Traffic Prediction

By: Madeleine Roberts, Mehul Gupta, Andrew Park
Group 2: timeseriesDISCOVERy

Problem Statement

Description of the Problem Context

In the United States, roads are a vital infrastructure component, heavily utilized for both public and private transportation. This immense reliance on road travel necessitates an efficient management of traffic flow, especially during peak congestion times, to mitigate traffic delays, reduce emissions, and enhance overall transportation efficiency. An analysis of traffic data, which includes observations on vehicle counts at specific junctions over a defined period, can provide valuable insights into traffic patterns and help in forecasting future congestion.

Problem Statement

This project aims to forecast future traffic congestion levels at specific road junctions by analyzing over 48,000 historical observations with various time series models. The insights derived from these predictions will enhance traffic management systems, improving infrastructure efficiency and reducing environmental impacts.

Source: Burrows et al 2021; Richter 2021; Shepardson 2024

Assumptions/Hypotheses and Questions

Assumptions/Hypotheses

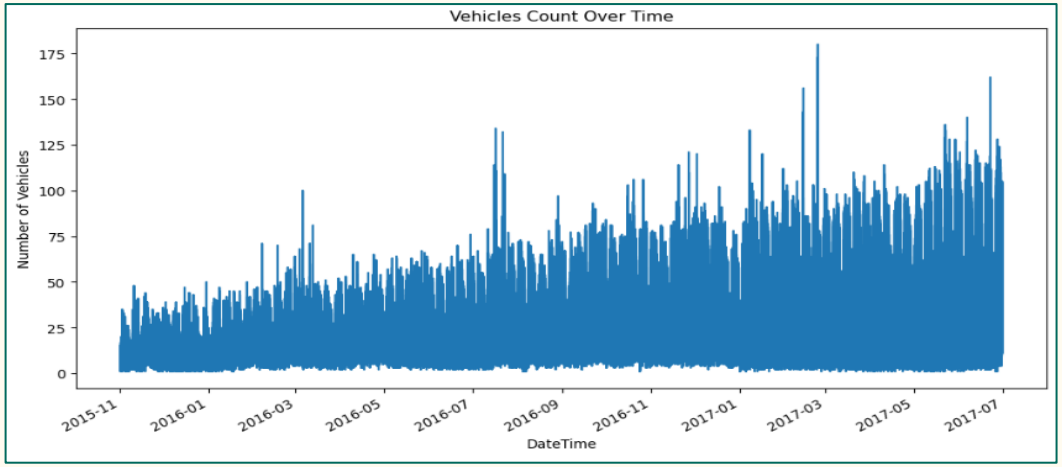
- Assume the data is stationary
- Assuming that there will be seasonality since it is traffic data
- STL might be best to handle this (multiple seasonality might be present)

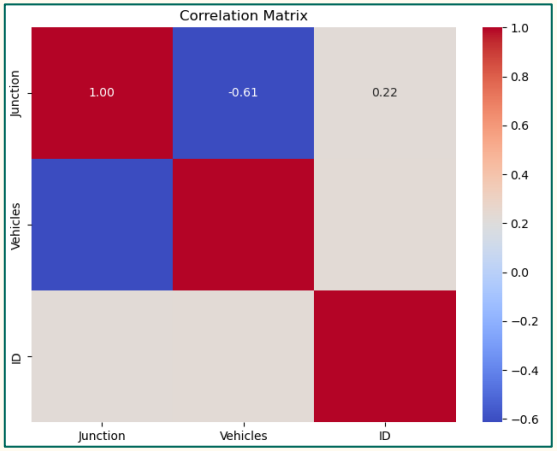
Questions

How can we predict and anticipate traffic congestion levels at each junction for future dates and times?

1. How do traffic volumes at each junction exhibit temporal trends and seasonal variations?
2. How can the dataset pinpoint peak congestion times at various junctions?
3. What are the differences between the three proposed models, and how can each contribute to our traffic pattern analysis and prediction?

Data Properties



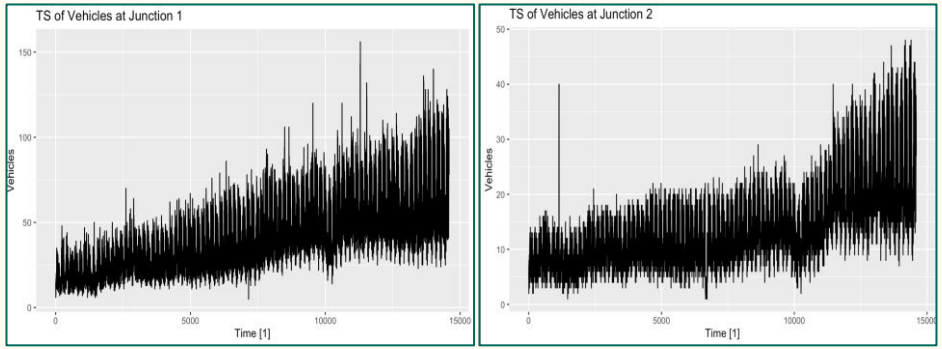


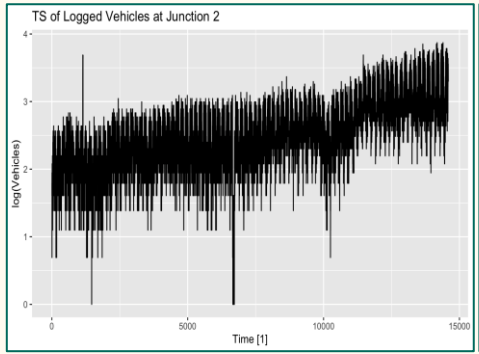
The plot of vehicle counts over time shows trends and possible seasonal variations, suggesting that the time series data of vehicle counts is non-stationary. This implies changes in mean and variance over time, which is typical for traffic data influenced by daily and seasonal patterns.

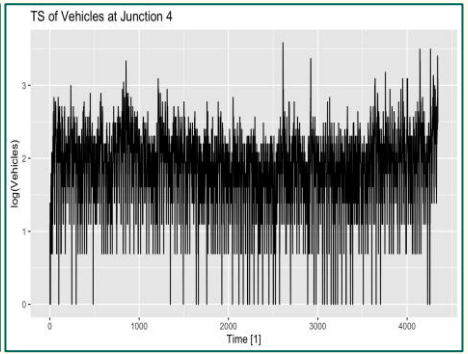
The correlation matrix reveals a modest correlation between the Junction and Vehicles (-0.61), suggesting that different junctions might have different traffic volumes. There is no significant correlation between the variables, which is expected since ID is likely just a sequential or coded identifier.

Plots by Junction

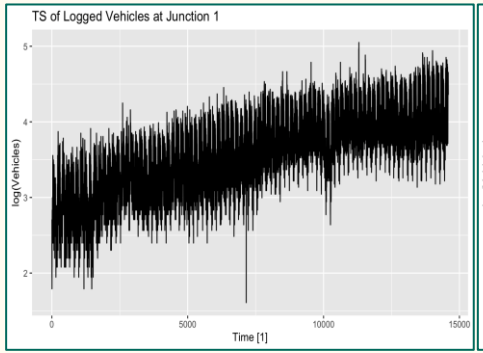
Vehicles

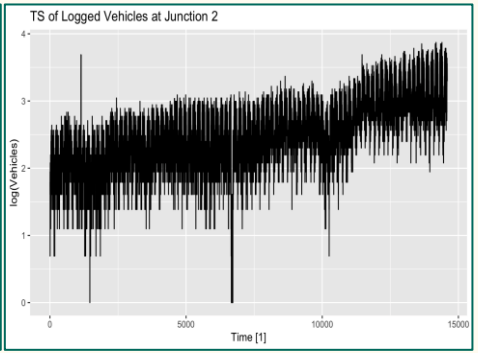


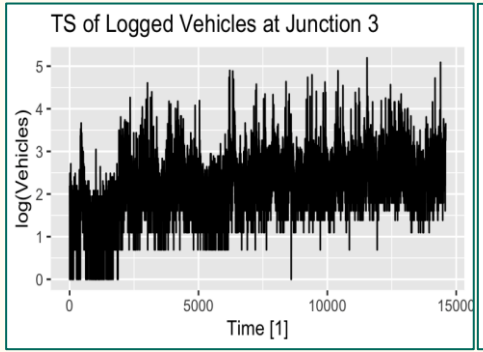


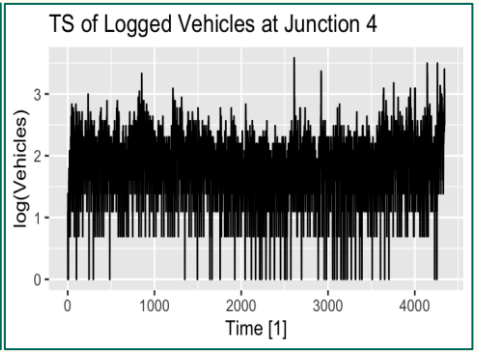


Logged Vehicles to Account for Heteroskedasticity







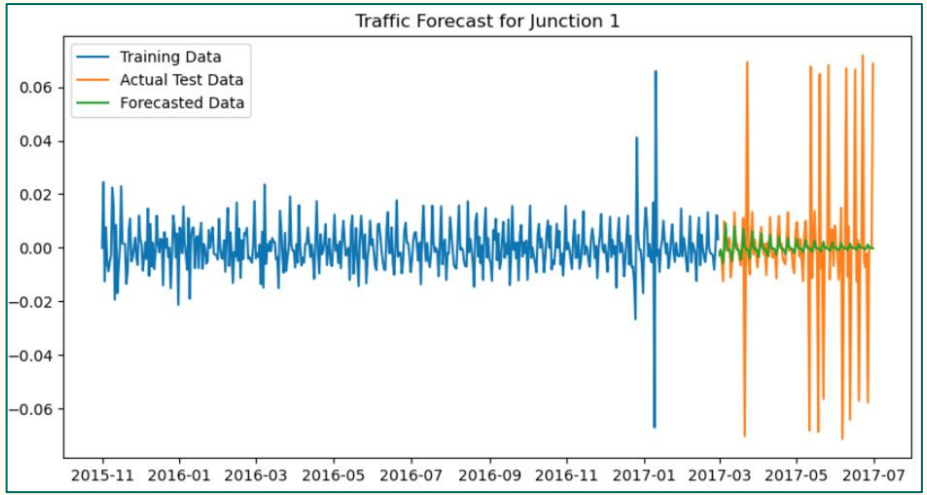


Proposed Approaches

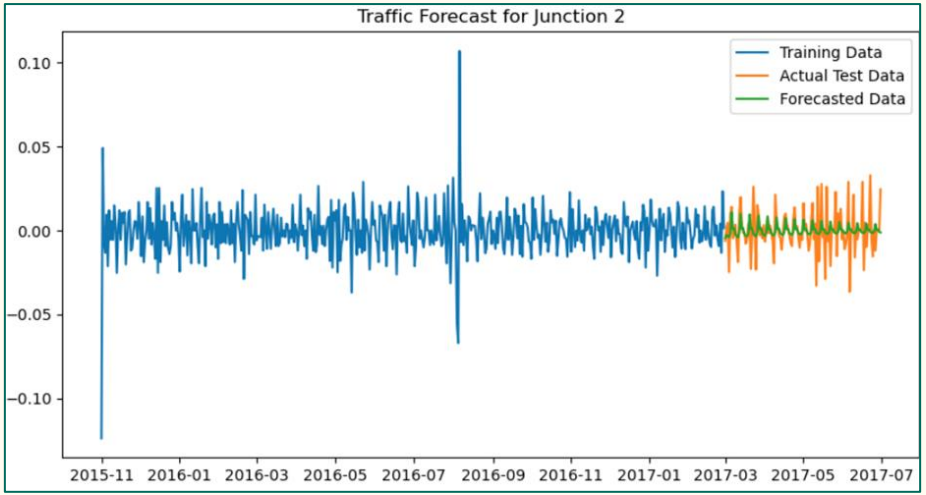
Three models have been developed:

1. **SARIMA** - Extends ARIMA model to account for seasonal patterns in the data
2. **Regression with ARIMA errors** - Utilize regression to incorporate explanatory variables and ARIMA to handle residual errors
3. **STL with Multiple Seasonal Periods** - Decomposes data into seasonal, trend, and remainder components, accommodating multiple seasonal patterns (hourly, daily, etc)

SARIMA

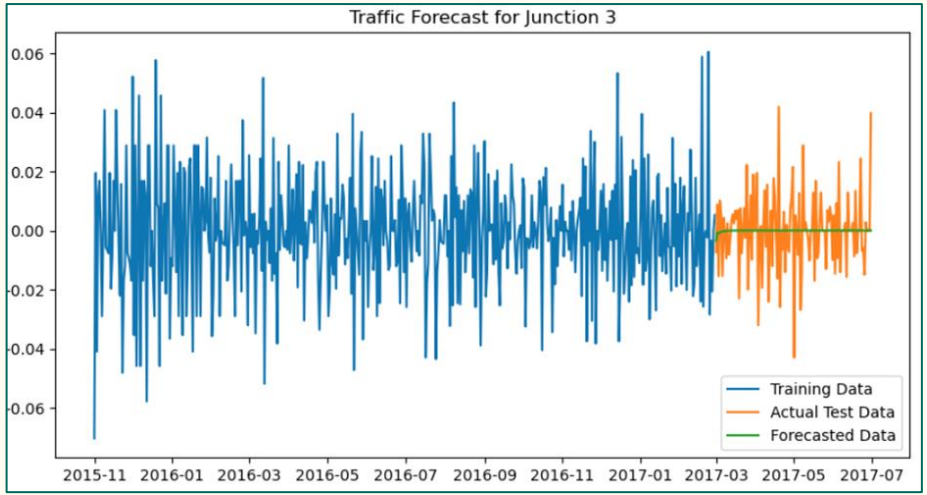


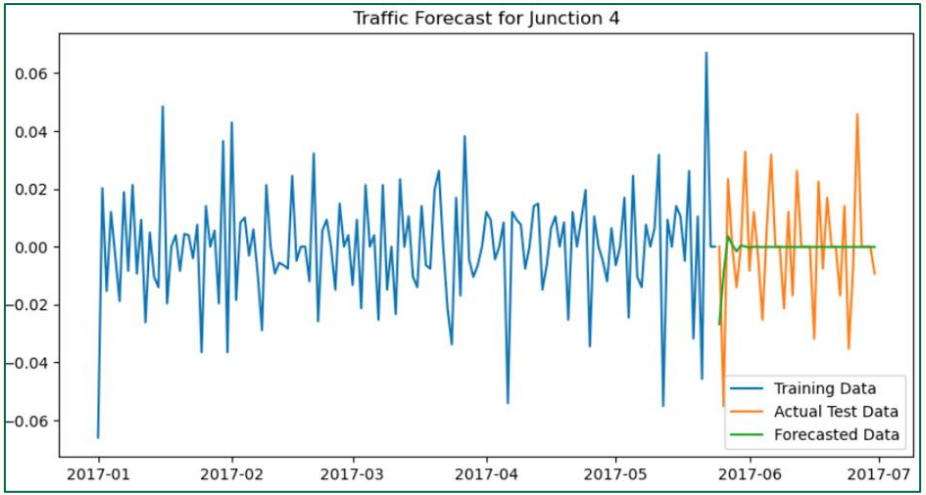
RMSE: 0.024647, AIC: -3513.79
86.9% of test data is within the confidence interval generated by the model: ARIMA (0, 0 ,1) (1, 0, 1) [7]



RMSE: 0.012079, AIC: -3006.18
97.5% of test data is within the confidence interval generated by the model: ARIMA (0, 0 ,1) (2, 0, 2) [7]

SARIMA

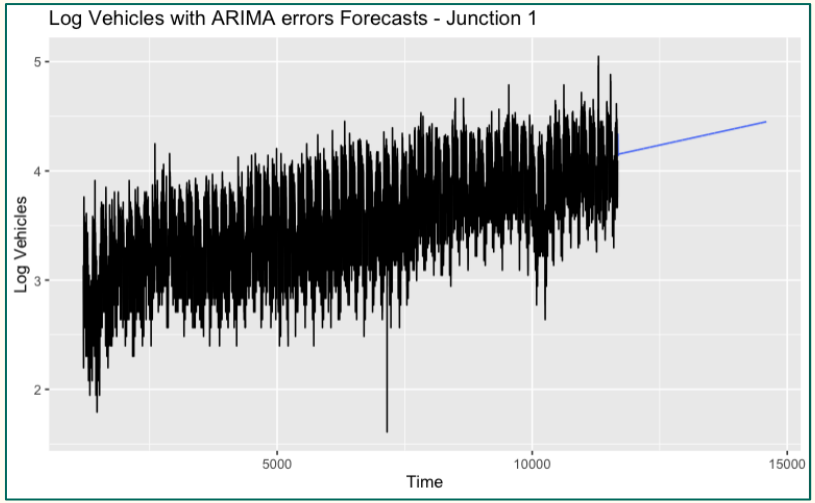




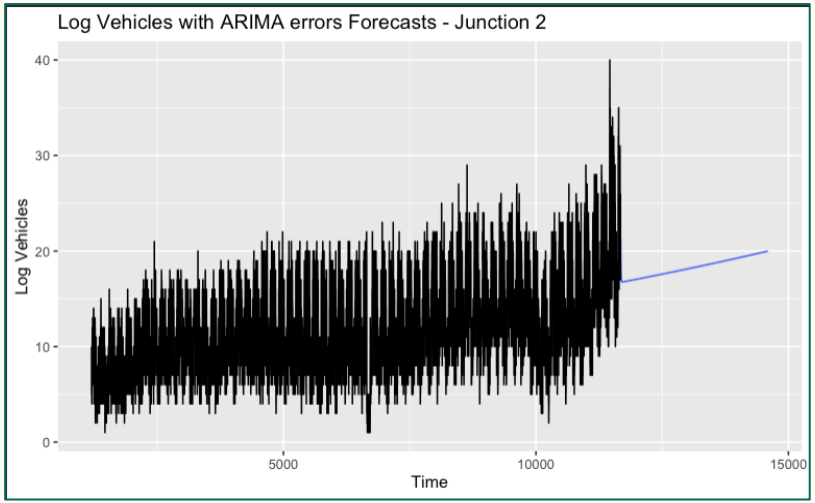
RMSE: 0.012884, AIC: -2550.37
97.5% of test data is within the confidence interval generated by the model: ARIMA (2, 0 ,1) (0, 0, 0) [7]

RMSE: 0.019267, AIC: -792.14
94.6% of test data is within the confidence interval generated by the model: ARIMA (2, 0 ,3) (0, 0, 0) [7]

Regression with ARIMA Errors

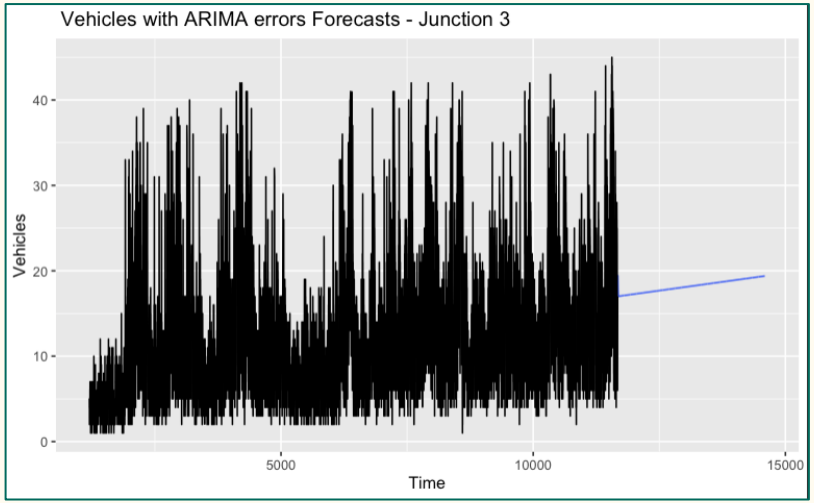


RMSE: 0.399, MAPE: 8.388
LM w/ ARIMA(2,0,4) errors
Ljung-Box Test: p = 5.753 e-10
AIC: -12585.74

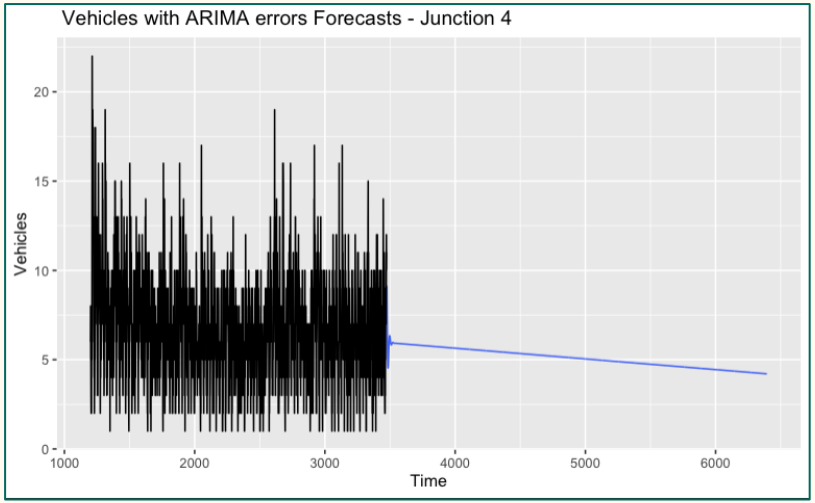


RMSE: 0.445, MAPE: 11.497
LM w/ ARIMA(1,0,5) errors
Ljung-Box Test: p = 0.145
AIC: 245.36

Regression with ARIMA Errors

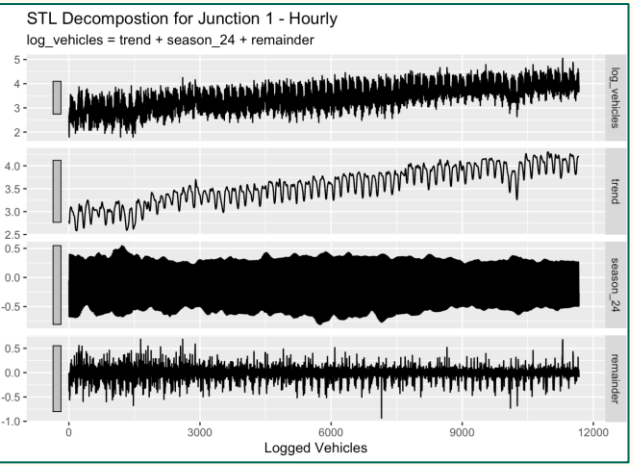


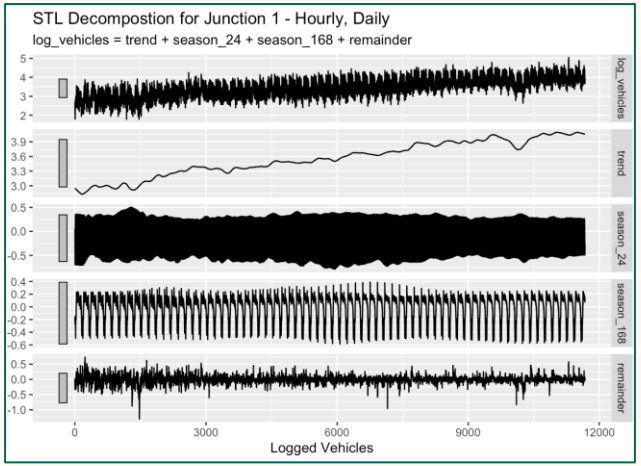
RMSE: 8.625, MAPE: 63.050
LM w/ ARIMA(3,0,3) errors
Ljung-Box Test: p = .115
AIC: 72547.15

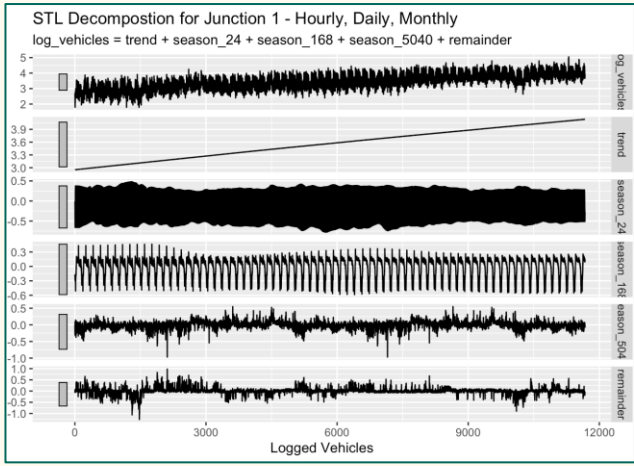


RMSE: 4.525, MAPE: 43.983
LM w/ ARIMA(2,0,3) errors
Ljung-Box Test: p = 0.847
AIC: 16241.73

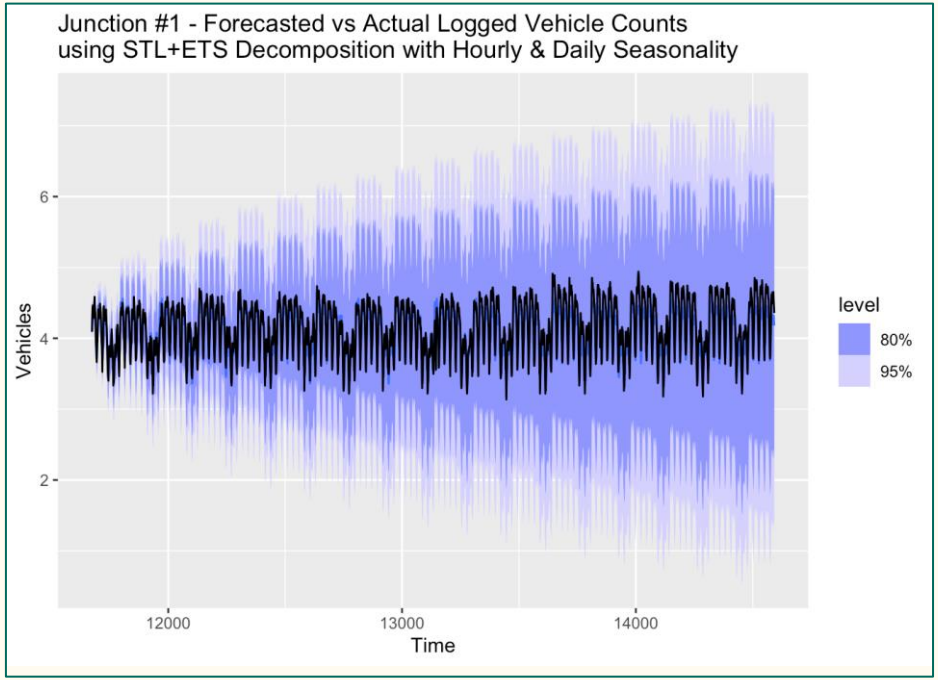
STL with Multiple Seasonal Periods - Junction 1

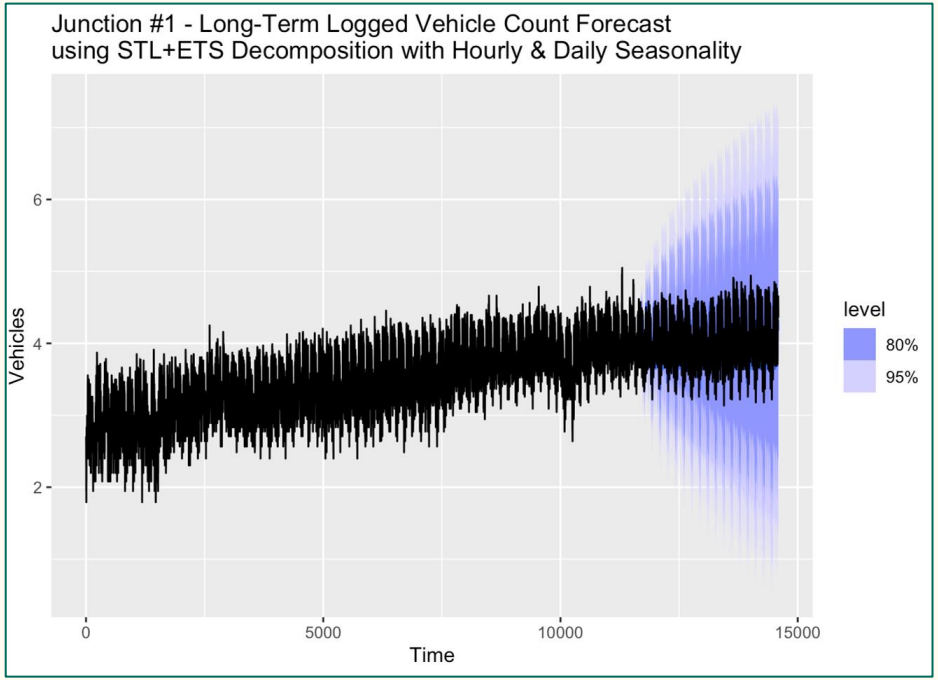






STL with Multiple Seasonal Periods - Junction 1 Forecast

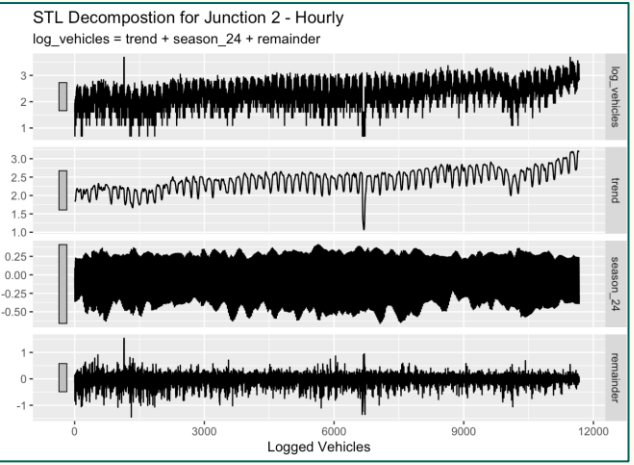


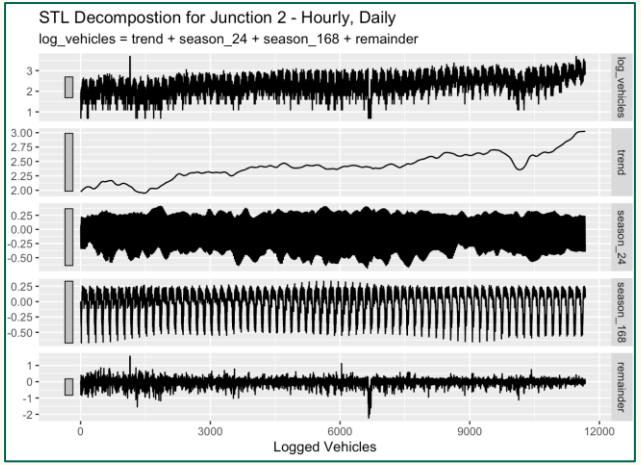


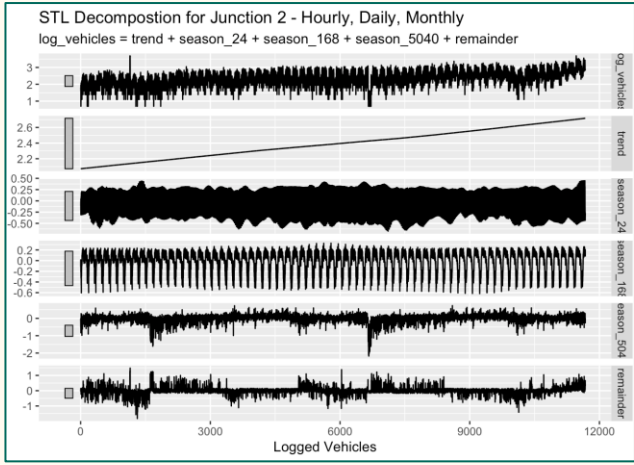
Logged Model with Hourly & Daily Seasonality

	ME	RMSE	MAE	MPE	MAPE	AIC
Test set	0.07827429	0.1504127	0.1173253	1.74012	2.778363	52344.89

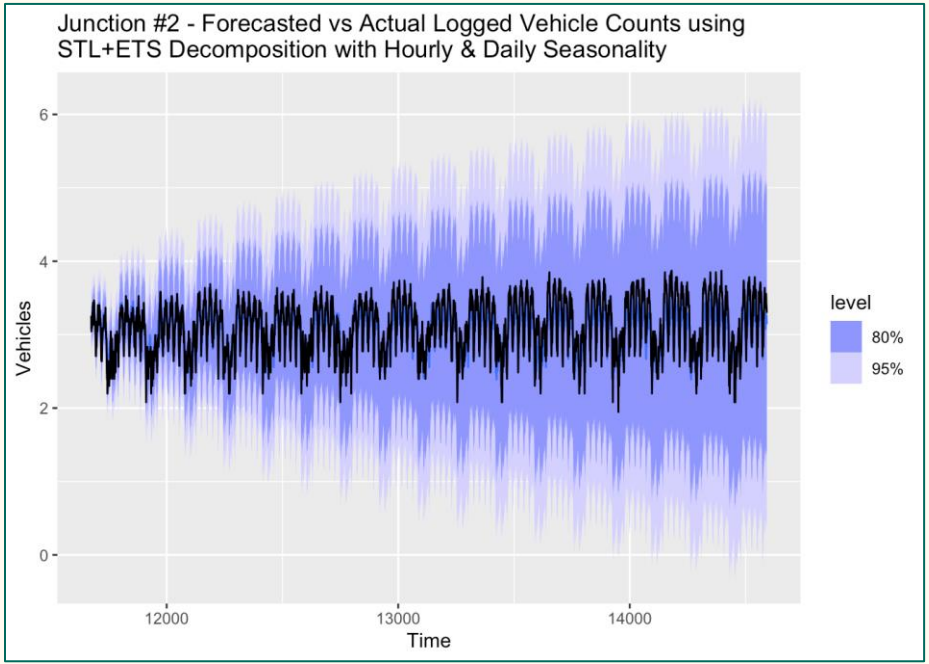
STL with Multiple Seasonal Periods - Junction 2

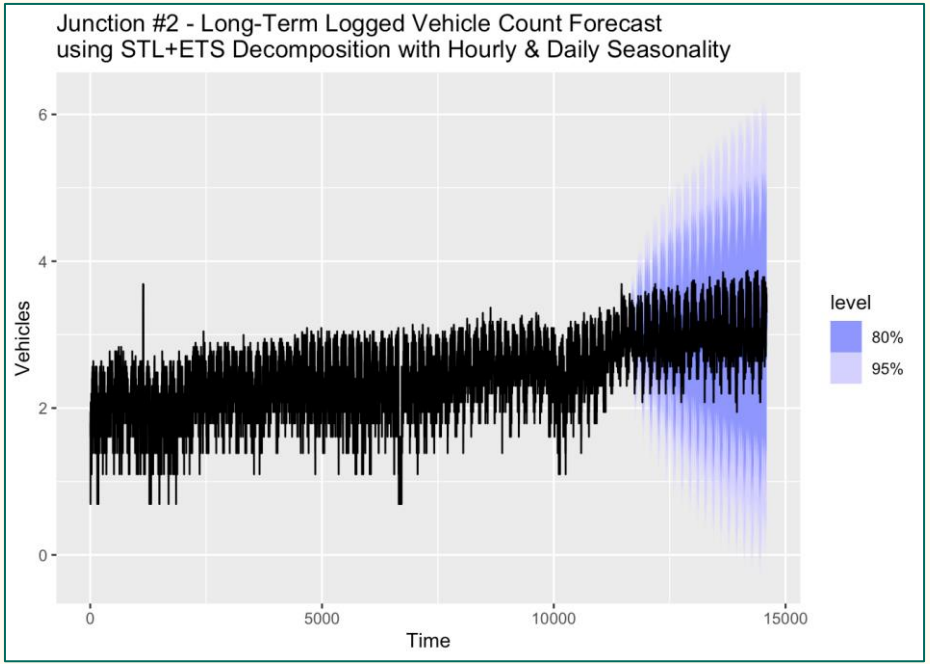






STL with Multiple Seasonal Periods - Junction 2 Forecast

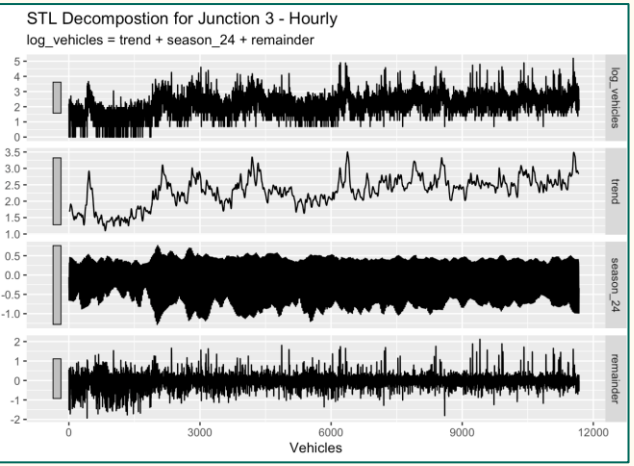


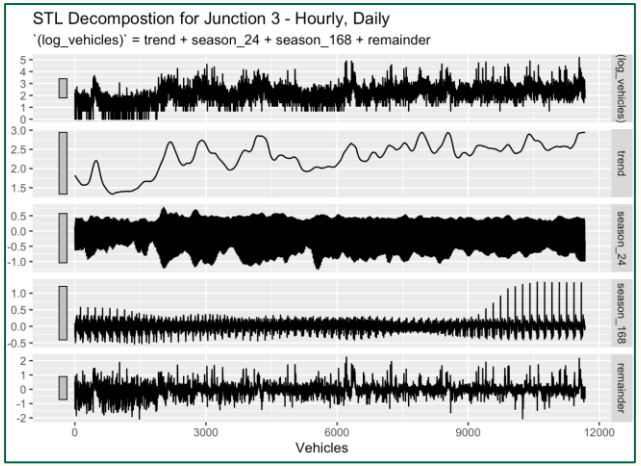


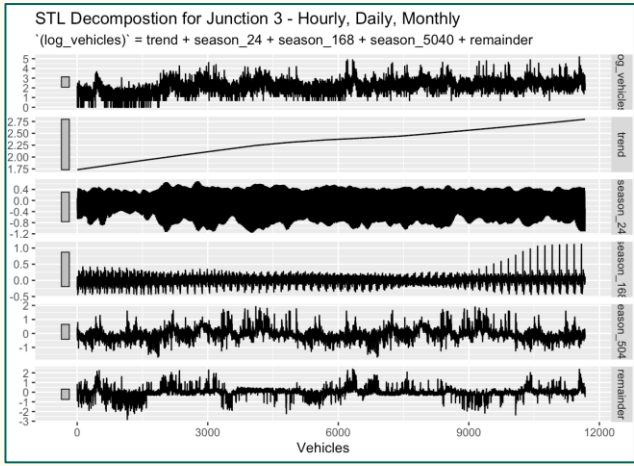
Logged Model with Hourly & Daily Seasonality

	ME	RMSE	MAE	MPE	MAPE	AIC
Test set	0.0733082	0.1948296	0.1549028	1.916589	4.968595	69896.44

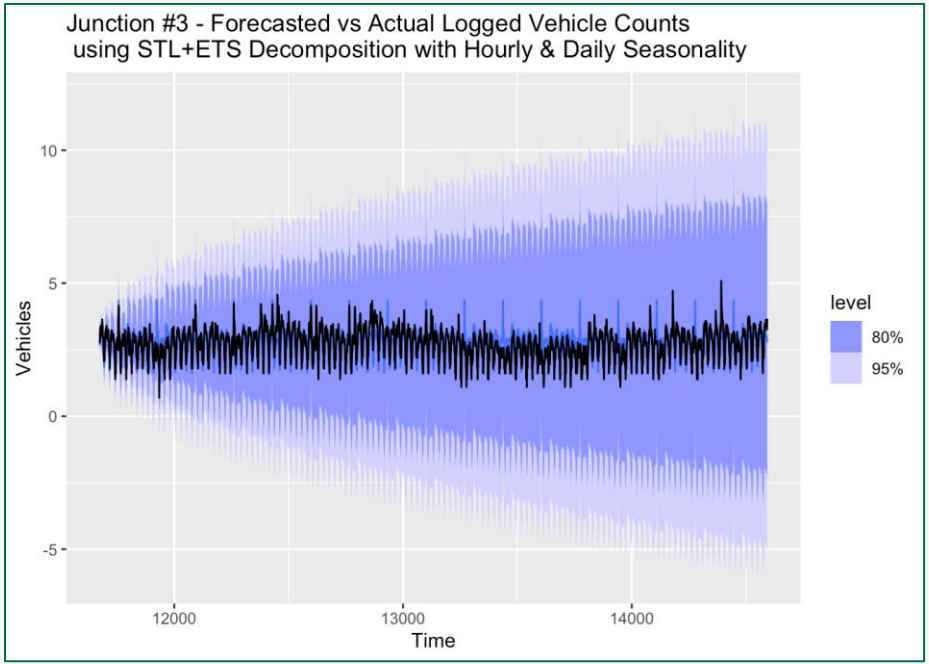
STL with Multiple Seasonal Periods - Junction 3

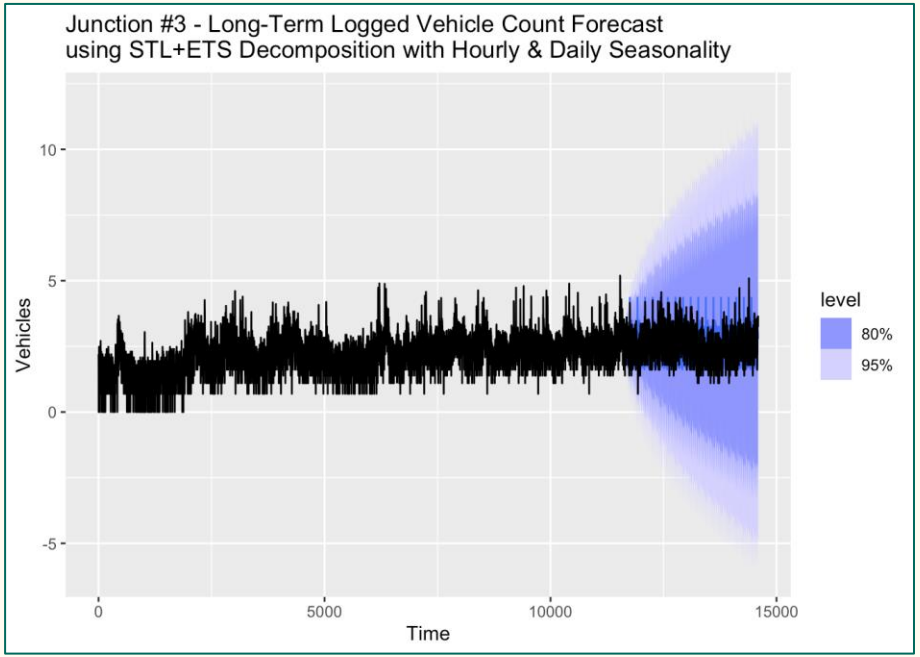






STL with Multiple Seasonal Periods - Junction 3 Forecast

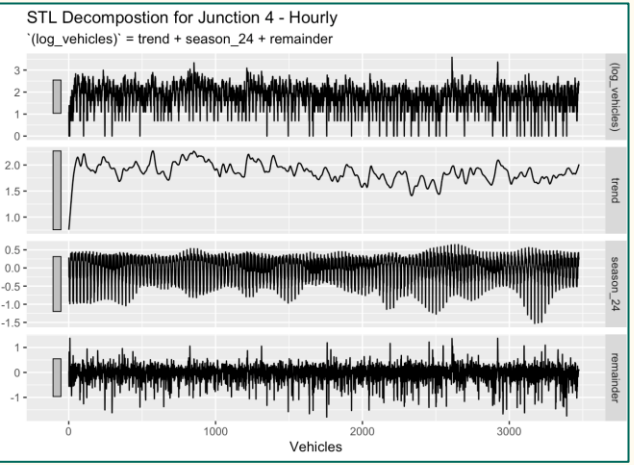


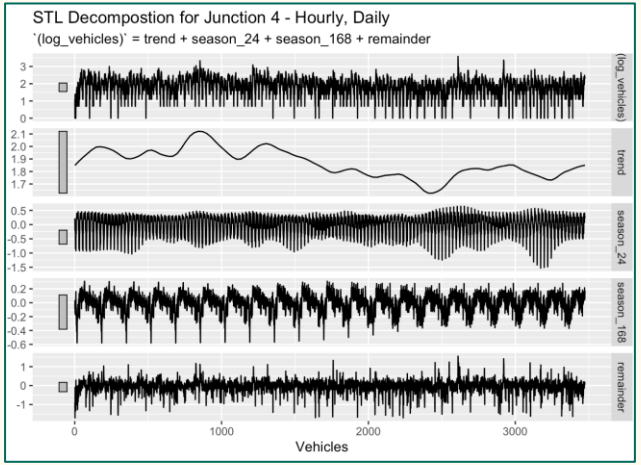


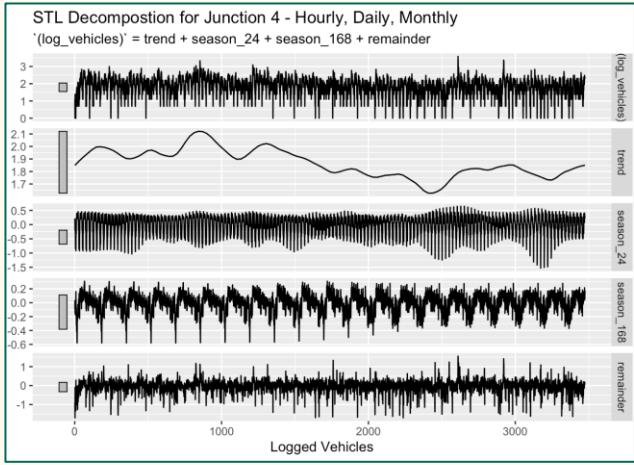
Logged Model with Hourly & Daily Seasonality

	ME	RMSE	MAE	MPE	MAPE	AIC
Test set	-0.03678734	0.3641142	0.2711315	-3.130344	10.77828	81354.95

STL with Multiple Seasonal Periods - Junction 4

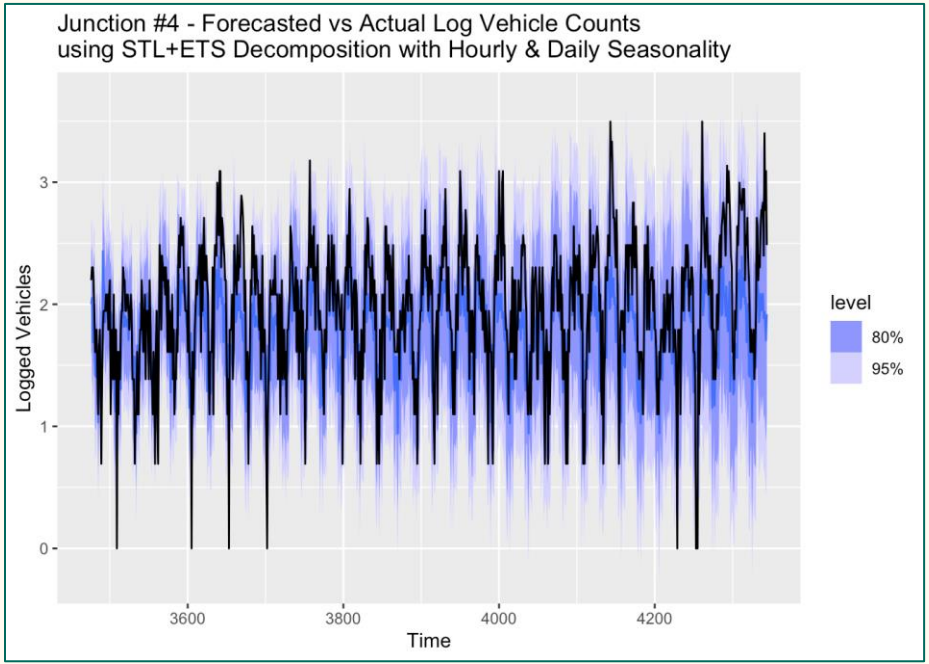


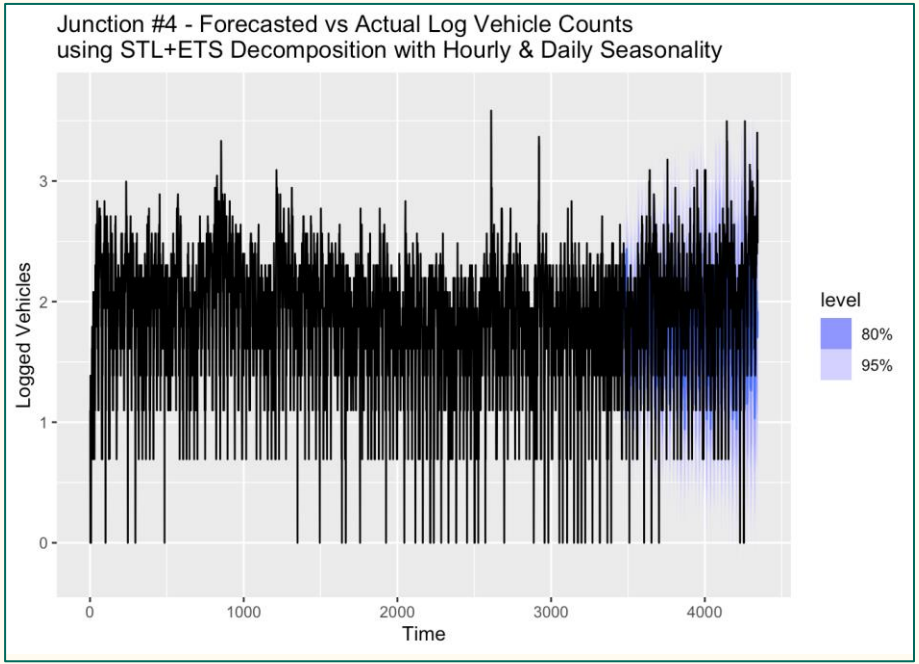




Notice there is not enough data at Junction 4 for STL to capture monthly seasonality

STL with Multiple Seasonal Periods - Junction 4 Forecast





Logged Model with Hourly and Daily Seasonality

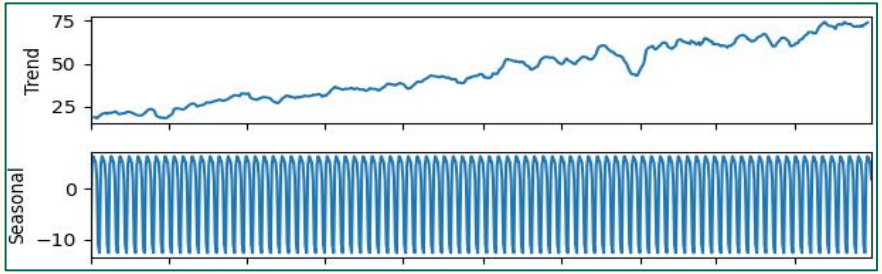
	ME	RMSE	MAE	MPE	MAPE	AIC
Test set	0.1880954	0.4710393	0.3729491	-Inf	Inf	20727.12

Results

- Junction 1: SARIMA (best RMSE)
- Junction 2: SARIMA (best RMSE and AIC)
- Junction 3: SARIMA (best RMSE and AIC)
- Junction 4: SARIMA (best RMSE and AIC)

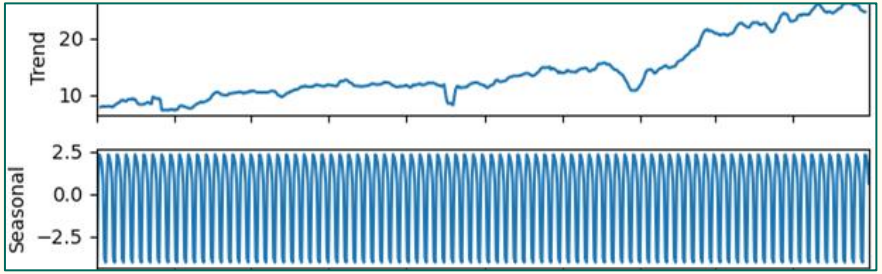
SARIMA models often have lower AIC values compared to STL with multiple seasonal periods due to their simpler structure and fewer parameters. While STL can provide a better fit for complex seasonal data, the increased complexity and higher parameter count usually result in a higher AIC, making SARIMA models preferable when simplicity and efficiency are prioritized.

Minor Question 1



Junction 1

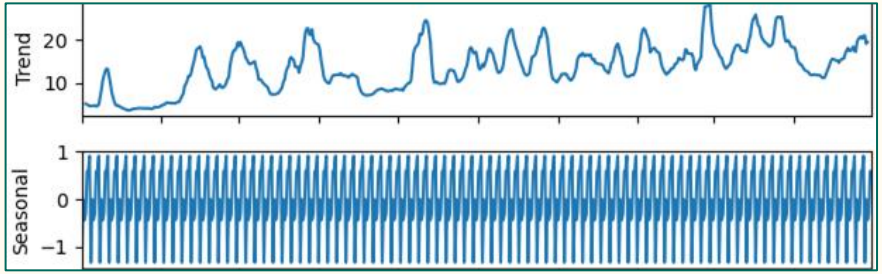
- **Trend Component:** Shows a long-term increase in traffic volumes, influenced by urban development and population growth.
- **Seasonal Component:** Features regular, repeating patterns indicative of daily commuting behaviors.



Junction 2

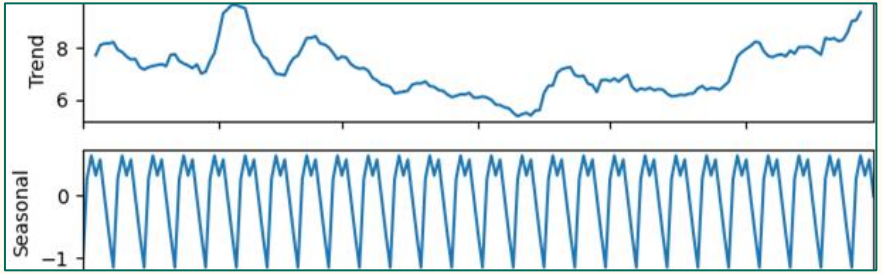
- **Trend Component:** Exhibits a smooth, steady increase in traffic volume due to area development or rising population density.
- **Seasonal Component:** Extremely regular and strong, reflecting consistent daily or weekly traffic cycles.

Minor Question 1



Junction 3

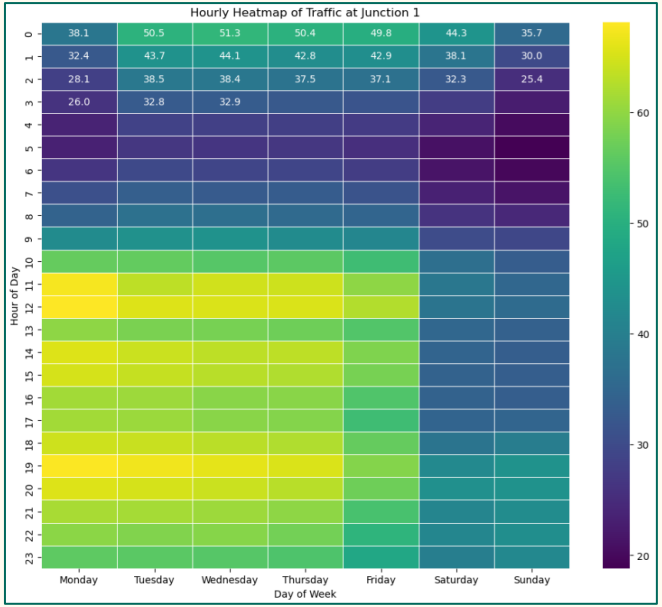
- **Trend Component:** Displays cyclic peaks and troughs influenced by periodic factors like seasonal businesses or school terms.
- **Seasonal Component:** Pronounced and regular, aligning with daily rush hours or weekday vs. weekend traffic.



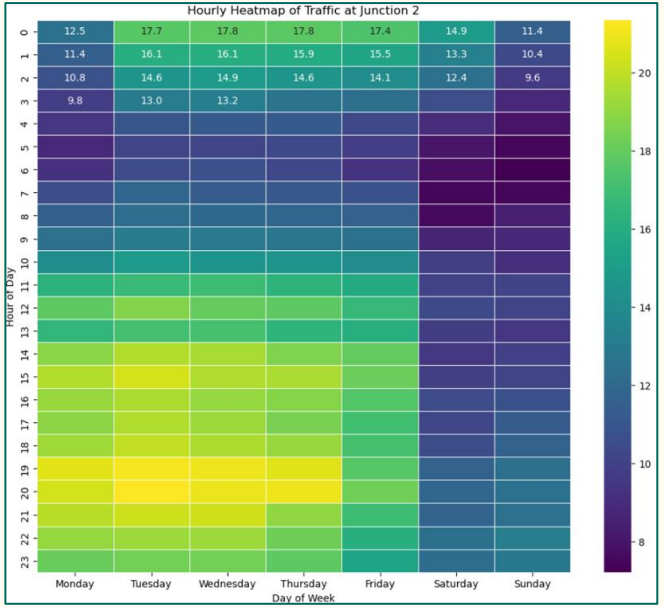
Junction 4

- **Trend Component:** Shows initial variability with several fluctuations due to external factors, followed by a gradual increase driven by area growth or infrastructural enhancements.
- **Seasonal Component:** Less frequent cycles suggesting annual influences such as tourism or major events, important for planning.

Minor Question 2

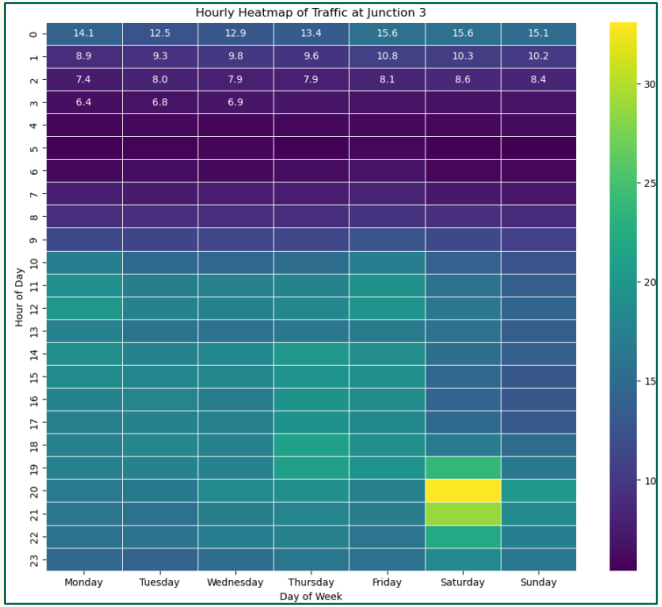


- **Weekday Traffic:** Peaks during 11-12 and 13-15 (work/school commutes and lunch hours) and 18-20 (end of the workday).
- **Weekend Traffic:** More evenly distributed throughout the day with lower volumes, especially early morning.

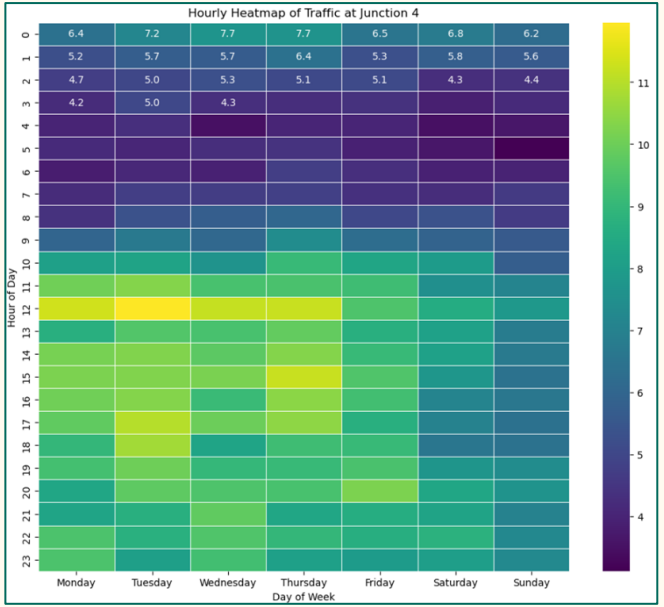


- **Weekday Traffic:** Peaks in the late afternoon to early evening (18-20) at the end of the workday and in the early afternoon (14-17).
- **Weekend Traffic:** More evenly distributed throughout the day with lower volumes, especially in the early mornings.

Minor Question 2



- **Peak Traffic:** Highest volumes occur in the early evening around 20:00.
- **Transport Preferences:** Generally lower traffic suggests a preference for public transport or other vehicles, with an exception of unusually high traffic at 20:00 on Saturdays.



- **Weekday Traffic:** Peaks late morning (11-12) and sees increases in the late afternoon to early evening (14-20), with a notable rise late evening (21-23).
- **Weekend Traffic:** More evenly distributed throughout the day.

Minor Question 3

There are three different models and each have their own way to contribute to traffic pattern analysis and prediction.

1. **SARIMA:** Enhances ARIMA by including seasonal trends, ideal for analyzing predictable fluctuations in traffic patterns
2. **Regression with ARIMA errors:** Merges linear regression and ARIMA to address non-stationarity and clarify time-vehicle count relationships
3. **STL with multiple seasonal periods:** Decomposes time series into trend, seasonal, and residual elements, perfect for complex, multi-seasonal traffic data and improving understanding of trends and seasonal variations

Data Processing

```
Missing Values:
Junction    0
Vehicles    0
ID          0
dtype: int64

Duplicate rows: 0
```

There are no missing values or duplicate rows in this dataset. In order words, this dataset is complete.

```
DateTime    object
Junction     int64
Vehicles     int64
ID           int64
dtype: object
```

There are four variables in the dataset: DateTime, Junction, Vehicles, and ID. For future analyses, DateTime has been converted to DateTime type instead of simple object/string.

Anomaly Detection and Imputations

- Log the data to account for different levels of variance (ACF and PACF)
- Performed Dickey-Fuller and KPSS tests and concluded that our models should account for differencing in order to make it stationary
- Detected and imputed outliers at each junction

Feature Engineering

- **Heteroskedasticity in Time Series:** Variables in time series data often exhibit heteroskedasticity, where the variability of the data increases with the value of the variable
- **Log transformation:** Taking the logarithm of a variable helps to stabilize this increasing variability, leading to a more consistent spread throughout the data
- **Improved model reliability:** This transformation simplifies the data structure, making statistical analyses and predictions more reliable

Obstacles and Future Work

Future Work:

- Real time data capture can help identify policy solutions, such as congestion pricing, to decrease traffic at peak times
- Generative AI and more advanced machine models can help quickly identify different elements of seasonality in the data, allowing us to fit more accurate models

Obstacles:

- Lots of data points at small intervals can make it difficult to model and forecast the data

Source: Sen et al 2013; de Palma & Lindsey 2011

References

Burrows, Michael, et al. “Commuting by Public Transportation in the United States: 2019.” *Census.Gov*, Apr. 2021, www.census.gov/library/publications/2021/acs/acs-48.html.

de Palma, Andre, & Lindsey, Robin. “Traffic congestion pricing methodologies and technologies.” *Transportation Research Part C: Emerging Technologies*, 19(6), 2011. https://doi.org/10.1016/j.trc.2011.02.010

Fedesoriano. “Traffic Prediction Dataset”. *Kaggle*, 2021. https://www.kaggle.com/datasets/fedesoriano/traffic-prediction-dataset?resource=download

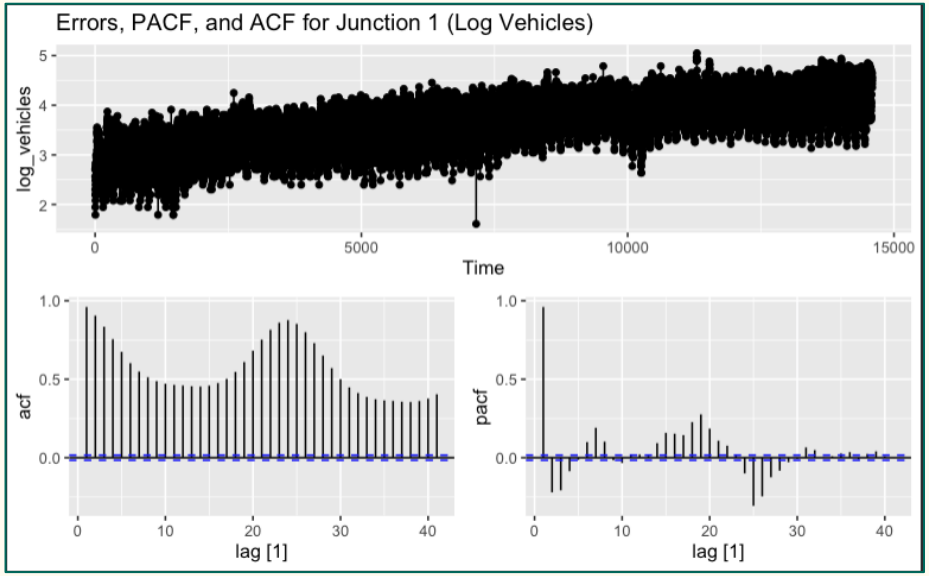
Richter, Felix. “Infographic: Cars Still Dominate the American Commute.” *Statista*, 19 May 2023, www.statista.com/chart/18208/means-of-transportation-used-by-us-commuters/.

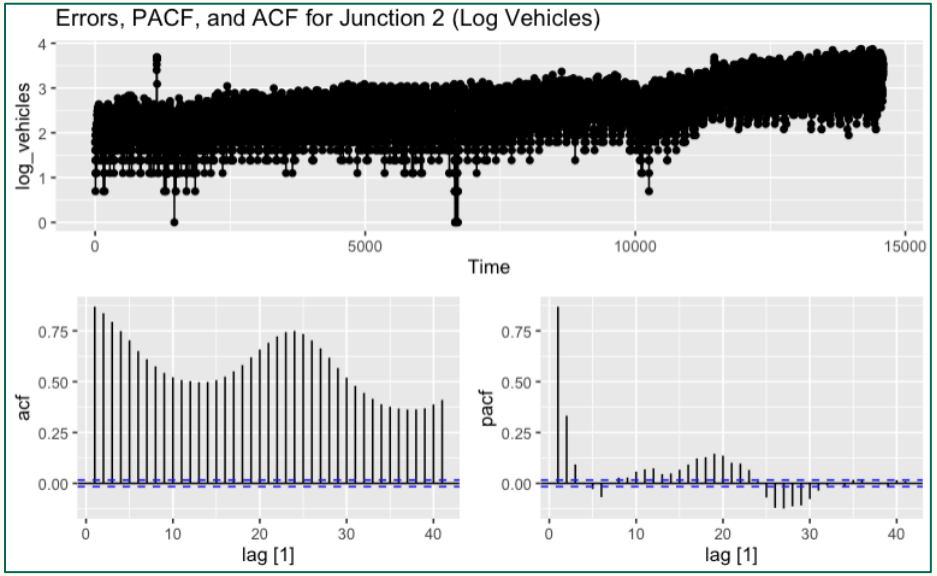
Sen, S., Joe-Wong, C., Ha, S., & Chiang, M. “Smart Data Pricing: Economic Solutions to Network Congestion.” 2013. https://www.cl.cam.ac.uk/teaching/1314/R02/sigcomm/sigcomm-ebook-2013paper3.pdf

Shepardson, David. “US Driving Hits New Record in 2023, Topping Pre-Covid Levels.” Reuters, 8 Feb. 2024, www.reuters.com/world/us/us-driving-hits-new-record-2023-topping-pre-covid-levels-2024-02-08/.

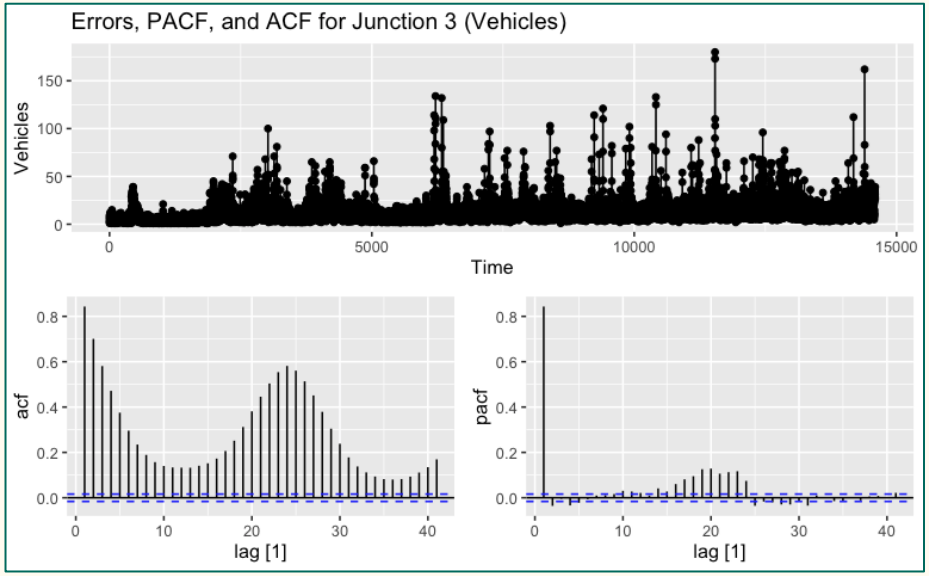
APPENDIX

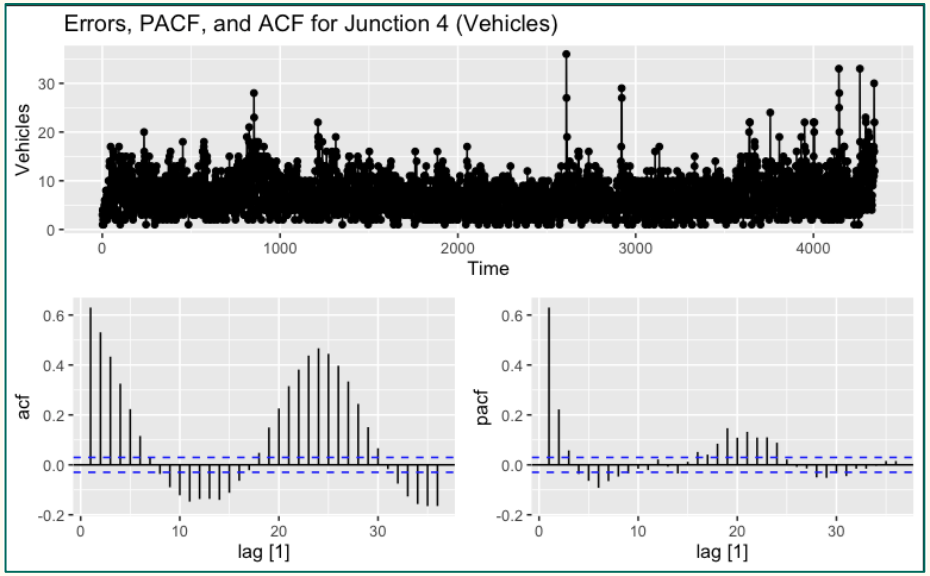
PACF & ACF


Errors, PACF, and ACF for Junction 1 (Log Vehicles)
log_vehicles
Time
acf
pacf
lag [1]
lag [1]


Errors, PACF, and ACF for Junction 2 (Log Vehicles)
log_vehicles
Time
acf
pacf
lag [1]
lag [1]

PACF & ACF


Errors, PACF, and ACF for Junction 3 (Vehicles)
Vehicles
Time
acf
pacf
lag [1]
lag [1]


Errors, PACF, and ACF for Junction 4 (Vehicles)
Vehicles
Time
acf
pacf
lag [1]
lag [1]