**Center for**
**Economic Theories and Policies**
Sofia University St. Kliment Ohridski
Faculty of Economics and Business Administration

ISSN: 2367-7082

Bulgarian Economic Papers

# Time Series Decomposition for Anomalous E-commerce Transactions

**Anton Gerunov**
**Ilia Atanasov**
**George Mengov**

**BEP 04-2020**
**Publication: November 2020**

# Time Series Decomposition for Anomalous E-commerce Transactions

*Anton Gerunov[1], Ilia Atanasov[2], George Mengov[1]*

*[1]Department of Industrial Economics and Management, Sofia University St. Kliment Ohridski and Centre for Modelling Socioeconomic Systems*

*[2]Department of Economics, Sofia University St. Kliment Ohridski*

**Abstract**:
Online trading is one of the pillars of the digital economy. The rapid increase of e-commerce transactions has increased the risk exposure of providers and made it virtually impossible to track consumer behavior by relying on human experts alone. Here we show how time series decomposition can be used to automatically detect suspicious transactions and flag them out for subsequent actions. The identified outliers have clear business meaning and can be interpreted as peaks in demand produced by idiosyncratic consumer behavior or by malicious activity. Either way, they deserve sufficient attention and active management.

**Key words**: Anomaly detection, time series decomposition, e-commerce, online trade

## 1. Introduction

Discovering potential anomalies at the individual transaction level comes with some challenges. The goal is to identify contextual anomalies, but a significant part of the context – the human being – is missing. Certain observations would be abnormal if they were made by one class of agents (e.g. clients), but completely normal if they were performed by another class of agents (e.g. employees). In this sense, it is always appropriate to work with data for the agent who performed the action, which can be human or automated (systems, interfaces, scripts, etc.). When identification of the agent is not possible due to missing data, a transaction-level anomaly approach can be used.

The analysis of trade data and the identification of extremely high or low values of consumer demand is a classic task in operational risk management. It enables the optimization of human and organizational resources, and inventory stock. Previously, when data sets were relatively small, this task was solved at an aggregated level, considering the total quantities of exchanged goods or services and making decisions about the necessary stocks and employees. In the era of large data sets, low computational costs, and flexible statistical algorithms, it becomes not only possible but even necessary to work at the level of the individual order or individual user (agent). This allows the risk of extreme behavior to be assessed at a much lower level and to take appropriate action to address it.

## 2. Methodology

The extreme values of certain observations show a significant discrepancy between current and expected realization and in this sense are potential anomalies. The detection of anomalous observation in relation to historical data is a de facto task for unsupervised machine learning. To solve it, one could apply classical statistical approaches (Rousseeuw & Hubert, 2018) as well as algorithms based on distance and density (Domingues et al., 2018).

In this publication we adopt a hybrid approach employing a time series analysis. First, we calculate the descriptive statistics of the data and then analyze the relationship between the expenditures made by the clients, the size of the order, and the value of the order. From the main sample we extract the quantities sold in each transaction and the value of the transaction. We combine the quantities sold on daily basis, thus forming a single time series of 302 aggregated daily observations. Then we do the same for the value of the transactions. As in both time series we observe seasonal cycles and trends, these are corrected with time series decomposition methods. The anomaly detection procedure is used on the residuals.

We consider two procedures for time series decomposition – the Seasonal and Trend decomposition (STL) using Loess (Cleveland et. al., 1990), and the Piecewise Median (also called the Twitter method), developed by Vallis et. al. (2014). The STL method removes an estimated trend component from the time series, splits the data into a sub-cycle series and then smooths them by using Loess. This process is then repeated until convergence on the decomposition is observed. The Piecewise Median method extracts the seasonal component by using STL and then removes the median and the seasonal component from the observations. The STL method performs better in circumstances where a long trend is present, because the Loess smoother is good at detecting trends. The Piecewise Median performs better when the cyclical component is more dominant than the trend.

After the trend and the cyclical components are removed from both series, we adopt a statistical approach for detecting anomalies. The general idea is that the residuals of the cleared time series can be viewed as distributions of observations. Then statistical methods for detecting outliers can be employed to identify the anomalous observations.

For the two distributions (one for quantities sold per day and one for the total value of sales per day), the inter-quartile range (IQR) defines two other distributions of observations, positioned between the first and the third quartile. Two limits are set above and below the IQR with the use of an IQR factor ($F_I$) determined by an expert. The IQR range is multiplied by the factor and the lower limit is set below the first quartile at a distance $F_I IQR$. The upper limit is set above the third quartile at the same distance. Any observation that is located beyond these limits (above the upper and below the lower) is considered anomalous. The IQR factor is calculated as $F_I = 0.15/\alpha$, where α controls how difficult is for an observation

to be identified as an anomaly. The relationship between α and the ease of detection is inverse – higher values of α lead to easier classification of an observation as an anomaly.

## 3. A Practical Example

By way of example here we use a known data set (Chen et at., 2012) and analyze it with methods of unsupervised self-learning. The data are about orders in an online store in the UK and is composed of 541,909 records from 01.12.2010 to 09.12.2011. Each observation contains the invoice number of the sale, code and description of the goods, ordered quantity, unit price, time-stamp of the invoice, customer number, and country.

*Table 1: Descriptive statistics of data for orders in the online store, N = 541,909*

| Variable | Mean | St. Dev. | Median | Min. | Max. | Skewness | Excess |
|---|---|---|---|---|---|---|---|
| *Quantity sold* | 9.55 | 218.08 | 3.00 | 0 | 80995 | -0.26 | 119767.61 |
| *Date and time of sale* | 11665.03 | 6677.69 | 11606.00 | 1 | 23261 | 0.02 | -1.20 |
| *Date of invoice* | 156.48 | 88.75 | 157.00 | 1 | 305 | -0.05 | -1.23 |
| *Unit price* | 4.61 | 96.76 | 2.08 | 0 | 38970 | 186.51 | 59004.96 |
| *Total amount of the order* | 17.99 | 378.81 | 9.75 | 0 | 168470 | -0.96 | 151196.04 |

*Note: Times are relative. Prices are in GBP.*

Table 1 presents a summary of the data. The online store sells 4,224 products, of which 4,070 have a unique code and the rest only a description. The average quantity of an order is 9.55 goods, but with a huge standard deviation (σ = 218). The smallest orders either fail to complete (0 items) or have 1, while the largest order contains almost 81,000 pieces, the latter probably a wholesale order. The average order is relatively small, GBP 17.99, but again with a high standard deviation (σ = 378.81). The total number of unique customers is 4,372, and orders come from 38 different countries, mostly from the European Union.

The dynamics of the total sold quantities is shown in Figure 1. It shows an upward trend with some periodicity. The latter probably come from the change of seasons as well as from other periodic time changes. Beyond that, there are clear peaks and troughs in demand. These are the potentially anomalous moments that should be adequately predicted and managed.

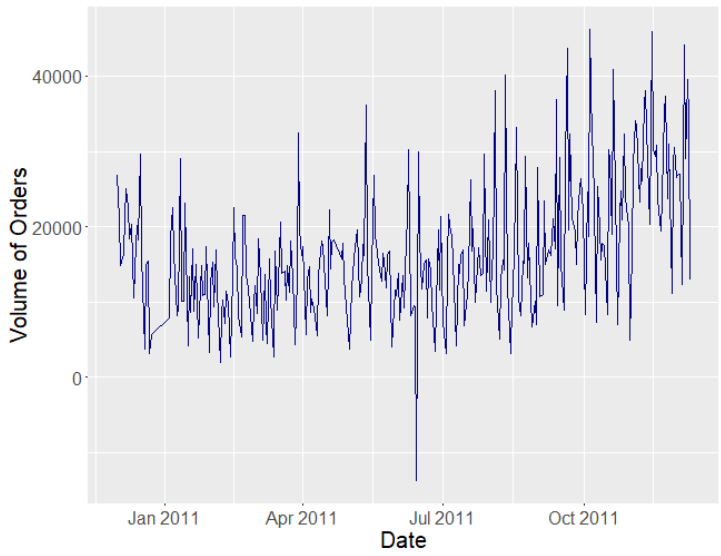


*Figure 1: Dynamics of ordered quantities from an online store.*

Figure 2 shows the dynamics of the total value of the quantities sold (daily turnover). It has a clear tendency to increase total daily value. We observe similar periodic changes as in the ordered quantities, which marks the strong and stable relationship between number of traded quantities and turnover observed most of the time. This is also evident in the large statistically significant relationship between them, Pearson's correlation being 0.88 (Figure 3).

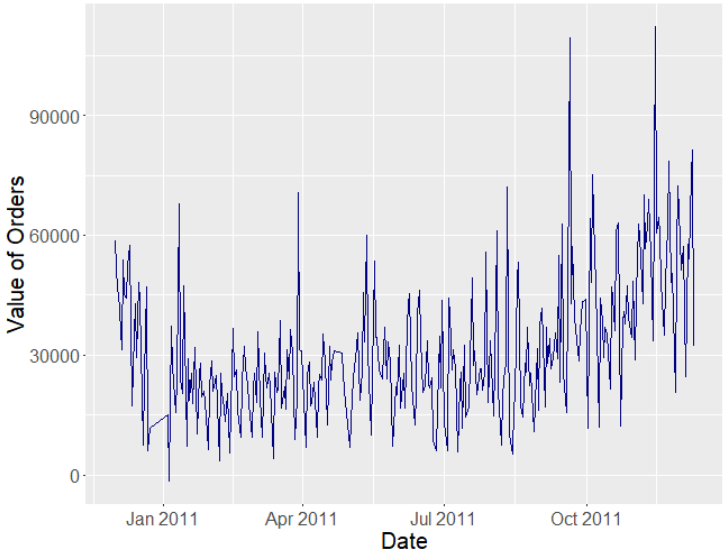


*Figure 2: Dynamics of order value from an online store. Value of orders is in GBP.*


*Figure 3: Correlation between volume and value of orders. Value of orders is in GBP.*

On the other hand, the relationship between the average number of purchased goods and the total amount of money spent by a customer is not strong. Visual inspection (Figure 4) does not suggest a clear tendency.

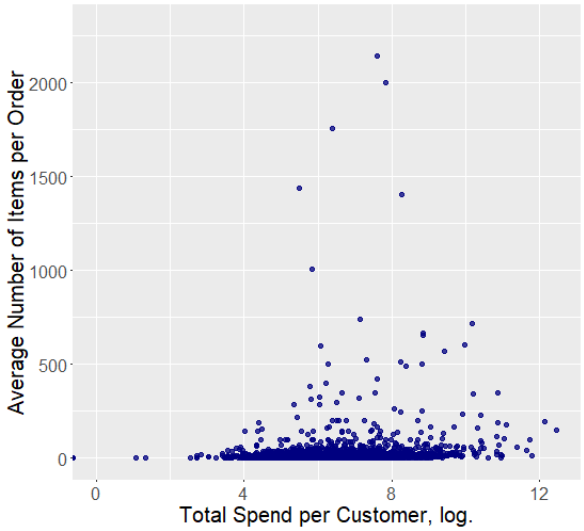


*Figure 4: Relationship between the average number of goods in an order and the total spending per customer (in GBP).*

We identified a small number of orders with an extremely high number of identical items, probably wholesale. We observe a similar trend in the relationship between the average price of an ordered product and the total spending by a given customer (Figure 5). The sample is

again dominated by many small orders (with a low average price), but consumers tend to buy a relatively large quantity of them.

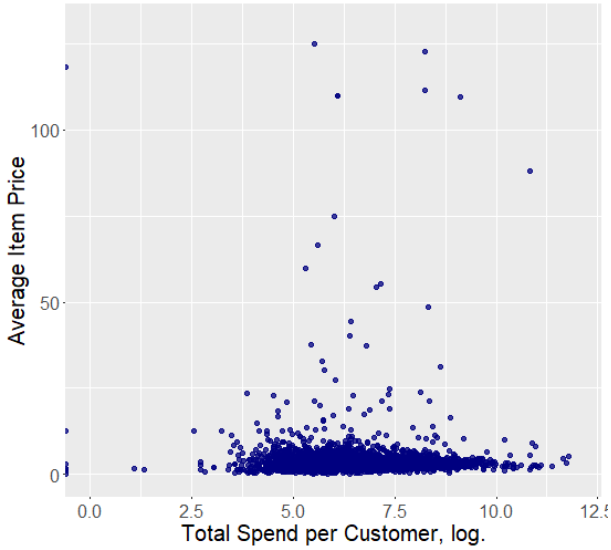


*Figure 5: Relationship between the average price of the goods in an order and the total spending per customer (in GBP).*

This combination of buying cheap items in large numbers leads to the fact that we do not see a clear relationship between average price and total turnover (r = -0.01). In general, the correlation matrix of the data does not show the presence of significant statistical links. This is an additional reason to use criteria-based methods to search for contextual anomalies instead of regression algorithms.

## 4. Potentially anomalous observations

As a first step in the analysis of anomalies at the level of an individual agent, it is appropriate to consider the general context. In this case, we have detailed time series of transactions and we can consider to what extent their behavior at the aggregate level is normal and whether we observe anomalies at this higher level. For this purpose, it is appropriate to use methods to search for anomalies within time series. It is known that a time series can be broken down into its main components: long-term trend, seasonal components and random residuals (see Cleveland et al, 1990). Here we use the procedure developed in (Gerunov, 2016). We identify potentially anomalous observations as unusual deviations in the order of random residuals in decomposition.

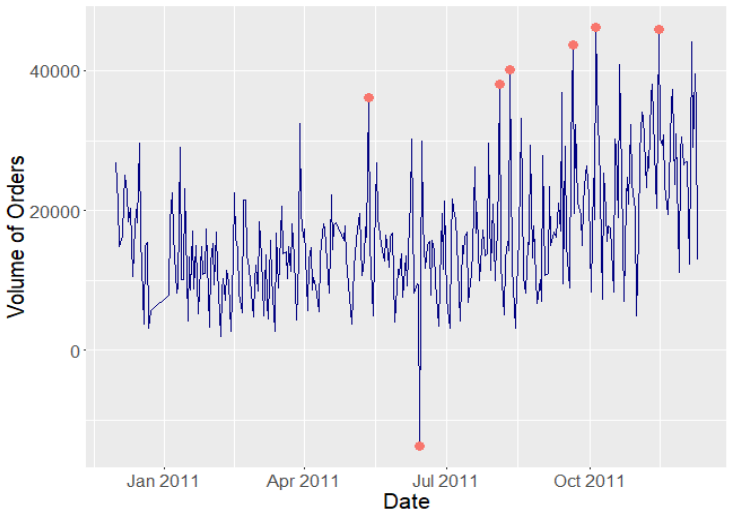


*Figure 6: Aggregated transactions in online store data. The red dots are potentially anomalous transactions.*

We transform the data set by aggregating all transactions for every date in the database and obtain two time series - one with the number of daily orders and one with the total value for a day (daily turnover). The number of new observations is N = 302. To identify potential anomalies, we use a time series decomposition method (Vallis et al., 2014), applying it first to the number of transactions and then to the amount of daily turnover. We identify seven potentially anomalous observations, 2.3% of the sample, with six of them unusually high and one unusually low (Figure 6). These are cases that demand special attention and probably organizational measures (increasing availability, ensuring the stability of the platform and the payment system, etc.). The automated algorithm reports four of these transactions (1.32% of the sample) as unexpectedly high turnover. These are risks with positive consequences and the organization could take steps to increase their beneficial effect. Otherwise they would be defined as missed business opportunities.

## 5. Conclusion

Here a concrete example was given how a time series decomposition can be usefully applied for identifying outliers in the behavior of the online trade data. In e-commerce, this method can be used to both manage peak demand, and also to mark potentially malicious activity such as fraud or service out of function. Automatic application of anomaly detection algorithms over troves of big data is now crucial for the smooth functioning of online trade.

## 6. Acknowledgement

This work was supported by Sofia University “St. Kliment Ohridski”, grant contract number 80-10-204/28.04.2020.

**REFERENCES**


Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, *19*(3), 197-208.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. J. (1990). STL: A seasonal-trend decomposition procedure based on Loess. *Journal of Official Statistics*, 6(1), 3–33.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition. *Journal of Official Statistics*, *6*(1), 3-73.

Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition*, *74*, 406-421.

Gerunov, A. (2016). Automating Analytics: Forecasting Time Series in Economics and Business. *Journal of Economics and Political Economy*, 3(2), 340-349.

Rousseeuw, P. J., & Hubert, M. (2018). Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(2), e1236.

Vallis, O. S., Hochenbaum, J., & Kejariwal, A. (2014). AnomalyDetection: Anomaly Detection Using Seasonal Hybrid Extreme Studentized Deviate Test. *R Package Version*, *1*.