Vance County EMS Station Analysis Report

Alayna Binder, Cindy Ju, Kathleen Zhang

2025-10-21

1 Background

Without an effective distribution of emergency medical resources, counties risk delayed response times that can mean the difference between life and death. 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.

2 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.

2.1 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.

3 Modeling

3.1 Data Simulation

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. The scenarios under consideration are in Table **??**. Dispatch rules governed how the simulation responded to different scenarios.

From our EDA, we determined that we wanted to consider load scenarios. Therefore, each incoming call first checks whether any ambulances are currently available. If at least one unit is free, the system assigns the unit with the smallest estimated travel time, resulting in 0 wait time. Once assigned, each unit stays occupied for the observed duration of the call, and then becomes available again. If all units are busy when the call arrives, we assign the unit that is scheduled to become free soonest. The wait time there is the time difference between the call time and the departure time, which is the time the assigned unit finally becomes free. 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. 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). Simulation results are in Table 2.

3.2 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:

$$\begin{aligned} \text{sim\_time}_{ij} &= \beta_0 + \sum_{k=1}^{4} \beta_k \, \mathbb{I}(\text{Scenario}_j = S_k) + u_i + \epsilon_{ij} \\ u_i &\sim \mathcal{N}(0, \sigma_u^2), \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2) \\ \text{where } \sigma_j^2 &= \sigma^2 \cdot \delta_j^2 \end{aligned}$$

sim_time_{ij} is the simulated response time (seconds) for the i-th Incident ID under the j-th scenario. u_i is the random intercept for the i-th Incident ID, accounting for the unique, unobserved baseline difficulty of that specific call. σ_u^2 is the variance of the random intercepts. σ_j^2 is the residual variance specific to the j-th scenario, which is modeled using our variance parameter δ_j^2 that scales the baseline residual variance (σ^2) for Scenario j.

3.3 Model Implementation and Evaluation

Linear mixed models were fit in R using the `nlme` package. We used the Restricted Maximum Likelihood (REML) method for estimation, which is standard practice for linear mixed models when comparing different variance structures. Additionally, the model was fit on a reduced dataset where incidents showing no variability in simulated response time across all five scenarios were dropped to ensure our coefficients primarily capture the effects of the scenario changes, not just baseline noise.

3.4 Model Evaluation

First, while modelling, 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. Next, after getting to our final model, we examined the residual plots to check model assumptions. A Q-Q plot can be found in Figure 6 and normalized Residuals vs Fitted Values plot can be found in Figure 7.

Due to anticipated heteroscedasticity, we fitted our model with a variance structure. We needed a normalized residuals plot to account for weighting due to our variance function, which applied weights to the residuals based on the Scenario to ensure the model fit is accurate despite the unequal variance. There did not appear to be any major patterns in the residuals, which are spread without

a lot of heteroscedasticity. Additionally, there did not appear to be significant deviation from the Q-Q plot, indicating the assumption of normality is satisfied.

We also looked at the Residuals vs Fitted Values plot 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. Simulated scenarios generally showed more vertical spread than S0, meaning after controlling for mean response time and an incident's baseline difficulty, the unpredictability of the response time is typically higher in alternate scenarios. S3, which had the fastest mean response time, had less switches than S1 and S2 and a tighter spread, indicating better consistency.

3.5 Model Results

To answer the research questions, we prioritized determining which of our 5 scenarios resulted in the shortest mean response time, as we felt that decreasing the average response time would be the best overall benefit to the community in terms of improving patient outcomes. Secondary considerations were the percent of simulated calls under 8 and 10 minutes, as more extreme wait times would lead to worse outcomes for patients, and another way to evaluate the best scenario was determining which one minimized "worse outcome" call times.

From our model results (Table 3), the mean simulated total response time in S3 is estimated to be 60.96 seconds faster than our baseline, Scenario 0, with this being a significant difference with a p-value of 0.0062, well under our threshold of 0.05. 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 indicates that the mean response time of S3 is estimated to be around 6 minutes, 48 seconds. All other scenarios had positive coefficients, indicating they resulted in longer mean response times than the baseline. Additionally, a boxplot revealed a wider spread of outliers in other scenarios compared to S3 (Figure 9). Therefore, we concluded that the S3 station placement and ambulance allocation was ideal.

4 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, performed 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.

One limitation to our analysis is travel-time realism. We used Google's best-guess 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.

5 Appendix

5.1 Tables and Figures

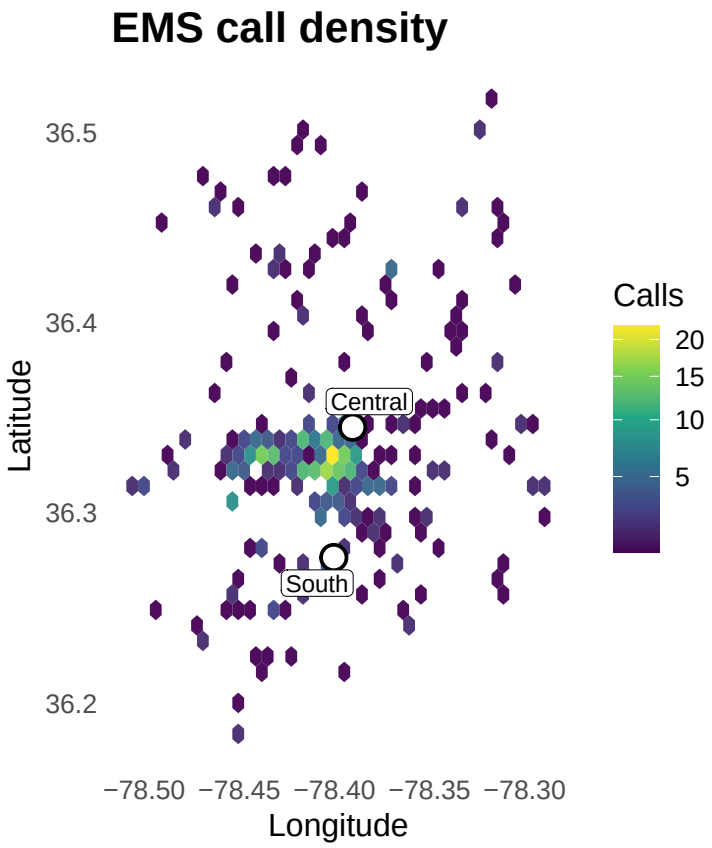

Figure 1: EMS call density across Vance County (geographically)

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

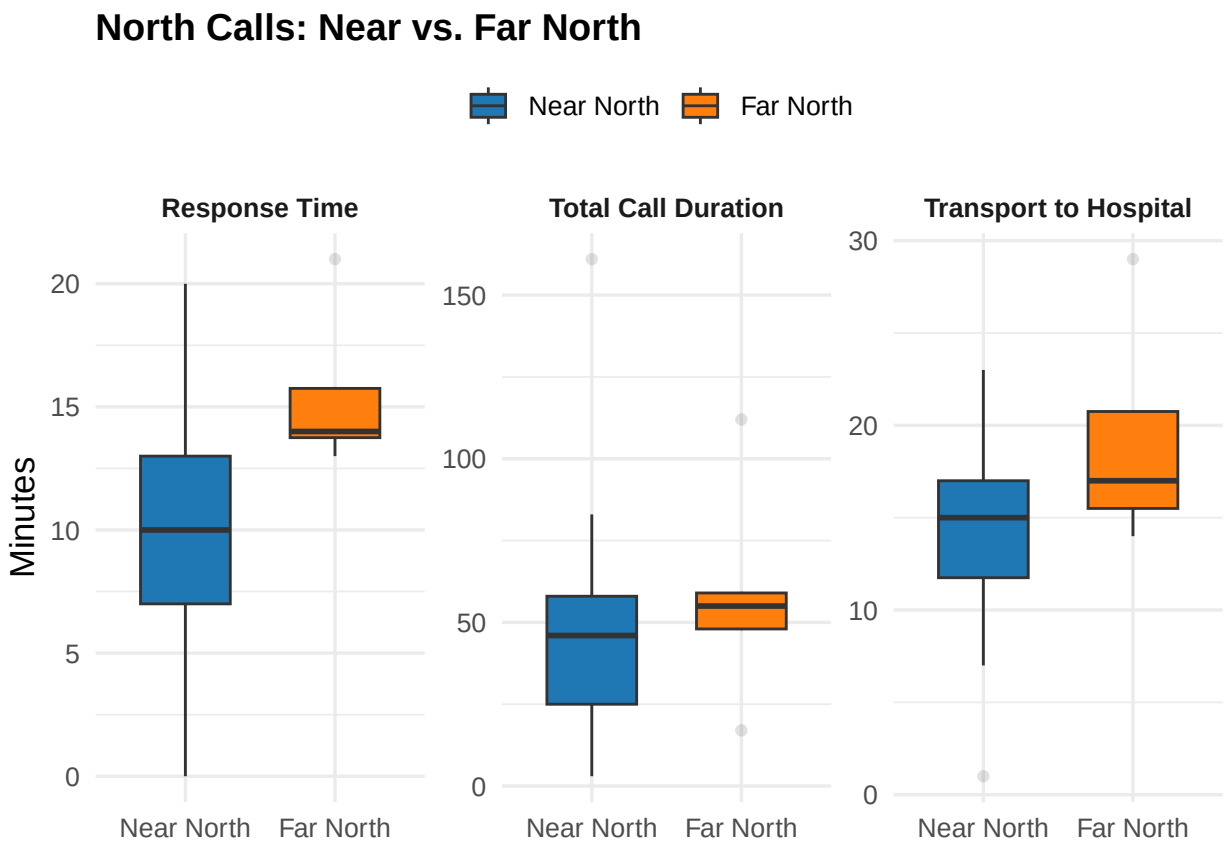

Figure 2: Comparison of Near vs. Far North demand

Table 2: Table of summary statistics for simulation response time results by scenario.

Scenario	n	mean_resp_s	median_resp_s	pct_under_8min	pct_under_10min
S3	472	371.3708	349.0	0.8050847	0.9088983
S0	472	388.6758	357.0	0.7838983	0.8792373
S4	472	391.1250	368.0	0.7669492	0.8898305
S1	472	411.8686	373.0	0.7245763	0.8347458
S2	472	429.2669	388.5	0.6906780	0.8220339

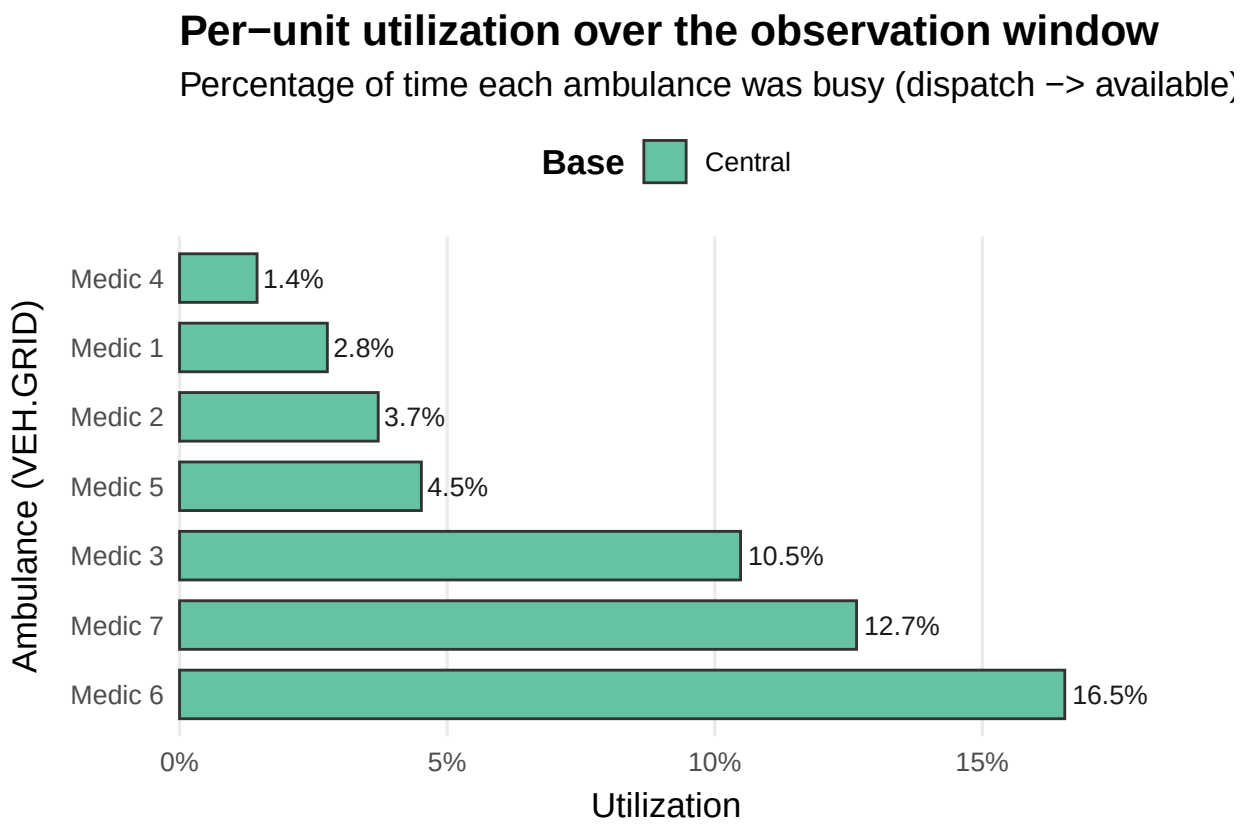

Figure 3: Percentage of ambulance utilization

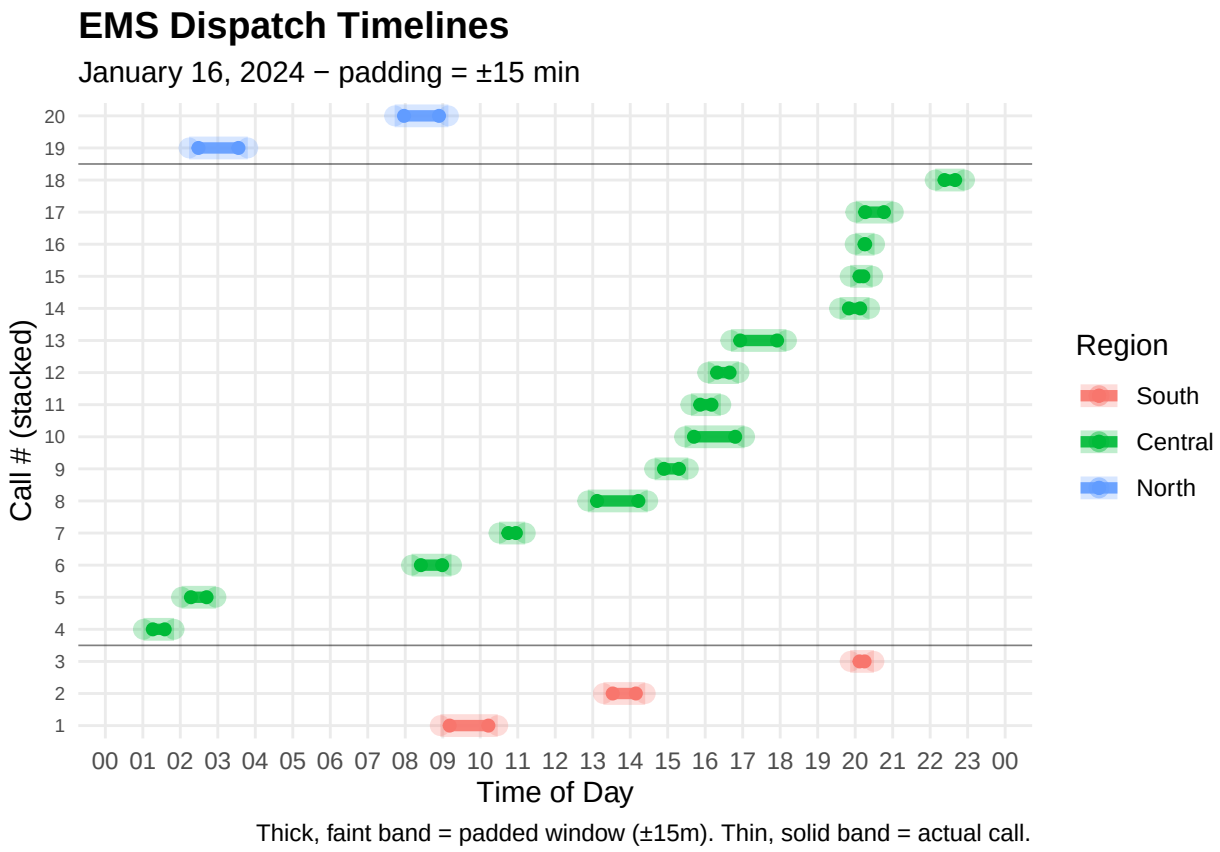

Figure 4: Example timeline of call status on January 16, 2024.

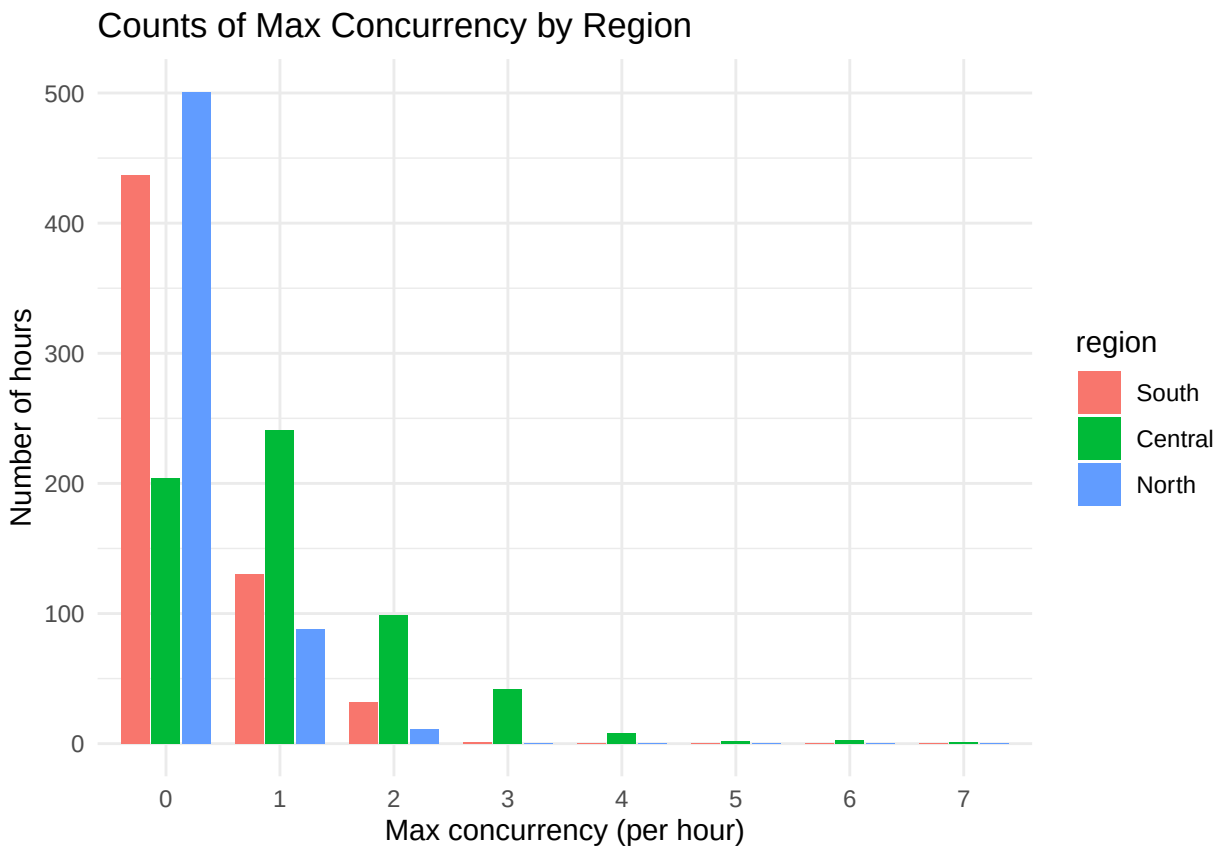

Figure 5: Shows the maximum concurrent number of calls within a given hour across the dataset. Most of the time there were no or very few concurrent calls, but Central was more prone to having concurrent calls.

Table 3: Table of fixed effects coefficients from the final model.

Term	Value	Std.Error	DF	t-value	p-value
(Intercept)	470.216418	23.81653	532	19.7432827	0.0000000
ScenarioS1	81.694030	27.17764	532	3.0059282	0.0027727
ScenarioS2	142.977612	25.66344	532	5.5712561	0.0000000
ScenarioS3	-60.955224	22.20271	532	-2.7453960	0.0062482
ScenarioS4	8.626866	23.70138	532	0.3639816	0.7160163

Table 4: Table of variance function residual standard deviation multipliers by scenario.

Scenario	SD multiplier	Var multiplier
S0	1.0000000	1.0000000
S1	0.9372167	0.8783751
S2	0.8215226	0.6748993
S3	0.5036209	0.2536340
S4	0.6546640	0.4285849

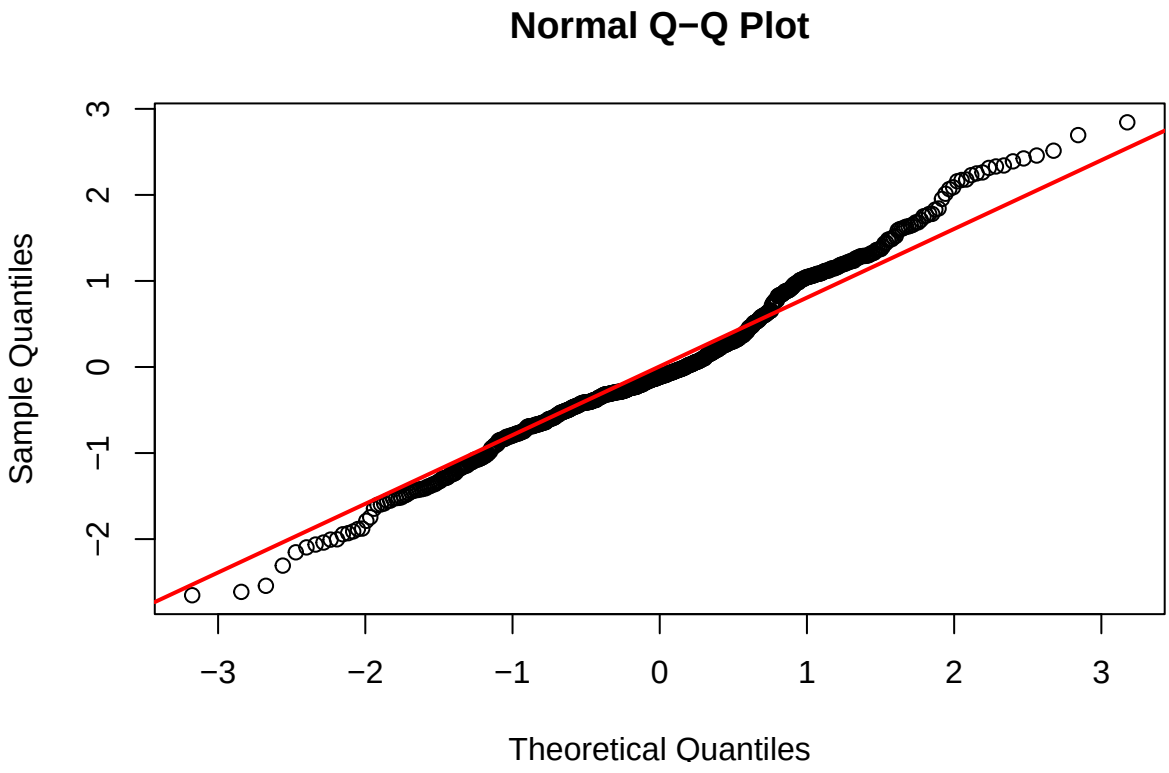

Figure 6: Q-Q plot of residuals. There is only a slight deviation in the Q-Q plot.

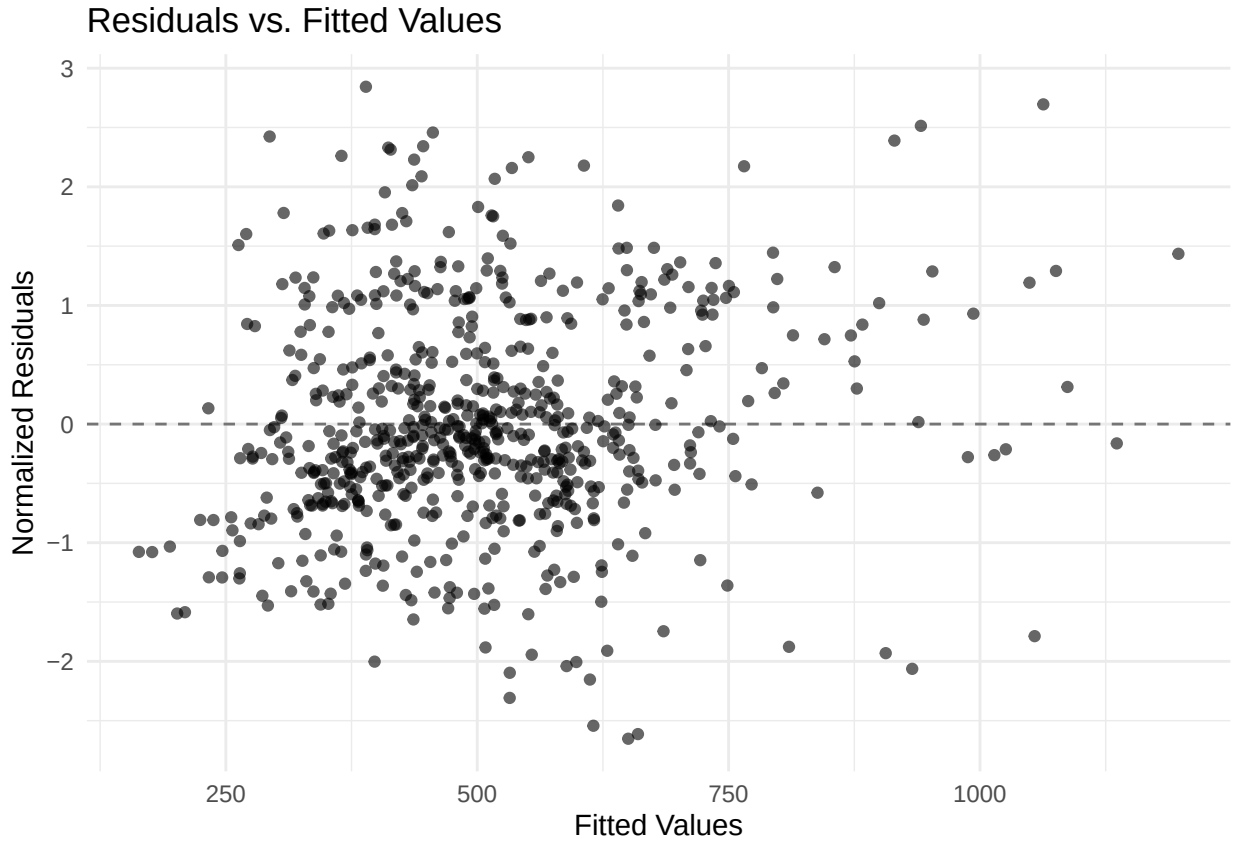

Figure 7: Residuals vs Fitted Values plot. There does not appear to be heteroscedasticity or a fan pattern.

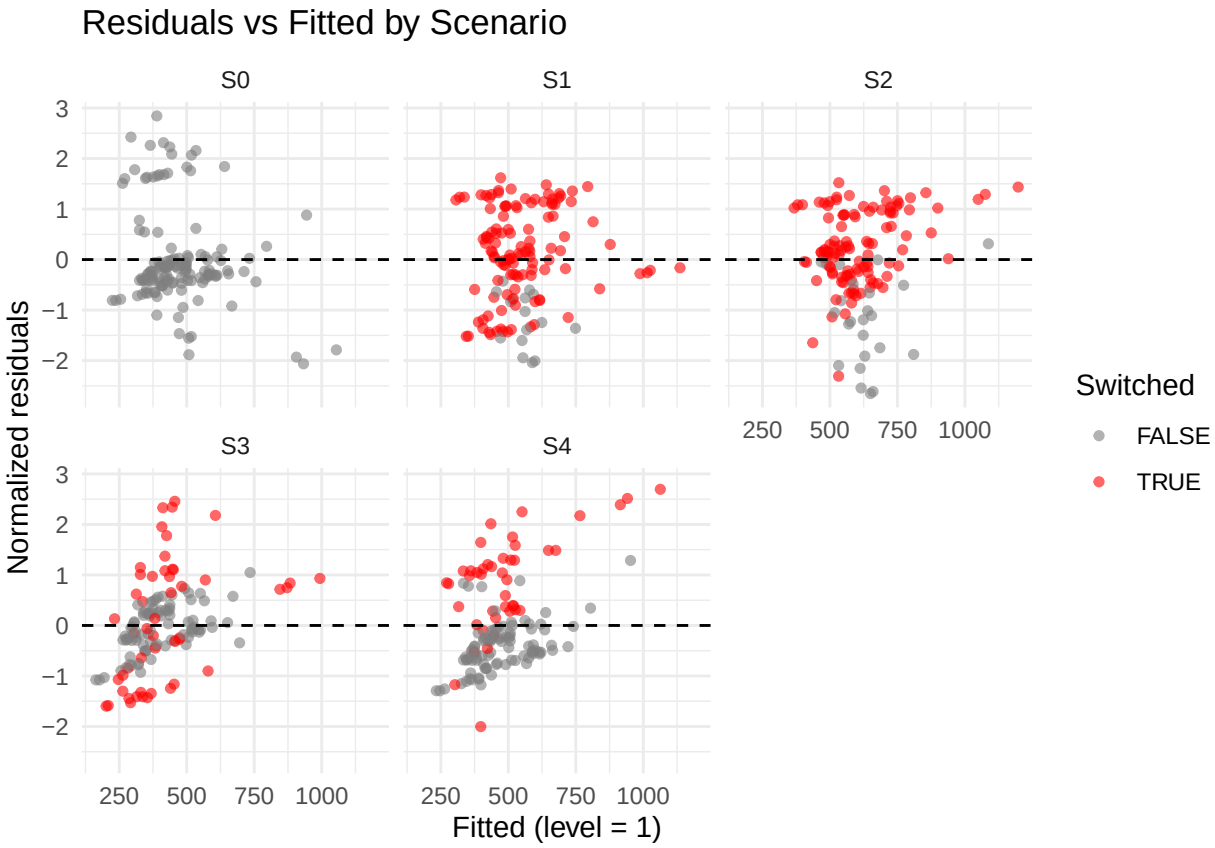

Figure 8: Residuals vs Fitted Values plot by scenario. There is variations in spread across the different scenarios.

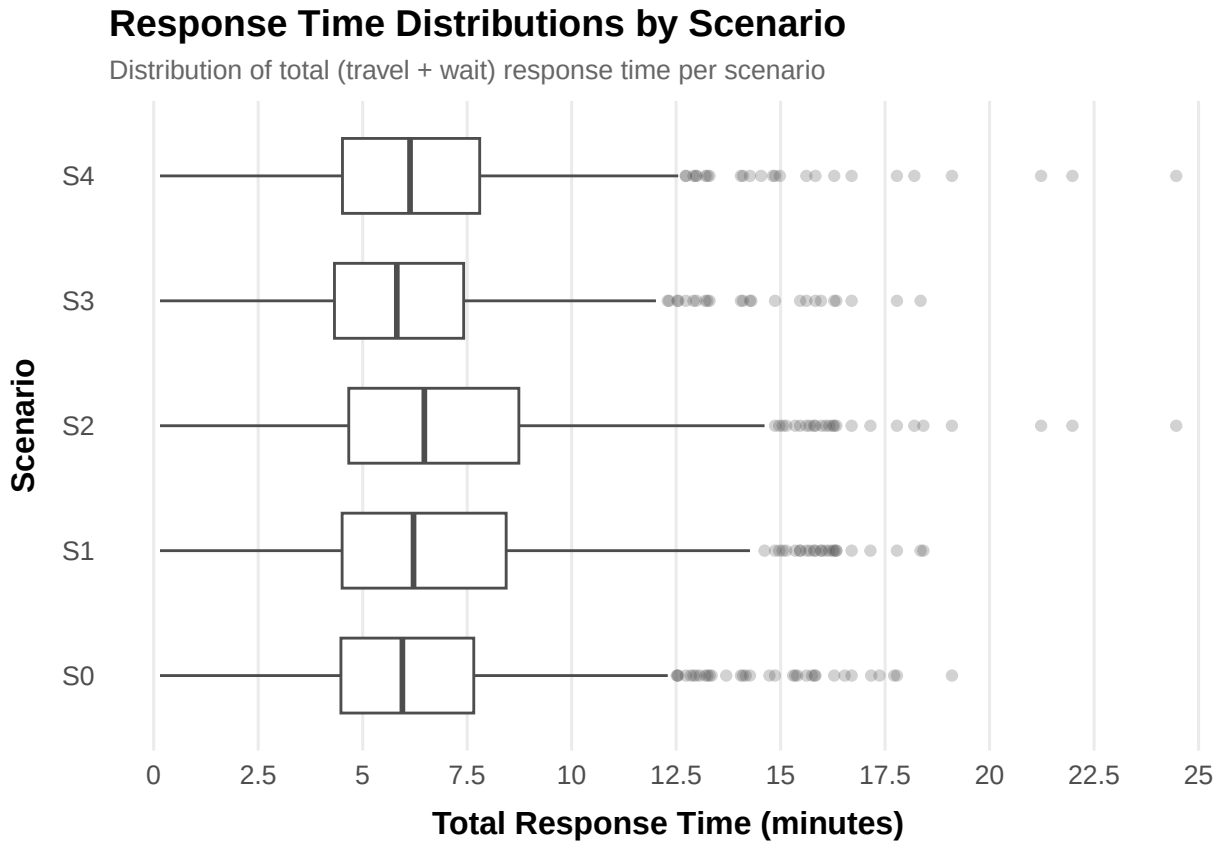

Figure 9: Plot of response time distributions by scenario. S3 had the lowest mean response time and a smaller spread of outliers