Vance County EMS Station Analysis Report

Alayna Binder, Cindy Ju, Kathleen Zhang

2025-10-21

Contents

1 Background

Without an effective distribution of emergency medical resources, counties risk delayed response times that can mean the difference between life and death. To improve coverage across Vance County, researchers analyzed data from the county's EMS system, which currently operates four ambulances from two stations located in the Central and South districts. Historical records show that residents in the North district face much longer response times, often averaging over 12 minutes, compared to 6 and 9 minutes in the Central and South regions.

Using a dataset of recorded EMS trips containing call locations, dispatch and arrival times, and Google API travel estimates, this analysis evaluates how different station configurations and vehicle allocations affect system performance. The county is considering establishing a new station in the North district, with two potential site options (Near North and Far North), and several ambulance distribution scenarios.

Our goal is to determine which North station location would most effectively reduce response times and how the four available ambulances should be allocated among the stations to balance coverage and minimize system strain. We aim to illustrate this through visual and numerical summaries that highlight tradeoffs in travel time and resource availability across the scenarios.

2 Data and Exploratory Analysis

2.1 Data

We were given 489 observations of calls occurring in Vance County between January 1st to January 25th in 2024.

One of the calls went to Duke Hospital, which we excluded. Additionally, in certain incidents where multiple patients were involved, they would be transported in the same ambulance. As we only care about the availability of ambulances and not how many patients were in a single ambulance, we cleaned the dataset so that every identical row represented a single, unique incident involving one ambulance. The identical condition ensured datapoints where multiple ambulances were sent out for one incident with multiple patients would still be in the dataset, as it would impact ambulance load.

2.2 Exploratory Data Analysis

To understand where and when ambulance demand occurs, our EDA began with plotting a call-density map (Figure 1) that showed a clear cluster around the Central station with spread in the South and scatter in the North, motivating scenarios that add northern coverage. We then compared Near North vs. Far North (Figure 2) calls, and found that using Google's UA travel times, Near North was closer for 90% of northern calls, cut response time by about 4 minutes on average relative to Far North, and had shorter total call duration with quicker transportation to hospitals.

3 Modeling

3.1 Simulating the Data

The objectives are to compare different scenarios to determine the optimal options for station placement as well as ambulance allocations, but we do not have data about ambulance response times under different station locations and ambulance allocations. As a result, we conducted a simulation that would simulate response times for each of the five scenarios for every single call in our cleaned dataset, resulting in 5 observations for every original call.

We developed Scenario Dispatch Rules that would govern how the simulation responded to different scenarios.

From our exploratory data analysis, we determined that we wanted to consider scenarios where an ambulance might not be available. Therefore, in our rules, each incoming call first checks whether any ambulances are currently available. If at least one unit is free, the system assigns the closest available unit — the one with the smallest estimated travel time, based on Google's best-guess ETA between that station and the call location, resulting in 0 wait time. If all units are busy, the call enters a queue and waits until the next ambulance becomes free. That waiting time is then added to the total response time. Once assigned, each unit stays occupied for the observed duration of the call, and then becomes available again for the next incident.

The wait time is determined by whether or not an ambulance unit is available at the moment an incident call comes in. This models the effect of queueing for service. The wait time is calculated based on two conditions:

If one or more units are already free when the call arrives, we assign the closest available unit, which is the one with minimum ETA. The unit is dispatched immediately at the call time, resulting in zero wait time.

Otherwise, if all units are busy when the call arrives, we assign the unit that is scheduled to become free soonest. The wait time is the time difference between the call time and the departure time, which is the time the assigned unit finally becomes free. This represents the queueing delay. In both cases, the unit's subsequent availability is updated by adding the observed service duration to its departure time, making the unit busy for the duration of the incident.

Finally, we calculate Total Response Time or simulated time as Wait Time + Travel Time.

We filtered the results to exclude extremely long simulated times (above 2000 seconds) to focus on more likely outcomes. We also added a variable, "switched", which determines if the ambulance assigned in a new scenario (S1−S4) is different from the baseline assignment (S0).