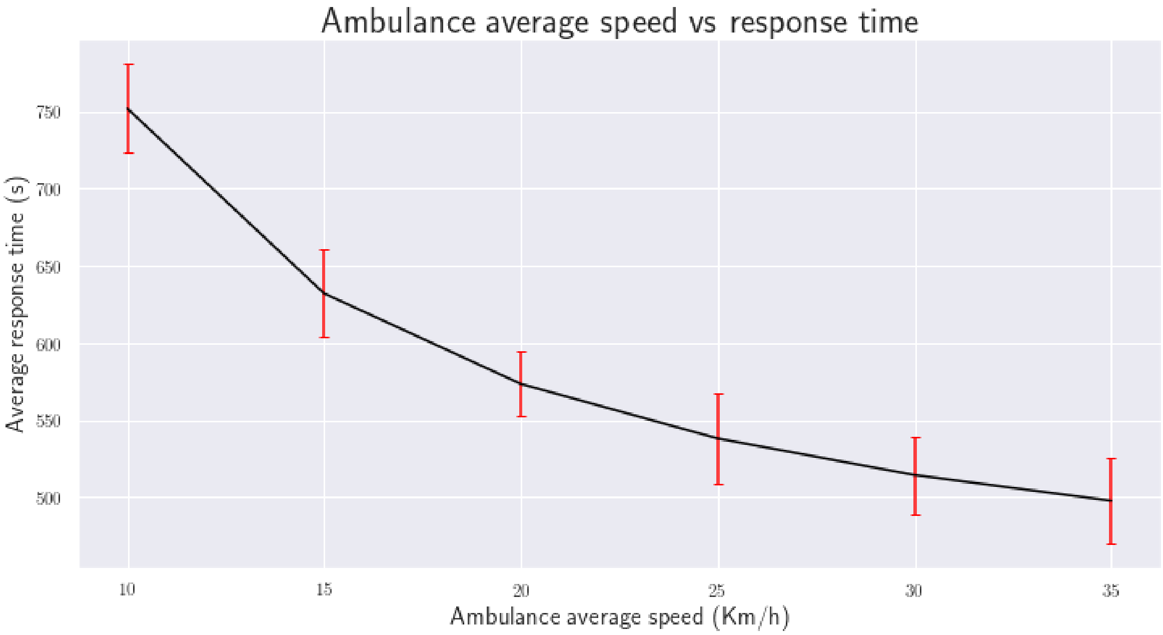
The distribution of response time is approximately normal with bounds between 550 to 600 seconds. Centered at 9 minutes and 33.52 seconds (573.52 s) with its 95% confidence interval of roughly 26 seconds size (±13 s)

5th percentile: 9 minutes and 19.34 seconds

95th percentile: 9 minutes and 44.11 seconds

4.1 Average speed of ambulances

Legally, ambulances in London are constrained with a speed limit of 20 mph (around 32 Km/h) [5]. To test the effect of varying speed, I generated 2000 requests for 100 simulations for each speed ranging from 10 to 35 Km/h

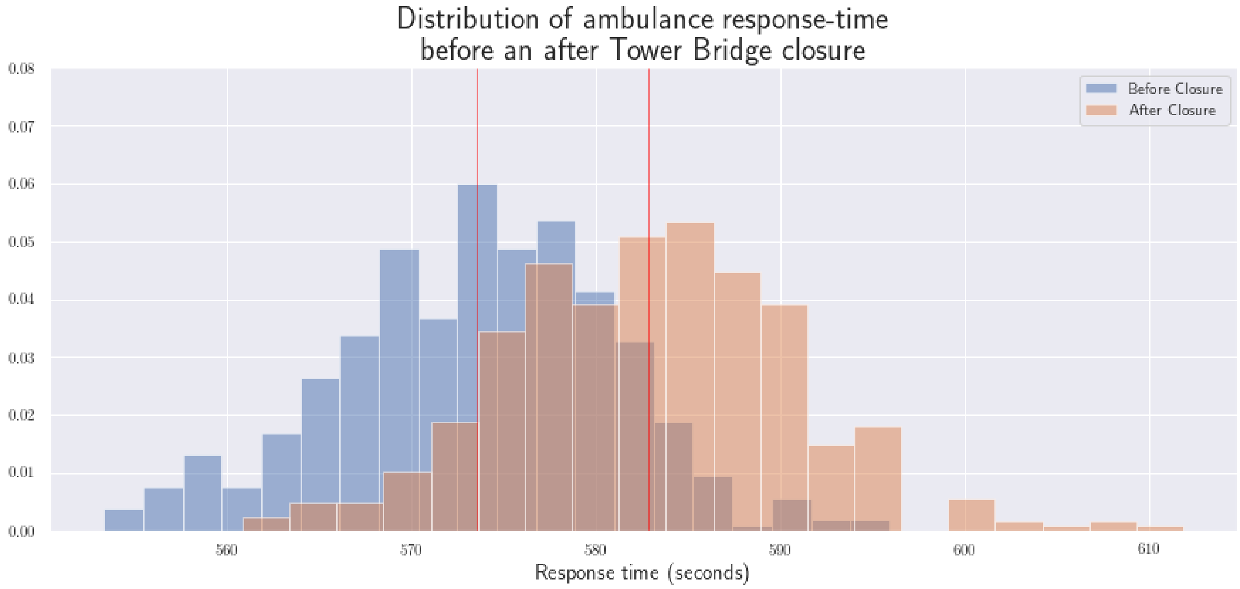


The decrease in average response time is gradual; hence, rising speed has a diminishing return. Starting with 10 Km/h, the response time is centered around 700 seconds, and then it decreases to around 570 seconds for an average speed of 20 Km/h to reach a maximum of 500 seconds at 35 Km/h speed. Notice that the standard deviation of these measurements are roughly the same across different speeds (±25s)

[5] Source: Policy Institute at King's, LAS Incident Response data 2016 https://www.kcl.ac.uk/policy-institute/assets/data-for-ambulance-dispatch.pdf

4.2 The closure of London Tower Bridge

One of the compelling reasons for choosing London was to test the effect of the change in infrastructure on the response time of ambulance. During the last quarter of 2016, the London Tower Bridge was coed due to construction and reopened on December 30th, 2016. This serves as a great natural experience to disentangle the causal effect on response time. The comparison was drawn between two networks (the first carrying travel times of the 3rd quarter of 2016, the second contains the data from the 4th quarter).



Both distributions approximately normal but centered 10 seconds apart as the response time across the network was 10 seconds faster before the closure of the bridge. The confidence intervals for both distributions have the same range ($\pm 13s$)

4.3 New ambulance base

Initially, the network simulation has a set of six hospitals scattered around the network of London. A heuristic for adding a new ambulance base was first to sort the nodes with the highest response time. Figure 4 illustrates a network colored based on the average response time for all the requests generated from a given node. We notice that the northeast regions have high response time compared to the rest of the network (Specifically, regions 155, 156, 921, 922, 646)

Region	Average response time	Distance from the mean
156	1125.10	96.17%
155	1123.67	95.92%
921	1111.25	93.75%
922	1099.49	91.7%
646	978.05	72.1%

Table 1: Top 5 regions with the highest response time

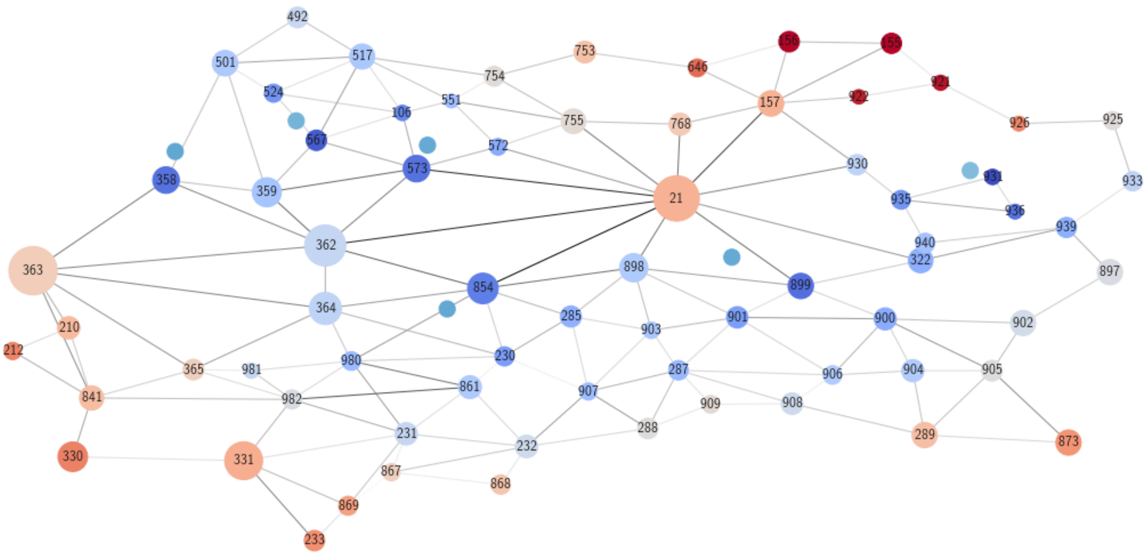
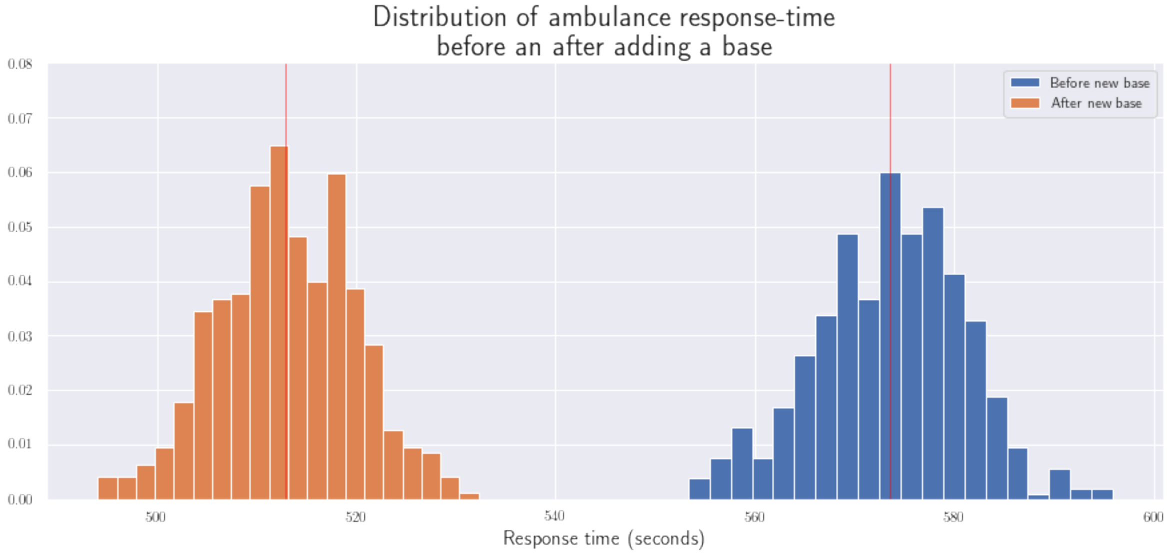


Fig 4. The average response time for each region in the network

Since the traveling time between the five regions is roughly the same, we can pick a node to assign it as a hospital node (say region 155 contains an ambulance base). By doing so, we expect to push down the average response time for the whole network.



The two distributions are distinct (not overlapping) as the addition of a base reduced response time by 60 seconds (from 573.52 to 512.96 s).

To give a perspective of the comparison, adding one base in region 155 had the same impact of changing the average speed of ambulances from 20 to 30 km/h

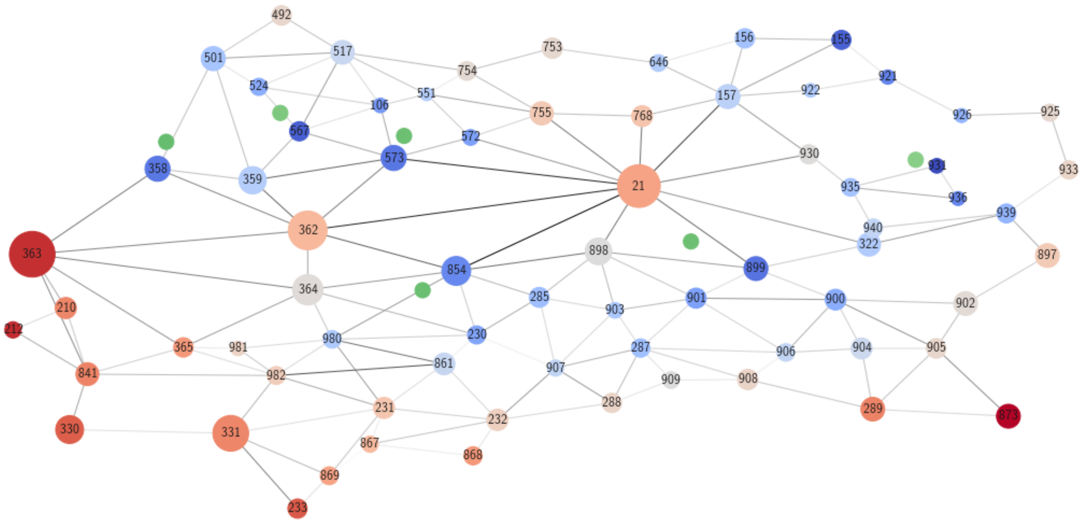


Fig 5. The average response time for each region in the network after adding an ambulance base in region 155

The difference is visible between the network in figure 4 and figure 5. Adding a base pushed down the average response time in the northeast region of London. Looking back at Figure 3, we notice that the crime rate in the northeast is not high relative to the rest of the network. However, the west side of London has a high crime rate, but the presence of three hospitals nearby mitigated the rise in response time due to high numbers of emergency calls.

In late 2017, the NHS set a goal to categorize emergency into three cases: [6]

- Category one: with an average time of seven minutes.
- Category two: with an average time of 18 minutes.
- Category three: with at least nine out of 10 times within 120 minutes.

The estimations of the analysis fall roughly within the range of category one. But further model-specification is necessary such as the type of emergencies, and the impact of sudden waves of emergency calls during a pandemic, etc.)

5 Conclusion

Response time has a crucial impact on the patient chances of receiving care in time. A slight improvement can prove life-saving for many emergency cases. Using simulation, we can not only judge the cost-effective strategy of optimizing response time (increase average speed or add an ambulance base), but we can only assess and device ways to mitigate the impact of a change in infrastructure (bridge closure as an example). In the ethical context, choosing between both strategies should account not only for the mean but also the spread of each distribution because using the utilitarian approach to optimize for the majority can hurt the outlier regions. (e.g., we don't aspire for a distribution where the region with the highest response time is too far from the mean). The findings mentioned that these gaps at most ±15 seconds, which is a reassuring indicator that the strategies are within a reasonable/feasible range to be included as decisions.

[6] NHS London Ambulance Services (2017). Meeting our targets Call prioritisation Retrieved from: https://www.londonambulance.nhs.uk/about-us/how-we-are-doing/meeting-our-targets/

Appendix

- LO/HC Application:

    #audience [HC]: The report breaks down the methodology and the findings in an easy-to-read manner as it would serve as a basis for a policy recommendation.

    #ethical_framing [HC]: Comparing strategies doesn't only concern improving numbers but also the implication of such measures. Ethical concerns are necessary in the context of emergency response as we can't decide on increasing average speed of ambulances without considering the risks.

    #purpose [HC]: In the introduction of the report, I stated the purpose of the simulation and the assumptions that it carries. Different simulation architectures can depends on the metric they're measuring, in this report, the purpose was to minimize average response time of ambulances.

    #descriptive_stats [HC]: Using descriptive statistics such as mean and standard deviation to compare between different strategies.

    #sampling [HC]: As a basis for Monte Carlo simulation, sampling with a weighted distribution reflects the reality (e.g., sampling requests based on crime rates or based on the average traveling time)

    #mc_modeling [LO]: Building an object-oriented program to model the ambulance network using Monte Carlo method. The methodology explains in details the process of imitating real-life settings.

    #mc_analysis [LO]: Although the system is stochastic as we cannot perfectly predict emergency requests. Simulating the system based on crime rate and travel time data gave a holistic approximation to real-life settings.

    #network_analysis [LO]: A big part of the project was to clean up raw data and construct a network with the relevant variables of interest. Besides, the network had edges that reflects travel time means and nodes with attributes.

    #interpret_results [LO]: Devising two scenarios for improving response time such as adding a base or increasing speed. Then the results are interpreted in the context of the network to address the feasibility of each strategy.

- Dataset sources:
    - Uber Movement (2020). Geometry of London regions (`london_lsoa.json`)
    - Uber Movement (2020). Mean travel time and standard deviation of third and forth quarter of 2016 London

References

[1] Crime rates in Greater London (2016). Metropolitan Police of London.

[2] Uber Movement (2018). Examining the Impact of the London Tower Bridge Closure

[3] Kepler (2018). Make an impact with your location data