Assignment 2

Devosmita Chatterjee *

Department of Mathematical Sciences, Chalmers University of Technology

*Correspondence:tel: +46 709669519, E.mail: chatterjeedevosmita267@gmail.com

1 I think you misunderstood what "meter-reset" means. Fourier transforms to detect seasonal behaviours could be relevant to district heating, because of the correlation between energy consumption and weather, but meter resets are a purely meter-technical occurence.

1.1 Look again at the definition of cumulative state meters and meter resets and reconsider the question.

According to my understanding,
"Cumulative state" water meter is defined as the total amount of water (in m^3) which has passed the meter since it was installed. "Cumulative state" water meter is strictly increasing.
"Meter-reset" is defined as state meters reset to 0. "Meter-reset" is done occasionally and then the meter will be strictly increasing normally after the reset.

- The state meter data series can normally look like '9837.1, 9837.4, 0.2, 0.7'. This is not an error.
- If the data series looks something like '9837.1, 9837.4, 98.6, 9837.8', then the datapoint '98.6' has no physical explanation and can be considered an outlier.
- If the data series looks something like '9837.1, 9837.4, 9837.2, 9837.8', then the datapoint '9837.2' is an error since state meter is strictly increasing.

1.2 Also, the plot of energy_consumption does not display seasonality, as far as I can tell, and the input data was purely random. Why did you include that plot as an example of seasonality?

I absolutely agree that the plot of energy_consumption shows no obvious seasonal pattern. Last time I mistakenly include that plot as an example of seasonality.
From the plot of energy_consumption as shown in figure below, we can visualize the facts and patterns from insights of the time series data in order to develop good and valid decisions and strategies that benefit district heating suppliers.

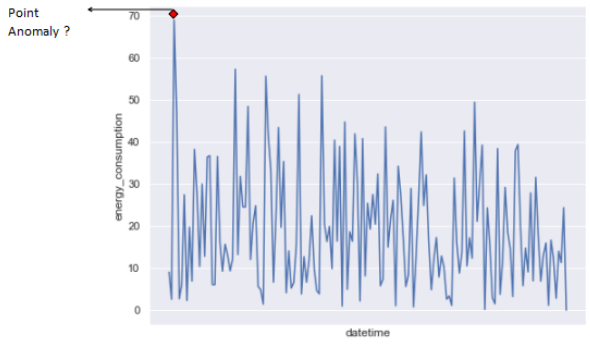
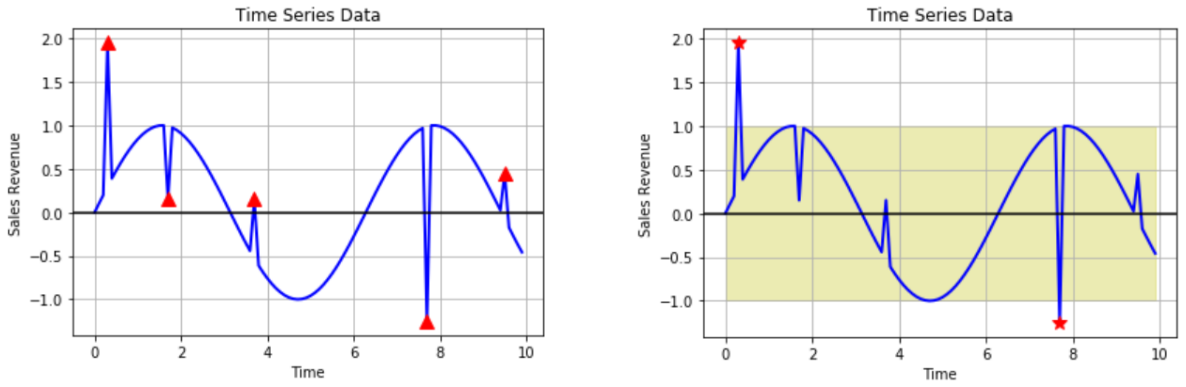


Figure 1: Plot of energy_consumption versus datetime.

2 The outlier detection method you used for finding the extremely low values of the meter-state series would not work on "real" meter-state data if the input was several years long, especially if a meter reset happened sometime during the series.

2.1 Why?

Yes I absolutely agree with your statement.
In the last code task, I used InterQuartileRange (IQR) outlier detection. IQR outlier detection method would not work for a large "real" meter-state data with meter-reset since this method can only filter out the observations at either extremes. The reason behind is that a time series data reflects the trends, cycles and seasonal patterns of the series and only an observation which is out of cycle is considered as an outlier that can not be captured by IQR outlier detection method.
Figure 2 shows an example time series data [1] which shows clearly that only extreme outliers are detected by IQR outlier detection method.



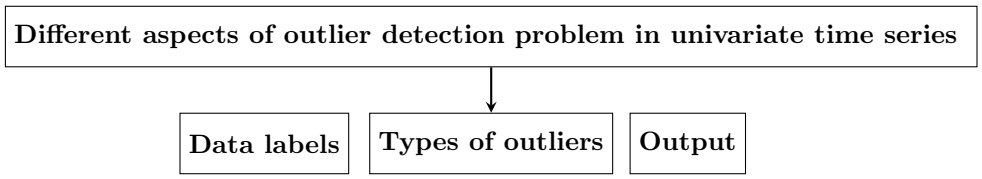
(a) Time series plot highlighting outliers.

(b) Outliers detected by IQR outlier detection.

Figure 2: Interquartile outlier detection in time series data [1].

2.2 What other methods could be used to detect incorrect datapoints in a meter-state series instead?

Meter-state series is an univariate time series data. Therefore, firstly I discuss the different aspects or properties [2] of outlier detection problem in univariate time series for selecting an appropriate outlier detection technique. Next I investigate some commonly used outlier detection techniques for univariate time series [3].



1 Data labels

Table : In the table, we present labeled data and unlabeled data.

Labeled data	**Unlabeled data**
Data is classified based on the training dataset which determines if it is a normal or anomalous data point.	Data is classified based on the properties of the given input dataset.
A labeled dataset with normal and anomalous points can be used by supervised learning. A labeled dataset with only normal points can be used by semisupervised learning.	An unlabeled dataset can be used by unsupervised learning.

2 Types of outliers

Table 2: In the table, we present different types of outliers

Types of outliers	**Definition**
Point outlier	Data point is significantly different from the rest of the data. Point outlier is detected by method with rare classification.
Contextual outlier	Data point is significantly different in a specific context. Contextual outlier is detected by methods that search for deviation.
Collective outliers	Collection of data points is significantly different from the rest of the data. Collective outliers are detected by methods that focus unusual shapes in the data.

3 Output

Table 2: In the table, we present different output produced by outlier detection techniques

Output	**Definition**	**Examples**
Scores	Scoring techniques assign an outlier score to each instance in the test data depending on the degree to which that instance is considered an outlier. Thus the output of such techniques is a ranked list of outliers. It allows an analyst to choose a domain specific threshold to select the most relevant anomalies.	Local Outlier Factor (LOF).
Labels	Labeling techniques assign a binary label (normal or anomalous) to each instance in the test data. It do not directly allow the analysts to make a choice, although this can be controlled indirectly through parameter choices within each technique.	Density Based Spatial Clustering of Applications with Noise (DBSCAN).

.

Table 3: In the table, we present some commonly used outlier detection methods for univariate time series data.

Outlier detection methods	**Nature of input data**		**Strengths**	**Weaknesses**
	Data distribution	**Applicability**		
Prediction Based Approach				
Auto Regressive Integrated Moving Average (ARIMA) - Supervised learning method predicts future data points and then detects outliers by evaluating the deviation of the predicted point to the observed one.	Data is normally distributed.	Based on future predictions.	1. Stable estimation of time-varying trends. 2. Few parameters.	1. Danger of overfitting if not handled with care. 2. Computationally expensive.
Clustering Based Approach				
Density Based Spatial Clustering of Applications with Noise (DBSCAN) - Unsupervised learning method groups together the points in clusters which are in high density regions whereas the other points are marked as outliers.	Underlying data distribution is unknown.	Based on the density in the data.	1. Number of clusters is not an input parameter. 2. Performs well with arbitrary shape clusters.	1. The data need to be scaled accordingly. Otherwise, choosing a meaningful distance threshold is difficult. 2. DBSCAN is sensitive to clustering parameters but selecting such optimal parameters can be difficult sometimes.
Proximity Based Approach				
Local Outlier Factor (LOF) - Semisupervised learning method identifies outliers by measuring the local deviation of a given data point with respect to its k-nearest neighbours	Underlying data distribution is unknown.	Based on k-nearest neighbors.	1. Only one input parameter. 2. Determines local outliers or outliers in local areas. A point that is at a small distance from a very dense cluster might be considered as a local outlier.	1. It is difficult to determine an appropriate value k for the k-nearest neighbors. 2. Time complexity of $\mathcal{O}(n^2)$ as compared to that of DBSCAN of $\mathcal{O}(n \log n)$.
Neural Networks Based Approach				
Autoencoder - Semisupervised learning method minimizes reconstruction error based on a loss function such as mean squared error and detects outlier which will be higher reconstruction error.	Underlying data distribution is unknown.	Based on dimenionality reduction.	1. Provides with multiple filters that can best fit the data. 2. Improves the performance of the data in some cases.	1. Training an autoencoder takes a lot of time. 2. It captures more information rather than relevant ones.

3 Standard outlier detection methods work poorly on district heating data, because the distribution of consumption is very far from normal. For many substations, $< 5\%$ of the hours in a year combine to $> 90\%$ of the consumption that year. What methods could be used to detect erroneous datapoints in such cases?

Unsupervised outlier detection methods seem to be the most reasonable choice when the data distribution is very far from normal since unsupervised outlier detection algorithms do not make assumptions about the distribution of data. Clustering based unsupervised approach such as Density Based Spatial Clustering of Applications with Noise (DBSCAN) could be used to detect erroneous datapoints in above cases.

References

[1] A. Bhattacharya, "Effective approaches for time series anomaly detection," Jul 2020.

[2] C. C. Aggarwal, *Time Series and Multidimensional Streaming Outlier Detection*, pp. 273–310. Cham: Springer International Publishing, 2017.

[3] M. Braei and S. Wagner, "Anomaly detection in univariate time-series: A survey on the state-of-the-art," 2020.