Predicting Freeway Traffic in the Bay Area

Jacob Baldwin, Ya-Ting Wang, Chen-Hsuan Sun
Department of Electrical Engineering, Stanford University

Motivation

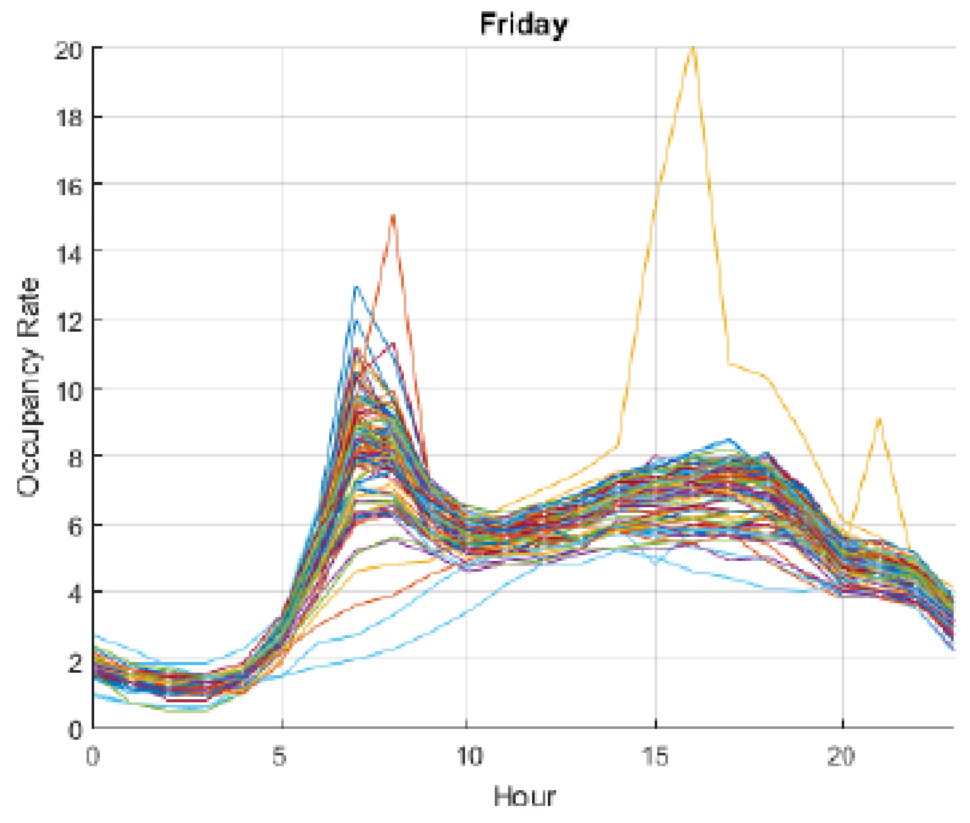
Prediction of traffic is an area where the application of machine learning could help to pinpoint where additional infrastructure may be the most beneficial. The aim is to build a model of traffic on freeways in the Bay Area based on the time of the day, day of the week, and the weather.

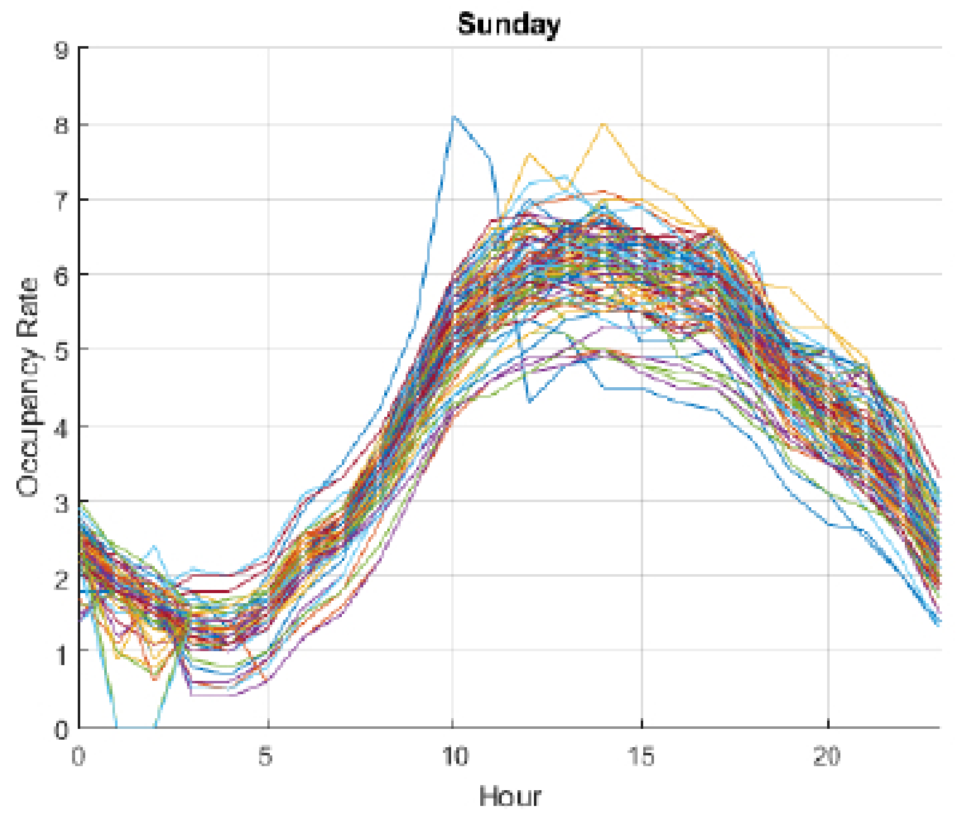
Data

The PEMS data from one of the sensors in I-280 are used to predict future traffic. Output metrics is the **aggregate occupancy rate** of lanes. Occupancy rate is a number between 0 and 1 describing how often the lane is occupied at specific **time points throughout the day**.

Weather data was extracted for each day of the year, including **precipitation** and **temperature.**

Feature1	Feature2	Feature3	Feature4	Output
Day of week	Hour	Avg. Temp	Precip	Avg. Occ





Models

Linear Regression (LR)

- **Treat all features as continuous (LR1)**

The algorithm fits θ such that $h(\theta) = \theta_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \theta_3 \cdot x_3 = y$ where x_1,x_2,x_3 correspond to features 1,2 and, and y refers to the occupancy rate.

- **Treat "Hour" and "Day" as categorical features (LR2)**

We effectively introduce more "features" to which we will fit coefficients - ie. hours 0-23 are each a feature, and weekdays 1-7 are also features.

- **Treat correlation between "Hour" and "Day" as a feature (LR3)**

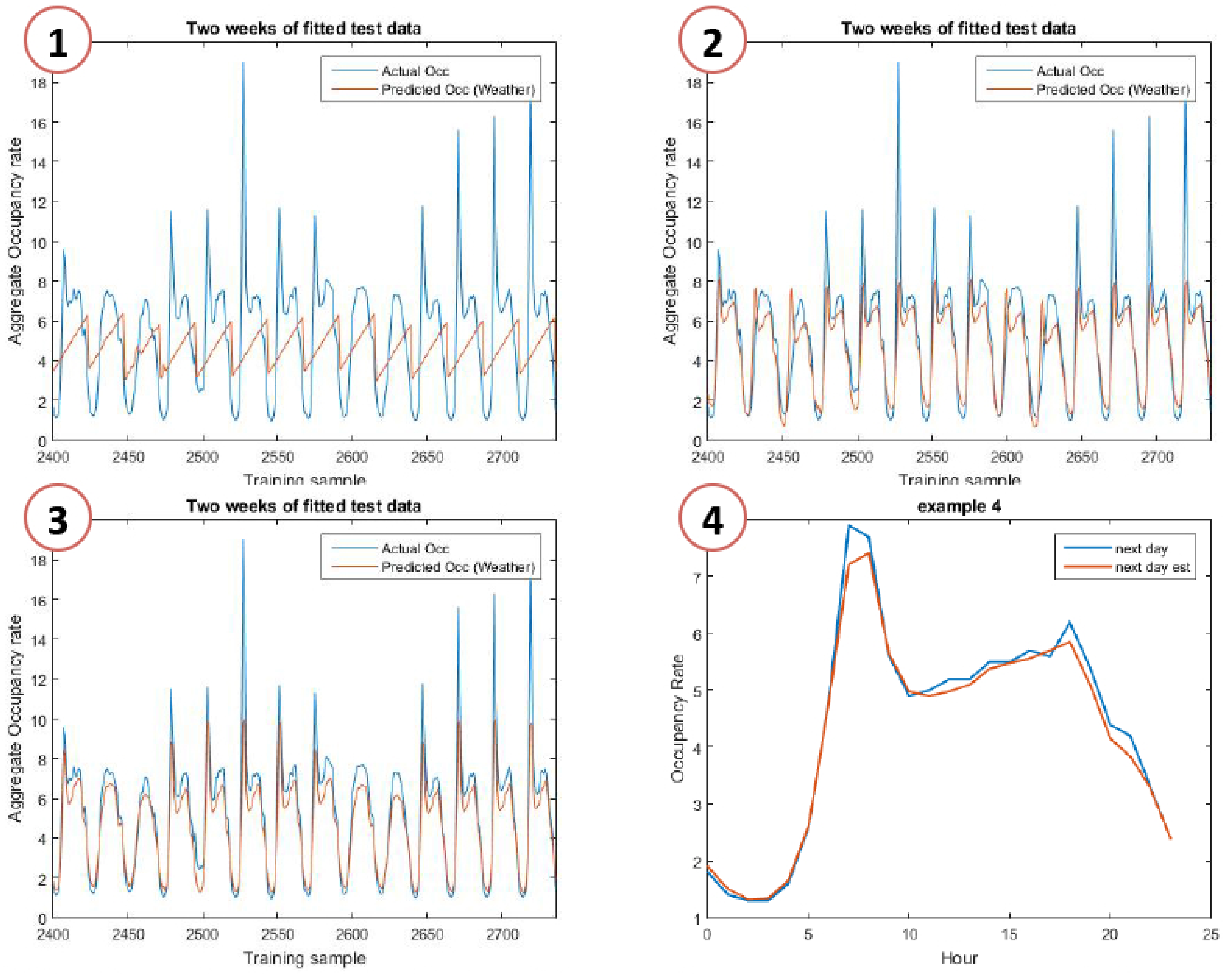
A set of categorical features of **Day*Hour** is introduced to model the interaction between day of the week and the hour of the day.

Functional Regression (FR)

- **Treat previous day as Features**

Five training time series closest to the testing data are chosen while outliers are removed. We weight the training data and made a prediction accordingly.

Results



We implemented 4 different models where it is clearly shown that LR3(Fig 3) and FR(Fig 4) model the occupancy rate more accurately.

	LR1	LR2	LR3	FR
Training Error	2.2650	1.3194	1.0356	X
Testing Error	2.2770	1.3271	1.0667	0.7709

Conclusion

- “Hour”,“Day”, and “Hour*Day” are the most relevant features for LR.
- Eliminating outliers in training data greatly improves FR performance.

Future Work

- Build a complete model of the the whole freeway to map out geographically related trends.