Vance County EMS Station Analysis Report

Alayna Binder, Cindy Ju, Kathleen Zhang

2025-10-21

1 Background

The current Vance County EMS configuration places all four ambulances at two existing bases in the Central and South districts, which has produced a systematic performance gap in the Northern district where calls incur materially longer response times. The county is considering establishing a third station in the North, at either a Near North or Far North site, together with a reallocation of the four ambulances across stations.

We analyzed the historical calls to evaluate the 4 new scenarios, guided by three key criteria: (1) expected travel time under each scenario, (2) system load, measured through unit availability and concurrency of busy ambulances, and (3) how often the nominal closest station is actually able to serve a call at dispatch.

We evaluated each call scenario using estimated travel times and the observed service durations and compare performance across the three criteria to determine which North location and ambulance allocation best improves coverage for the county.

1.1 The Data

The dataset contains 489 EMS incidents recorded between January 1–25, 2024. One record corresponding to a transport to Duke Hospital was excluded, since it falls outside the county of interest and was abnormally long. In incidents with multiple patients transported in the same ambulance, duplicate records appear. Because the analysis concerns ambulance availability rather than patient count, we collapsed these duplicates so that each row represents a unique ambulance-level incident, while preserving separate rows when distinct ambulances responded to the same scene, since those affect system load.

For each incident, the data include the call location (latitude/longitude), dispatch, arrival, hospital, and clear timestamps, the assigned base under the current system, and Google-estimated travel times from each candidate station (South, Central, Near North, Far North). The Google estimates are provided under several traffic assumptions (optimistic, best-guess, pessimistic); we use the best-guess estimates in the simulation to reflect a typical operating regime. Hospital destination and service-duration fields are retained only to compute how long a unit remains busy after dispatch, which enters the system-load and queueing components of the analysis.

1.2 Exploratory Data Analysis

We began by examining the spatial distribution of incidents. As shown in Figure 1, calls were densely concentrated near the current Central station, with a smaller secondary cluster around the South

station. Calls also extended into the northern part of the county in a dispersed band, indicating sustained call activity in a region with no current station presence and motivating evaluation of a northern site.

Next, we compared the two candidate North locations using unadjusted Google travel times for calls originating in the North. As shown in Table 1, the Near North site was closer for over 90% of northern calls. Figure 2 further showed shorter response time, shorter total call duration, and faster hospital transport relative to Far North, providing initial evidence in favor of Near North conditional on establishing a station in the North.

We then examined unit-level workload. Figure 3 shows that centrally based units dominated utilization, with Medics 6 and 7 alone accounting for roughly one-third of observed busy time, while South-based units were seldom in service. This asymmetry reflects the earlier spatial concentration and indicates that the current layout leaves some units persistently near capacity.

To assess temporal pressure, we examined concurrent activity. Figure 4 illustrates episodes where multiple Central calls were active while a North call occurred simultaneously, implying that the nearest units were already engaged. We therefore quantified concurrency over the full window. As shown in Figure 5, Central exhibited both the highest typical concurrency and the widest tail. Although concurrency in the North was rare overall, 17.3% of North calls occurred during hours with at least three active Central calls, meaning that when North demand does arise, it often coincides with a period when nearby resources are already heavily committed.

2 Modeling

2.1 Simulating the Data

The objectives are to compare the five candidate scenarios to determine the optimal options for station placement and ambulance allocation. Because we do not observe response times under unimplemented layouts, we conducted a simulation that generates response times for each scenario for every call in the cleaned dataset, resulting in five simulated observations per original incident.

We constructed a set of Scenario Dispatch Rules to govern how incidents are handled under each configuration. From exploratory analysis, we determined that the simulation must allow for the possibility that no ambulance is free when a call arrives. Under the rules, each incoming call first checks whether any units are available. If at least one unit is free, the system assigns the closest available one, using Google's best-guess estimated travel time from station to call location; this yields a wait time of zero. 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 dispatched, a unit remains occupied for the observed duration of the incident and then becomes available for future calls.

Wait time is therefore determined entirely by availability at arrival. If one or more units are free, the closest (minimum ETA) unit is dispatched immediately with zero wait. If all units are busy, the unit scheduled to become free soonest is assigned, and the wait time is the difference between the call time and that unit's release time. In both cases, subsequent availability is updated by adding the observed service duration to that unit's departure time.

Total simulated response time is computed as wait time + travel time. We excluded simulated values above 2000 seconds to remove extreme, low-plausibility outcomes. We additionally created

a "switched" indicator flagging whether the assigned station under a given scenario (S1–S4) differs from the baseline assignment (S0).

2.2 Model Selection and Rationale

We chose to use a linear mixed model with a fixed effect on Scenario, which allowed us to calculate and test the difference in the mean simulated response time between our five Scenarios (S0 through S4), as one of the primary objectives was determining changes in response time across different scenarios.

We also included a random intercept for Incident_ID, allowing every unique incident to have its own baseline average response time that differs from the overall mean, even before accounting for the Scenario.

We fit a preliminary model with just these two components, but it had a lot of heteroscedasticity in the residuals plot. Therefore, we added a variance function that allows the remaining unexplained variability in response time to be different for each Scenario. If a scenario has a high residual variance, it means that even after controlling for the mean, its response times are highly unpredictable or erratic. Meanwhile, if a scenario has a low residual variance, it indicates highly consistent service performance, which is generally desirable in a real-world context.

Thus, our final model was:

$$\text{sim\_time}_{ij} = \beta_0 + \sum_{k=1}^{4} \beta_k I(\text{Scenario}_j = S_k) + u_i + \epsilon_{ij}$$

$$u_i \sim N(0, \sigma_u^2),\ \epsilon_{ij} \sim N(0, \sigma^2)$$

where $\sigma_j^2 = \sigma^2 \cdot \delta_j^2$

sim_time_{ij} is the simulated response time for the i-th Incident ID under the j-th Scenario.

u_i is the random intercept for the i-th Incident ID. σ_j^2 is the residual variance specific to the j-th Scenario, which is impacted by our variance parameter δ_j^2 that scales the baseline variance for Scenario j.

2.3 Model Implementation and Evaluation

Linear mixed models were fit in R using the `nlme` package. Additionally, we ended up dropping data where simulated time was the same across all scenarios in order to focus on the effect of changes resulting from the different Scenarios.

Dropping the data where simulated time was the same across all scenarios resulted in a lower AIC and BIC of the model fitted on the reduced dataset compared to the non-reduced dataset.

We also examined a QQ plot which appears approximately normal. Our Normalized Residuals vs Fitted Values plot of the total response time appears randomly distributed.

We also looked at the Residuals vs Fitted by Scenario, where the points in red indicate calls where the assigned station under the simulation was different from the assigned station in our original dataset.

2.4 Model Results

The intercept term is the estimated mean simulated total response time for the reference group, Scenario 0, which is the current distribution of ambulances and stations. The mean response time in S0 is approximately 470.22 seconds or 7 minutes, 50 seconds. This value is highly significant, with a p-value < 0.001 .

The mean simulated total response time in Scenario 3 is estimated to be 60.96 seconds faster than in Scenario 0, with this being a significant difference.

3 Conclusion, Shortcomings and Future Work

After evaluating five ambulance deployment layouts across the county by replaying the same 489 incidents under each scenario with a dispatch rule (sending the closest available unit, if none are free, dispact the next free unit when it is available), our analysis showed that Scenario 3, which locates one ambulance in the Near North, two ambulances in the Central, and one ambulance in the South, perfomed best. Across mean and median response times, Scenario 3 reduced average response time the most and produced the highest share of calls meeting eight and ten minute targets. These results are consistent with intuition, where a more balanced coverage of the county opposed to a centralized fleet lowers typical responses and extreme delays.

However, one limitation to our analysis is the travel-time realism- we used Google's unadjusted ETAs and treated them as fixed, so rush hour, weather, and road disruptions are not modeled, which underestimates our variability. Additionally, there was no priority handling and all calls were treated the same, so our model is limited in analyzing how our layouts affect critical cases vs. low-priority calls. In the future, additional work would involve adding time-of-day factors to travel by multiplying ETAs by peak/off-peak multipliers, and running two queues for emergency and non-emergency, allowing emergencies to jump ahead of any waiting non-emergency calls when waiting for the next dispatch.

4 Appendix

4.1 Tables and Figures

Figure 1: Density of Calls in Vance County

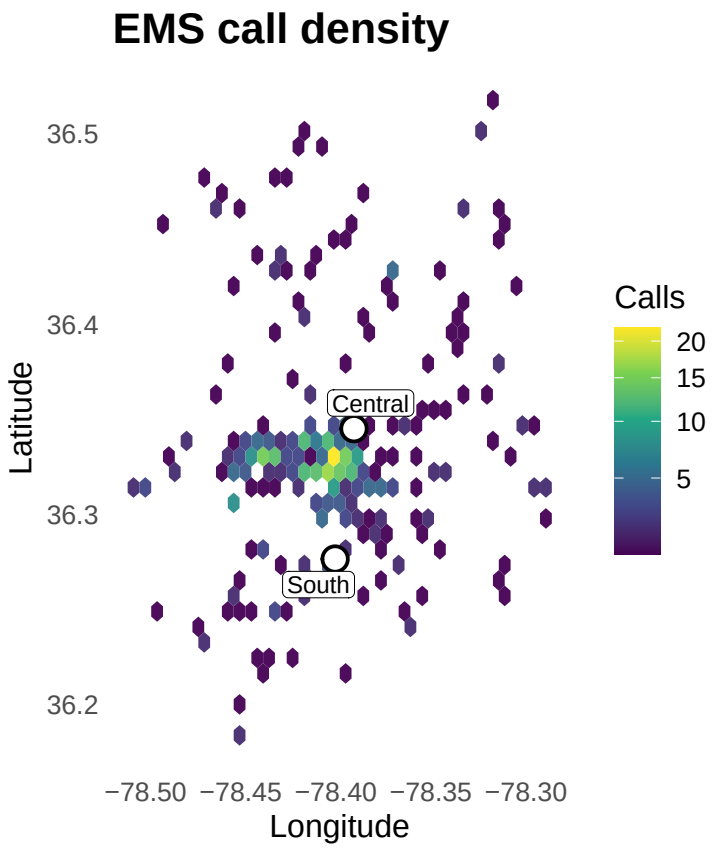

Table 1: Proportion of North Calls Closer to Near vs Far North

Closer Station	Count	Proportion
Near North	47	0.904
Far North	5	0.096

Figure 2: Comparison of Near vs. Far North Demand

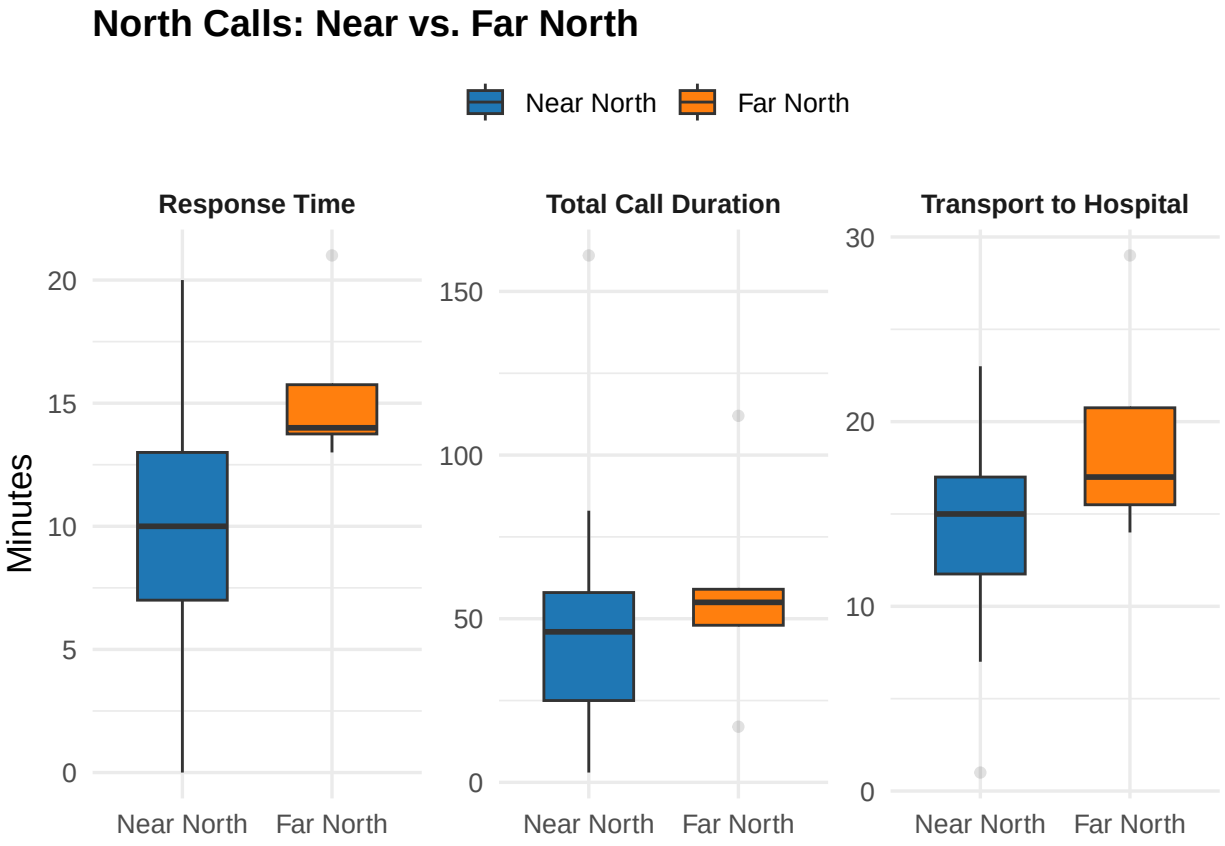

Figure 3: Percentage of Ambulance Utilization

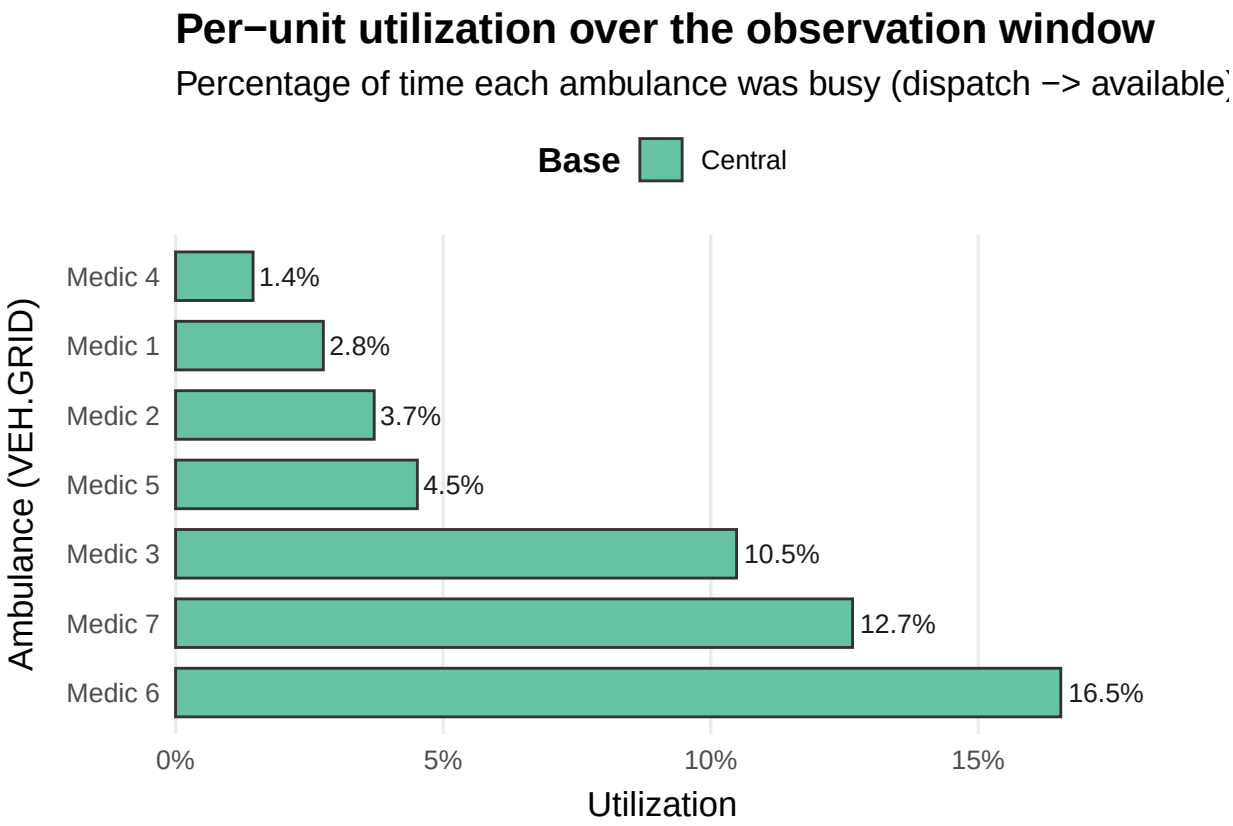

Figure 4: Examining Overburdening Issue

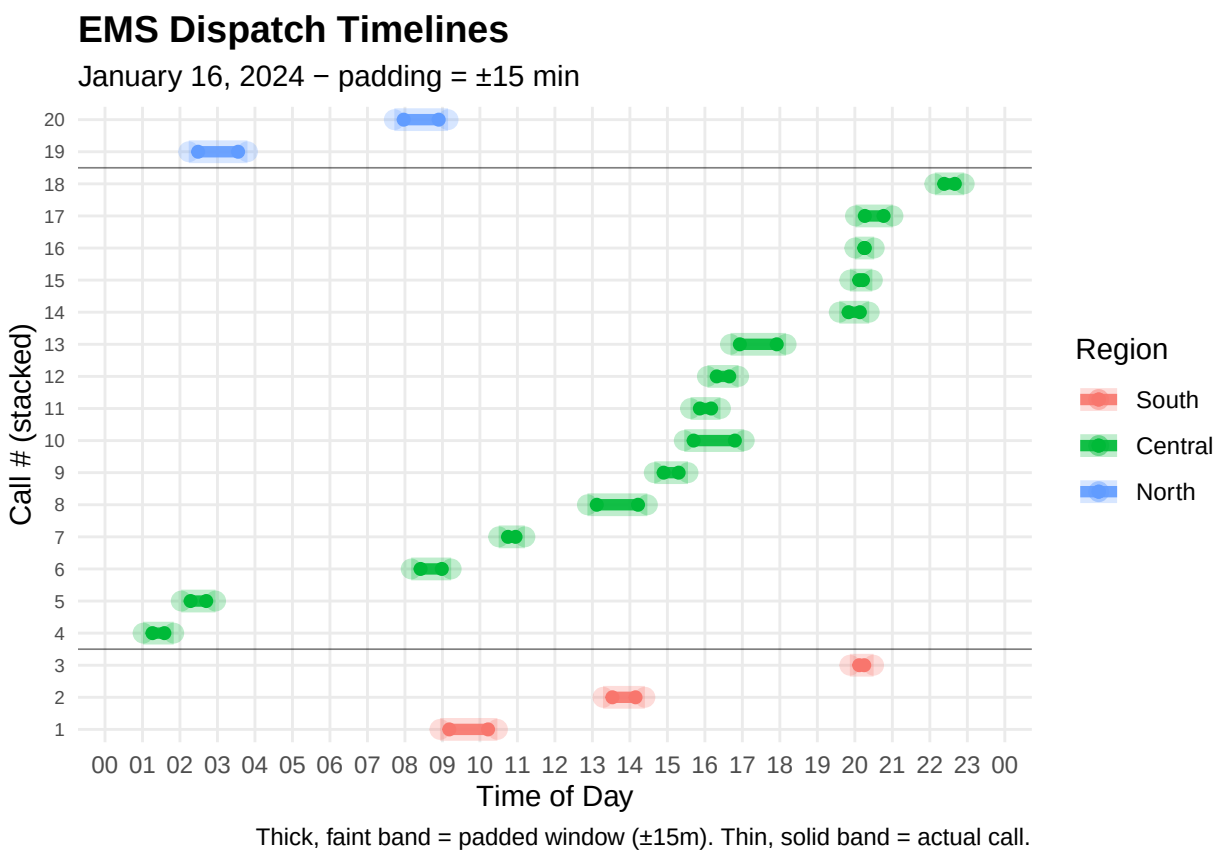

Figure 5: Overburdening by Region