**Pattern Prediction and Anomaly Detection in Yahoo! Production Traffic Time-series Data**

**Project Reasoning**

My reasoning for choosing this project is two-fold:

On one hand, this project validates my expertise and skills to the sorts of positions I find compelling. As my career plan after graduation with a Ph.D. in Industrial Engineering and a minor in Computer Science, I would like to work as a data scientist. It seems logical that working on a portfolio project founded solidly in time-series data analytics and visualization reinforces my future career goals.

My second reason for choosing this project is its high correlation with my current Ph.D. research domain. Due to my interest in analyzing various types of time-series problems, this project brings me another opportunity to deal with a new data analysis challenge. Meanwhile, I would respond to the suggestions in your challenge to address an issue of greater general interest.

**Significance**

Nowadays, automatic anomaly detection is a critical issue in the world in which the streams of data makes it impossible to recognize the outliers manually. Therefore, developing novel and efficient automatic framework to predict the time-series data pattern and detect anomalous behaviors is always worthy of investigation.

**Project Overview**

I plan to use publicly available Yahoo! Webscope data sets including historical production traffic information and recognize the abnormal states. The application of this abnormality detector is to predict the traffic pattern prediction and detect the anomalous states based on sudden changes in the behavior of the data. The dataset consists of real and synthetic time-series with tagged anomaly points. The dataset tests the detection accuracy of various anomaly-types including outliers and change-points. Moreover, the dataset includes time-series with varying trends, noise, and seasonality.

**Breakdown of Proposed Project**

- **Python:** used for data wrangling and exploration, simple analysis, and regressions**.**
- **R & Rstudio:** used for autocorrelation analysis and the assembly of the complete machine learning training set. It may or not be used for the final time-series analysis. It depends on the algorithms that will be used to train and test the data.

**Data Sets**

This dataset consists of four benchmarks that are downloaded from the Yahoo Computing Systems Data Center. These datasets include the production traffic to some of the Yahoo! Properties. The A1 and A2

datasets include single time-series data with labeled anomalies. Besides, the A2 dataset contains time-series with random seasonality, trend, and noise. Furthermore, the A3 and A4 datasets consist of multiple time-series data channels. The A3 benchmark only contains outliers while the A4 benchmark also contains the anomalies. Both of these datasets include seasonality and trends.

**Exploratory Analysis**

**Autocorrelation analysis**

Using Q-statistic, we test the possibility of autocorrelation among various lags of different data sets. Running the Box-Pierce test with significance level 0.05 for each benchmark results in a very small p-value which reveals the existence of autocorrelation among the lags of these data channels. Figures 1-4 represent the periodical trends in the time series distribution of different datasets. Likewise, the ACF plots in these figures verify the non-stationarity of given data channels, as the spike values decrease slowly through incremental lags.

**Results**

According to the simple interpretation of the exploratory graphs, and the very small p-values resulting from the Box-Pierce autocorrelation test, there are autocorrelations between values of different data channels at sequential time stamps.

One approach to develop a pattern prediction model and detect abnormal states is to fit an Autoregressive Moving Average (ARIMA) model to the time-series data of channels in different datasets. This model realized the linear part of the existing autocorrelation in the data. Then, the residuals of this linear model can be further analyzed to recognize the nonlinear correlations in the trend of time-series datasets. The wavelet neural network can be used at this stage of the program to reach the nonlinear relationships in the data. Finally, the residual of this hybrid approach can be further analyzed to detect the abnormal states based on the high deviations of the observed data from the prediction values. Another approach to deal with multi-variate time-series anomaly detection is to develop a Hidden Markov Model (HMM) combined with Viterbi Algorithm to detect abnormalities.

These are the ideas that I am proposing to do for my portfolio project.

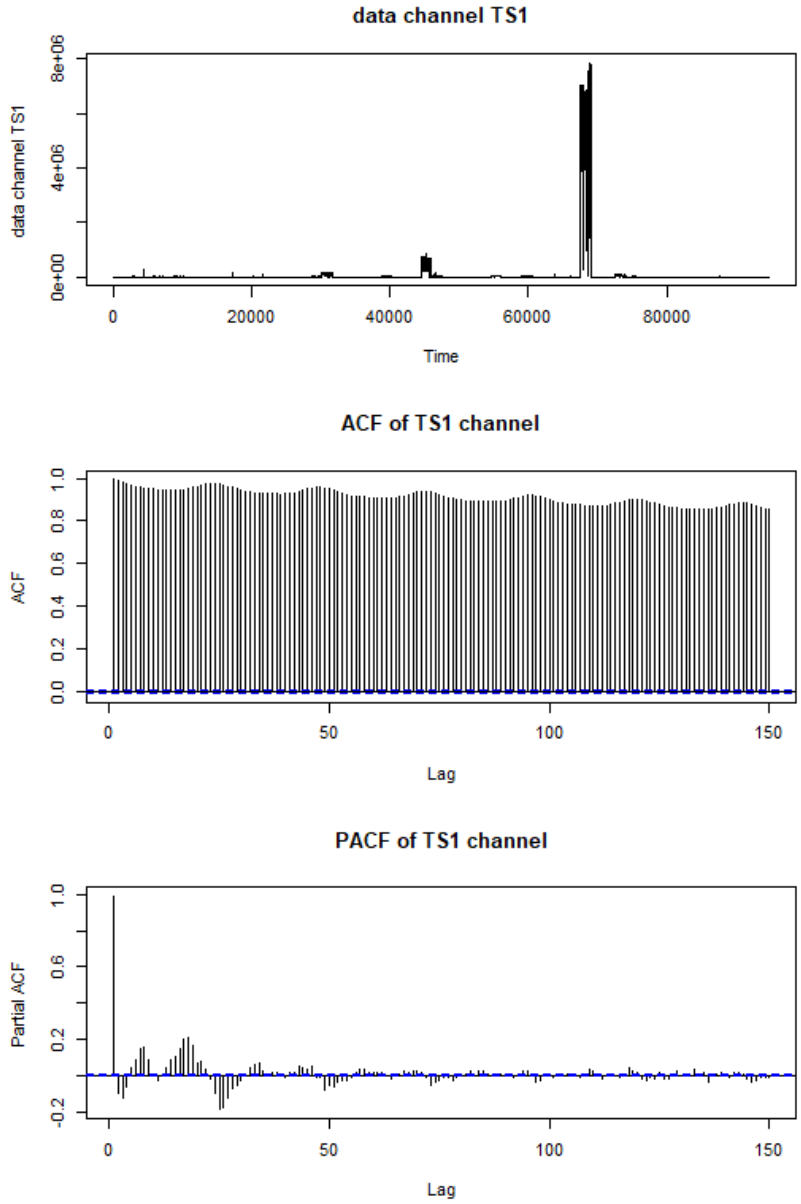


*Figure 1. The value data channel of A1 dataset*

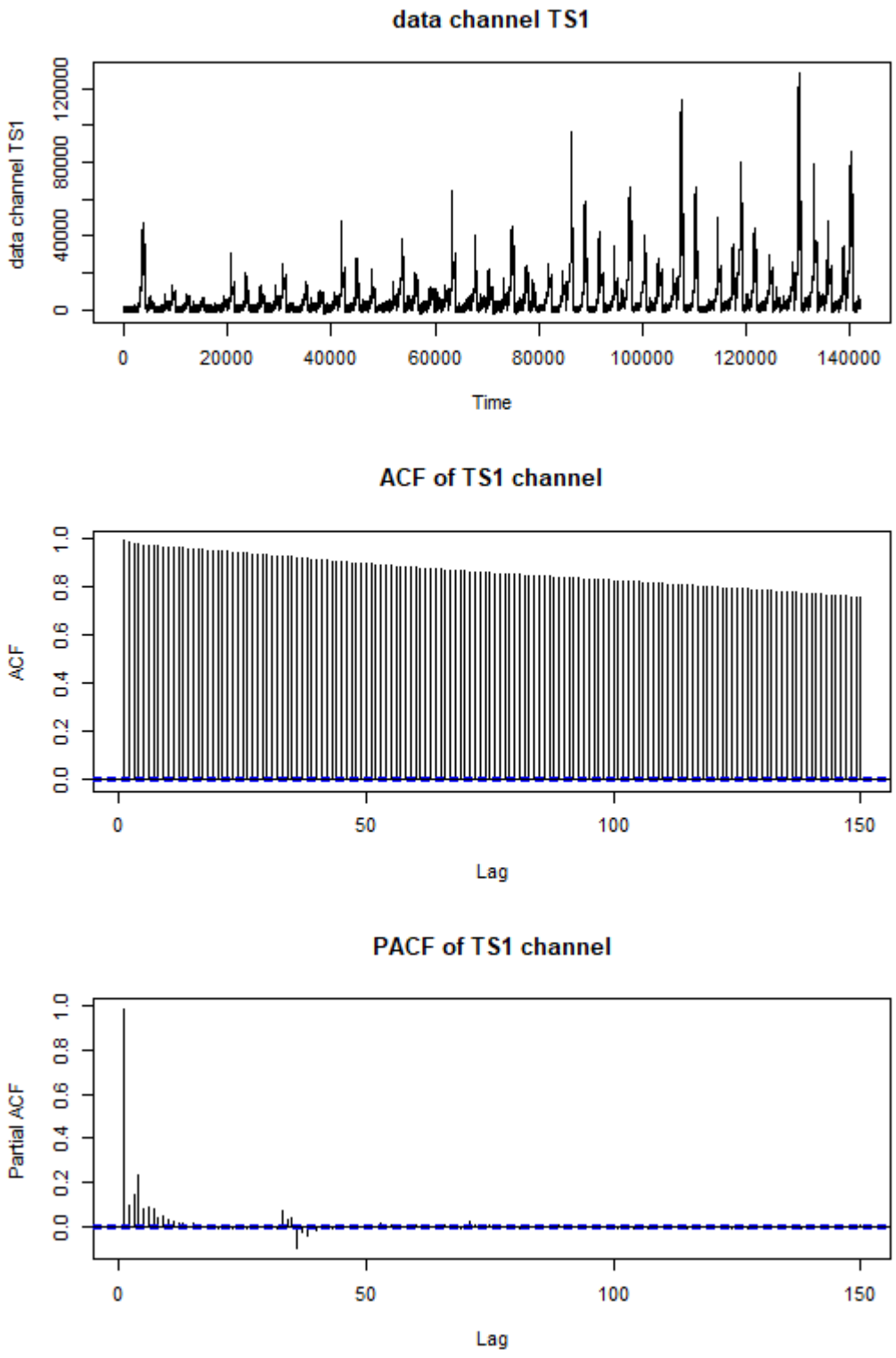


*Figure 2. The value data channel of A2 dataset*

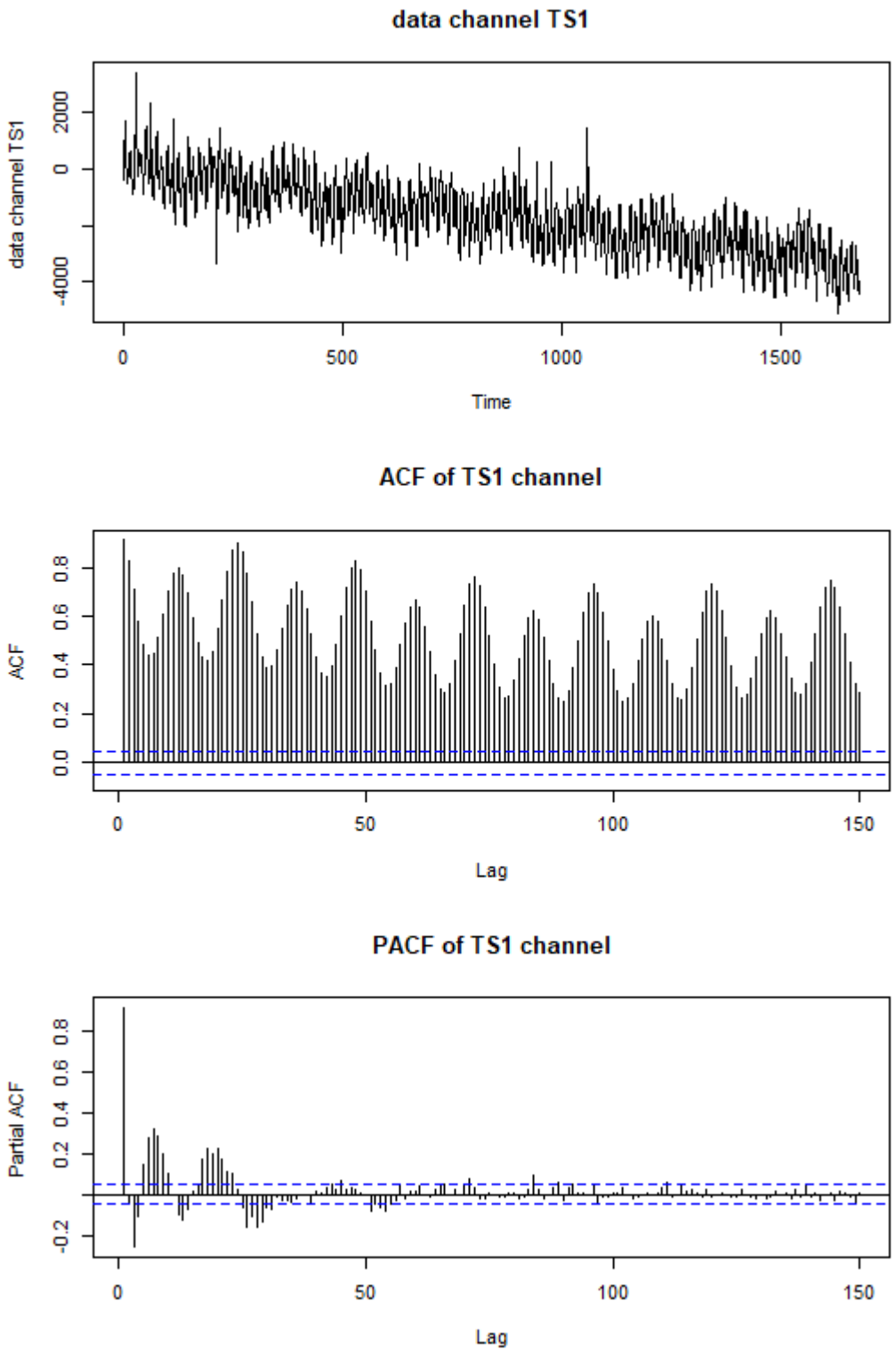


*Figure 3. The TS1 data channel of A3 dataset*

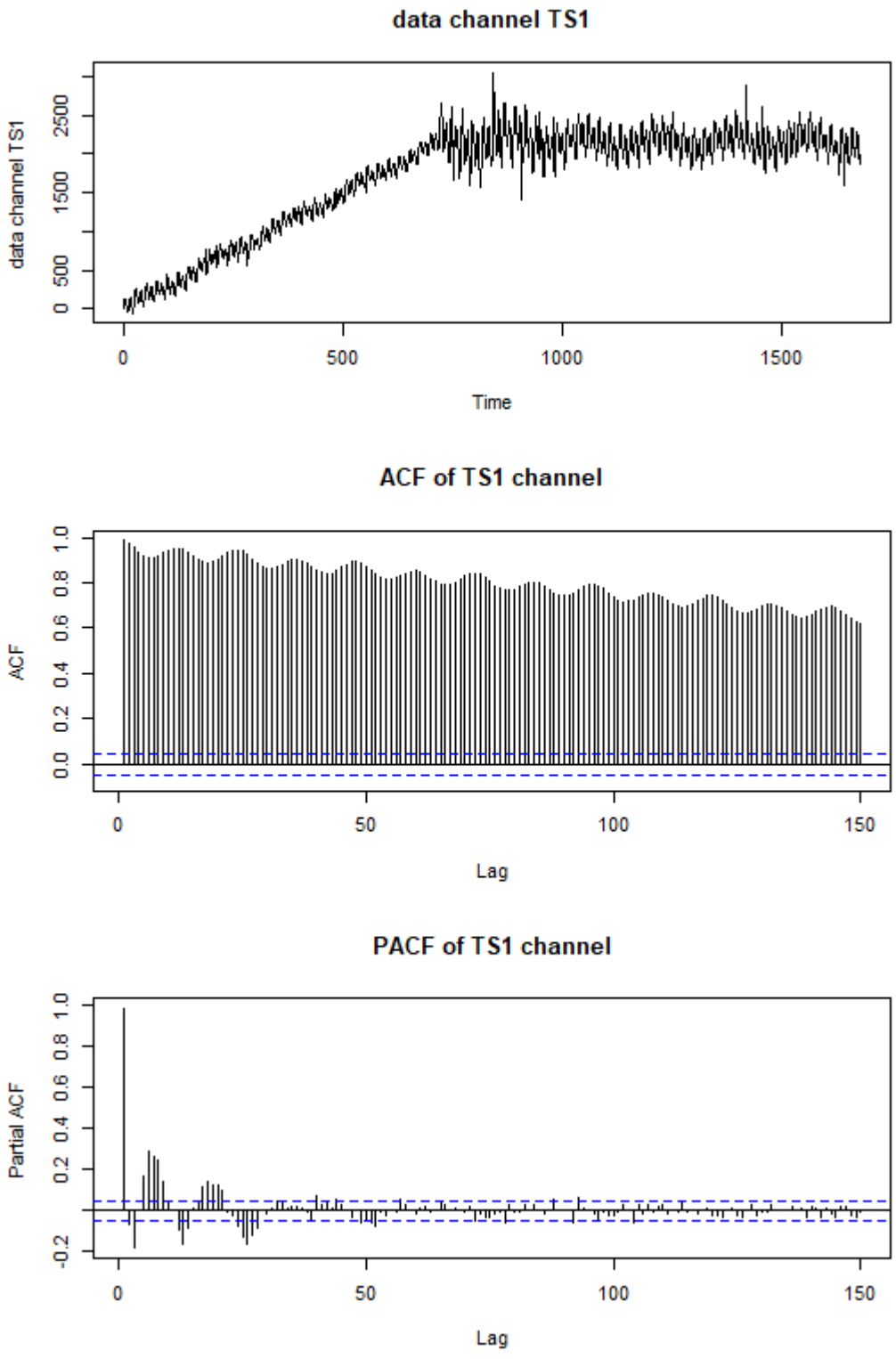


*Figure 4. The TS1 data channel of A4 dataset*