



Portland Traffic Analysis: Final Report

Xu, Suru Nelson, Austen
suru@pdx.edu ajn6@pdx.edu

March 5, 2025

Contents

| | |
|--|-----------|
| 1 Executive Summary | 2 |
| 2 Introduction and Background | 2 |
| 3 Data | 2 |
| 3.1 Data Quality | 2 |
| 4 Methods | 3 |
| 4.1 Generalized Additive Model (GAM) | 3 |
| 4.2 Maximum Mean Discrepancy (MMD) | 4 |
| 5 Results | 4 |
| 5.1 Generalized Additive Model (GAM) | 4 |
| 5.2 Maximum Mean Discrepancy (MMD) | 4 |
| 6 Conclusions | 4 |
| A Median Peak Time and Volumes | 4 |
| B GAM Figures | 5 |
| C Data Availability | 11 |
| D Data Issues Deep Dive by Locations | 12 |
| D.0.1 Troutdale | 12 |
| D.0.2 Vista Ridge Tunnel | 13 |
| D.0.3 I5 Bridge | 13 |
| D.0.4 Hoyt | 14 |
| D.0.5 Iowa Street | 14 |
| D.0.6 Lents | 15 |
| D.0.7 Glenn Jackson Bridge | 15 |
| D.0.8 Fairview | 16 |
| D.0.9 Wilsonville | 16 |
| D.0.10 Stafford | 17 |
| D.0.11 North Plains | 17 |
| D.0.12 Beaverton Bethany | 18 |



1 Executive Summary

2 Introduction and Background

The primary objective of this project is to analyze highway traffic count data to identify changes in traffic patterns pre- and post-pandemic, with a specific focus on the I-5 and Glenn Jackson bridges. The Oregon Department of Transportation (ODOT) seeks a comprehensive, data-driven analysis using historical traffic data from Automatic Traffic Recorders (ATR). The report will address the following key questions:

- How have peak traffic periods changed in terms of structure (flatter, sharper, later, or earlier throughout the day) at these key locations?
- How has daily traffic demand shifted across different days of the week at these bridges?
- Are there distinct trends between these two major highway crossings?

The data provided

3 Data

ATR data provides 24/7 traffic volume at specific highway locations across Oregon and has been curated by ODOT's monitoring unit for high reliability. The data consists of vehicle counts over hourly time periods at 12 different highway locations, each consisting of 2 opposite directions (either NB/SB or EB/WB). Some locations have many more years available than others, but in total there are about 3.25 million hours of data recorded over the 24 location and direction combinations.

Many trends and area specific patterns can be easily identified by simply looking at the median values in the data. Table 1 presents an approximation to weekday peak-time and peak-volume based on cubic periodic spline interpolation of hourly medians.

General visible trends seem to be a slightly later peak traffic time in the mornings and location dependent changes in afternoon peak times. The peak volume is lower post-Covid all across the board, with very few exceptions outside of data availability issues.

3.1 Data Quality

Overall the provided dataset is high quality and quite complete. For a visual representation of the time coverage for each location / direction combination, see figure 13. The only issues with data quality we discovered in our analysis fall into two main categories:

1. Daylight Savings (DST) Inconsistencies

The raw provided data is given in Coordinated Universal Time (UTC) format, which does not observe any DST seasonal changes. Filtering the data for duplicated hours results gives occasional duplicates at midnight in April (before 2006) and March suggesting that some timestamps were recorded or converted improperly. A visual of one of the examples where this happens is shown in where counts from March through November are clearly shifted exactly one hour forward during the early morning hours when the traffic is most consistent. For most of the analysis we discuss this shouldn't pose a huge issue, but table 1 approximates what time peak rush hour happens based on hourly medians and occasionally shows a drastic near hour shift of the peak time (see Vista Ridge EB 2018, Wilsonville SB 2024, and Interstate Bridge NB 2024). This makes it difficult to make conclusions about shifts in the traffic structure without more careful analysis. Fortunately, the morning traffic for almost every location is extremely consistent and predictable and identifying and correcting these errors should be possible if that is of interest.

2. Under-Reported and Malformed Data

During the initial interview, the client warned that some of the data is lower than it should be as a result of sensor issues. The client provided us a report of which locations and time periods have such issues and specifying that some data was estimated. One example of estimated data is I5 bridge SB for the year 2018. During this time there was a construction project so traffic could potentially be lower, but reported (potentially estimated) values are so low that the period is not useful for analysis. This is immediately clear in the table 1 for the Interstate Bridge 2018 median entry. This is a primary location and time period of focus for our analysis, so this data is omitted from the models we create.

Another example of malformed data is periods of time with many zero values. In the reliability report it is stated that the 2024 Vista Ridge Tunnel data is reliable, but from 10-07 to 12-23 there are 1440 hourly entries with 0 and 1872 non-zero hourly entries.



One suggestion to address some of these issues would be to add an additional column(s) to the curated data indicating low reliability or estimated entries. This would enable easy filtering when creating models and comparing model results with or without estimated or unreliable data.

A complete listing of potential issues we found in the analysis can be found in appendix D.

4 Methods

4.1 Generalized Additive Model (GAM)

One of the most simple but ubiquitous models in data science is linear regression, but as the name suggests isn't suitable for capturing non-linear relationships in data on its own. A common technique to address this constraint is to choose a suitable non-linear mapping such that the mapped inputs can be modeled with linear regression. GAMs are a class of Generalized Linear Models which provides a framework for selecting such a mapping as the sum over *smooth* functions (often called smoothers). The assumptions for using this kind of model go hand in hand with its additive and smooth properties:

1. The target (output) variable *varies smoothly* in response to the explanatory (input) variables.
2. The function we are modelling can be decomposed into a sum of functions.

Isolating the data to a single location / direction and only considering weekday values, the target variable we model is the hourly traffic counts and explanatory variables which this target varies smoothly with are the hour of the day and month of the year. Then we can model the changes in hourly traffic by creating additional interaction terms between hour of day and categorical variables of interest such as before and after Covid or day of the week.

One of the primary reasons to use GAMs is that due to the additive nature the parameters associated with variables of interest give easily interpretable ways to explain dependency effects between variables. The non-linear smoothers we use are splines. Spline fitting is similar to polynomial fitting, but are generally easier to work with for a variety of reasons (see Runge effect, sklearn example).

Our GAM model is as follows:

$$f(h, m, p) = s_1(h) + s_2(m) + s_3(ph)$$

where h is the hour of the day, m is the month of the year, and p is 0 before Covid or 1 after. The functions s_i are the splines we need to determine. Each s_i can be decomposed into

$$s_i(x) = w_i^T \hat{s}_i(x)$$

where \hat{s}_i is the nonlinear mapping of the explanatory variable x onto a chosen spline basis. `scikit-learn` is a popular Python machine learning library which has APIs to easily create the \hat{s}_i mappings from the desired parameterizations (polynomial order and knot locations). Once the \hat{s}_i are determined, the problem is traditional (regularized) linear regression to solve for each w_i and the resulting s_i give interpretable *partial dependence* functions for their associated variables.

More advanced GAM specific packages such as R's `mgcv` and `gam` or Python's PyGAM have optimization algorithms to determine the appropriate parameters for the smoothing functions automatically. If our model was more complex or included more explanatory variables and interactions, it may be necessary to use such hyper-parameter tuning. After some experimentation we found the periodic cubic splines with knots:

- 15 uniform knots for \hat{s}_1
- 4 uniform knots for \hat{s}_2
- knots located at (0, 4, 6, 8, 10, 15, 17, 19, 24) for \hat{s}_3

generalized well for our data. The knots for \hat{s}_3 were chosen because we are primarily interested in the effect on peak traffic periods, so we choose the knots to be more densely located during these times to give the model more ability to express differences in the interaction while limiting negative effects of over-paramaterizing the model.

4.2 Maximum Mean Discrepancy (MMD)

One of the most fundamental tasks in traffic data analysis is identifying distributional changes over time, particularly before and after major events such as the COVID-19 pandemic. Traditional statistical methods often rely on parametric assumptions, but real-world traffic data is complex, high-dimensional, and often does not follow simple parametric distributions. A common approach to address this issue is to use Maximum Mean Discrepancy (MMD), a kernel-based statistical test that quantifies differences between two distributions without assuming any specific parametric form.



5 Results

5.1 Generalized Additive Model (GAM)

5.2 Maximum Mean Discrepancy (MMD)

6 Conclusions

A Median Peak Time and Volumes

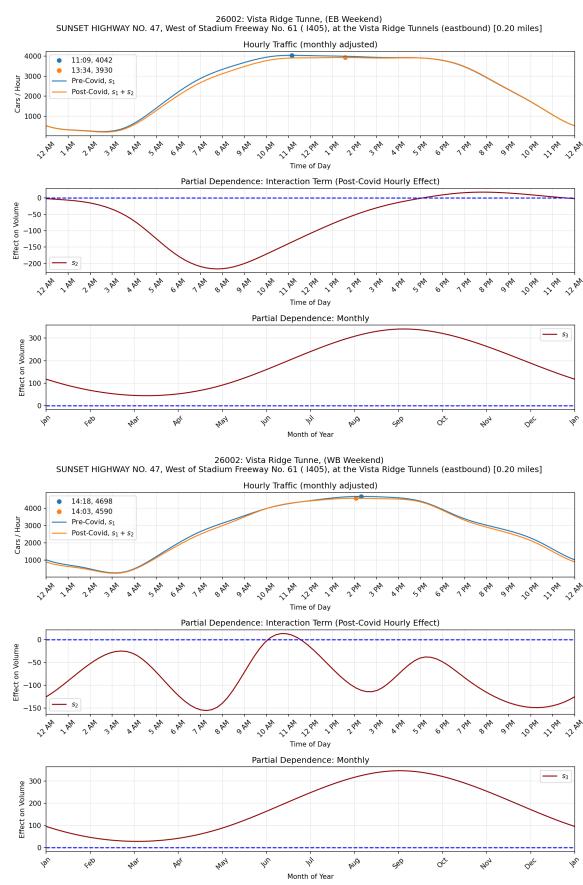
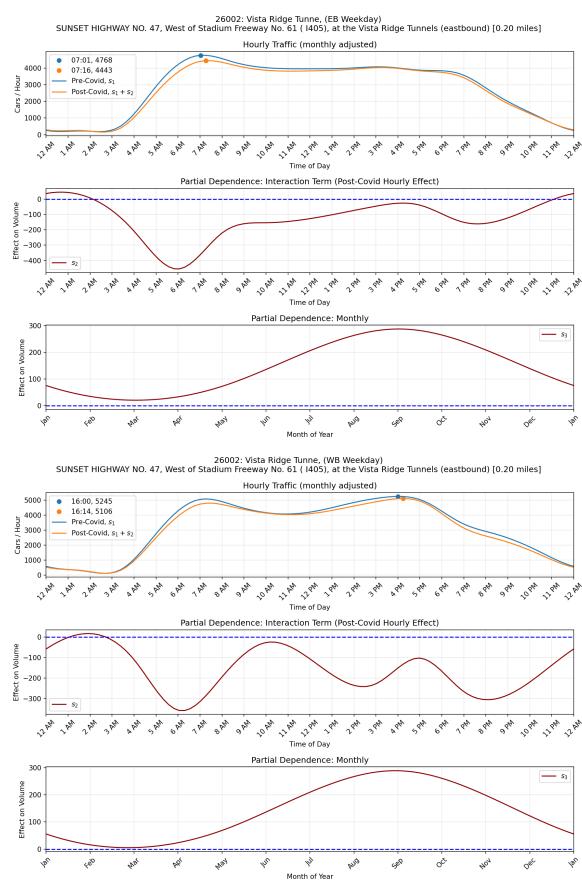
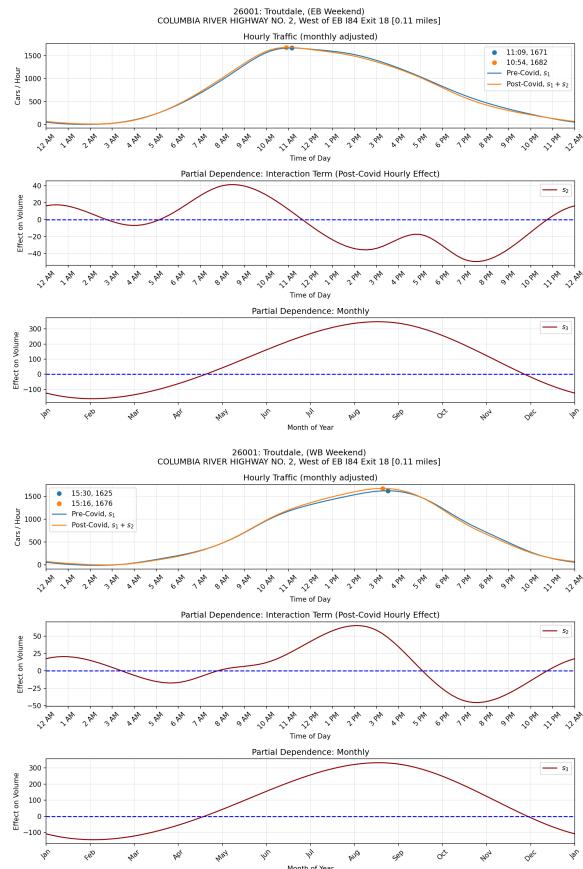
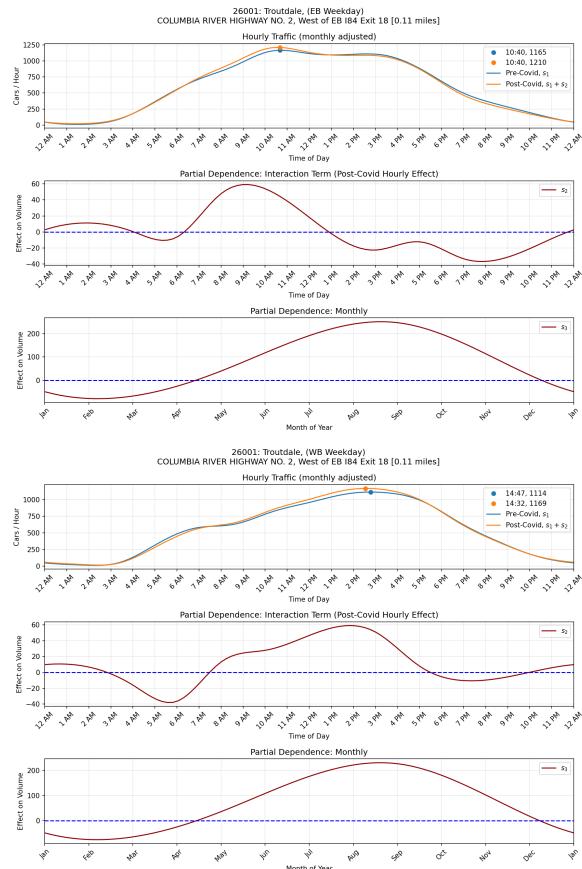
Table 1: Weekday Peak Time Table

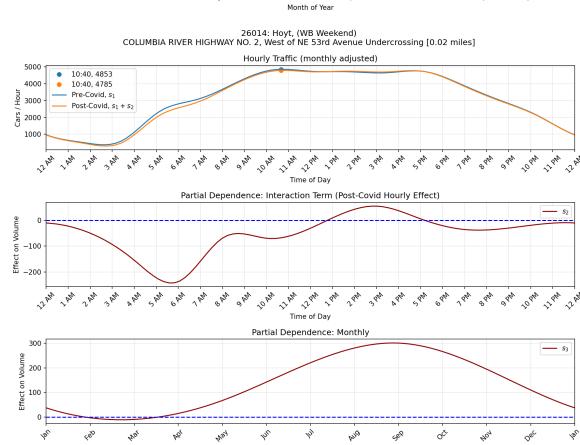
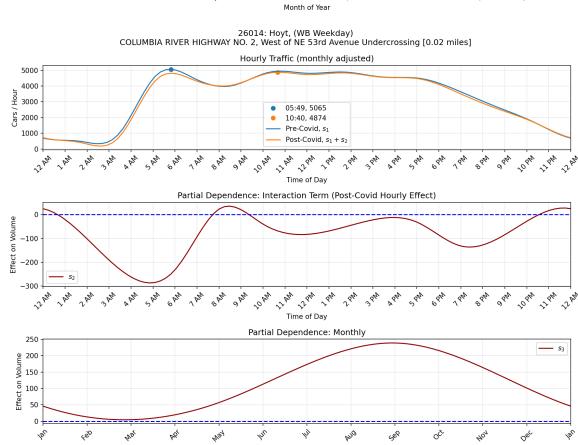
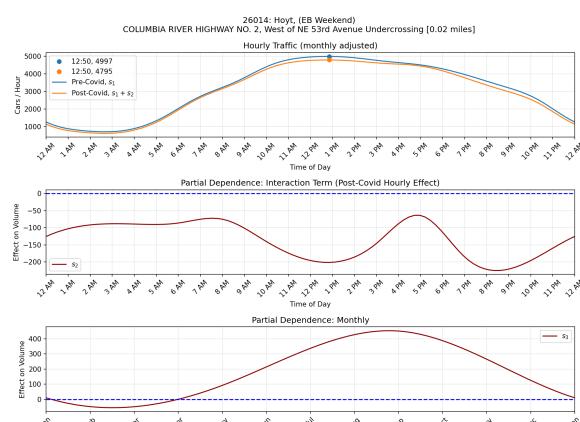
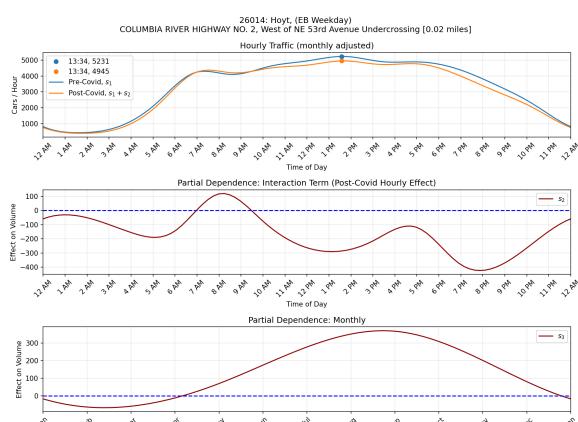
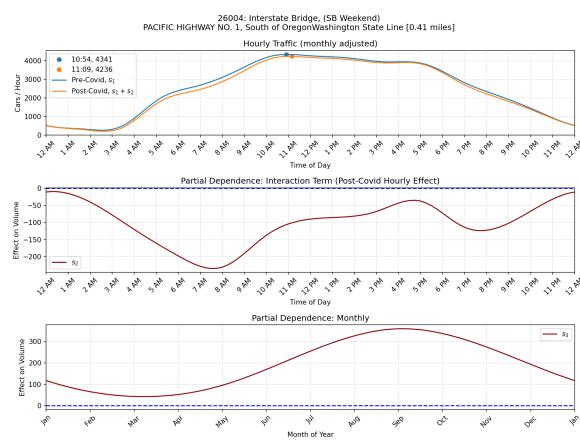
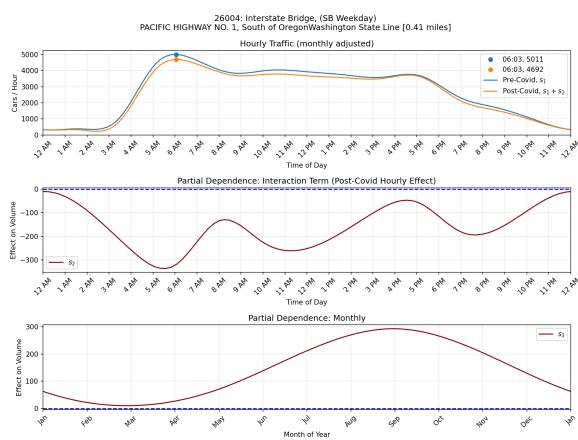
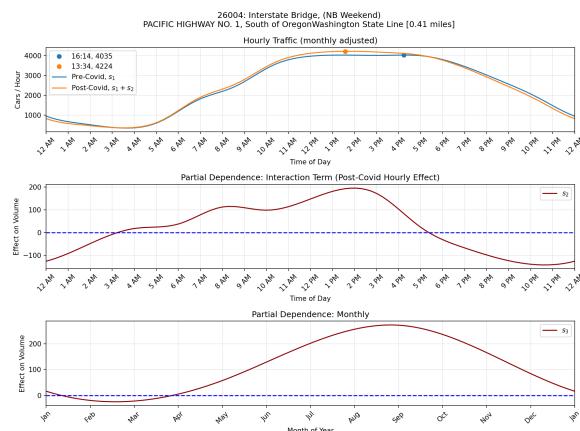
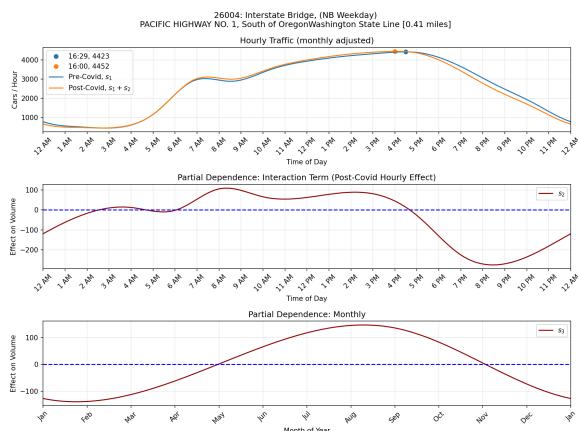
| Location Info | | | Median Peak Time | | | | Median Peak Volume | | | |
|---------------|----------------------|-----|------------------|-------|-------|-------|--------------------|------|------|------|
| ID | Name | Dir | 2018 | 2019 | 2023 | 2024 | 2018 | 2019 | 2023 | 2024 |
| 26002 | Vista Ridge Tunne | EB | 06:41 | 07:35 | 07:21 | 07:30 | 5062 | 4961 | 4735 | 4623 |
| | | WB | 07:12 | 07:18 | 07:27 | 07:24 | 5669 | 5623 | 5244 | 5239 |
| 3016 | Stafford | NB | 14:02 | 14:05 | 14:08 | 14:11 | 3288 | 3424 | 3195 | 3315 |
| | | SB | 06:20 | 06:18 | 06:52 | 06:52 | 3370 | 3390 | 3271 | 3342 |
| 26001 | Troutdale | WB | 14:34 | 14:48 | 14:48 | 14:54 | 1188 | 1252 | 1219 | 1246 |
| | | EB | 10:34 | 10:26 | 10:34 | 10:03 | 1224 | 1313 | 1213 | 1282 |
| 26024 | Glenn Jackson Bridge | NB | 15:11 | 15:11 | 15:17 | 14:45 | 7184 | 7177 | 6781 | 6664 |
| | | SB | 06:15 | 06:18 | 06:49 | 06:20 | 7511 | 7430 | 6334 | 6328 |
| 26014 | Hoyt | EB | 13:56 | 13:53 | 13:48 | 13:48 | 5597 | 5596 | 5307 | 5255 |
| | | WB | 05:40 | 05:34 | 06:00 | 06:00 | 5870 | 5831 | 5443 | 5411 |
| 3011 | Wilsonville | SB | 15:03 | 14:51 | 14:40 | 15:40 | 3749 | 3756 | 3602 | 3621 |
| | | NB | 05:57 | 05:54 | 06:06 | 06:03 | 3885 | 3895 | 3823 | 3924 |
| 26016 | Iowa Street | NB | 06:29 | 06:29 | 07:09 | 07:07 | 5707 | 5678 | 5356 | 5319 |
| | | SB | 16:35 | 16:35 | 16:32 | 16:32 | 5391 | 5380 | 4838 | 4793 |
| 26022 | Lents | NB | 06:20 | 06:20 | 07:01 | 06:58 | 5066 | 5071 | 5198 | 5155 |
| | | SB | 16:29 | 16:35 | 16:32 | 16:32 | 5524 | 5578 | 5287 | 5214 |
| 34010 | Beaverton-Bethany | WB | 07:15 | 07:15 | 16:24 | 16:24 | 5040 | 5196 | 4509 | 4552 |
| | | EB | 16:03 | 16:03 | 06:49 | 06:52 | 4574 | 4498 | 4255 | 4283 |
| 26028 | Fairview | EB | 16:26 | 16:35 | 16:18 | 16:32 | 4468 | 4522 | 4375 | 4367 |
| | | WB | 06:29 | 06:32 | 06:41 | 06:35 | 4069 | 4106 | 3939 | 3834 |
| 26004 | Interstate Bridge | NB | 16:26 | 16:12 | 16:03 | 16:52 | 4636 | 4807 | 4637 | 4383 |
| | | SB | 06:29 | 05:34 | 06:09 | 06:00 | 1906 | 5647 | 5262 | 5021 |
| 34007 | North Plains | WB | 16:03 | 16:03 | 15:52 | 15:58 | 1102 | 1135 | 1086 | 1137 |
| | | EB | 06:35 | 06:38 | 06:49 | 06:52 | 927 | 950 | 809 | 823 |

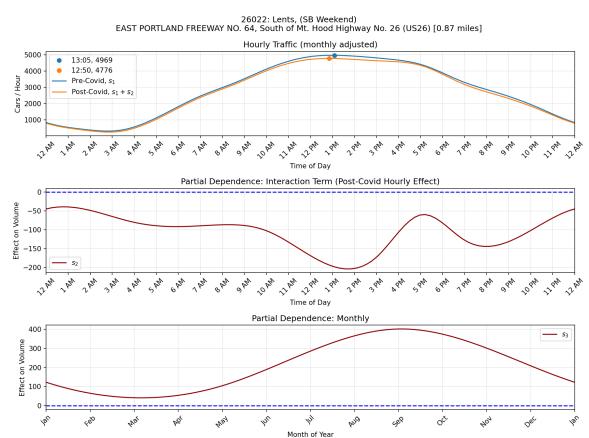
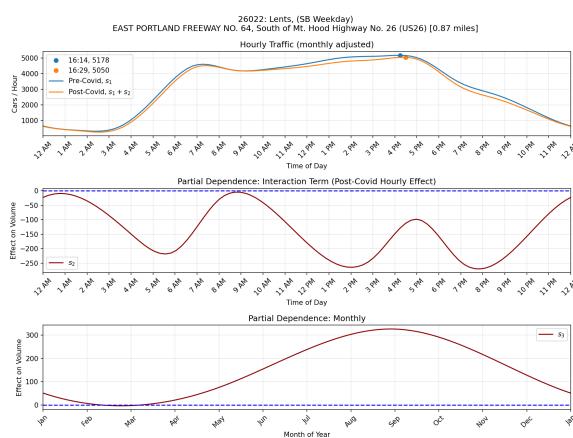
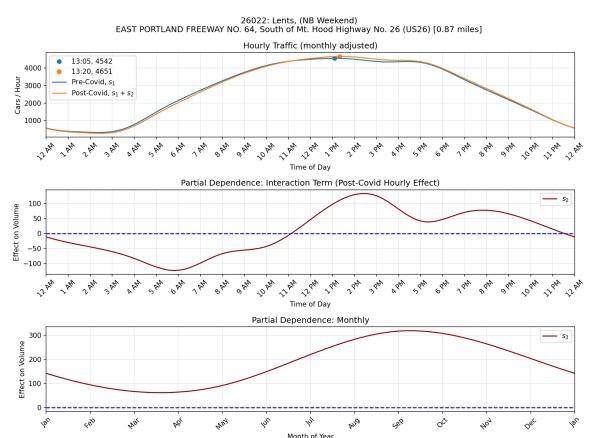
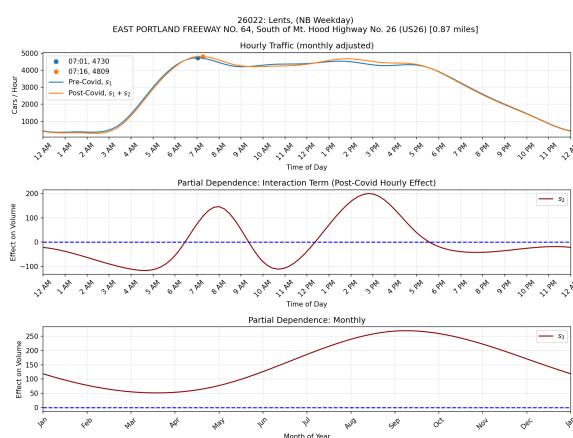
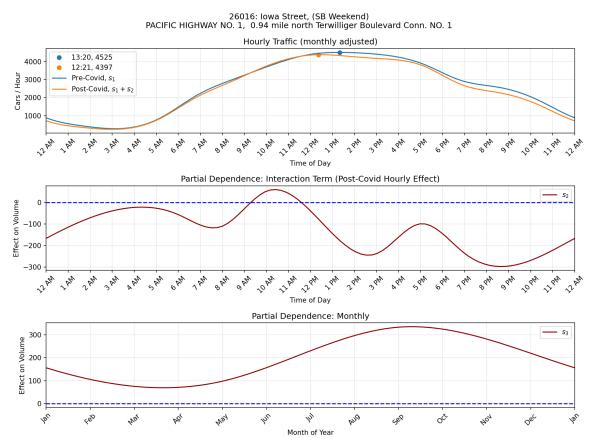
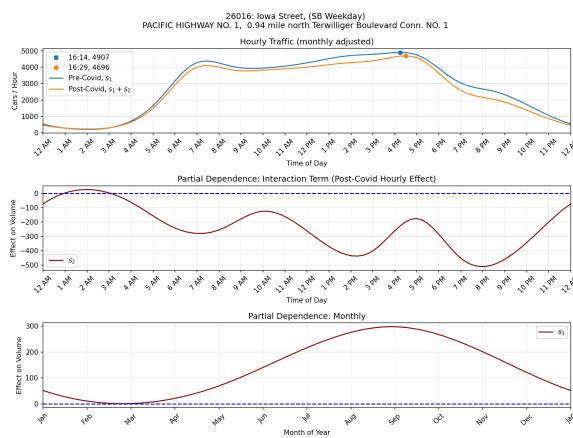
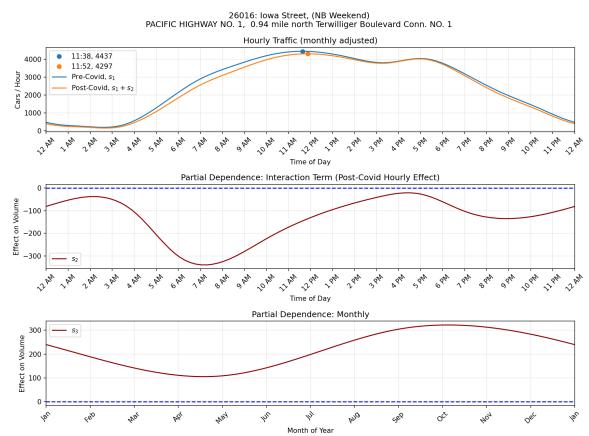
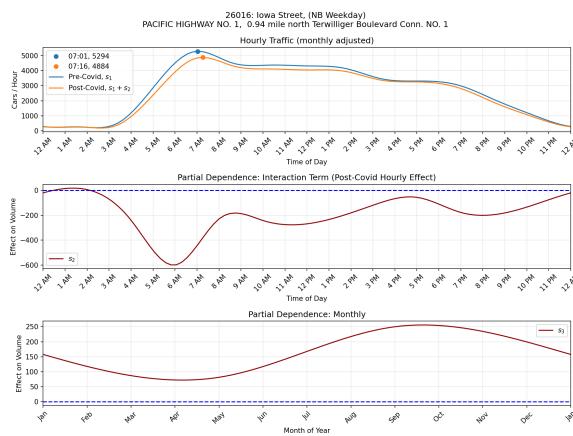
Hourly median traffic volume for each entry (location, direction, year) for weekdays is interpolated with a cubic spline to find approximate peak traffic time.

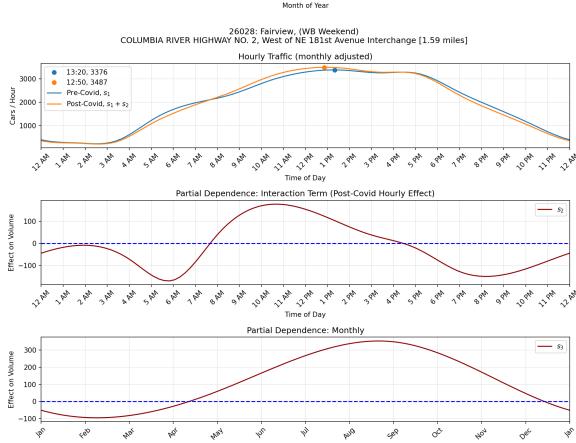
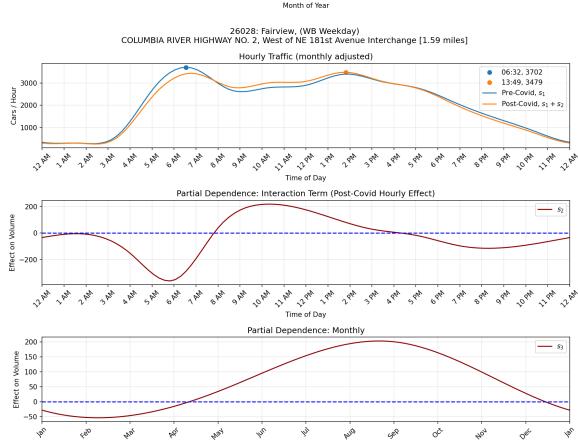
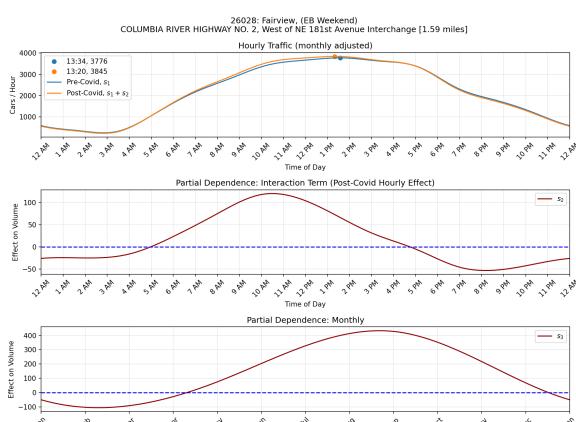
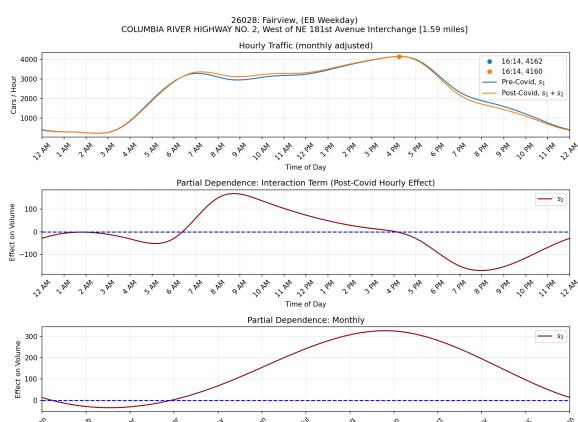
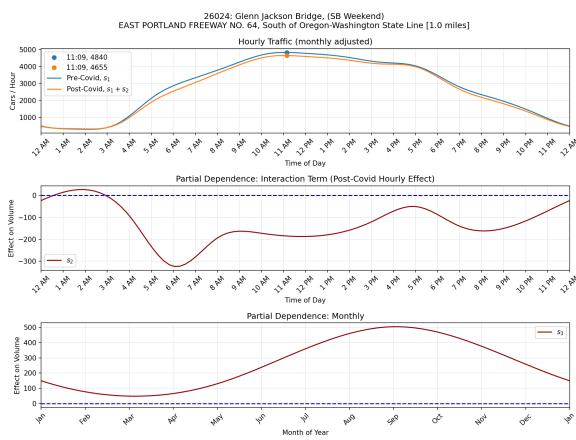
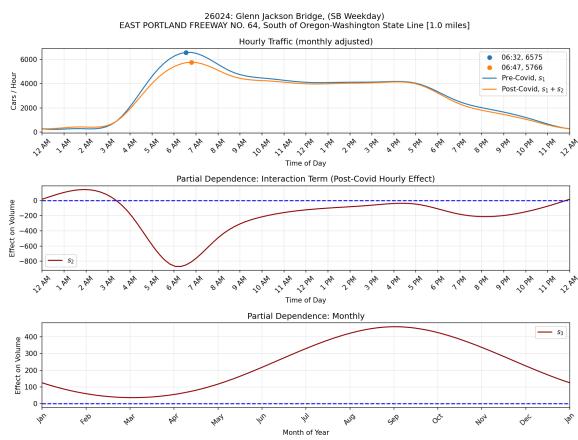
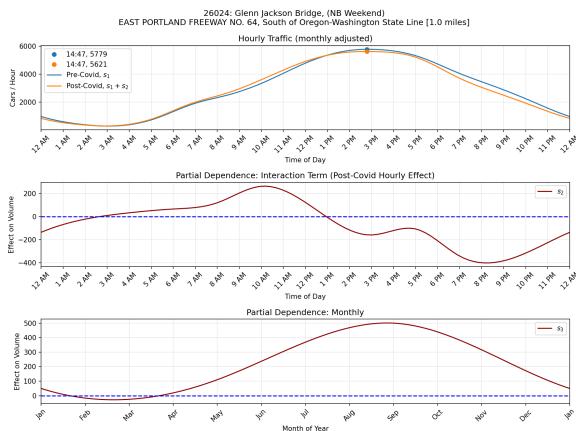
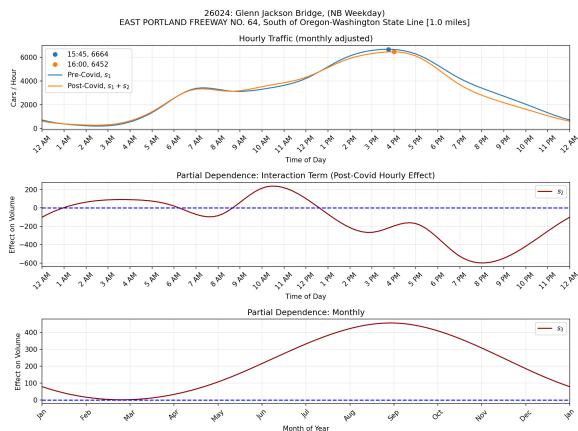


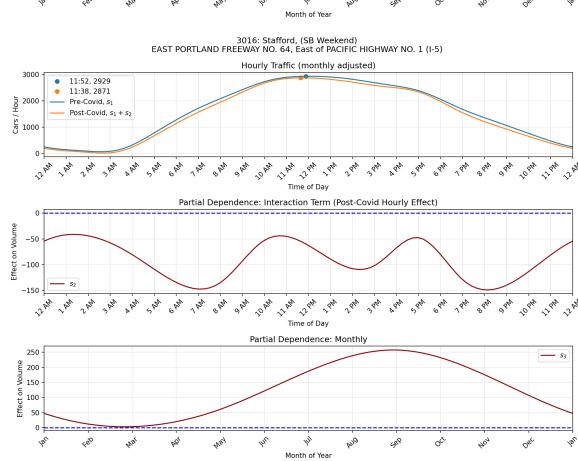
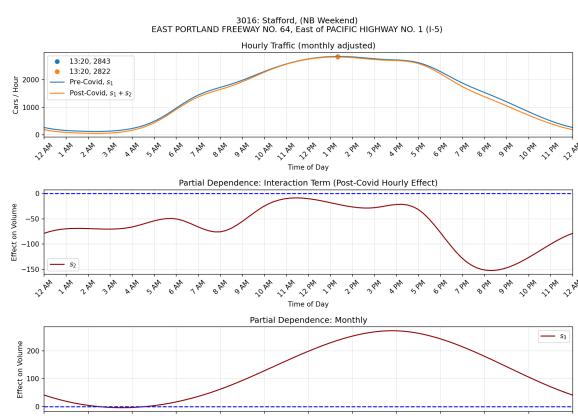
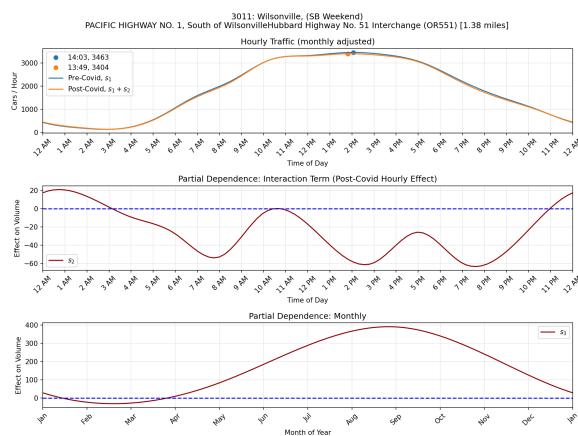
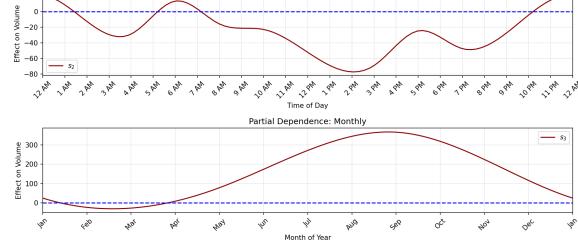
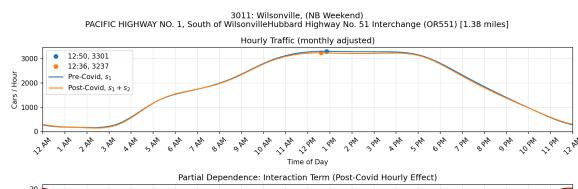
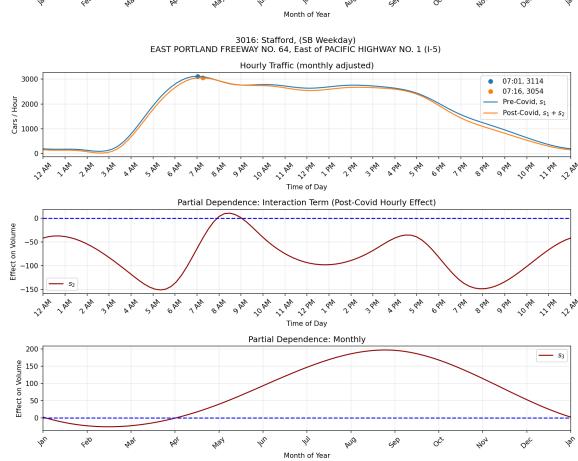
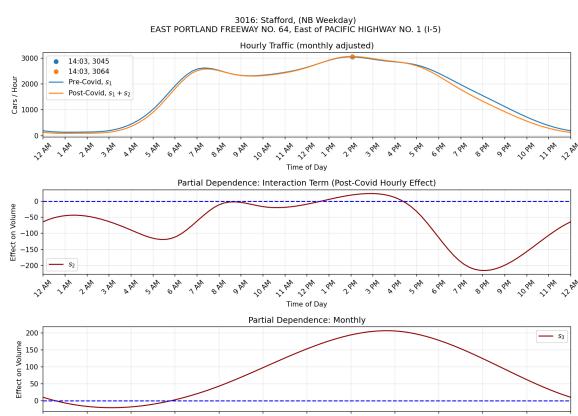
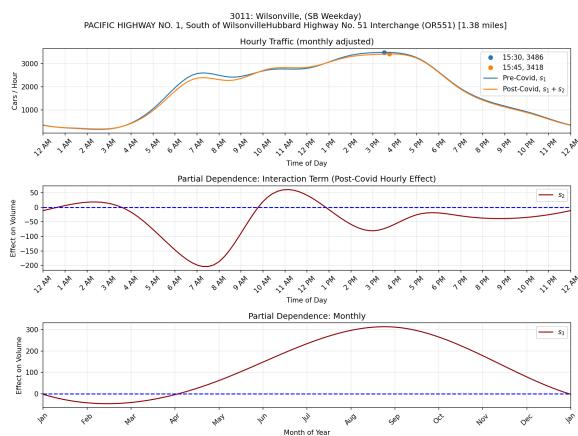
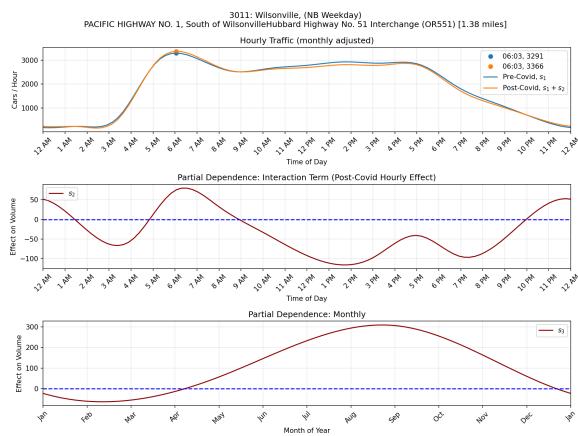
B GAM Figures

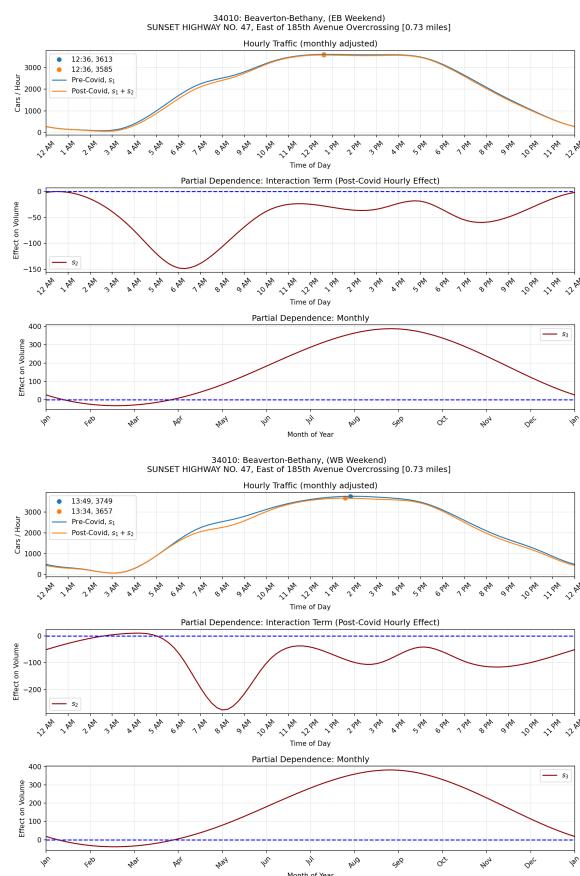
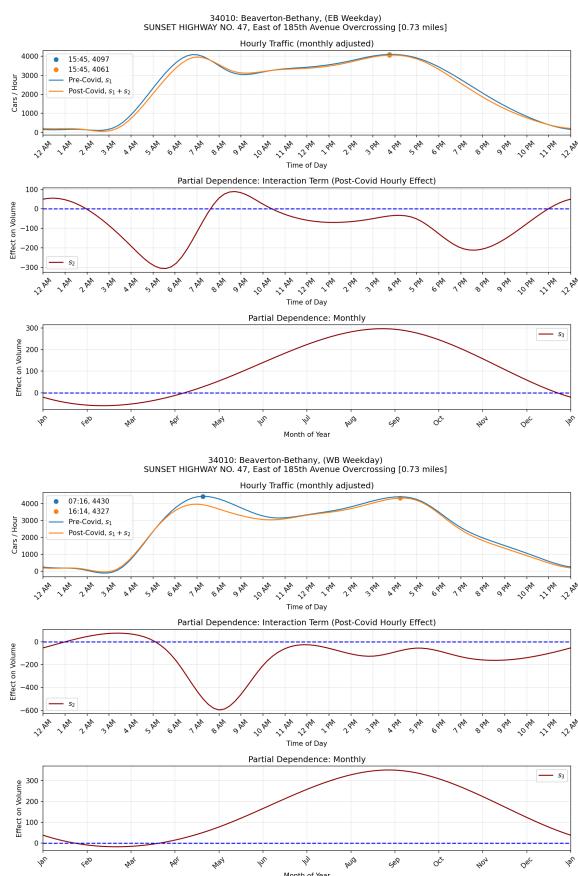
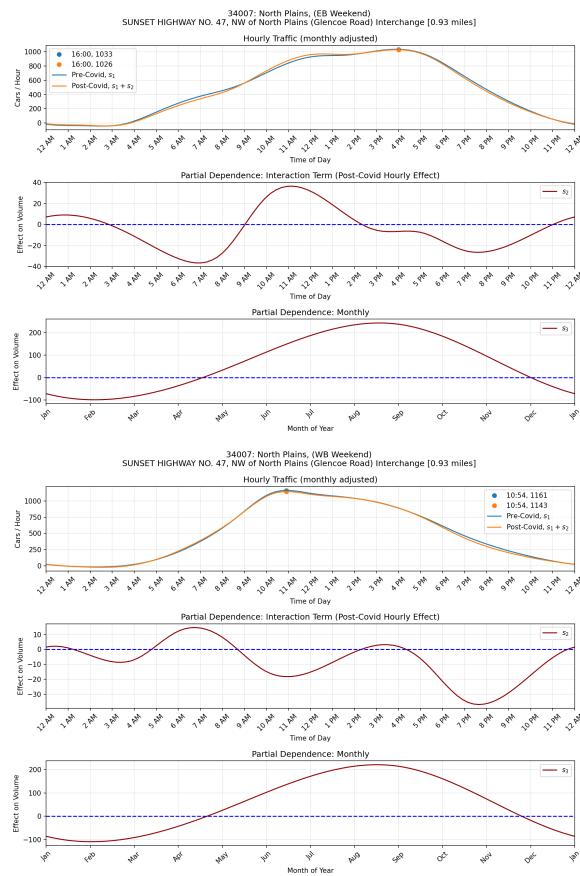
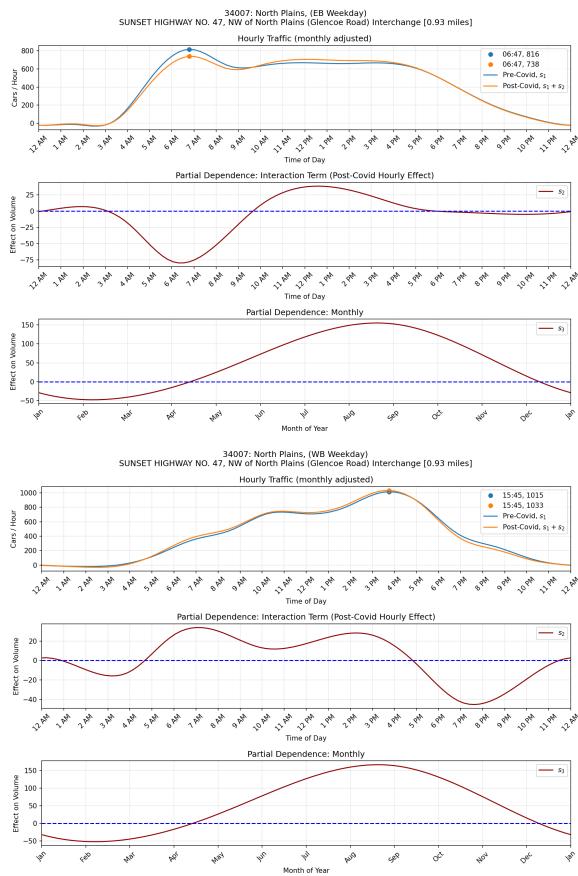














C Data Availability

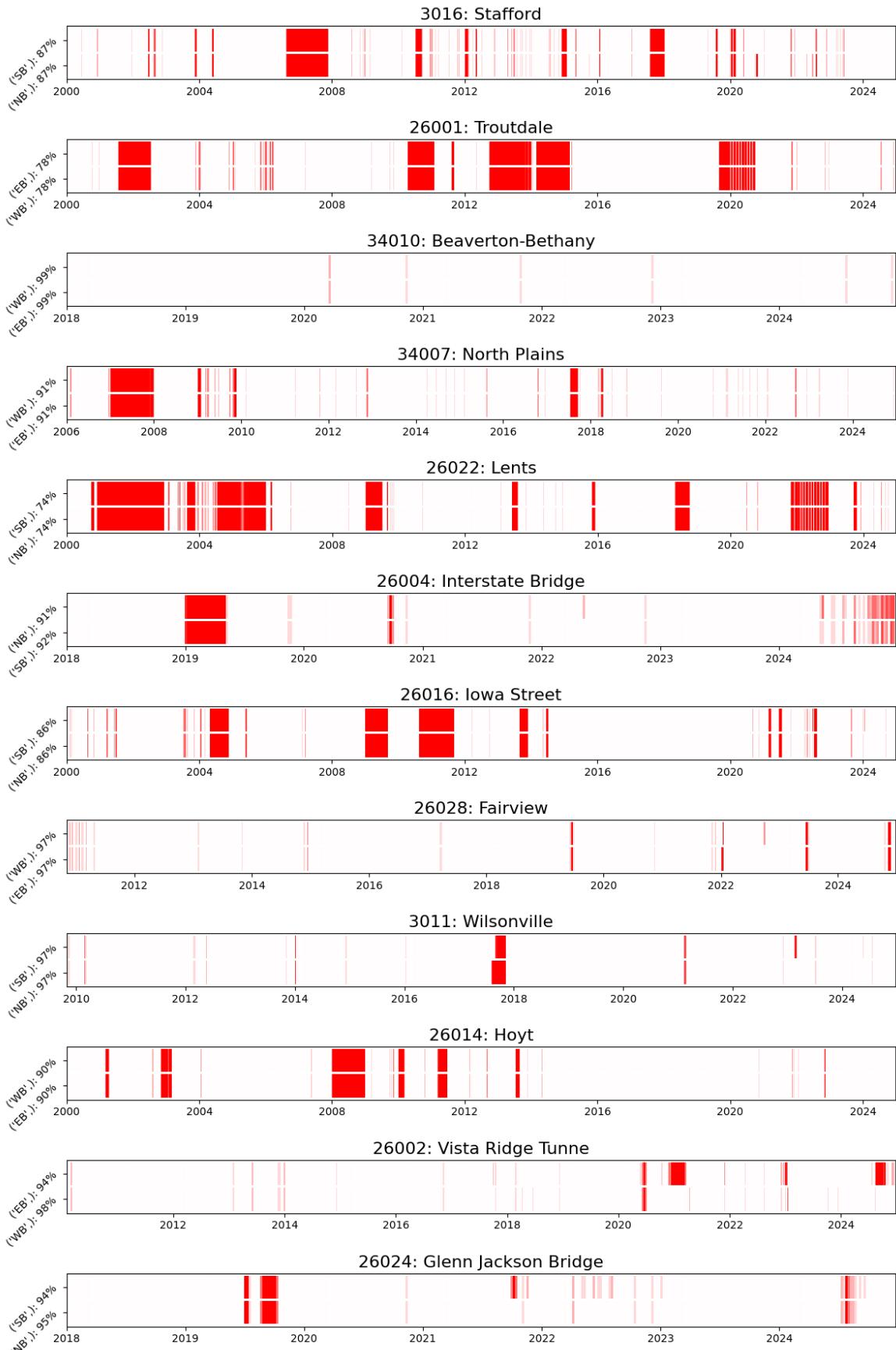


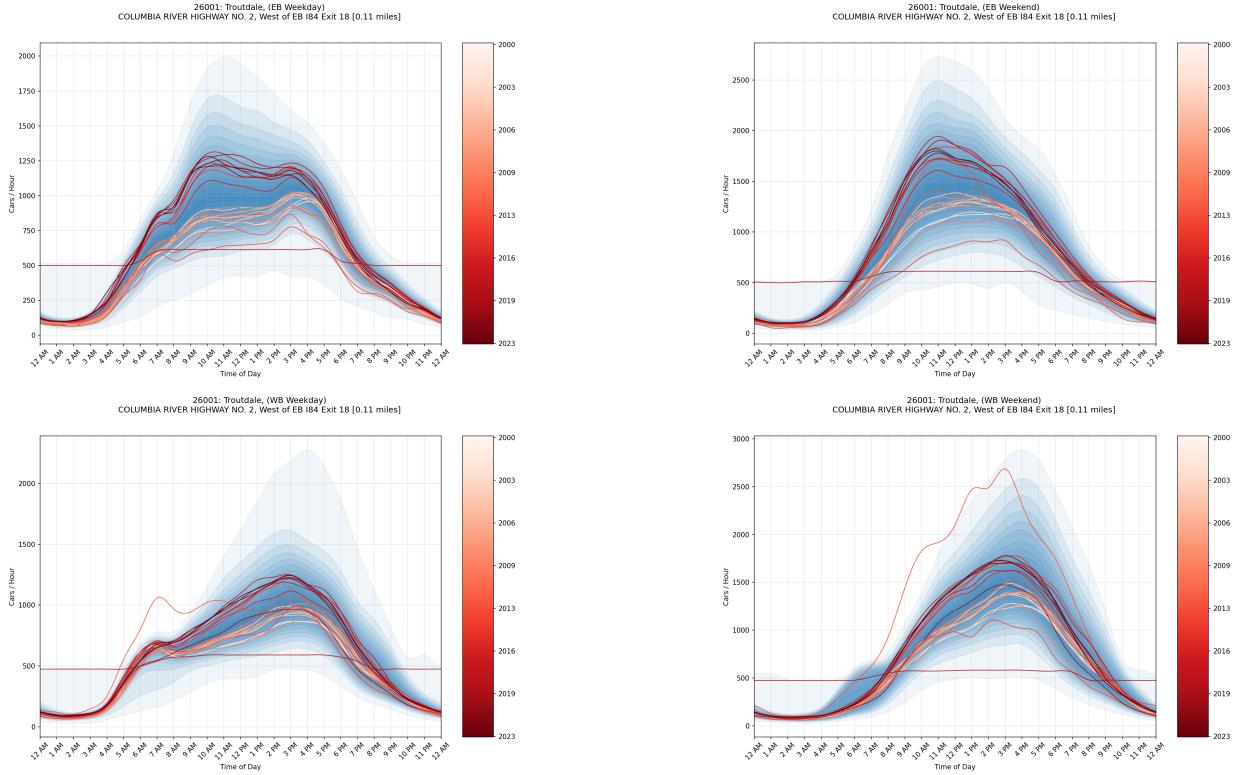
Figure 13: Red indicates missing data, opacity indicates how much of that week is missing. This doesn't account for "0" data which is technically missing and misrepresented.



D Data Issues Deep Dive by Locations

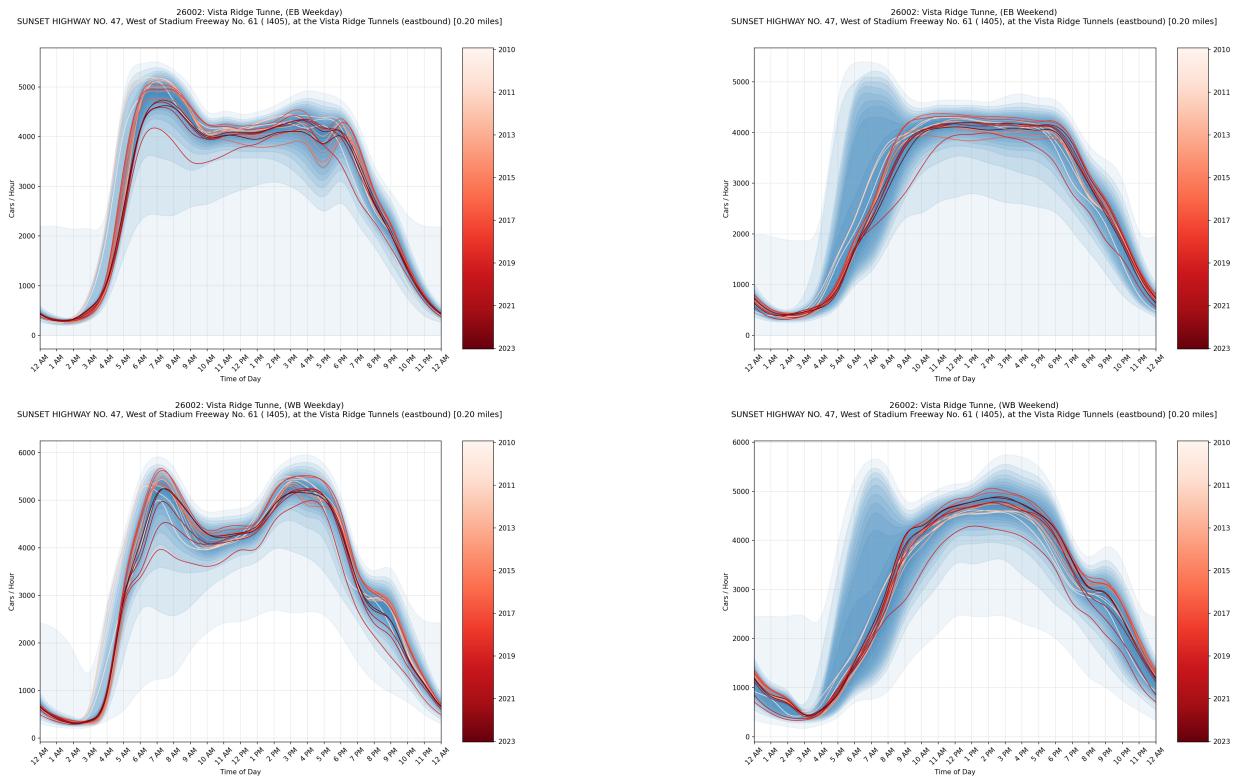
These plots provide a visual for similar data presented in table 1 but includes every year available in the dataset with quantile shading. Each line is the periodic cubic spline interpolation of the hourly medians/quantiles for each (location, direction, weekday/weekend) tuple. Some strange data issues become apparent in these plots, but they are almost entirely only in years we aren't concerned with in this analysis.

D.0.1 Troutdale

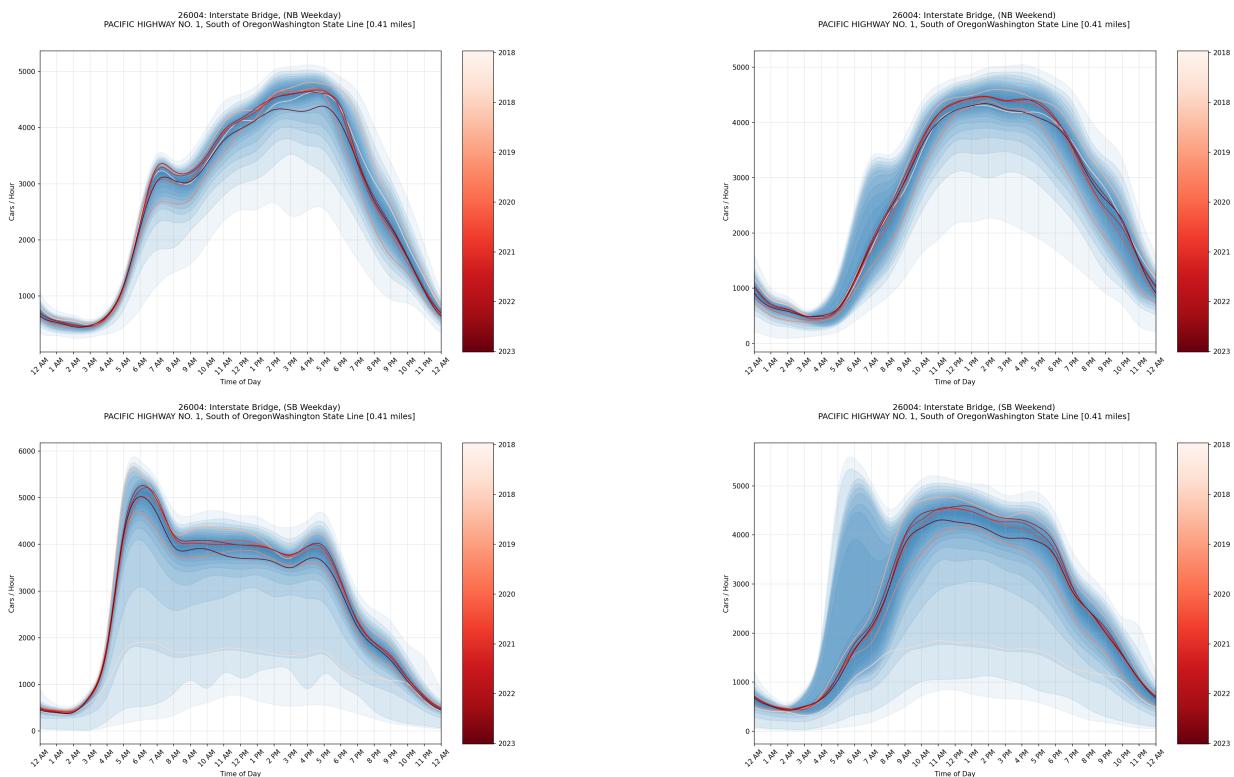




D.O.2 Vista Ridge Tunnel

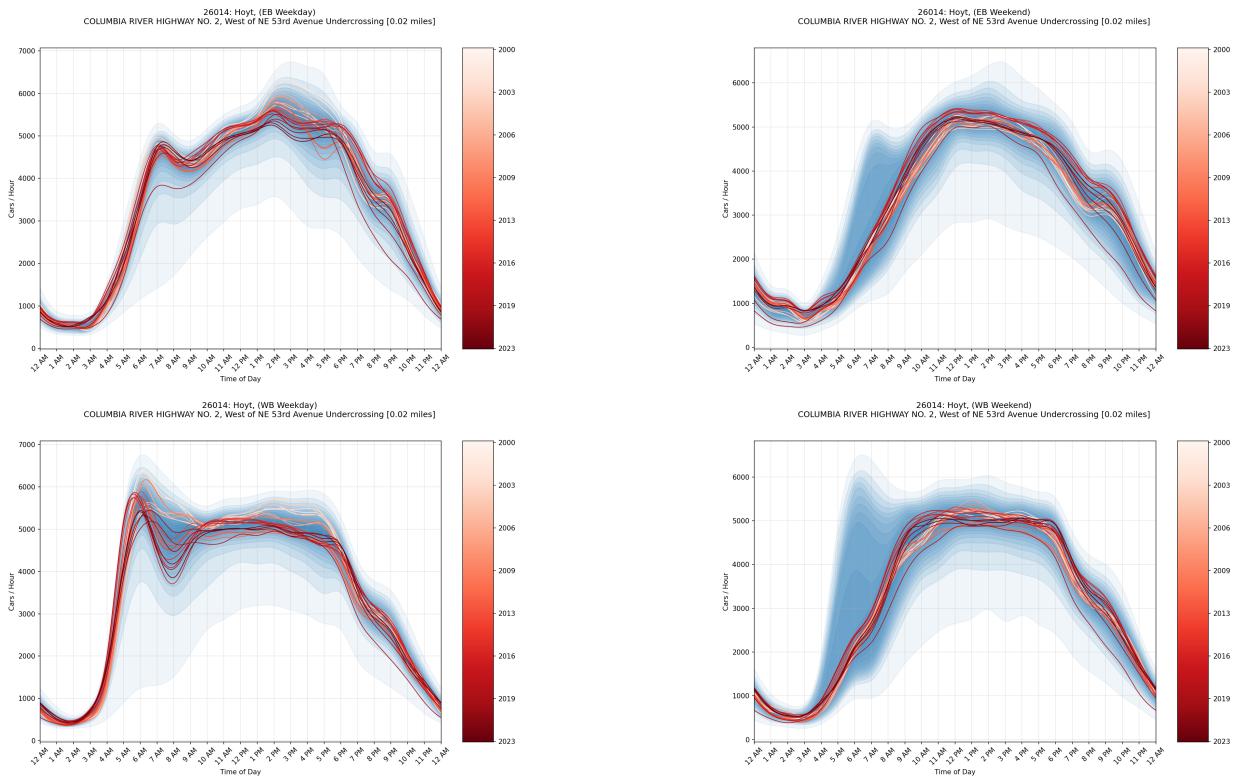


D.O.3 I5 Bridge

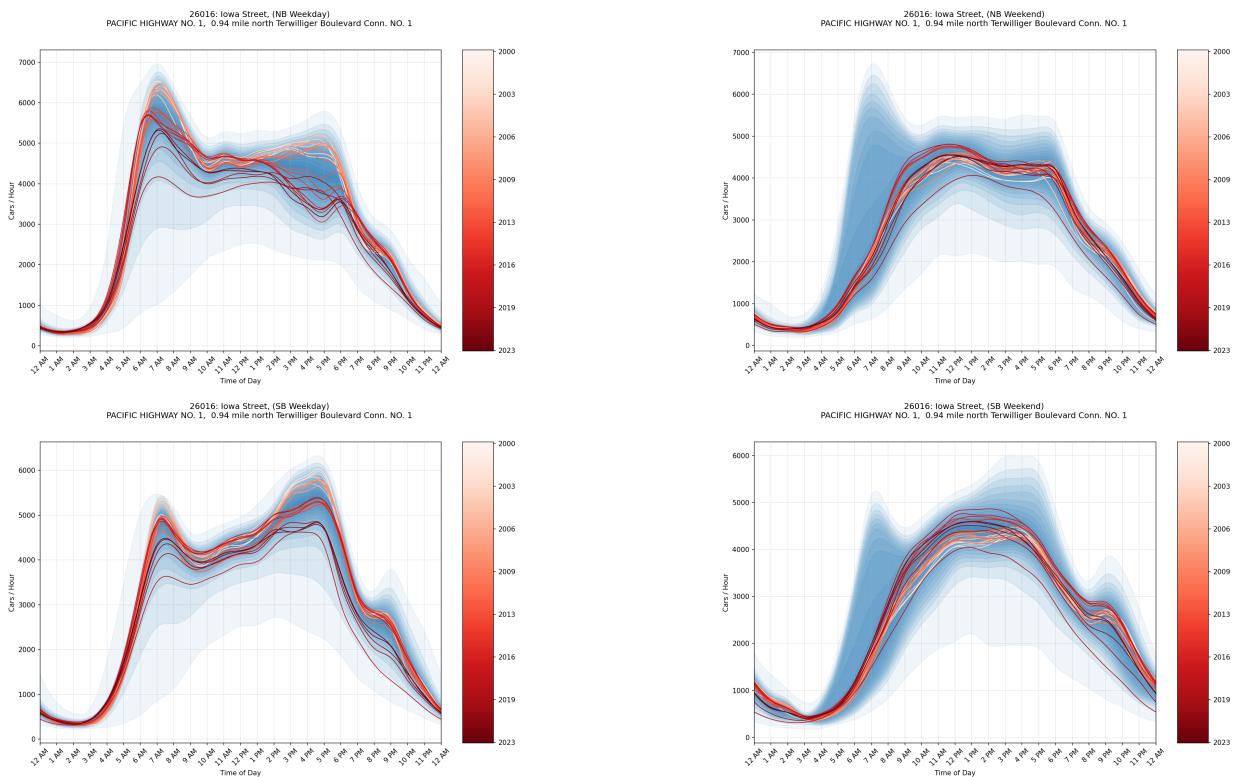




D.O.4 Hoyt

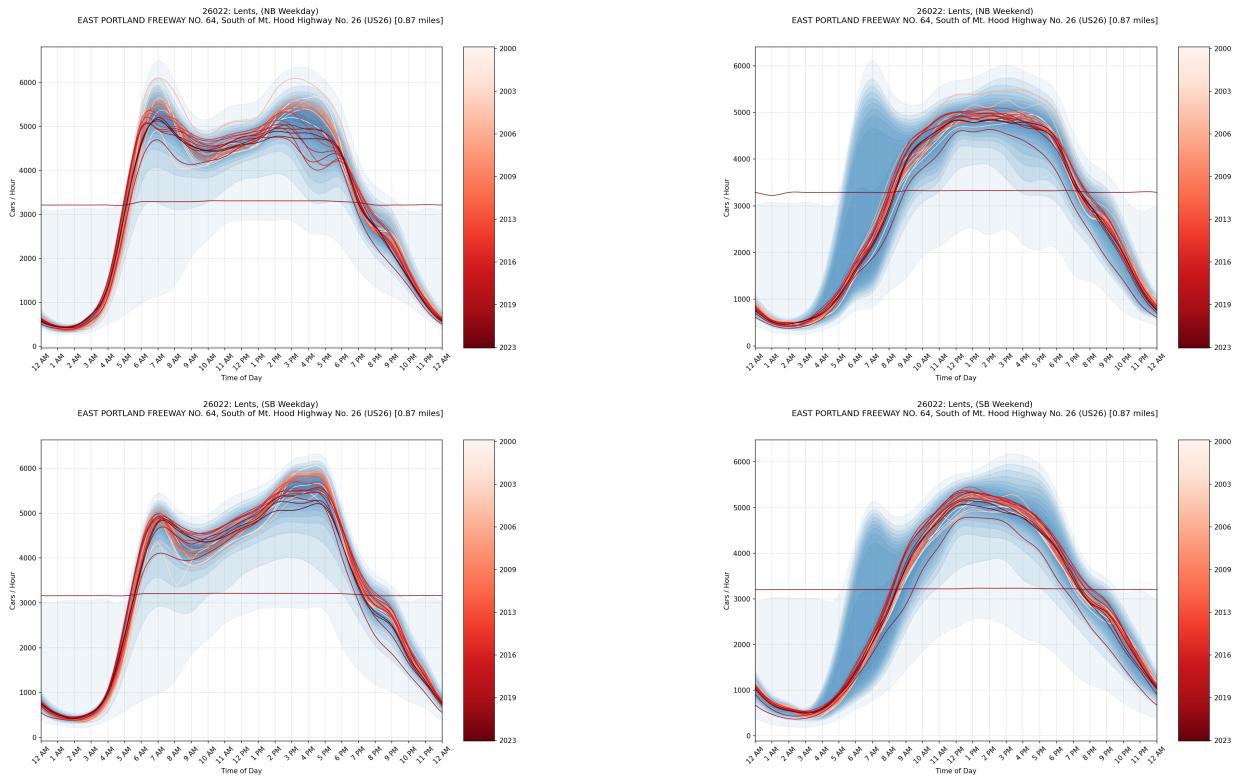


D.O.5 Iowa Street

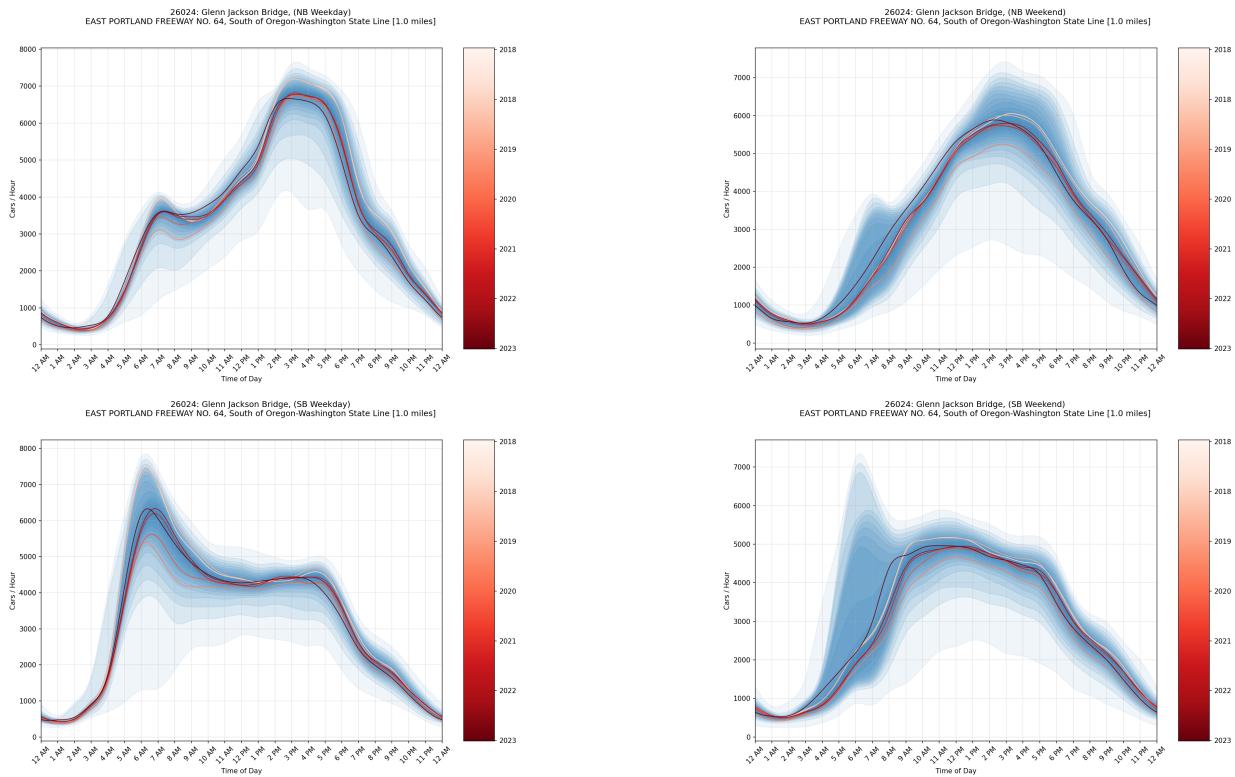




D.O.6 Lents

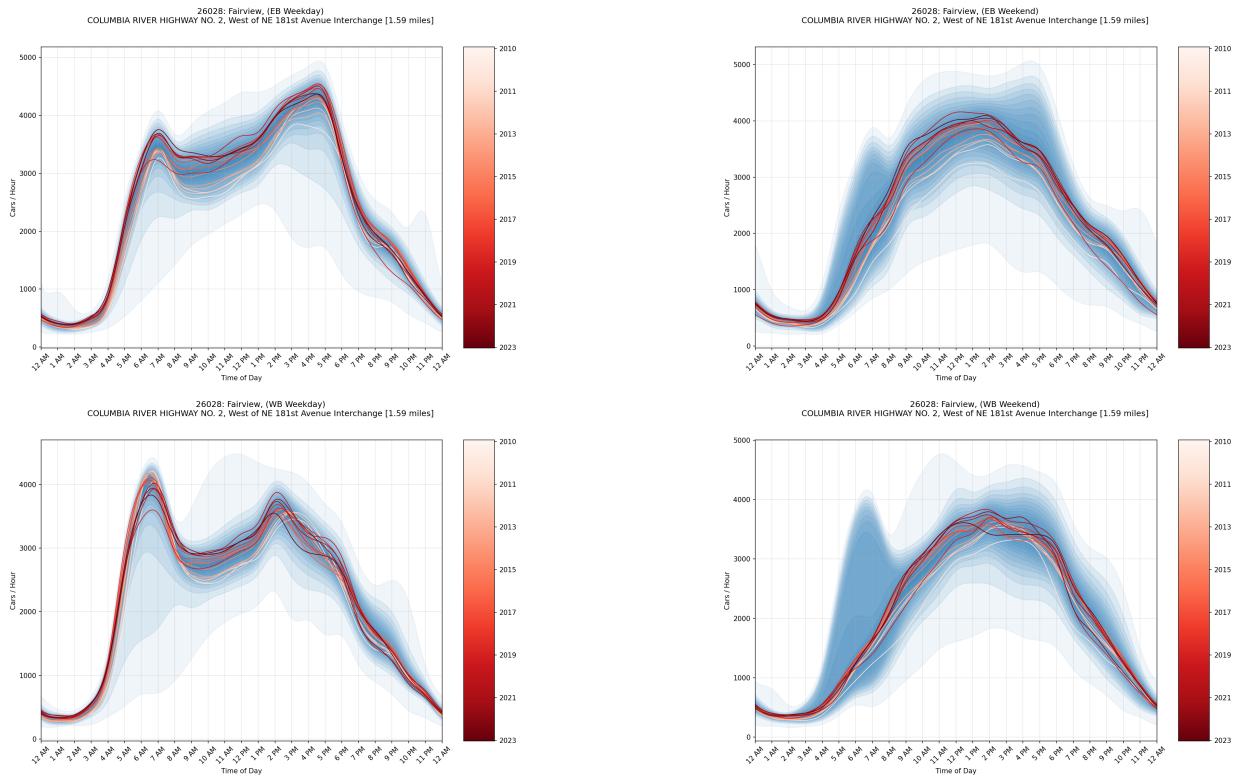


D.O.7 Glenn Jackson Bridge

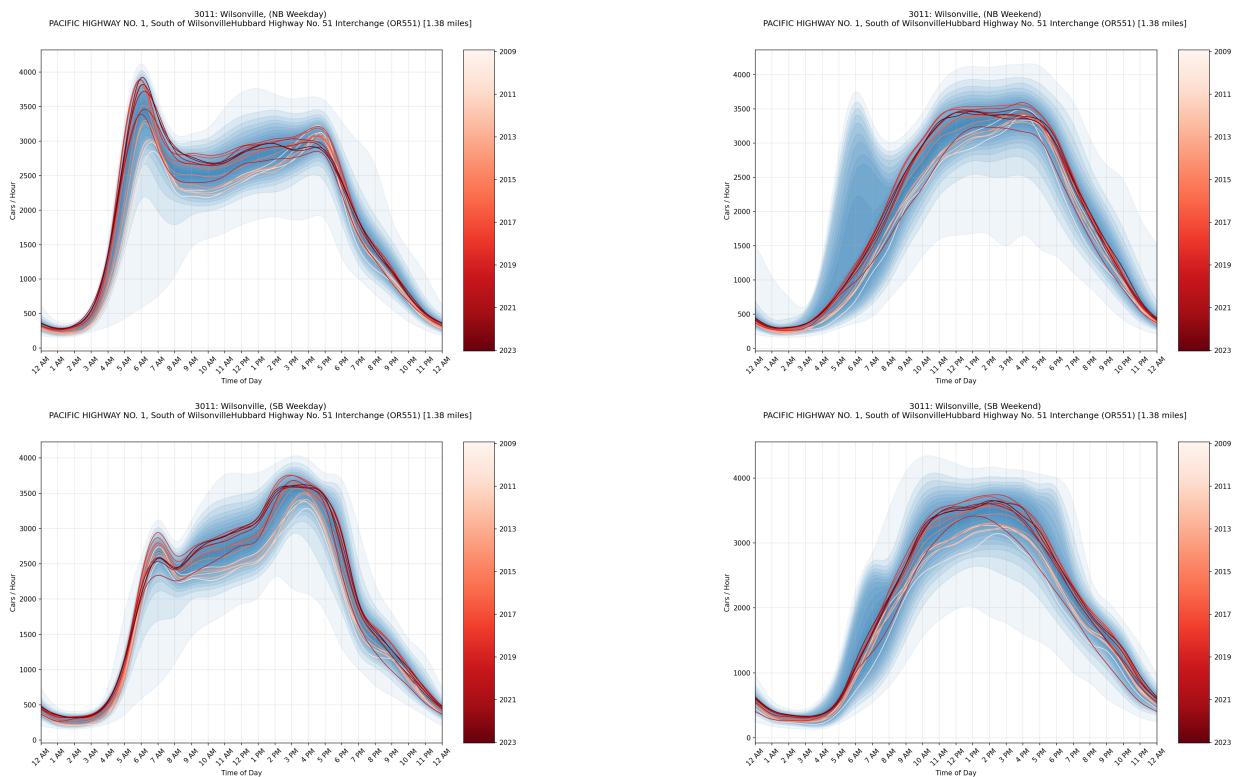




D.O.8 Fairview

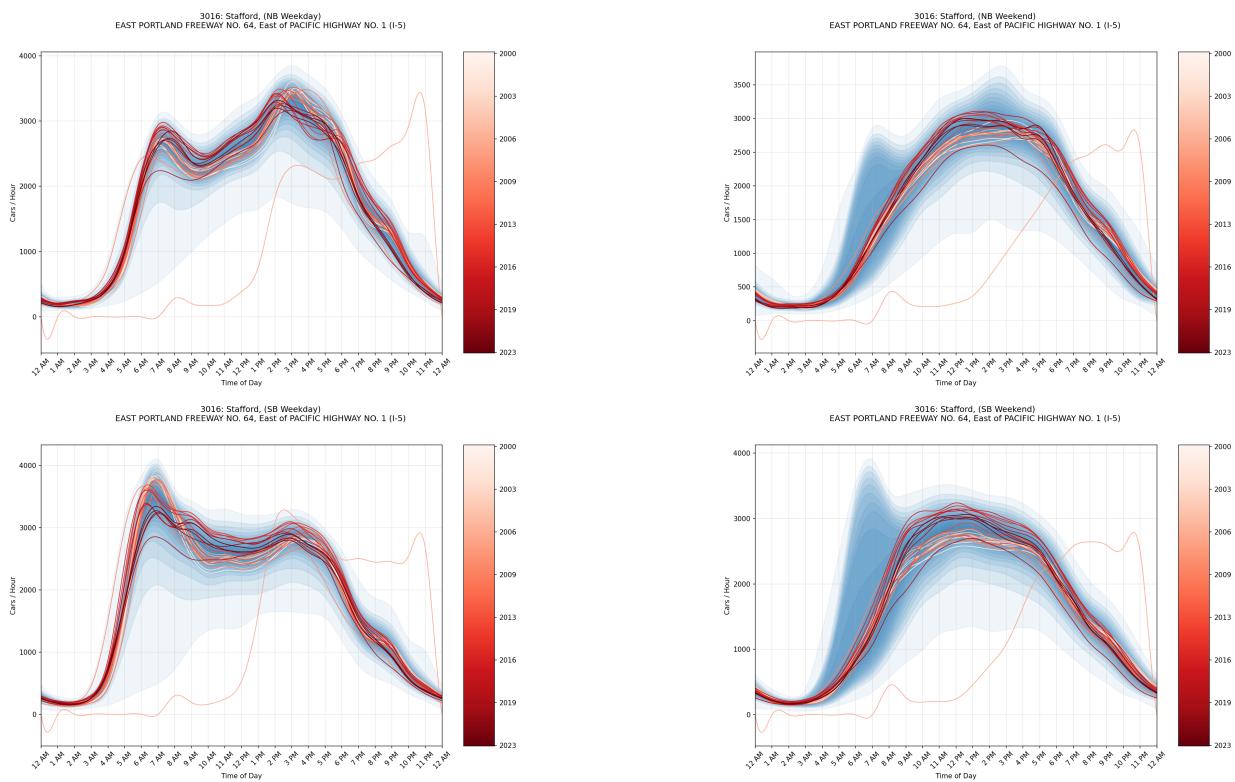


D.O.9 Wilsonville

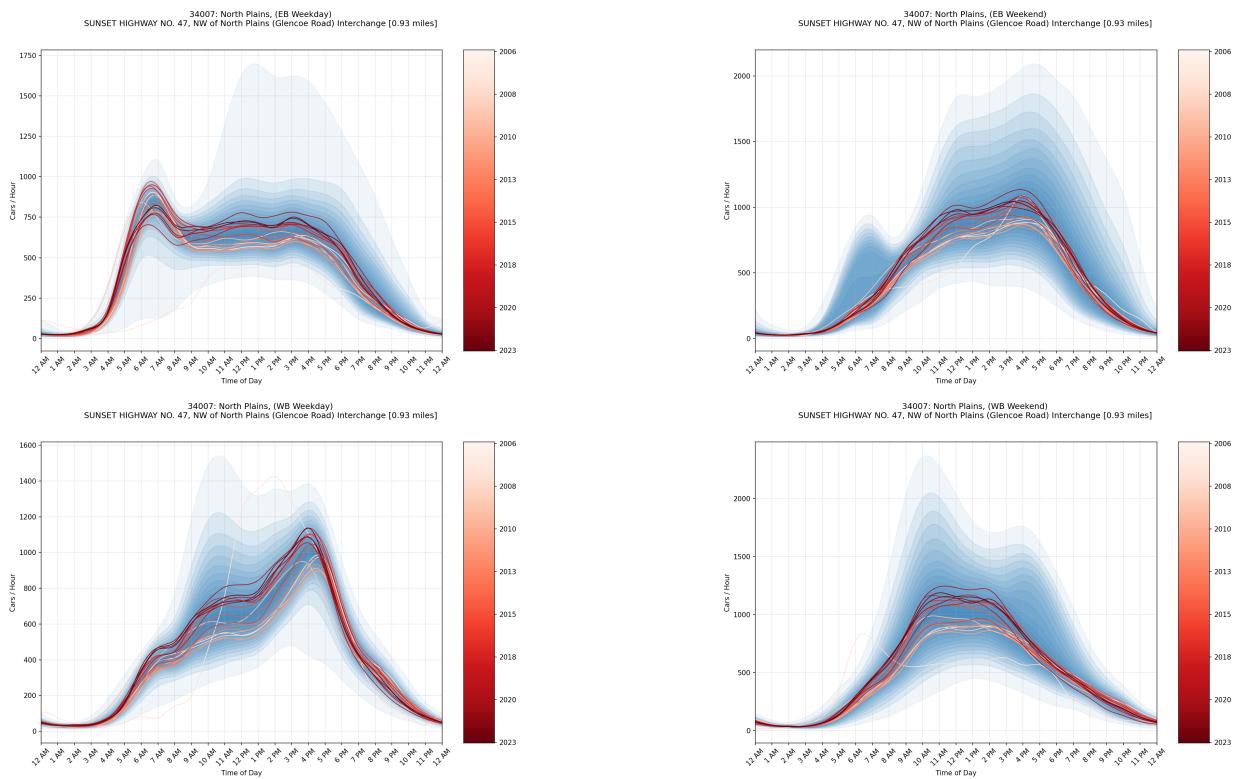




D.O.10 Stafford



D.O.11 North Plains





D.0.12 Beaverton Bethany

