Mini-Project (ML for Time Series) - MVA 2024/2025

Alexis Marouani alexis.marouani@polytechnique.edu
Grégoire Béchade gregoire.bechade@polytechnique.edu

December 16, 2024

1 Introduction and contributions

The paper we studied aims to introduce a novel method to perform anomaly detection in large time series, based on the introduction of a "normal behaviour" of the time series. Anomaly detection can be defined in several manners. It can be interpreted as "outlier detection", where the objective is to detect single anormal points (which can for instance correspond to sensor failures). The paper focuses on the second type of task, which aims to detect anormal subsequences of the time series. In this case, several consequent points behave anormaly (for instance, an anormal heartbeat or an anormal day of electricity consumption). Two definitions of anomaly sequence tear apart the community. The authors decided to consider that an anomaly sequence is a sequence that can be observed several times in the time-series. The method aims to define the expected behavious of the time series, and defines anomalies as subsequences that are too far from this expected behaviour. The real contribution of this new method is not an improvment in terms of accuracy, but in computational time. Indeed, the approximation of the expected behaviour prevents to compute the distances between all pairs of subsequences.
We decided to reproduce the algorithm described in the paper, to study the impact of the variation of some parameters and test it on new data.
We decided to split the work as follows : TODO. As the source code was not available, we reimplemented everything ourselves. Experiments: TODO. Improvment of the original method : TODO.

The Introduction section (indicative length : less than 1 page) should detail the scientific context of the article you chose, as well as the task that you want to solve (especially if you apply it on novel data). **The last paragraph of the introduction must contain the following information**:

- Repartition of work between the two students
- Use of available source code or not, percentage of the source code that has been reused, etc.
- Use of existing experiments or new experiments (e.g. test of the influence of parameter that was not conducted in the original article, application of the method on a novel task/data set etc.)
- Improvement on the original method (e.g. new pre/post processing steps, grid search for optimal parameters etc.)

2 Method

Classical methods for anormal sequences detection rely on the fact that anormal sequences are considered as unique sequences in the time-series. These methods label as anomalies the sequences that are far away from every other subsequence of the time series, for a given metric (euclidian, DTW for instance). However, as explained in the article, one can be looking for anomalies that are not unique, like anormal heartbeats in a ECG. The method proposed introduces the **normal model** (N_M), which is a sequence of fixed length (here $3 \times l$, with l the length of the anomalies we are looking for), that represents the "normal behaviour" of the model. The proposed method to determine the N_M is :

1. Extract the subsequences of length $3 \times l$ (randomly or with motif selection).
2. Perform a hierarchical clustering of the subsequences.
3. Select the cluster c in $\mathbb{C}$ (the set of clusters) that maximises the following quantity : $N(c) = \frac{frequency(c)^2 \times coverage(c)}{\Sigma_{x \in \mathbb{C}} dist(center(c), center(x))}$
   With $frequency(c)$ the number of sequences in the cluster, $coverage(c)$ the time lag between the first and the last sequences in the cluster, and $center(c)$ the barycenter of the cluster. This quantity aims to select the cluster with the most sequences, and that is close to all the other clusters.
4. Finally, N_M is defined as the barycenter of the cluster c.

The distance of a subsequence to the normal model is then defined as the minimal (euclidian) distance between a subsequence os size l and all the subsequences in the normal model of size l (as a recall, the normal model is of size $3 \times l$). The authors describe several metric to label a subsequence as *anormal*. Selecting the k subsequences with the biggest minimal distance to the normal model, or selecting the ones that are above a certain threshold can be methods to determine the anomalies.

The key point of the method is that instead of comparing an anomaly to all the subsequences of the time-series and then comparing it to a very large number of identical pattern, the clustring step enables to drastically reduce the number of comparisons. Indeed, each sequence of size l is compared to $2 \times l$ sequences in the normal model, instead of $T - 2 \times l$ in the whole time series (with the condition of non overlap).

We decided to implement the algorithm from scratch to try to reproduce the results from the paper.

We randomly selected TODO INSERT NUMBER OF SEQUENCES subsequences of size TODO INSERT SIZE from the original time-series. A hierarchical clustering was then performed using the scipy.cluster.hierarchy.fcluster function, with the distance criterion. TODO INSERT NUMBER clusters were identified, and the most representative cluster was selected using the criterion described above. The centroid of the cluster was then computed, and led to the following normal mode : TODO INSERT FIG.

To evaluate our model, we computed the score of each subsequence of the origin time-series, defined as the minimal distance between this subsequence and each subsequence in the normal model, and selected the TODO INSERT NUMBER most anormal subsequences.

The Method section (indicative length : 1 to 2 pages) should describe the mathematical aspects of the method in a summarized manner. Only the main steps that are useful for understanding

should be highlighted. If relevant, some details on implementation can be provided (but only marginally).

3 Data

We decided to run the algorithm on the NYC taxis dataset, which represents the number of calls to taxis in New York City every 30 minutes between 01/07/2014 and 30/01/2015, available here. This dataset seemed performant because of its big size : 10320 points. Its mean is 15137.57, with a standard deviation of 6931.92.
The figure 1 shows a day of the time-series.

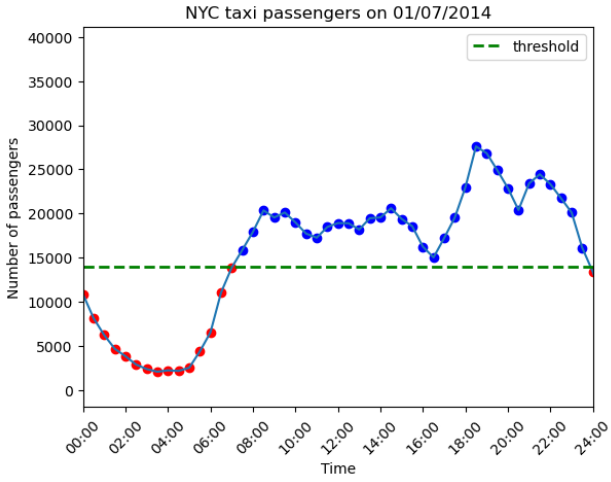
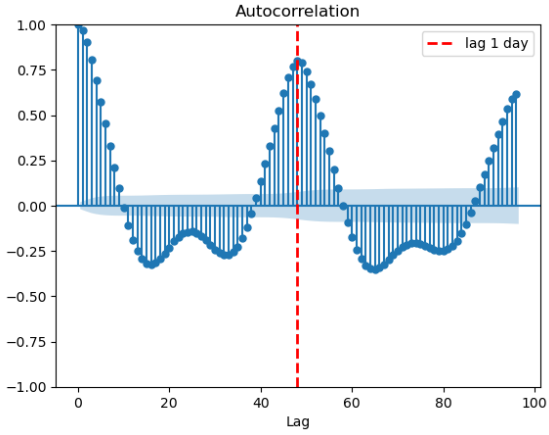
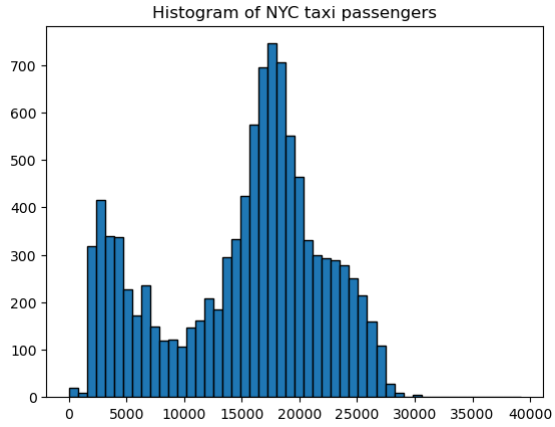


Figure 1: A day of the NYC taxis dataset



(a) Autocorrelation plot



(b) Histogram of values

Figure 2: Data analysis

Augmented Dicky-Fuller and KSS tests showed p-values of respectively 2.5e-19 and 0.06, which confirms the stationarity of the time-series. The periodicity of the time-series was confirmed by an autocorrelation plot (see Fig.2a), with a peak every 48 points, corresponding to 24 hours time-lag. An histogram of the values taken by the time-series shows a bimodal distribution with few

outliers, separated by the 14000 value (see Fig.2b). Values below 14000 correspond almost exactly to hours between 00:00 and 07:00, which is coherent with the physical meaning of the time-series (see figure 1).

The Data section (indicative length : 1 page) should provide a deep analysis of the data used for experiment. In particular, we are interested here in your capacity to provide relevant and thoughtful feedbacks on the data and to demonstrate that you master some "data diagnosis" tools that have been dealt with in the lectures/tutorials.

4 Results

The Result section (indicative length : 1 to 2 pages) should display numerical simulations on real data. If you re-used some existing implementations, it is expected that this section develops new experiments that were not present in the original article. Results should be discussed not only based on quantative scores but also on qualitative aspects. In particular (especially if your article focuses on black box methods), please provide some feedbacks whether the method was adapted to the data or not and whether the hypothesis behind the approach you used were validated or not.