Agnieszka DURAJ[1], Magdalena LUDWICKA[1]

Politechnika Łódzka, Instytut Informatyki, Wydział Fizyki Technicznej, Informatyki i Matematyki Stosowanej (1)

doi:10.15199/48.2019.01.10

Modelling volatity of time series data containing outliers observations with ARCH effect

Streszczenie. *Przedmiotem niniejszej pracy jest analiza porównawcza wybranych modeli służących do opisu zmienności szeregów czasowych, w tym wyjątków. Artykuł koncentruje się na dynamicznych właściwościach szeregów czasowych, na ogół na heterogeniczności warunkowej wariancji w czasie. W niniejszym artykule opisano powszechne metody wykrywania wartości odstających, modelowania i prognozowania szeregów czasowych. Na podstawie badań przeprowadzonych przez RF Engle, TB Bollerslev, J. Caiadoin, zbadano wybrane ARIMA, ARCH i GARCH. Zwrócono uwagę na efekt ARCH w szeregach czasowych i jego wpływ na zmienność modelowania finansowych szeregów czasowych, które zawierają odstające. Badania wykazały, że typowymi cechami finansowych szeregów czasowych są tak zwane pogrupowane wariancje. Dlatego wykorzystanie modeli ARIMA do prognozowania było niewystarczające, modele ARCH i GARCH prezentowały dobre właściwości statystyczne do modelowania danych szeregów czasowych.* ***Modelowanie zmienności w czasie szeregów danych zawierających obserwacje efektu ARCH***

Abstract. *The subject of this work is a comparative analysis of selected models used to describe the volatility of time series including exceptions. This paper is focus on the the dynamic properties of the time series, generallyon the heterogeneity of conditional variance over time. This paper describes common approaches to detecting outliers, modelling and forecasting time series. Based on the researches performed by R. F. Engle, T. B. Bollerslev, J. Caiadoin, were examined selected ARIMA, ARCH and GARCH.An attention was paid to the ARCH effect in time series and its impact on the modelling volatility of financial time series, which contain outliers. The studies showed that the typical features of financial time series are the so-called grouped variances. Therefore, using ARIMA models for forecasting was insufficient, ARCH and GARCH modelsshowed good statistical properties for modelling time series data.*

Słowa kluczowe: wykrywanie wyjątków, strumienie danych, modele ARCH, GARCH.
Keywords: outlier detection, data stream, ARCH and GARCH models

Introduction

Modeling and forecasting the volatility of time series has become a popular research area. Currently, this topic attracts the attention of scientists, researchers and economists who are focused on creating models, analytical tools to describe mathematically changes in real world. Process of data gathering is very complicated and exposed to many distortions. Nevertheless, modelling of data becomes more and more popular. It is worth mentioning that government, economic and engineering data are usually published as time series. Time series are the measurements of variable taken at regular intervals over time. Observations are made sequentially over time at regular intervals, such as daily, weekly, monthly, quarterly or hourly.

Some time series data are incomplete, contain many missing observations, which may affect the estimation of the model and the result of its estimation. Commonly dataset has missing values that impacts on the model correctness. What is more, data is susceptible to change due to noise and exceptions. Statistical noise is an unexplained variability within a given data sample, which impairs the model's correctness. An outlier is an observation that is inconsistent with the remainder of the given data set. Another definition says that an outlier is an important kind of deviation, which occurs in the data set as a result of measurement error. Outliers existence also in case of heavy-tailed distribution of population. [2]

An outlier can be perceived as a valid data point or a representative of a noise. Identification of the outlier's nature positively impacts on the process of fitting the best model to the given sample. Using the visualization methods is the best solution to discover outliers as these observations lie outside the overall pattern of distribution. Usually the distribution of observations differs from the normal distribution, which is perceived as one of the characteristics of the time series. [3]

The rise and the fall of values are linked to a number of reasons, such as example: economic growth, international trends, general business conditions, product demand. Outliers lie outside the overall pattern of a distribution, so can be discovered using the visualization methods. Charts like an example: scatter diagram or histogram, may be used to illustrate the distribution in each dataset and indicate this way the outliers. It is easy to compare examined data distribution and normal distribution via scatter diagram. What is more, visualization of the data sets is a good example of a tool to distinguish between two main types of exceptions, namely, univariate and multivariate. A univariate outlier is a data point that is significantly different from the other values of only one variable. A multivariate outlier is a combination of unusual scores that may be observed for at least two variables. [2]

Outliers can take different forms, depending on the distribution of the examined dataset: [6]

- an extreme value, very high or small in comparison to others observation
- a contaminant, understood as observation from other distribution
- a legitimate, allowed but also surprising observation
- an incorrectly measured observation

Review of literature

Usually models are used to explain the cause of changes visible in the economy and predict its changes in the future. Analytical tools are used specially to predict future values of the time series data. Process of data gathering is complicated and can be exposed to many distortions.

Many studies focus on testing the dynamic properties of time series, generally on volatility clustering. Detecting this effect may ease prediction of the unpredictable time series behaviors. C. Hafner presented the financial time series as the accumulation of independent, identically distributed, random variables. He discovered that usually the size of changes is similar in the whole examined period, namely large changes are followed by large changes, small changes are followed by small changes. This effect is visible for time series data and impacts directly on the entire model. Model correctness may be affected also by outliers' appearance. The occurrence of outliers in a given time series was carried out by A. J. Fox in 1972. An article entitled "Outliers in Time Series" mentions unexpected, extraordinary observations and methods for detecting outliers in each set. It was the first publication describingthe problem of exceptions. Furthermore, A. J. Fox defined two types of the outliers: additive and innovative.

In the following years, number of researches were publish which try to identify and explain the role of outliers in statistical analysis. J. W. Osborne and A. Overbay summarized in their publication the potential causes of occurrence the extreme scores, outliers, in a data set. The reasons for the occurrence of outliers were indicated, namely: data errors, deliberate or justified incorrect reporting, sampling error, standardization error.[7]

Both the origin of outliers and the impact on the entire data set have been examined to this day. Some researchers preferred visual inspection of the data. E. N. Lornez in 1987 said that outlier detection is only a special case of the data examination for influential data points.Guttmanand Tiao (1978), Miller (1980) and Chang (1982) presented in their work that the effect of outliers' occurrence is a partial autocorrelations and autoregressive moving average parameters.[2]

This topic is investigated by many authors, because the problem of detecting outliers is one of the most import research problem. Despite the many studies conducted, there is no unambiguous way to detect outliers in a dataset. A rich overview of the topic is presented by S. Panigrahi and H.S. Beher in [9]. One of the proposed methods by polish researches is detecting outliers by using linguistic summaries, it is described in detail in [15]. The second currently investigated method is use a multicriteria optimization. More details about this method can be found in [14][16][17].

Methods used for Time series volatility modelling

The question, at is the best way to describe the data that makes up the outliers is still open. Engle in 1982 presented Autoregressive Conditional Heteroskedasticity (ARCH) model.[1] He focused on the modelling financial time series that exhibit time varying conditional variance. In 1986, Bollerslev created a generalized arch (GARCH), which is the ARCH model extended with a function that is estimating stochastic volatility. Both mentioned models are generally used in various branches of econometrics, especially in a financial time series analysis.

J. Caiado in publication [8] examined the volatility of daily and weekly returns of the Portuguese Stock Index PSI-20 using simple GARCH, GARCH-M, Exponential GARCH (EGARCH) and Threshold ARCH (TARCH) models. Based on the research presented in the mentioned article, there were discovered significant asymmetric shocks to volatility in a daily stock returns and smaller shocks, when monthly returns were examined. [8]

Time series showing nuclear energy production examination

The process of outlier analysis was carried out on the information found on the U.S. Energy Information Administrationpage.Analysis of data regarding production areas, electricity or nuclear energy production, what is the topic of this essay, is not often examined. Nevertheless, nuclear electricity is perceived as one of the most efficient source of energy and the demand for nuclear energy is still growing. Data that is examined in this research shows the value of Nuclear Electricity Net Generation in Million Kilowatthours from February 1973 to March 2018. This data set was downloaded from U.S. Energy Information Administration, report was entitled: Monthly Energy Review June 2018. As was indicated by J. Caiado in [8], examination of the monthly returns relates to less significant asymmetric shocks to volatility than in case of a daily stock returns. Therefore, due to the benefits associated with the analysis of returns, instead of the value of nuclear electricity production, new variables were created, namely simple rates of returns. The advantage of using return rates instead of the given instruments values is the normalization of all variables. Most classical statistics assume the normality of the distribution of the examined feature.

After the normalization, missing values in a given sample, was identified and removed. The examined data, showing nuclear energy production changes, is presented using selected graphical methods. Outliers lie outside the overall pattern of a distribution, so it is easy to spot them using histograms or scatter diagrams. In figure 1 is showed the timeplot illustrating the nuclear energy production changes over the time. It is visible that the variance of the model is not constant in the examined period.

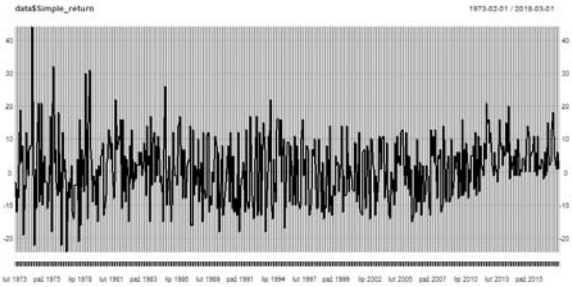

Fig. 1 Timeplot showing nuclear electricity production changes – data converted to simple returns

Taking into consideration problems with time series characteristics, like outliers appearance, volatility clustering, not constant variance over the time, different approaches may be considered. The question means what is the best model for describing time series data and forecasting future values of these series.

Based on publication of J. Caiado, [8], it was decided to test different ARCH and GARCH models.What is more selected ARIMA models were created. Then the forecasts were created and the quality of all models were checked. Usingthemost popular error measures and penalized-likelihood criterion (AIC), five models were selected and presented below. Akaike Information Criterion (AIC) is criterion for choosing between statistical models with a different number of variables explaining the model. [12]

Mentioned above models are:

ARIMA(0,1,1)
ARCH(1)
ARMA(1,2) with GARCH(1,1)
Threshold ARCH
GJR-GARCH

Autoregressive Integrated Moving Average (ARIMA) is one of the most popular models for forecasting a time series. Time series data can be transformed into "stationary data" by differencing, nonlinear transformations or deflating. ARIMA models can be used to check stationarity of the examined data set. In this model, it is assumed that autocorrelations, correlations with its own prior deviations from the mean, remain constant over time and a random variable in this model is combination of signal and noise. ARIMA model consists of three elements [9]: autoregressive process – AR, moving average process – MA, degree of integration – I.

The problem of autocorrelated errors in a random walk model was fixed in two different ways: by adding a lagged value of the differenced series to the equation or adding a lagged value of the forecast error. The best way to correct an autocorrelation is addition to the model the AR term in case of positive autocorrelation and in the opposite situation, negative autocorrelation, it is usually good to adding an MA term. Usually in analysis of the business and economic time series, negative autocorrelation often arises, as an artifact of differencing. [5]

ARIMA(0,1,1) with constant is a simple exponential smoothing with growth. It is abasic model that considers a parameterization of the conditional variance of the time series,which is based on the order one lag on squared past recent perturbations.

Engle's ARCH test, which is a Lagrange multiplier test, was executed. It used to assess the significance of ARCH effects. This tests indicated the occurrence of the ARCH effect. Therefore, the results of the ARIMA estimation are very weak. ARIMA models are not able to consider the effect of variance grouping.

ARCH (Autoregressive Conditional Heteroscedastic process) is a model based on the autoregressive process with conditional heteroscedasticity, in which the variance of the random component in the autoregressive model is explained by the appropriate equation. The ARCH (q) model was introduced by Engle in 1982. It allows to describe the inhomogeneity of a random component in time, namely heterogeneity of conditional variance over time [1]. The disadvantage of the model is assumption that positive and negative shocks have the same effect on variability, because they depend on the square of previous shocks. In fact, the value of production may react differently to positive and negative shocks in economy. The same problem is related to the GARCH modelsassuming the positive and negative error terms have a symmetric effect on the volatility. [10][11][12]

As the ARCH effect is found to be significant, it was decided to test also other models based on the ARCH. The Threshold GARCH (TGARCH) model was created by Jean-Michel Zakoian in 1994. This model's specification is based on the ARCH model with conditional standard deviation instead of conditional variance.[4][5]

GJR-GARCH (1,1) models were introduced independently by Glosten, Jaganathan and Runkle (1993) to take into account the leverage effect. This is a GARCH model with the addition of a dummy variable which is multiplied by the square of the error term of time spent in the conditional variance equation. [10][11]

Results of prediction using all listed above methods are shown in table 1. Criterions measures, AIC and BIC values are very low in case of modelling using ARCH and GARCH models, as well ARMA(1,2) with GARCH(1,1) model.

Table 1 Quality measures of models using selected methods

Model	AIC	BIC	log likelihood (n)
ARIMA(0,1,1)	3968.76	3977.99	-3.643514
ARCH(1)	7.343482	7.367257	-3.666206
ARMA(1,2) with GARCH(1,1)	7.315322	7.354947	-3.648436
Threshold ARCH	7.654514	7.702063	-3.816187
GJR-GARCH(1,1)	7.309168	7.356717	3.643514

Table 2 shows the percentage change of the nuclear enegy production value in next 15 months, including mean squared prediction error for the best fitted model – GJR-GARCH (1,1).

After completing the process of outlier detection and correction of the data obtained using an iterative method, the GJR-GARCH model is best suited to forecasting examined time series. The model with the lowest value of AIC and BIC is the best fitted, in this research it is a GJR-GARCH (1,1) model. What is worth to be mentioned, all selected ARCH and GARCH models, which including modelling of the ARCH effect have good statistical properties. ARIMA models, which does not include volatility clustering, has bad statistical properties, AIC and BIC indicators have high values.

Conclusion

In this study, volatility of time series data showing nuclear energy production changes has been tested by using the ARIMA, ARCH and GARCH models. Examination of the simple returns showing nuclear energy production changes indicated the need to use models dedicated for ARCH effect modelling, as this phenomenon is visible for the residuals.

Table 2 Prediction results

	meanForecast	meanError	standardDeviation
1	1.364039	5.813043	5.813043
2	1.444946	5.837170	5.812185
3	1.444946	5.836312	5.811331
4	1.444946	5.835458	5.810481
5	1.444946	5.834607	5.809634
6	1.444946	5.833760	5.808790
7	1.444946	5.832917	5.807950
8	1.444946	5.832077	5.807114
9	1.444946	5.831240	5.806281
10	1.444946	5.830407	5.805451
11	1.444946	5.829577	5.804625
12	1.444946	5.828751	5.803802
13	1.444946	5.827928	5.802983
14	1.444946	5.827109	5.802167
15	1.444946	5.826293	5.801355

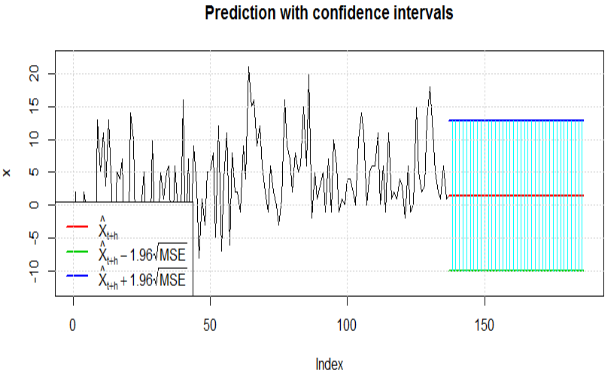

Fig. 2 Forecast of daily return volatility using the GJR-GARCHmodel.

GJR-GARCH(1,1) model is the most appropriate model to describe the persistence of volatility showing nuclear electricity production changes – data converted to simple returns from February 1973 to March 2018. We will also check outlier detection methods for data obtained in [16, 17].

Authors:. *Agnieszka Duraj, PhD. Politechnika Łódzka, Instytut Informatykii, ul. Wólczańska 215, 90-924 Łódź, E-mail: agnieszka.duraj@p.lodz.pl; Magdalena Ludwicka., Politechnika Łódzka, Instytut Informatykii, ul. Wólczańska 215, 90-924 Łódź.*

REFERENCRES

[1] R. F. Engle, Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation,

Econometrica, Vol. 50, No. 4, pp. 987-1007, 1982. [Online]. Available: https://www.jstor.org/stable/1912773

[2] I. Chang, G. C. Tiao C.Chen, Estimation of Time Series Parameters in the Presence of Outliers, Technometrics, USA, 1988, vol. 30, no. 2

[3] V. Barnett, T. Lewis, Outliers in Statistical Data, John Wiley & Sons, New York, 1994.

[4] F. M. Longin, The threshold effect in expected volatility: A model based on asymmetric information, The Review of Financial Studies 10 (3), 837-869, 1997.[Online]. Available: https://longin.fr/Recherche_Publications/Articles_pdf/Longin_The_threshold_effect_in_expected_volatility.pdf

[5] C. M. Hafner, Nonlinear Time Series Analysis with Applications to Foreign Exchange Rate Volatility (1 ed.). Physica-Verlag Heidelberg, New York, 1998.

[6] R. F. Engle, A. J. Patton, What good is a volatility model?, Quantitative Finance Volume 1 (2001), 237–245.[Online]. Available:http://www.stern.nyu.edu/rengle/EnglePattonQF.pdf

[7] J.W. Osborne, A. Overbay, The Power of Outliers (and Why Researchers Should Always Check for Them), 2004. [Online]. Available: https://www.researchgate.net/publication/242073851_The_Power_of_Outliers_and_Why_Researchers_Should_Always_Check_for_Them.

[8] J. Caiado, Modelling and forecasting the volatility of the portuguese stock index PSI-20, MPRA Paper 2077, University Library of Munich, Germany, 2004.

[9] E. M. Knorr, B. Math, Outliers and Data Mining: Finding exceptions in Data, University of Waterlo, University of British Columbia, 2012. [Online]. Available: https://www.cs.ubc.ca/grads/resources/thesis/Ma y02 /Ed_Knorr.pdf

[10] U.E.S. Kumara, W.A. Upananda, M.S.U. Rajib, Do Dynamic Properties of Stock Return Vary Under Hostile Environment? A Study During and After the Ethnic Conflict in Sri Lanka, Ruhuna Journal of Management and Finance, Volume 1, Number 2, 2014, [Online]. Available: https://www.researchgate.net/publication/269103861_Do_Dynamic_Properties_of_Stock_Return_Vary_Under_Hostile_Environment_A_Study_During_and_After_the_Ethnic_Conflict_in_Sri_Lanka

[11] A.K. Mittal, N. Goyal, Modeling the volatility of indian stock market, IJRIM, Volume 2, Issue 1, 2012. [Online]. Available:

[12] http://euroasiapub.org/wp-content/uploads/2016/09/1-1-42.pdf

[13] N. Hamzaoui, B. Regaieg, The Glosten-Jagannathan-Runkle-Generalized Autoregressive Conditional Heteroscedastic approach to investigating the foreign exchange forward premium volatility, International Journal of Economics and Financial Issues. [Online]. Available:https://www.econjournals.com/index.php/ijefi/article/viewFile/2740/pdf

[14] S. Singh, L. K. Tripathi, Modelling Stock Market Return Volatility: Evidence from India, Research Journal of Finance and Accounting, Vol. 7(13), pp 93-101 (2016). ISSN: 2222-2847, 2016. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2862870

[15] A. Duraj A., P.S. Szczepaniak, Information Outliers and Their Detection in: M. Burgin and W. Hofkirchner (Eds.): Information Studies and the Quest for Transdisciplinarity World Scientific Publishing Company, Vol.9, Chapter 15, pp. 413—436

[16] A. Duraj, Outlier Detection in Medical Data Using Linguistic Summaries, 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications INISTA 2017, Gdynia, Poland, 3-5 July 2017.

[17] Ł. Chomątek, A. Duraj, Multiobjective Genetic Algorithm for Outliers Detection, 2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications INISTA 2017, Gdynia, Poland, 3-5 July 2017, pp. 379-384

[18] A.Duraj, Ł.Chomątek, Outlier Detection Using the Multiobjective Genetic Algorithm, Journal of Applied Computer Science, Vol.25, No 1, pp.29-4

[19] Lebioda, M.; Rymaszewski, J.; Korzeniewska, E., Simulation of Thermal Processes in Superconducting Pancake Coils Cooled by GM Cryocooler, MICROTHERM' 2013 - MICROTECHNOLOGY AND THERMAL PROBLEMS IN ELECTRONICS Book Series:Journal of Physics Conference Series Volume:494Article number:012018 Published:2014

[20] Rymaszewski, Jacek; Lebioda, Marcin; Korzeniewska, Ewa, Simulation of the loss of superconductivity in a three-dimensional model of the metal-superconductor connection, PRZEGLAD ELEKTROTECHNICZNY, Volume:88, Issue:12B, Pages:183-186, Published:2012