Reproducible Research: a Hello World Example

The existence of capacity drop phenomenon

Zuduo Zheng

Introduction

This is a “Hello World” example on producing reproducible research in Traffic Engineering. This example is based on a paper co-authored by Professor Simon Washington and me in 2012: *Zheng and Washington (2012) On selecting an optimal wavelet for detecting singularities in traffic and vehicular data, Transportation Research Part C: Emerging Technologies p18-33.* In this paper, before we focuse on how to select a good mother wavelet for detecting singularities in traffic and vehicular data, we put significant effort on objectively comparing performances of several data processing techniques that are commonly used by researchers in our field. If you are from traffic flow community, you likely know that capacity drop is a big deal to us, as evidenced by numerous studies/models in the literature dedicated to this phenomenon (Some references on capacity drop can be found at the end of this document). Many researchers believe that the capacity of a road bottleneck can drop about 10% at the onset of traffic congestion, and some researchers even reported that the bottleneck capacity could drop as much as more than 30%. Yes, capacity drop does exist, as consistently reported in the literature. However, in analysing capacity drop, many things could go wrong, e.g., the accuracy of the bottleneck location, the data missing and noise, the sample size, causality, and other exogenous (but unavailable to researchers) factors, just to name a few. By demonstrating that sometimes the **capacity drop** could be a mere artifact caused by inappropriate data processing techniques, one of the motivations of this paper was to caution researchers in our field to think it more carefully before jumping into any conclusion about capacity drop. Unfortunately, for various reasons such motivation has not been effectively reflected in this paper’s title, and partially due to this reason this particularly intended contribution of the paper has been largely unnoticed by the community so far (Evidence: Its Google Scholar cites is only 9 including 3 self-citations, as I’m writing).

To generate a PDF, Word or HTML document of this example, you need to install

- RStudio (Version 0.99.903 or later);
- Package *Knitr* if it’s not already installed in RStudio by default;
- Package *Markdown* if it’s not already installed in RStudio by default;
- Package *TTR*.

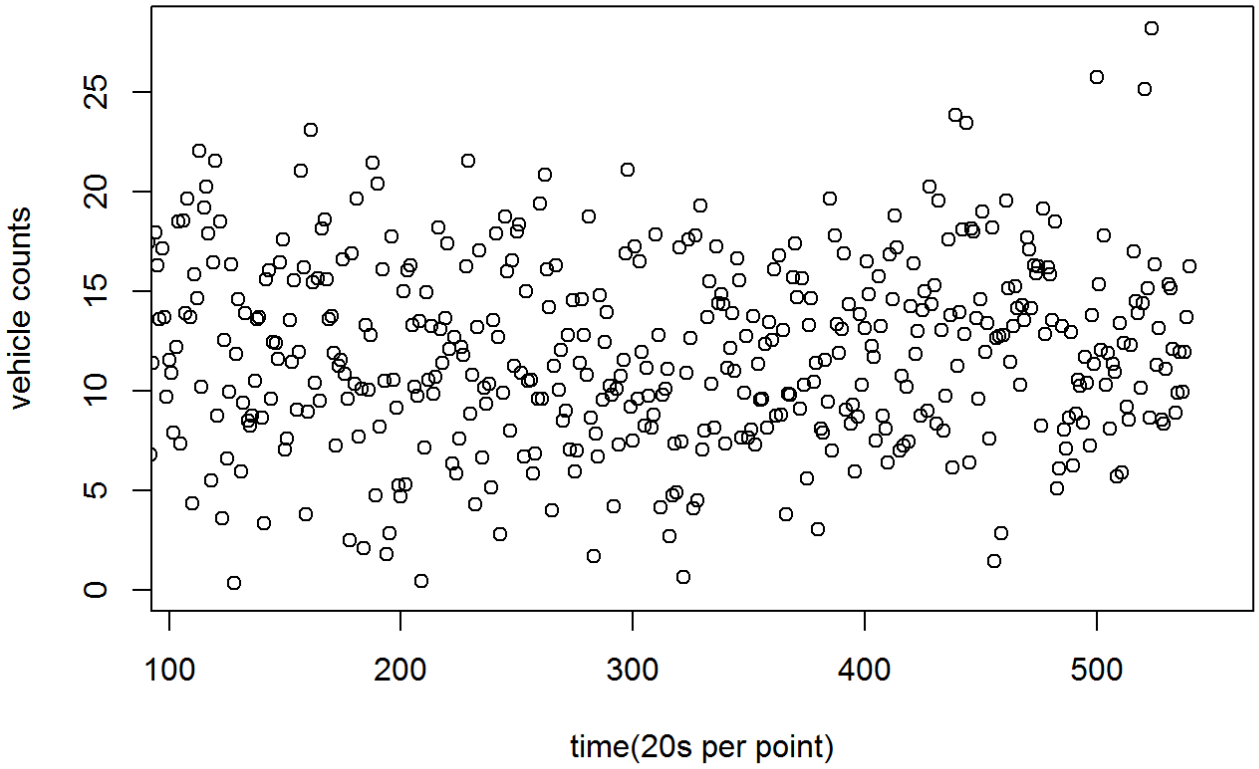
Then open the file with the extension of *Rmd*, click the button *Knit to HTML* (or other types) in **RStudio**.

For the remaining of the document, the numerical experiment that is used for testing performances of popular techniques for analyzing traffic data is first presented, and then the false capacity drop phenomenon caused by these commonly used data processing techqniues is demonstrated. Note that since this is meant to be a “Hello World” example, to reduce its complexity, some of the techniques discussed in Zheng and Washington (2012) are not covered in this example, such as second-order difference of cumulative data, short-time Frourier transform, etc.

The Numerical Experiment

The performances of popular techniques for analyzing traffic data are tested using numerical simulation, with the intent to uniquely compare these techniques' performances and to underscore their advantages and disadvantages. Through such a comparative analysis, undesirable consequences of using these popular data processing techniques are demonstrated.More specifically, to mimic vehicle counts in rush hours collected at a loop detector near a bottleneck, a time series is randomly generated, representing a sample of traffic data with a mean of 12 vehicles and a standard deviation of 5 vehicles. The simulation period is 3 hours with a time resolution of 20 s (so the average flow is 2160 vehicles per hour). In total, 540 data points are generated. Note that the simulated data are stochastic and exhibit white noise properties except for a non-zero mean, as shown in Figure 1. Please note that the data generated in this document is not the same as in Zheng and Washington (2012) because the analysis for the paper was done by using **Matlab**, and also a different seed for generating random numbesr was used.

Figure 1 Vehicle counts from the numerical experiment



The False Capacity Drop Phenomenon Caused by Data Processing Techniques

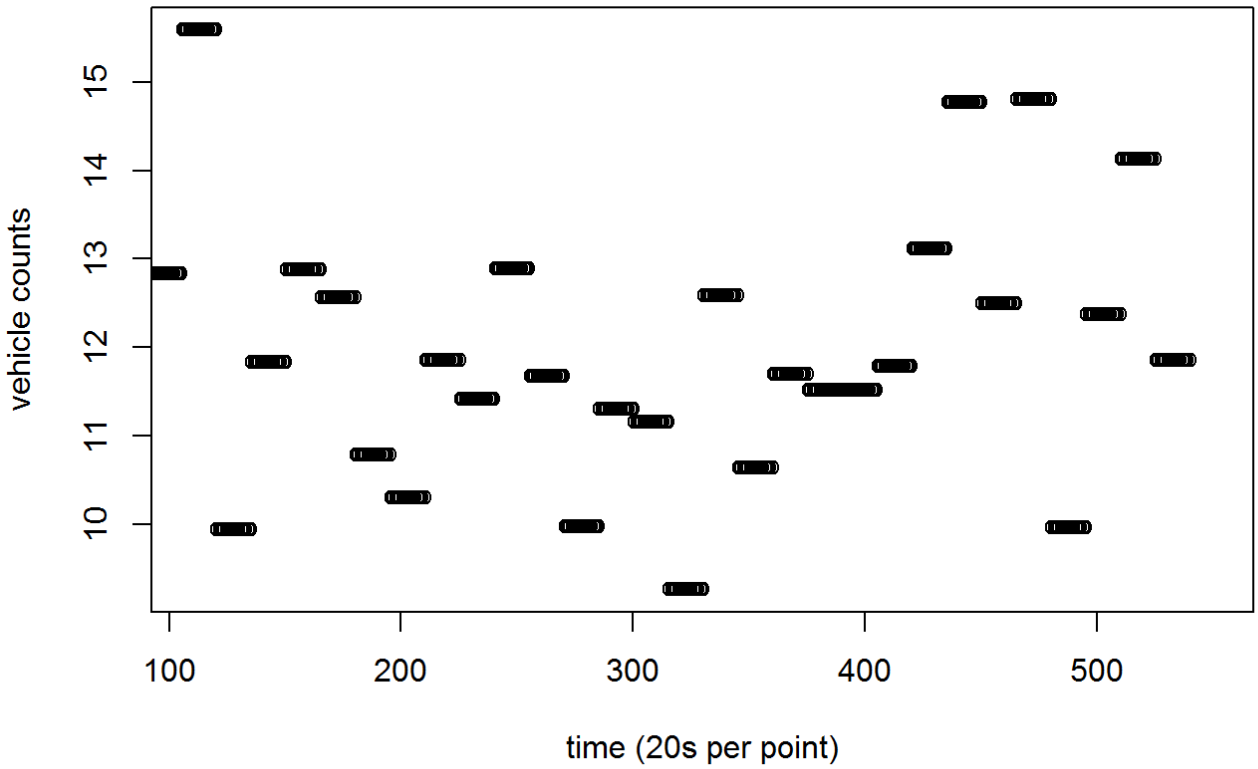
Commonly-used data processing techniques in Traffic Engineering considered in this example include averaging, moving average, and oblique cumulative curve.

Averaging

To dampen or attenuate statistical noise in the traffic data, the simplest and one of the most commonly used techniques in the time domain is to aggregate data over a certain time interval (e.g., 5 or 15 minutes), which

may be sufficiently effective to reveal long-term trends in traffic patterns. However, the shortcoming of this technique should be apparent: fine resolution information will be smoothed out and inaccurate or distorted information may be obtained because of the decreased time resolution. For example, depending on the starting point of averaging and the window size, data belonging to a single event may be divided into two or more groups and aggregated in different ways. The vehicle counts in the numerical experiment are averaged for each 5 minute period as shown in Figure 2. Based on this figure, one might conclude that the lane capacity dropped from 14 vehicles per 20 s to 11 vehicles per 20 s at the time of about 33 minutes from the start of the simulation. In other words, one might claim that the capacity of this bottleneck dropped by about 20 percent. Recall, however, these data are simulated and randomly generated-thus no authentic capacity reduction exists.

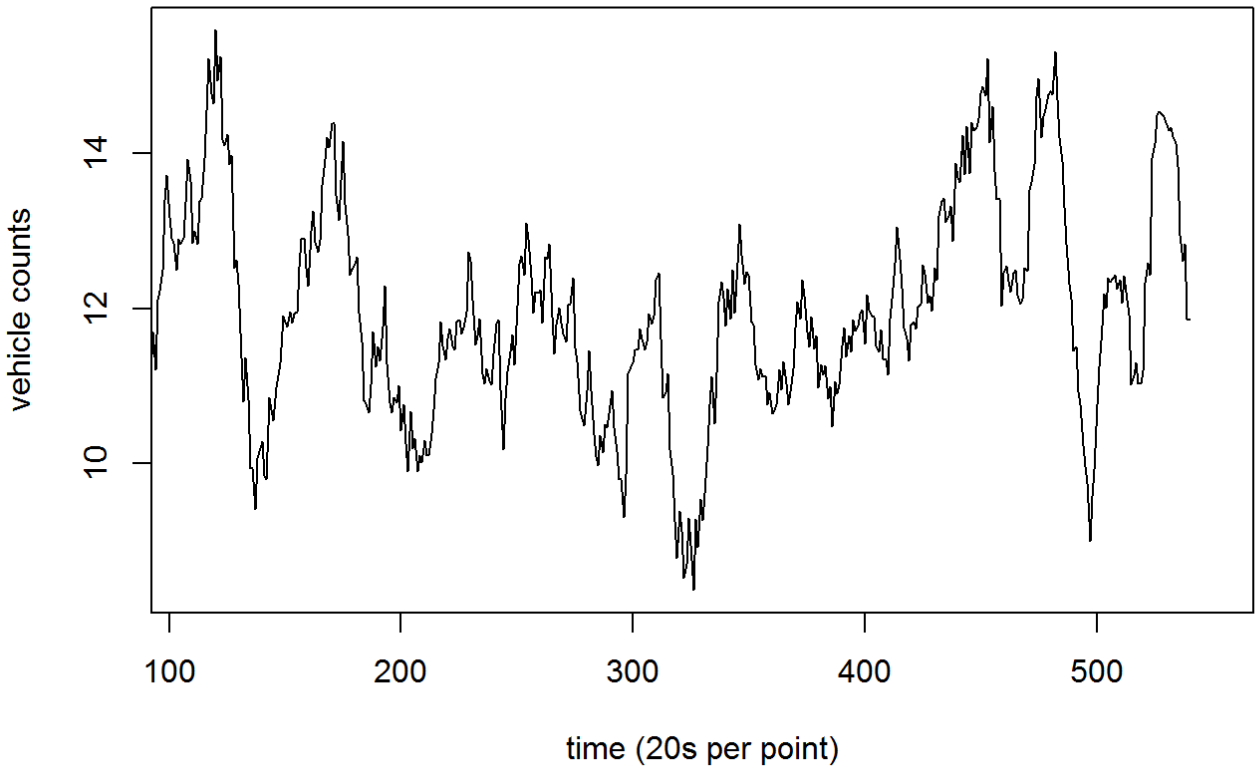
Figure 2 The averaged vehicle counts



Moving average

To preserve the time resolution, a moving average and its variants (e.g., weighted moving average) are often used to process time series traffic data. Generally, a moving average out-performs simple aggregation. However, a moving average suffers from some of the same issues confronted by simple averaging. Much of the most interesting information contained in the traffic data, i.e., singularities, is attenuated when a moving average is applied. Figure 3 demonstrates that a clear pattern appears once the vehicle counts in the numerical experiment are averaged using a 5-minute moving window.Based on such pattern, one might (again, falsely) claim the occurrence of capacity drop. Note that for simplicity, points within the window are equally weighted.

Figure 3 The moving-averaged vehicle counts



Oblique cumulative curve

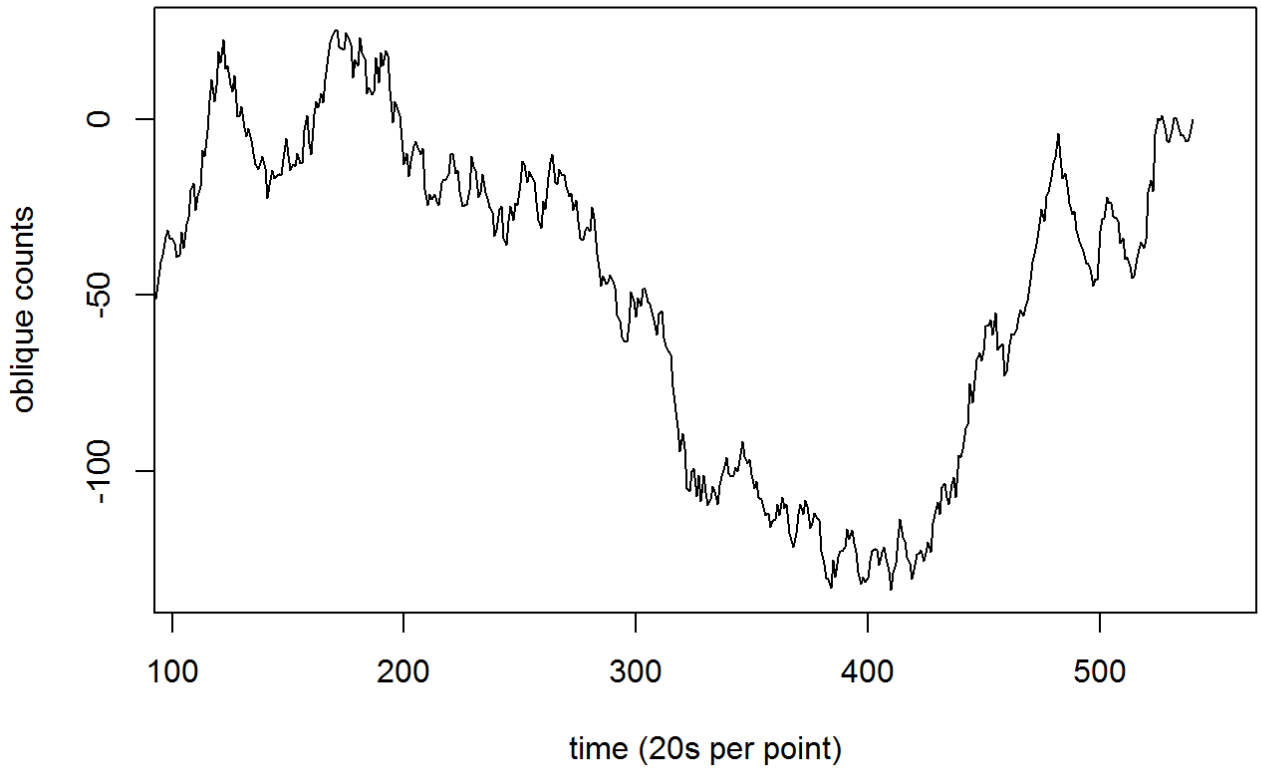
Oblique cumulative curve is another widely used analysis method intended to attenuate noise and reveal the underlying data trend. As defined in Eq. (2), an oblique curve is constructed by taking the difference between the cumulative measurement at time t, and a background reduction. This method is simple to use and often reveals interesting patterns that are invisible through plotting of the original data and as a result is gaining in popularity.

$$\check{x}(t) = X(t) - X_0 \times (t - t_0) \quad (2)$$

where $X_t = \sum_{j=t_0}^{t} x_j$; X_0 is a scaling factor and t_0 is the starting time.

In addition to the issues (e.g., subjective and labor intensive) identified previously, this method purports to provide a superior visual glimpse of the underlying phenomenon without providing solid theoretical support for such a claim. Thus, the soundness of its output requires additional exposition. The oblique curve of the data from the numerical experiment (Figure 4) shows different slopes for different segments of the signal, which implies different traffic flow rates. Based on curves like this one, researchers often measure characteristics of the bottleneck, e.g., capacity reductions, and activation and deactivation times. Intuitively, such different slopes represent random rather than structural differences in these data, as was the case previously.

Figure 4 The oblique cumulative curve of the simulated data



Conclusion

This document is a simple example for demonstrating how reproducible research can be carried out by using open-source computational tools widely available for and accessible to researchers in Traffic Engineering and beyound. By executing this script, any third party can easily get the identical document as the one you are reading now. More importantly, how the analysis has been exactly implemented in this document is crystal clear, that is, the analysis upon which the conclusions are drawn is reproducible. Thus, any person can scrutinize the analysis itself and then make her/his own decision on whether the conclusions reported are trustworthy or not. Furthermore, the codes can be easily extended/adapted for further analysis.

More resources on how to conduct reproducible research can be found from my webpage (http://www.connectedandautonomoustransport.com/).

References

Banks, J. H. (1991). Two-capacity phenomenon at freeway bottlenecks: A basis for ramp metering?. Transportation Research Record, 1320, 83-90.

Chen, D., Ahn, S., Laval, J., & Zheng, Z. (2014). On the periodicity of traffic oscillations and capacity drop: the role of driver characteristics. Transportation research part B: methodological, 59, 117-136.

Chung, K., Rudjanakanoknad, J., & Cassidy, M. J. (2007). Relation between traffic density and capacity drop at three freeway bottlenecks. Transportation Research Part B: Methodological, 41(1), 82-95.

HALL, F. L. & AGYEMANG-DUAH, K. (1991). Freeway capacity drop and the definition of capacity. Transportation Research Record, 1320, 1-98.

Parzani, C., & Buisson, C. (2012). Second-order model and capacity drop at merge. Transportation Research Record: Journal of the Transportation Research Board, (2315), 25-34.
Treiber, M., Kesting, A., & Helbing, D. (2006). Understanding widely scattered traffic flows, the capacity drop, and platoons as effects of variance-driven time gaps. Physical Review E, 74(1), 016123.
Zheng, Z., & Washington, S. (2012). On selecting an optimal wavelet for detecting singularities in traffic and vehicular data. Transportation Research Part C: Emerging Technologies, 25, 18-33.