

ANOMALY DETECTION IN TIME SERIES DATA

N G Dulanjalee S Niralgama

199542 R

Degree of Master of Financial Mathematics

Department of Mathematics

University of Moratuwa

Sri Lanka

May 2024

ANOMALY DETECTION IN TIME SERIES DATA

Niralgama Gamaethige Dulanjalee Samanthika Niralgama

199542 R

Thesis/Dissertation submitted in partial fulfillment of the requirements
for the

Degree of Master of Science in Financial Mathematics

Department of Mathematics

University of Moratuwa

Sri Lanka

May 2024

Declaration

I declare that this is my own work and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date:

The above candidate has carried out research for the Masters Dissertation under my supervision.

Name of the supervisor:

Signature of the supervisor:

Date :

Abstract

Fraudulent activities have been affecting the world's day-to-day business, resulting losses over millions to its users. Due to the advancement of technology, fraudsters have many platforms to stay undetected and fraud detection has become critical for organizations and stakeholders such as bankers, investors, credit card users, financial auditors and insurance agents. In addition to detecting fraud, anomaly detections such as diagnosing certain sicknesses, identifying climate change disasters, early warnings on floods and volcano activations are important for many stakeholders. Time series data, in any field can provide significant information when it comes to detecting anomalous activities. Bayesian network, neural network, logistic method, K-nearest neighbor and regression are the most popular data mining approaches in anomaly detection that are being introduced so far. Finding anomalies among time series data is the key approach of detecting fraudulent activities and identifying striking real-world scenarios such as economic recessions. The existing methods and approaches have limitations such as manual thresholds to determine anomaly points, lack of theoretical foundation and lack of attention to class imbalance problem and multivariate data. Approaches such as interquartile rang (IQR) methods assumes that time series data are normally distributed. This assumption does not comply with practical data as they could be positively or negatively skewed. Therefore, to overcome these limitations, this research project introduced a framework to detect anomalies in both univariate and multivariate time series data using extreme value theory. Results of this framework demonstrates that this method's succeeds identifying the anomalies better than the methods implemented in the two R software packages, anomalize and timetk for detecting anomalies in univariate timeseries data. Further, this framework visualizes the identified anomalies which can be utilized by all stakeholders in for both univariate and multivariate anomalies.

Keywords: anomaly detection, fraud detection, univariate time series, multivariate time series, extreme value theory, data visualization.

Acknowledgement

I want to express my profound gratitude Dr Priyanga D Thalagala, who served as my supervisor and helped me write this thesis. Her guidance, patience, insightful feedback, continuous checks on the progress of my thesis and support to fill the gaps in my knowledge is invaluable. Even when I faced challenges, she provided unwavering support, helping me navigate the complexities of my research.

Many thanks to my husband, who has helped me throughout this research work, for his continuous inspiration, encouragement and for patience, late-night pep talks, and belief in my abilities kept me going when the research felt overwhelming. Without his constant dedication and assistance in every single moment, this thesis would not have been possible.

And most importantly, heartiest thanks to my parents for their unwavering encouragement and belief in me, which fueled my determination. From my childhood dreams to this very moment, they have been my unwavering cheerleaders. This achievement is as much theirs as it is mine.

Table of Contents

Declaration	I
Abstract	II
Acknowledgement.....	III
Table of Figures	VI
Chapter 01	1
Introduction	1
1.1 Anomaly detection in time series context	1
1.1.1 Importance of anomaly detection	1
1.1.2 Significance and motivation for this research	2
1.2 Different types of anomalies in time series data	2
1.3 Extreme value theory (EVT)	3
1.4 Methods of anomaly detection in time series data	5
Chapter 02	7
Literature Review.....	7
2.1 Background study on fraud detection and anomaly detection	7
2.2 Background study on anomaly detection in multivariate time series data.....	8
2.2 Extreme value approaches to anomaly detection	9
2.3 Methodology developed in “anomalize” package in R.....	11
2.4 Methodology developed in “timetk” package in R.....	11
2.5. Methodology developed in “HDoutliers” package in R	12
2.6 Research Gap	13
Chapter 03	14
Methodology	14
3.1 Anomaly detection in time series data	14

3.1.1 Applying the framework for univariate time series analysis.....	15
3.1.2 Applying the framework for multivariate time series data	16
Chapter 04	17
Results and Discussion.....	17
4.1 Anomaly detection in univariate time series in bitcoin data	17
4.2 Anomaly detection using anomalize and timetk packages in Bitcoin data	20
4.3 Anomaly detection in univariate time series in USA unemployment data	23
4.4 Anomaly detection in multivariate time series in crypto currency data	29
4.5 Anomaly detection in multivariate time series in economic data in USA	35
Chapter 05	42
Conclusion and Further Work.....	42
5.1 Anomaly detection in univariate time series in bitcoin closing price	42
5.2 Anomaly detection in univariate time series USA unemployment data	44
5.3 Anomaly Detection in multivariate time series cryptocurrency data.....	49
5.4 Anomaly Detection in multivariate time series in economic data in USA	50
Chapter 06	52
References	52
Annexture	58

Table of Figures

Figure 1. 1: The PDF of GEV for different shape parameter values	4
Figure 3. 1: Novel Framework to identify anomalies in time series data	14
Figure 4.1. 1: Closing Price of Bitcoin	17
Figure 4.1. 2 : Close Rate of Bitcoin data (after transforming to a stationary data set)	17
Figure 4.1. 3 : Fitted ARIMA model for the Bitcoin Stationary data set.....	18
Figure 4.1. 4: Residuals vs Date plot, the ACF and the Distribution of the Residuals of the fitted ARIMA model	18
Figure 4.1. 5: Ljung Box Test on the residuals of the ARIMA model for Bitcoin stationary data	19
Figure 4.1. 6: Identified Anomalies (in red) on the stationary Bitcoin closing price change rate data set	19
Figure 4.1. 7: Anomalies identified using the EVT based threshold, on the original data set (on Closing price of Bitcoin)	20
Figure 4.2. 2: Anomaly detection using timetk method in R of the closing price of bitcoins	21
Figure 4.2. 1: Anomaly detection using anomalize method in R of the closing price of bitcoins	21
Figure 4.2. 3: Residuals of the anomalize method's STL decomposition	22
Figure 4.2. 4: Showing when the price increments started occurring in 2017	22
Figure 4.3. 1: Quarterly unemployment rate in USA.....	23
Figure 4.3. 2: Fitted ARIMA model for first order differenced unemployment rate.	24
Figure 4.3. 3: Unemployment rate after first order differencing.....	24
Figure 4.3. 5: Exploring the residuals of the fitted ARIMA model for first order differenced unemployment rate.....	25
Figure 4.3. 4: Ljung Box test for the residuals of the fitted ARIMA model.....	25

Figure 4.3. 6: Identified anomalies of unemployment rate - via HDoutliers package	26
Figure 4.3. 7: Anomalies identified by the HDoutliers package mapped to the stationary dataset of unemployment data.....	26
Figure 4.3. 8: Anomalies identified by the HDoutliers method, mapped to the original dataset.....	27
Figure 4.3. 9: Anomalies identified by the Anomalize package	28
Figure 4.3. 10: Anomalies identifies by the timetk package.....	29
Figure 4.4. 1: Bitcoin closing prices from 2016 to 2019	29
Figure 4.4. 2: Lite-coin closing prices from 2016 to 2019.....	30
Figure 4.4. 3: Bitcoin rate data - after transformation.....	30
Figure 4.4. 4: Litecoin data after transformation	31
Figure 4.4. 5: Cross correlation of the stationary data sets	31
Figure 4.4. 6: Lead-lag correlations of the two-dimensional dataset.....	32
Figure 4.4. 7: Significant plot of cross-correlation matrix.....	32
Figure 4.4. 8: Fitted VARMA model for the two-dimensional dataset	33
Figure 4.4. 9: Multivariate portmanteau test on the residuals of the fitted VARMA model.....	33
Figure 4.4. 10: Significant plot of the cross-correlation matrix of the residuals of the fitted model	34
Figure 4.4. 11: Anomalies identified with Alpha values 0.06 to 0.19 - marked in black	34
Figure 4.4. 12: Anomalies identified when alpha values were set to 0.195.....	35
Figure 4.5. 1: GDP, Unemployment rate and CPI data in USA from 1947 to 2023..	36
Figure 4.5. 2: Stationary data set (GDP rate, Differed Unemployment Rate and CPI Rate in USA).....	37
Figure 4.5. 3: Correlations between GDP Rate, Differed Unemployment Rate and CPI Rate	38
Figure 4.5. 4: Significant plot of the cross-correlation matrix of the stationary multivariate data.....	38

Figure 4.5. 5: Fitted VARMA model results of the stationary multivariate data	39
Figure 4.5. 6: Lead-lag correlation of the residuals	40
Figure 4.5. 7: Significant plot of cross-correlation of the residuals.....	40
Figure 4.5. 9: Anomalies identified at alpha 0.2 – marked in black	41
Figure 4.5. 8: Anomalies identified at alpha 0.05 - marked in black.....	41
Figure 5.1. 1: Identified anomalies using the EVT based anomalous threshold calculation	43
Figure 5.1. 2: Anomalies identified by the anomalise method	44
Figure 5.1. 3: Anomalise identified by the timetk method.	44
Figure 5.2. 1: Anomalies identified by the introduced framework.....	45
Figure 5.2. 2: Anomalies detected by the anomalise method	46
Figure 5.2. 3: Anomalies identified by the timetk method	46
Figure 5.2. 4: Results of the introduced framework with alpha is 0.2	48
Figure 5.2. 5: Results of anomalise method with alpha 0.2.....	48
Figure 5.4. 1: Multivariate Anomaly detection results with alpha 0.2.....	51