Anomaly Detection with Mean-flow Interpolation of Temporally Sparse Data

Harris Hardiman-Mostow, Tufts University

Supervised by Dr. James Murphy, Tufts University Department of Mathematics

with Marshall Mueller (Tufts University) and Opemipo Esan (Pennsylvania State University)

Abstract

Motivated by the NSF 2020 Algorithms for Threat Detection challenge, we propose a novel algorithm, ADMITS, which reconstructs sparse periodic time-series signals and predicts anomalous observations based on the reconstruction. In situations where data is very sparse, reconstruction via traditional interpolation or dictionary methods may fail to produce desirable results. ADMITS remedies this by making a simple assumption of the underlying periodic structure of the signal, and can thus make accurate reconstructions and anomaly classifications from very limited data. We demonstrate competitive anomaly detection performance on a range of different sampling levels compared to interpolative and dictionary-based methods with exceptional performance in the low sampling regime.

Introduction

Anomaly detection[1] is a subfield of machine learning that is focused on identifying when data deviates from normality, where "normal" is defined in a precise mathematical sense. In our case, a traffic anomaly may occur due to a traffic jam or holiday, for example. Applications are wide-ranging, including in medicine, cybersecurity, and banking. In this project, we investigate anomaly detections methods with multiple sets of traffic data collected and distributed by the NSF. Our challenge was to identify when a traffic "anomaly" occurs within the data.

The main challenge presented by this NSF data was that it was purposefully incomplete – traffic sensors recorded hourly traffic flow data for two years, but only one to twenty percent of the data was given to us by the NSF. This was to simulate a common issue that occurs in data science: sparse data. Hence, our goal of detecting anomalies will be predicated on estimating these missing data points.

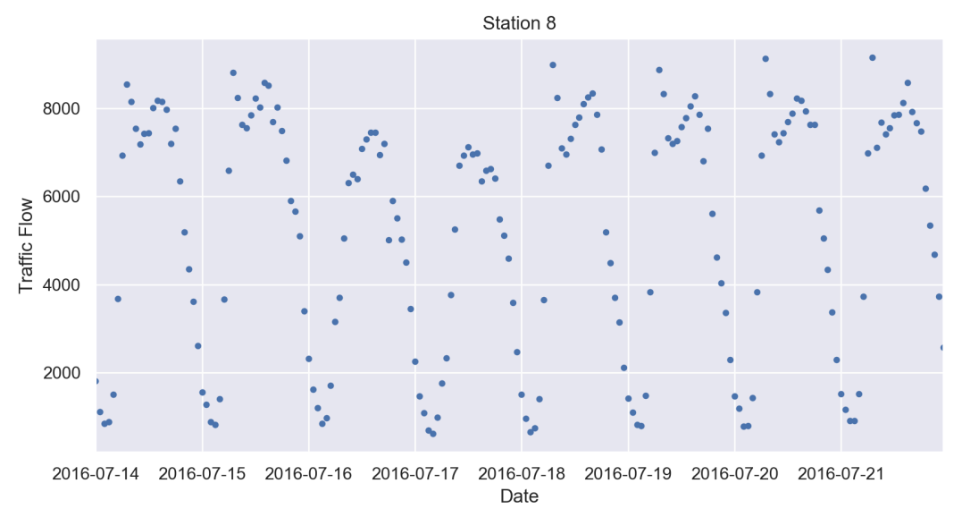


Figure 1: A complete data set example

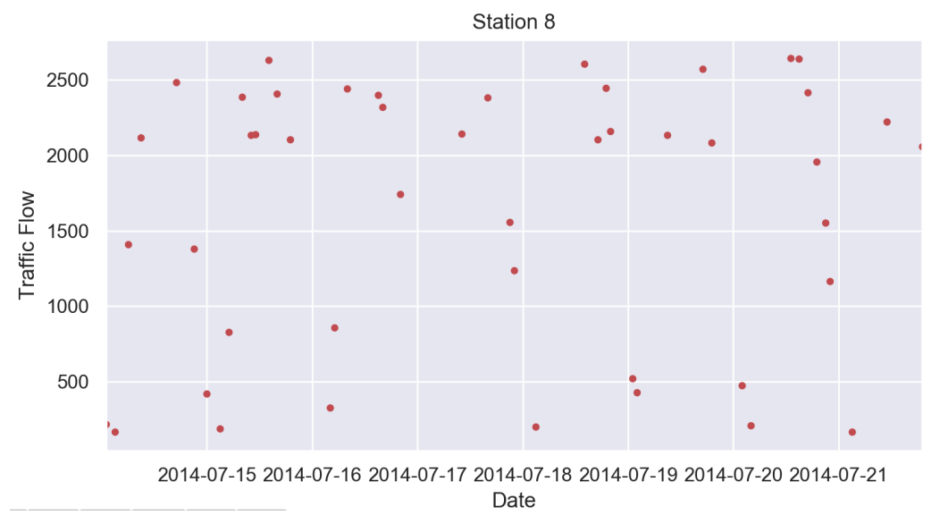


Figure 2: The sparse data we would like to analyze

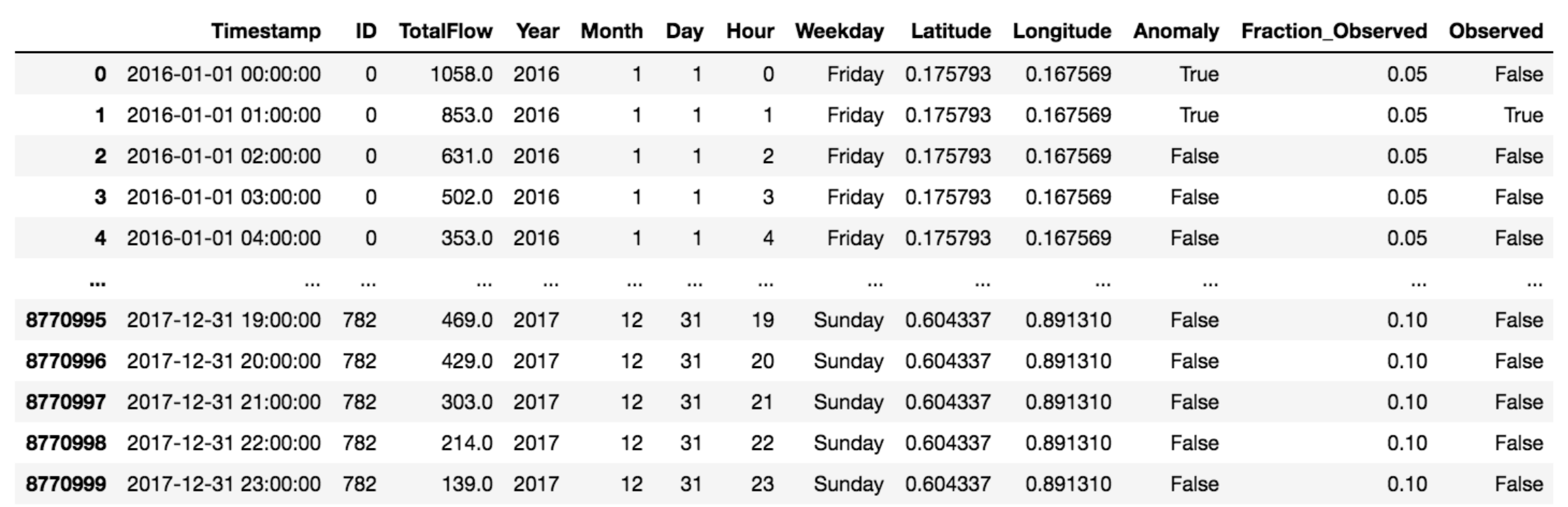
A traffic anomaly is defined in the following way by the NSF: "For a given sensor *s*, hour *h*, and weekday *w*, the n^{th} observation of traffic flow $d^{(n)}{}_{(shw)}$ is anomalous if

$$\left|d_{shw}^{(n)} - \mu_{shw}\right| \geq 3\sigma_{shw} \qquad (1)$$

Put into words, a traffic observation is anomalous if it is three or more standard deviations from the sample mean for that station, hour, and day.

Data

We are provided three datasets, labeled "City 1", "City 2", and "City 3", in the following format:



	Timestamp	ID	TotalFlow	Year	Month	Day	Hour	Weekday	Latitude	Longitude	Anomaly	Fraction_Observed	Observed
0	2016-01-01 00:00:00	0	1058.0	2016	1	1	0	Friday	0.175793	0.167569	True	0.05	False
1	2016-01-01 01:00:00	0	853.0	2016	1	1	1	Friday	0.175793	0.167569	True	0.05	True
2	2016-01-01 02:00:00	0	631.0	2016	1	1	2	Friday	0.175793	0.167569	False	0.05	False
3	2016-01-01 03:00:00	0	502.0	2016	1	1	3	Friday	0.175793	0.167569	False	0.05	False
4	2016-01-01 04:00:00	0	353.0	2016	1	1	4	Friday	0.175793	0.167569	False	0.05	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
8770995	2017-12-31 19:00:00	782	469.0	2017	12	31	19	Sunday	0.604337	0.891310	False	0.10	False
8770996	2017-12-31 20:00:00	782	429.0	2017	12	31	20	Sunday	0.604337	0.891310	False	0.10	False
8770997	2017-12-31 21:00:00	782	303.0	2017	12	31	21	Sunday	0.604337	0.891310	False	0.10	False
8770998	2017-12-31 22:00:00	782	214.0	2017	12	31	22	Sunday	0.604337	0.891310	False	0.10	False
8770999	2017-12-31 23:00:00	782	139.0	2017	12	31	23	Sunday	0.604337	0.891310	False	0.10	False

Methodology and Results

We want to estimate the missing data points:

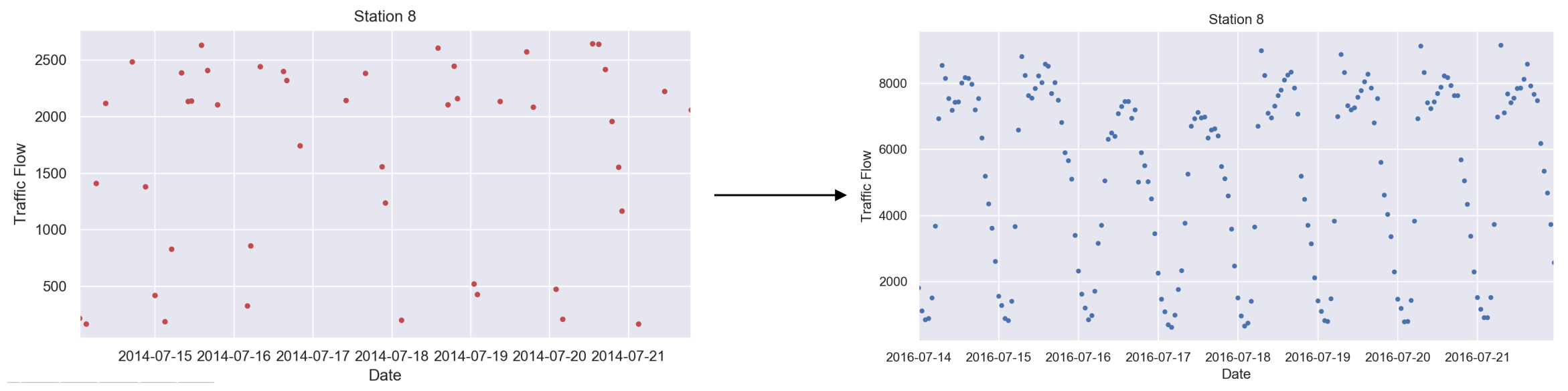


Figure 3: An illustration of what we are given, and where we would like to go!

From here, we can construct the following image, which allows us to explicitly classify anomalies:

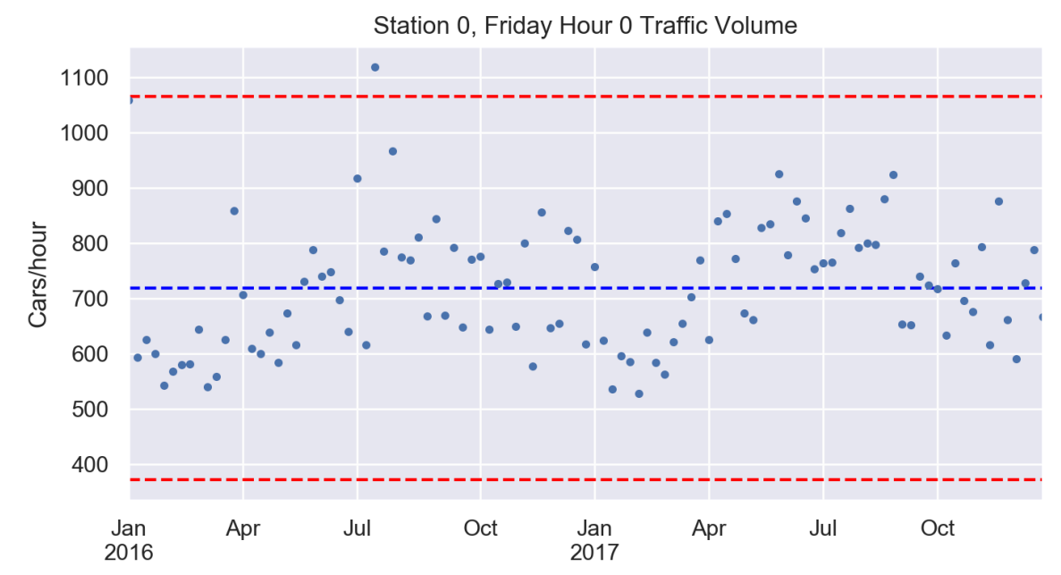


Figure 4: Anomaly detection plot. Note the anomaly on July 4.

Our evaluative metric is the F1 score, which is calculated as follows; *tp, fp,* and *fn* mean True Positive, False Positive, and False Negative, respectively:

$$F_1 = \frac{\mathrm{tp}}{\mathrm{tp} + \frac{1}{2}(\mathrm{fp} + \mathrm{fn})} \qquad (2)$$

The proposed anomaly detection method operates in two steps. First, we reconstruct the signal relying on known a priori periodic structure inherent to the signal. The regularity of the signal helps to overcome the limitations of existing methods in the low sampling rate regime. Then, anomalies are predicted from the reconstructed flow according to the ATD classification routine, namely Equation (1). However, the rigid anomaly detection rule invites some uncertainty near the decision boundary; so, an optional dictionary reconstruction step can be used as a discriminator on the predictions to better resolve uncertain classifications.

To reconstruct a signal with almost all samples missing, one must trade-off the lack of data for stronger assumptions about the signal. For example, sparsity methods assume that a linear combination of a few basic signals are representative of the measured signal. Instead, we assume that the signal has well-defined periodicity across multiple scales in time. In our case, traffic signals exhibit regular daily patterns (e.g. 0800 traffic is similar each Monday). This leads to a structure for how traffic flow during a given week should behave. Indeed, if we can learn the structure of a typical weekly traffic signal, then we can use observed values to predict unobserved values.

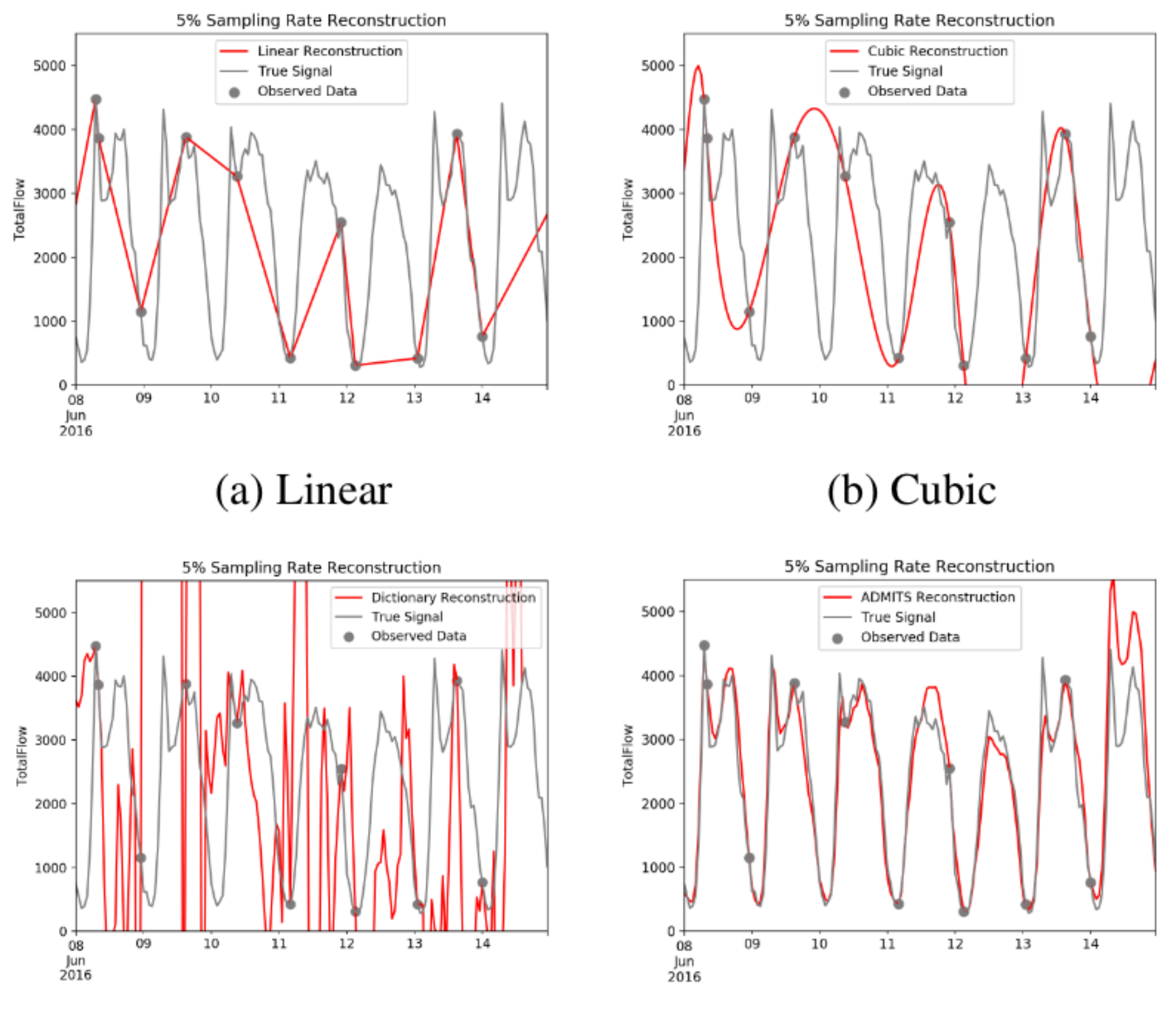
We estimate the structure of a typical weekly flow signal by assuming that the means $u_{w,h}$ for each *(w, h)* pair are linearly related to each other; more precisely, there exists a map $c((w_1, ,h_1), w_2,h_2())$ such that for $(w_1, ,h_1) \neq (w_2,h_2)$:

$$u_{w1,h1} = c((w_1, ,h_1),(w_2,h_2)) * u_{w2,h2} \qquad (3)$$

Methodology and Results (continued)

For signals obtained at a sufficiently low sampling rate we may find that the estimations in are inaccurate. Each pair will have few – or no – samples, and any anomaly present will significantly affect the sample means and thus the estimated map between them. One way to mitigate this is to find a "nearest neighbor" observed signal from a sensor with a higher sampling rate (if one exists). That signal's map *c* can then be used to influence or adjust the map *c'* of the lower-sampled signal by taking a weighted average of the maps. The nearest neighbor signal is chosen to be a signal that has similar sample means to the current signal.

To demonstrate the efficacy of the proposed method, we compare ADMITS to a patch-based dictionary reconstruction algorithm, linear and cubic interpolations, and a baseline detector that applies the ATD classification rule (1) on the sample data without any reconstruction. The experimental comparisons were performed on the "City1" data set provided by the ATD challenge via Caltrans PeMS, which contains contiguous hourly traffic flow data at 500 sensors over all of 2016 and 2017. We conduct experiments at sampling rates [0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8]. At each sampling rate, 100 independent randomized trials were performed. For each trial a different 10% contiguous chunk of the data is used for training the dictionary, while the remainder of the data is sampled at each of the rates on a per-sensor basis. Once reconstructed, F1 scores are computed using the explicit anomaly detection rule expressed in (2).



(a) Linear (b) Cubic

(c) Dictionary (d) ADMITS

Figure 5: Reconstruction methods at 5% sampling rate over a random 1-week period

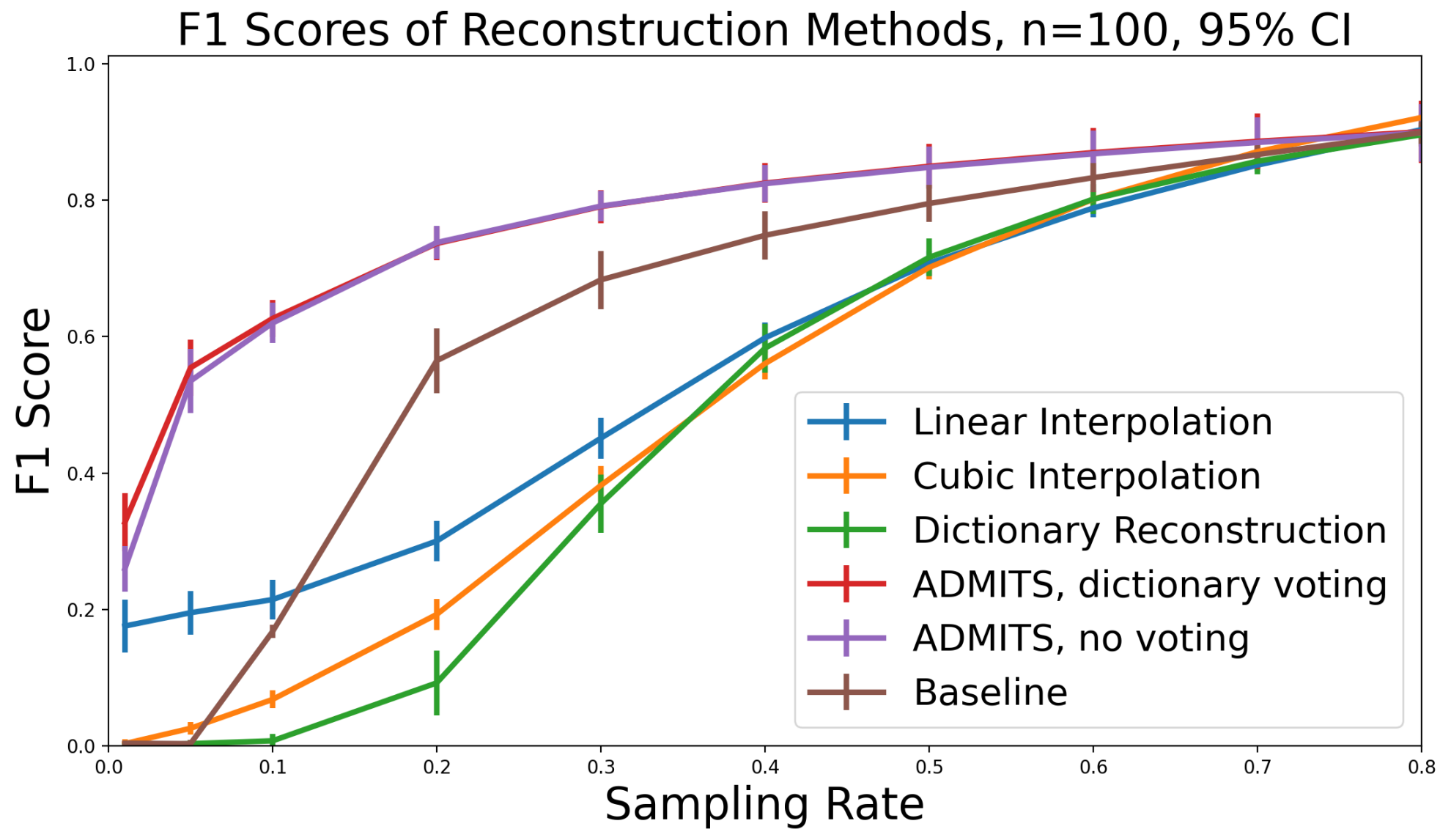


Figure 6: F1 score comparison of methods across sampling rates

Conclusion

We have introduced and demonstrated the efficacy of a reconstruction and anomaly detection algorithm that relies on known structure in data to reconstruct signals more accurately when very few samples are present. The main drawback to our method is the heavy assumptions required in order to make the reconstruction. Efforts to reduce these would increase the generality of the method. A structured dictionary reconstruction scheme that makes stronger assumptions about which elements to use in reconstruction may be worth investigating. Another approach could employ the notion of nearest neighbors in flow to construct a graph in which graph signal processing-based reconstruction techniques could be applied to better leverage all the observed values measured at every sensor.

The results of this project demonstrate several techniques which are applicable to any setting where sparse data is present and anomalies must be identified. For example, medical imaging is an area with active anomaly detection research, which has been leveraged to detect tumors with image processing algorithms of images of patients' organs captured by powerful cameras. However, these are expensive and difficult to collect in large datasets – hence, analyzing traffic data provides an easy way to experiment and draw conclusions about anomaly detection methods.

References

1. Chandola, V., A. Banerjee, and V. Kumar. "Anomaly Detection: A Survey." *ACM Computing Surveys.* 2009.
2. Bretherton, C. "Interpolation and Smoothing." 2015.
3. Aheron, M., M. Elad, and A. Bruckstein. "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation." *IEEE Transactions On Signal Processing*, Vol. 54, No. 11. 2006.
4. Mairal, J., M. Elad, and G. Sapiro. "Sparse representation for color image restoration." *IEEE Transactions on Image Processing* 17, no. 1 (2007): 53-69.
5. Zelnik-Manor, K., and P. Perona. "Self-Tuning Spectral Clustering." *NIPS Proceedings.* 2005.
6. Olshausen, B., and K.J. Millman. "Learning Sparse Codes with a Mixture-of-Gaussians Prior." 1999.

Acknowledgements

I would like to thank Professor James Murphy, for his mentorship and guidance. I would also like to thank Marshall and Ope, for being intelligent and hard-working collaborators, and to Dr. Anne Moore, for organizing the Summer Scholars programming (even in the middle of pandemic!).

Finally, thank you to the Summer Scholars Program and thank you to the T-Tripods Institute (NSF HDR Grant 1934553) for partial support of this work. The traffic data collection and challenge was funded by NSF DMS 1924513.