**Submission for the 2022 IAOS Prize for Young Statisticians**

**Use of Machine learning technologies for quality improvement**

**of statistical data**

Ms. Alina Korobchuk

Deputy Chief of the IT Department

Main Statistical Office of Brest region, Belarus

20 March 1994 (27 years)

<https://brest.belstat.gov.by/>

Brest, Belarus

alinakorobchuk94@mail.ru

**Abstract**

Statistical information is one of the main components of the state information resource. Nowadays it is the subject of higher requirements. This research work proposes the approach for quality improvement of statistical data by using Machine learning technologies. Various forecasting methods can be applied to primary statistical data so that it can be possible to evaluate deviations of real value from a forecasted value based on the historical data. In case of a large deviation, the data should be returned for correction. This work includes 3 predictive models (ARIMA, regressive model, and recurrent neural network), which are based on industry statistics data. The time series of heat energy production of one anonymized enterprise in the Brest region was used as an example. The values for one period were predicted and also the quality of built models was assessed and the best one was selected.

**Introduction**

Statistical information is one of the core components of the state information resource. Important political decisions are made on the basis of it. It gives an adequate assessment of key economic and social indicators, which cover all the sides of state development. Nowadays society and economics impose higher requirements on the quality of official statistical information. In this context, the necessity of quality improvement of statistical data with the use of new technologies such as Data Mining (DM) and Machine Learning (ML) appears.

The use of these technologies in statistical organizations leads to quality improvement of statistical data (increase the accuracy, level of detail, and processing speed of data), as well as the creation of innovative statistics based on such technologies and also new sources of data (innovative projects, products, applications, and services).

DM and ML technologies are based on big amounts of data that are rapidly expanding. The term “Big Data” can be defined as structured and not structured data of a very big volume and variety and also a group of technologies and methods of its processing that provide the organization of qualitatively new useful information.

Big data technology can be applied in different spheres including the sphere of government statistics. These technologies allow to unleash the potential of massive data arrays, to receive qualitatively new useful information on the basis of it, and also to find hidden patterns and facts, to receive predictive results.

Today Statistical Offices of the European Union and European countries develop experimental statistics which suppose the use of new Machine learning methods, Artificial Intelligent, Data mining technologies, and new sources of data. For instance, Center for Big Data Statistics (CBDS) operates in the Office of Statistics in the Netherlands. CBDS develops innovative products and does research. The purpose of operating this center is the processing speed increase, accuracy, quality of official statistics, and also creation of new projects [1].

According to the innovative strategy in the field of data, Federal Statistical Office (FSO) Switzerland seeks to apply additional analysis and forecasting methods by using modern statistical methods of Machine learning and Data mining. Five pilot projects have been chosen to implement this strategy and are in progress: Area Statistics Deep Learning (ADELE), Automation of NOGA coding (NOGAuto), Machine Learning SoSi (ML\_SoSi), and others [2].

At the beginning of 2021 FSO established a Data Science Competence Centre which benefits all government departments of Switzerland. Data science and machine learning methods can help the community of Official Statistics to better inform the Government, businesses, the research community, and the public in general about the economic, demographic, social, and environmental situation [3].

Statistical services of the European Union (Eurostat), Estonia, the Netherlands, Switzerland, and other countries develop Experimental statistics. Experimental statistics are supposed to produce innovative statistics using new data sources and new ML and DM methods such as clusterization, classification, regression, time series analysis, neural networks, and others [4]. Experimental statistics aims at the quality improvement of official statistics data (increase the accuracy, level of detail, and processing speed of data) and also at the creation of new statistics (innovative projects, products, applications, and services).

After analyzing the methods of processing statistical data we found out that data validation of statistical data involves a great deal of work. In some cases, the staff is required to manually check the data again, that is why there is a necessity for additional checks. And it is a real problem. After analyzing the use of machine learning methods by statistical services, we suggest using these methods so as to improve the quality of basic statistical data producing official statistics. Methods of Machine learning are able to train a model on the basis of historical data and then to predict this value for the appropriate period. If the predicted value differs extensively from the real one, then data should be returned to the adjustment to a respondent.

In this work three predictive models of Machine learning were built: ARIMA, regressive model, and recurrent neural network on the basis of anonymized basic data of production statistics of the Brest region of the Republic of Belarus.

Data, used in this work, contains information about the production volume of goods that have been released by enterprises of the Brest region beginning in January 2016 until November 2021. In total, obtained data contain 2 493 codes of industrial products and 504 industrial enterprises of the Brest region which can’t be identified.

The time series of production of heat energy of one anonymized enterprise of the Brest region was chosen as an example. The analysis was made and predictive models were built based on it. Then the quality of models was evaluated and the best model was chosen. The working paper is structured as follows: The methodology is explained in Section 2. The results are shown in Section 3.

The paper concludes in Section 4. The links are presented in section 5.

# **Methods**

The methods used for building prediction models are described in this section: ARIMA (Autoregressive integrated moving average), regressive model, built with the use of the library Fbprophet for Python, and recurrent neural network. The approach to the quality evaluation of built models is described in this section as well. All the models were built using the programming language Python, while solving the problem we were using such Python packages as Pandas, Numpy, Statsmodels, SciPy, Matplotlib, Fbprophet, and Tensorflow.

1. **ARIMA**

ARIMA is an abbreviation for AutoRegressive Integrated Moving Average. ARIMA type models are a generalization of the ARMA class model. If we take an autoregressive model of order p (AR(p)) and the moving average of order q (MA(q)) and add everything in their right parts, the result will be the model ARMA(p, q), the formula (1):

, (1)

Wold's theorem states that any stationary time series can be described by the ARMA (p, q) model with the correct selection of the values ​​of the parameters p, q [5].

Formally, the stationarity hypothesis can be tested using the Dickey-Fuller test, which will be used in this work.

When working with non-stationary time series, a number of standard tricks are used to make them stationary. If the variance changes monotonically in time in the time series, a special transformation is applied to stabilize the variance.

Taking the logarithm belongs to the family of Box-Cox transformation, formula (2).

, (2)

It is a parametric family of functions, in which the parameter λ defines how series will be transformed: λ = 0 – taking the logarithm, λ = 1 – series identity transformation and for other values of λ – power transformation. The value of a parameter can be picked up so as the variance is more stable over time.

One more important trick, which allows to make the series stationary, is differentiation, transition to pairwise differences of neighboring values ​​according to the formula (3):

, (3)

For non-stationary time series, it often turns out that the resulting series after differentiation is stationary. Such operation allows to stabilize series average and get rid of the trend and sometimes get rid of seasonality. Moreover, differentiation can be used repeatedly: from a series of first differences, if we differentiate it again we can arrive at a series of second differences and etc. The length of the series will be shortened every time but it will be stationary in this.

Seasonal differentiation also can be used, transition to pairwise differences of neighboring seasons. If the length of season period is s the new series is given by the differences according to the formula (4):

, (4)

Time series have main components:

* *The trend* is a smooth long-term change in the level of the series. This characteristic can be received if we observe series for quite some time.
* *Seasonality* is cyclic changes in the level of a series with a constant period.
* *A cycle* is a change in the level of a series with a variable period.
* *An error* is an unpredictable random component of a series.

To understand if a predictive model has small flaws residuals should be analyzed.

Model residuals are the difference between the fact and the prediction, formula (5):

, (5)

Residuals estimate an error, noise component, which cannot be observed.

Desired property of residuals is non-autocorrelation, that is, no dependence on previous observations. The hypothesis of non-autocorrelation can be tested using the correlogram, as well as using the Ljung-Box Q-test [5].

1. **Regressive model**

The second method of forecasting is using the Fbprophet package of the Python programming language. The library makes it possible to improve the forecast by changing the parameters and does not require analysts to have deep knowledge of the structure of predictive models.

In fact, Prophet is an additive regression model, consisting of the following components, formula (6):

y(t) = g(t)+s(t)+h(t)+t (6)

Seasonal components s(t) are responsible for modeling periodic changes associated with weekly and annual seasonality. Weekly seasonality is modeled using dummy variables. Annual seasonality is modeled by the Fourier series.

Trend g(t) is a piecewise linear or logistic function. Logistic function of the view allows to model growth with saturation, when as the indicator increases, its growth rate decreases.

The h(t) component is responsible for user-specified anomalous days, including irregular ones.

The error t contains information that is not taken into account by the model [6].

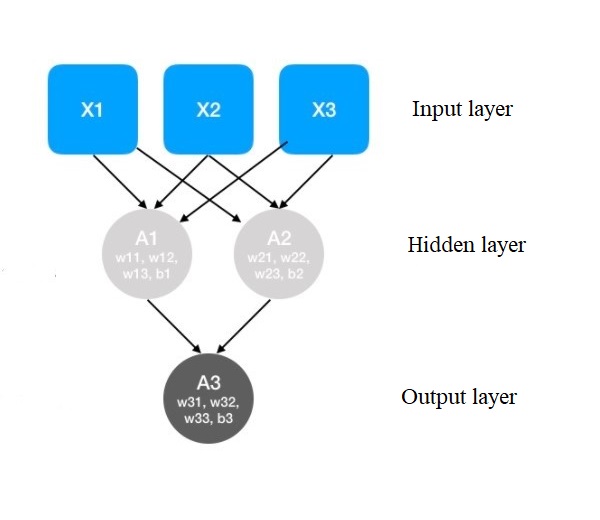
1. **Recurrent neural network**

Artificial neural network (ANN) is a mathematical model, the principle of which corresponds to the principle of the way the biological neural networks function. The main elements of the model ANN are artificial neurons, which are capable of sending an output signal obtained by processing a set of input signals [7].

An important step of forecasting using the ANN model is the process of training the network, which consists of determining and "fitting" the weights of each signal at the input of the neuron in such a way that the value at the output of the network corresponds to the target value. The algorithms of ML can be considered as functions, which select the values ​​of internal variables in such a way so as to select the values ​​of internal variables in such a way that the correct output values ​​correspond to the corresponding input values. In other words, training is the process of solving an optimization problem to minimize the error between the values ​​of the target (actual) parameter and the output parameter of the network.

A neural network is a collection of layers, each of which consists of pre-known mathematics (formulas) and internal variables. Each layer of the network is made up of nodes called neurons.

In figure 1 the model of the neural network is shown, which takes three parameters as input x1, x2, x3. A1, A2, A3 are neurons of the network. This network has one hidden layer with two neurons and one output layer with a single neuron. Internal variables in this network are weights w\* and bias b\*. It is the values ​​of these variables that are adjusted in the learning process to obtain the most accurate results of comparing input values ​​with output ones.



**Figure 1: The model of neural network**

Neurons of this network can be represented by the following formulas:

*,* (7)

*,* (8)

*,* (9)

There are many types of neural network architectures. For solving the problem of forecasting Recurrent neural network (RNN) was used. RNN is a type of ANN, which is well suited for solving problems related to time series. RNN processes the temporal sequence of data step by step, iterating through its elements and keeping the internal state obtained during the processing of the previous elements. To build a neural network, we used the tensorflow library.

1. **Quality evaluation of built models**

The quality of models was evaluated using MAPE and MAE.

MAPE (mean absolute percentage error) is the average absolute forecast error. Let is indicator and is the model prediction corresponding to this value. Then is forecast error, a is the relative forecast error, the formula (10).

(10)

MAPE is often used to assess quality, since this value is relative and it can be used to compare quality even on different data sets.

Moreover, it is useful to look at the absolute error MAE (mean absolute error), to understand how wrong the model is in absolute terms, the formula (11) [6].

(11)

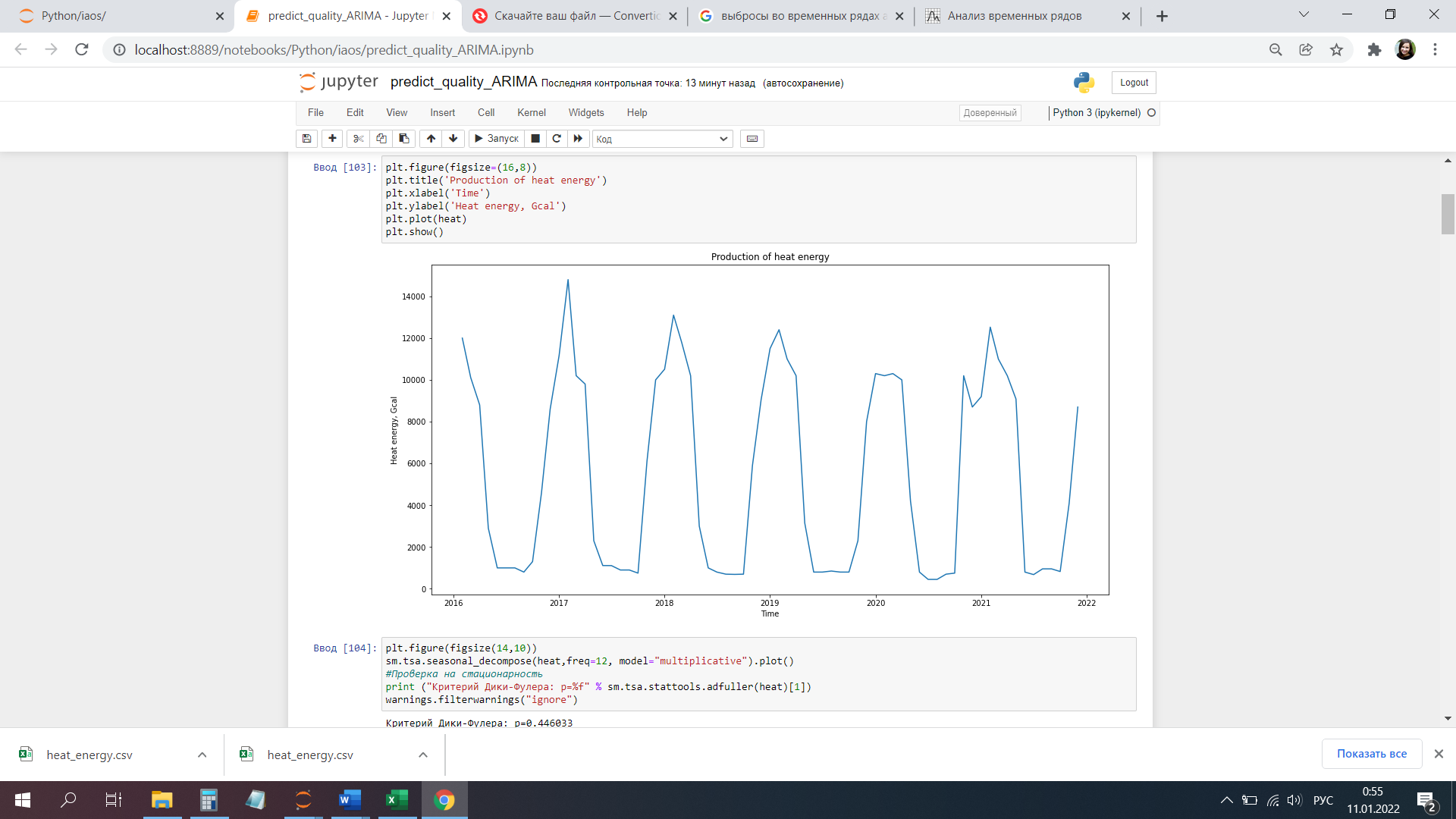
**Results**

Models were built with the use of the programming language Python as noted above.

In the process of solving the problem, we use such packages of Python as:

1. Pandas – a high-level library for data analysis.
2. Numpy – the library of Python, which adds the support for large multidimensional arrays and matrices together with a large library of high-level (and very fast) math functions for operations with these arrays [8].
3. Statsmodels – the library with tools for statistical modelling.
4. SciPy – the library used for solving mathematical and scientific problems.
5. Matplotlib – the library for plotting data.
6. Fbprophet – the library for time series prediction Facebook Prophet.
7. Tensorflow – the library for Machine learning, building and training of neural networks.
8. **ARIMA**

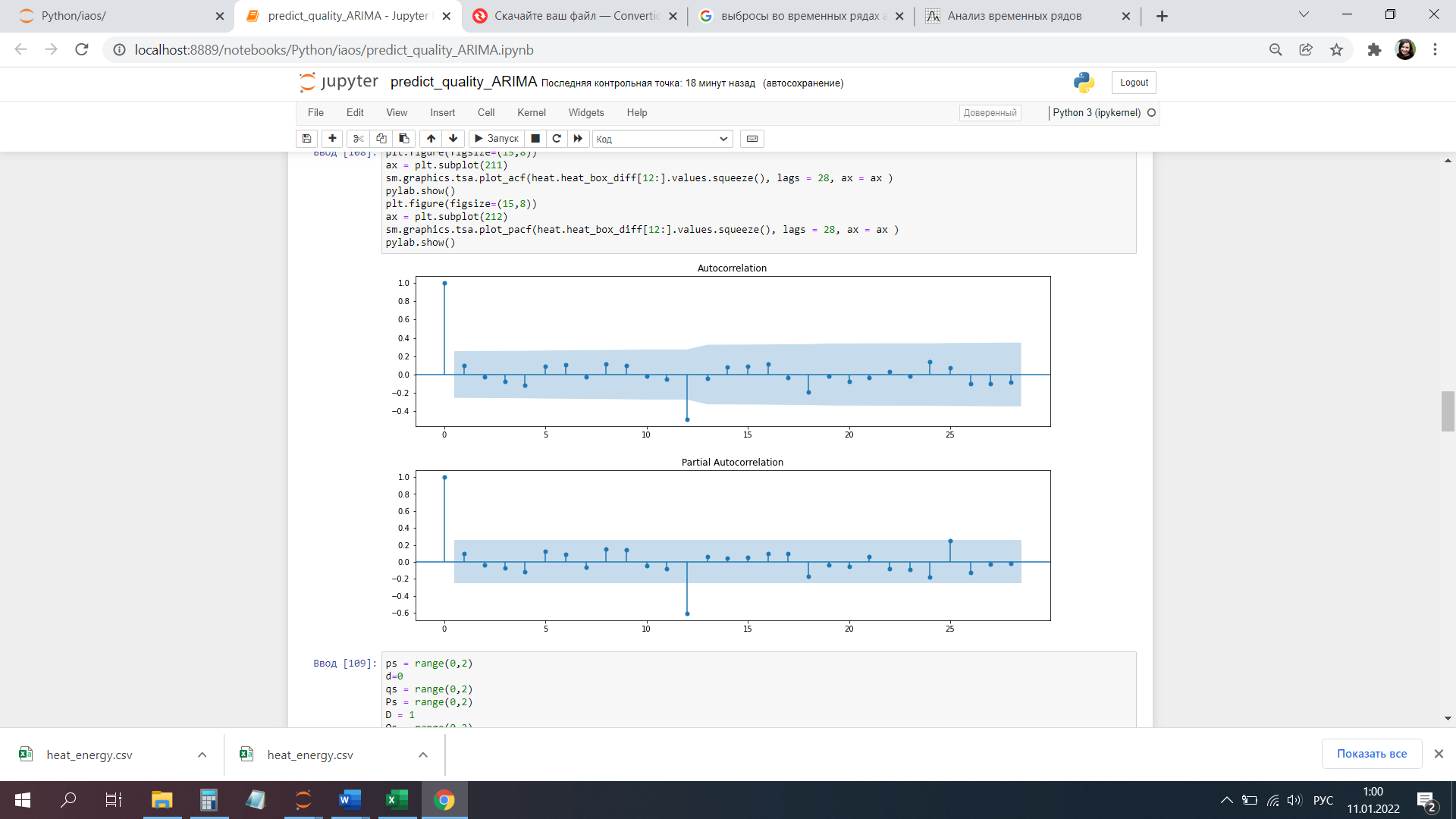
As noted above, the time series of the volume of heat energy production of one anonymized enterprise was chosen from all the data array as an example. Time series is shown in figure 2, on the abscissa, time series deposited along the x-axis, the volume of heat energy production deposited along the y-axis.



**Figure 2: Production of heat energy**

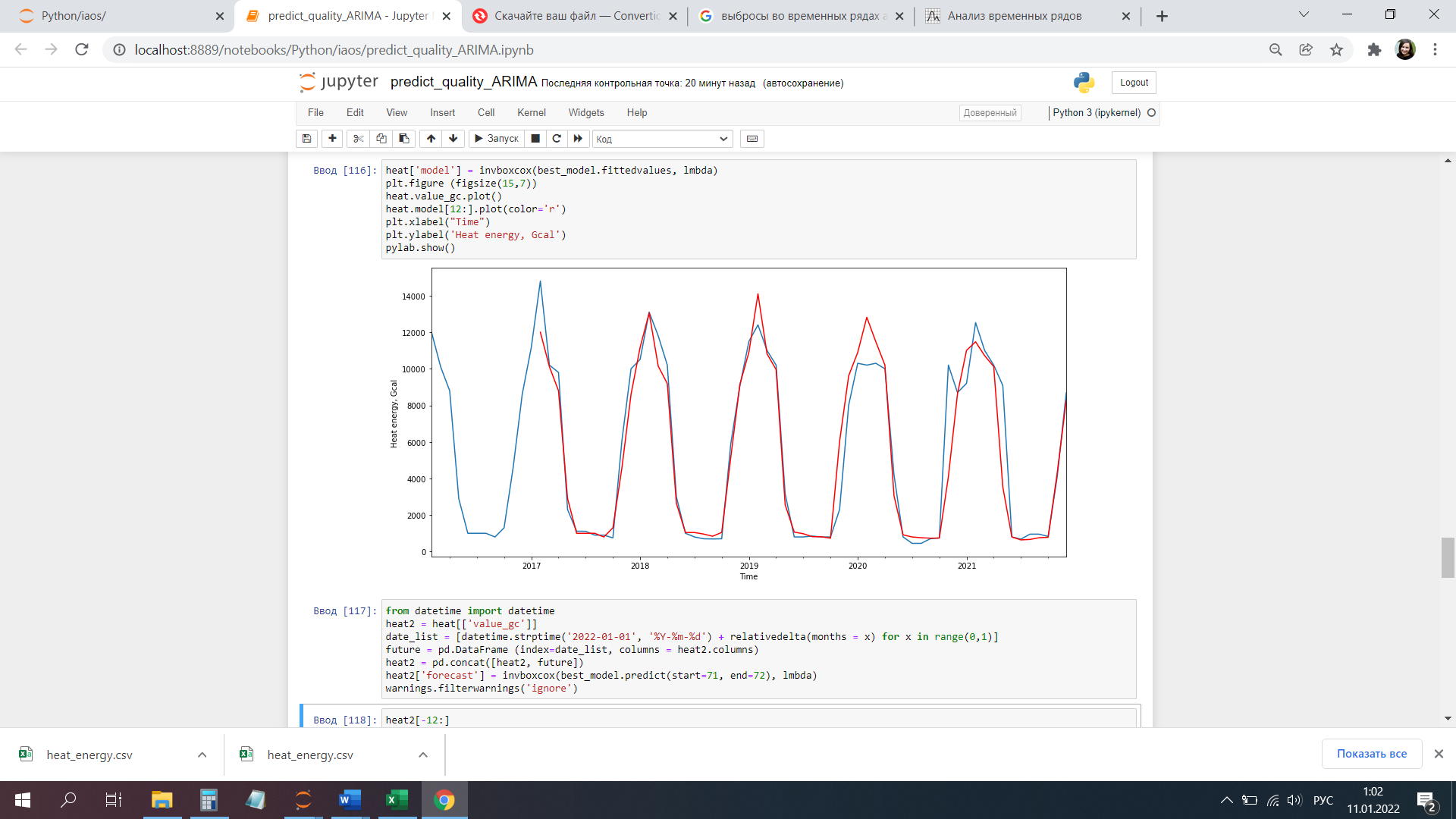
The time series was converted to stationary by Box-Cox transformation and seasonal differentiation. Dickey-Fuller test was calculated and the decomposition of the time series was done.

Then autocorrelation function and partially autocorrelation function were built, with the help of which seasonal model parameters for SARIMA were selected, as shown in figure 3.



**Figure 3: Autocorrelation function and partially autocorrelation function**

The next step was the analysis of model residuals and inverse Box-Cox transformation. We looked at how the model describes the data, as shown in figure 4.



**Figure 4: Model validation**

Then the value as of December 2021 was predicted and it was 9 848 Gcal. The real value of heat energy volume is 8 700 Gcal.

For the quality evaluation of the built model 12 last observations were cut off, then the forecast for 12 months into the future was done. Then we calculated prediction errors MAE and MAPE, which were 831,7 Gcal and 13,8 % accordingly. Real and forecasting values are shown in table 1.

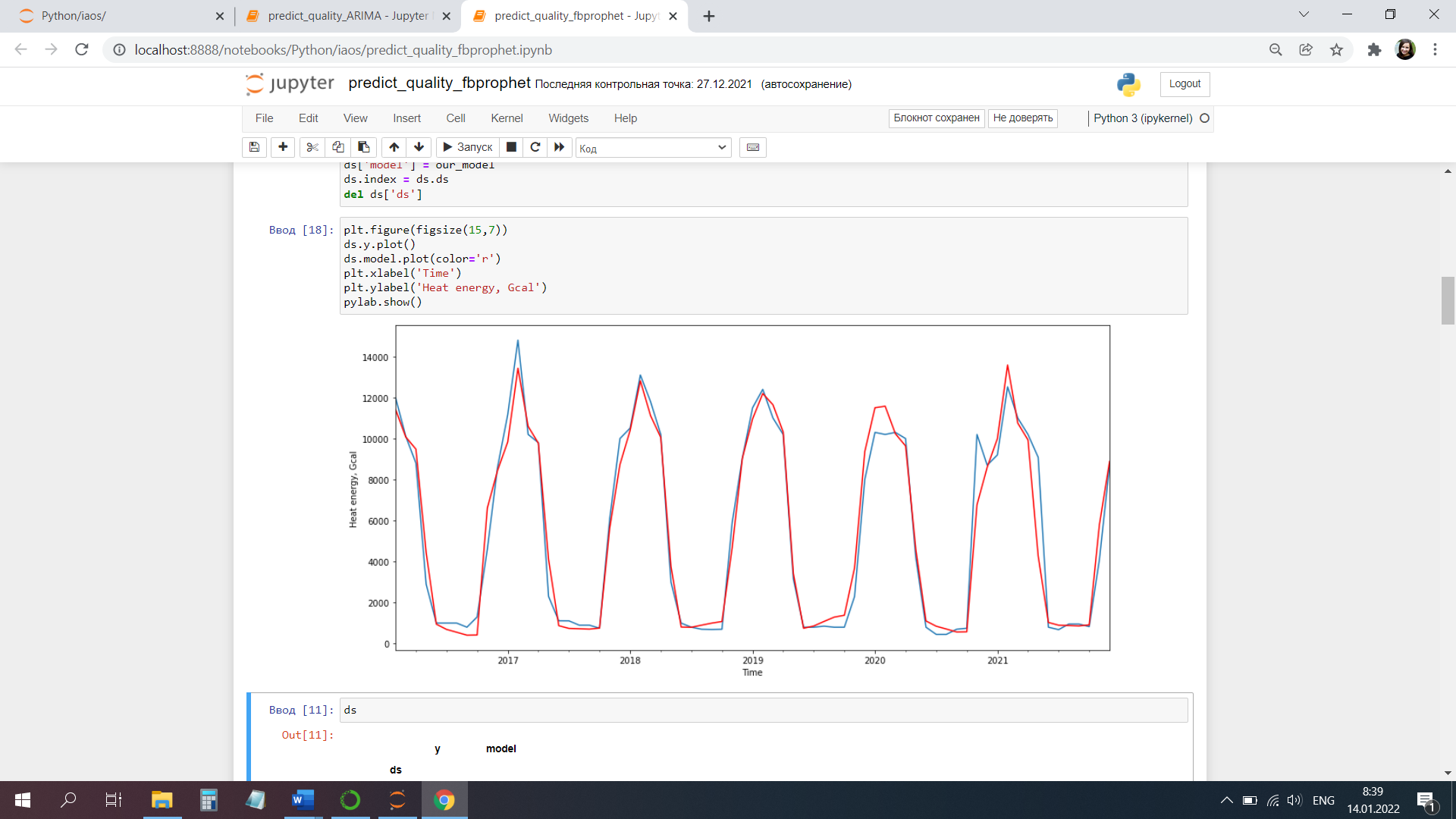
**Table 1: Real and forecasting values of the model**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 01-01-2021 | 01-02-2021 | 01-03-2021 | 01-04-2021 | 01-05-2021 | 01-06-2021 | 01-07-2021 | 01-08-2020 | 01-09-2021 | 01-10-2021 | 01-11-2021 | 01-12-2021 |
| Real values  (Gcal) | 9 200 | 12 524 | 11 000 | 10 200 | 9 085 | 800 | 680 | 954 | 954 | 827 | 4 090 | 8 700 |
| Forecasted values  (Gcal) | 11 007 | 11 476 | 10 715 | 10 120 | 3 537 | 800 | 639 | 663 | 759 | 780 | 4 302 | 8 273 |

As can be seen from the table, the model predicted values close enough. On the date 01-06-2021, the predicted value is equal to the real one. But a large discrepancy is seen on the date 01-05-2021 in connection with anomalous real value for this period. After recalculation of errors without this predicted value, we got values MAE and MAPE equal to 402,9 Gcal and 9,47 % accordingly.

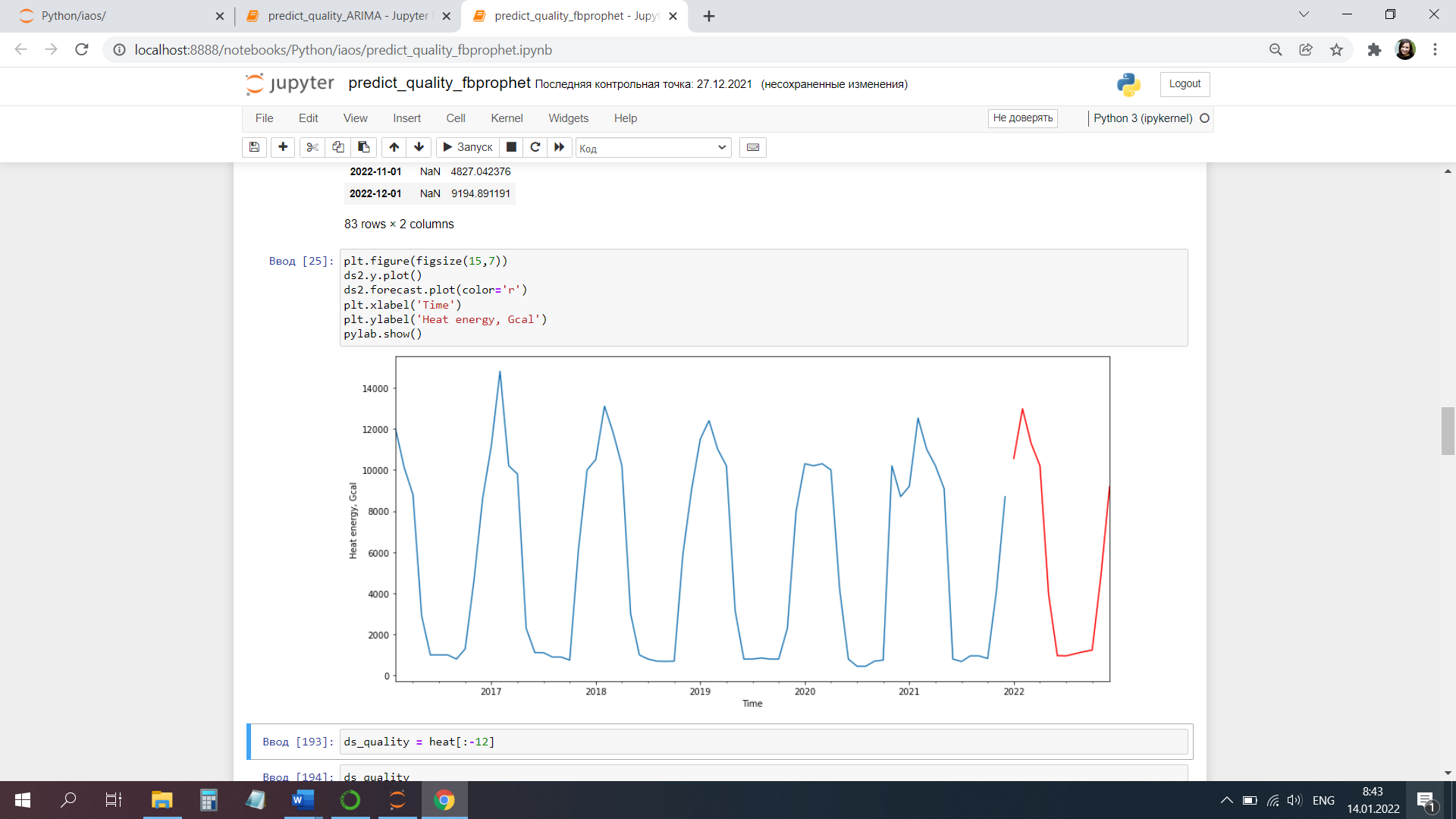
1. **Fbprophet**

To build the prediction with the use of Fbprophet library the data should be converted to the necessary type and the library should be included. Then we built a forecast with the use of the function “Predict”. We viewed how the model described the data, as shown in figure 5.



**Figure 5: Model validation**

Building a forecast for 1 year into the future, we got the plot, shown in figure 6.



**Figure 6: A forecast for 1 year into the future**

But we are interested in the value as of December 2021 and it is 10 556 Gcal.

For the quality evaluation of the built model 12 last observations were cut off, then the forecast for 12 months into the future was done analogically. We calculated errors MAE and MAPE without an anomalous value on the date 01-05-2021, got values MAE and MAPE 555,4 Gcal and 15,5 % accordingly.

We built also a forecast based on data converted with Box-Cox transformation to improve the quality of the model, after that the data were converted back. The results of observations are listed in table 1. We got errors MAE and MAPE 481,7 Gcal and 12,5 % accordingly excluding anomalous value on the date 01-05-2021.

**Table 2: Real and forecasting values of the model**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 01-01-2021 | 01-02-2021 | 01-03-2021 | 01-04-2021 | 01-05-2021 | 01-06-2021 | 01-07-2021 | 01-08-2020 | 01-09-2021 | 01-10-2021 | 01-11-2021 | 01-12-2021 |
| Real values  (Gcal) | 9 200 | 12 524 | 11 000 | 10 200 | 9 085 | 800 | 680 | 954 | 954 | 827 | 4 090 | 8 700 |
| Forecasted values  (Gcal) | 8 844 | 12 543 | 9 689 | 8 951 | 2 803 | 841 | 705 | 660 | 680 | 748 | 5 024 | 7 983 |

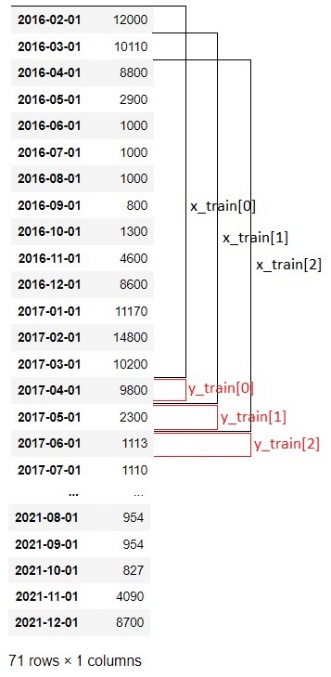
This model predicts the value as of December 2021, which is equal to 10 664 Gcal.

1. **Recurrent neural network**

In this work, a neural network was trained based on one feature, such as the volume of heat energy production of one anonymized Brest region enterprise. Dataset was divided into training and test samples.

Data scaling is an important step before training ANN. One of the common ways of scaling is standardization, which can be implemented by subtracting the mean and dividing by the standard deviation for each feature. Standardization should be implemented only with the use of training data.

We prepared the data for the model with one-dimensional input. The last 14 observations are given to the input of the model and this model should be trained to predict an indicator at the next time step. The preparation of the data for the model with one-dimensional input is schematically shown in figure 7.

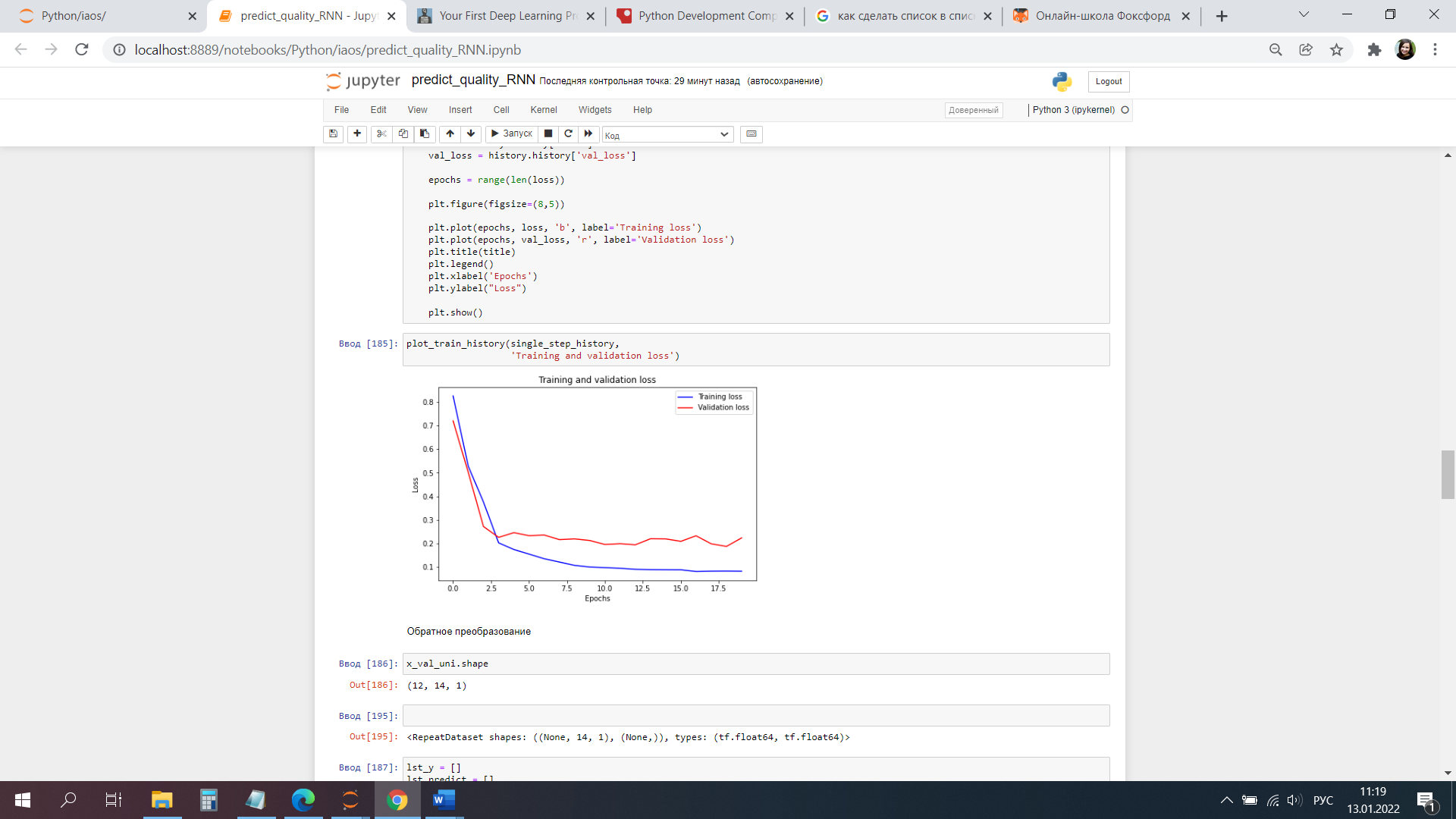


**Figure 7: Preparation of data for the model**

Hyperparameters are variables that we need to set before applying a learning algorithm to a dataset. The challenge with hyperparameters is that there is no magic number that works everywhere. The best numbers depend on each task and each dataset [9].

Therefore, Optimizer hyperparameters (Mini batch size, Number of epochs, Learning rate) and Model specific hyperparameters (Number of hidden units, First Layer, Number of layers) were defined empirically.

Then the model of recurrent neural network was trained. Loss curves at training and validation steps look like in figure 8.

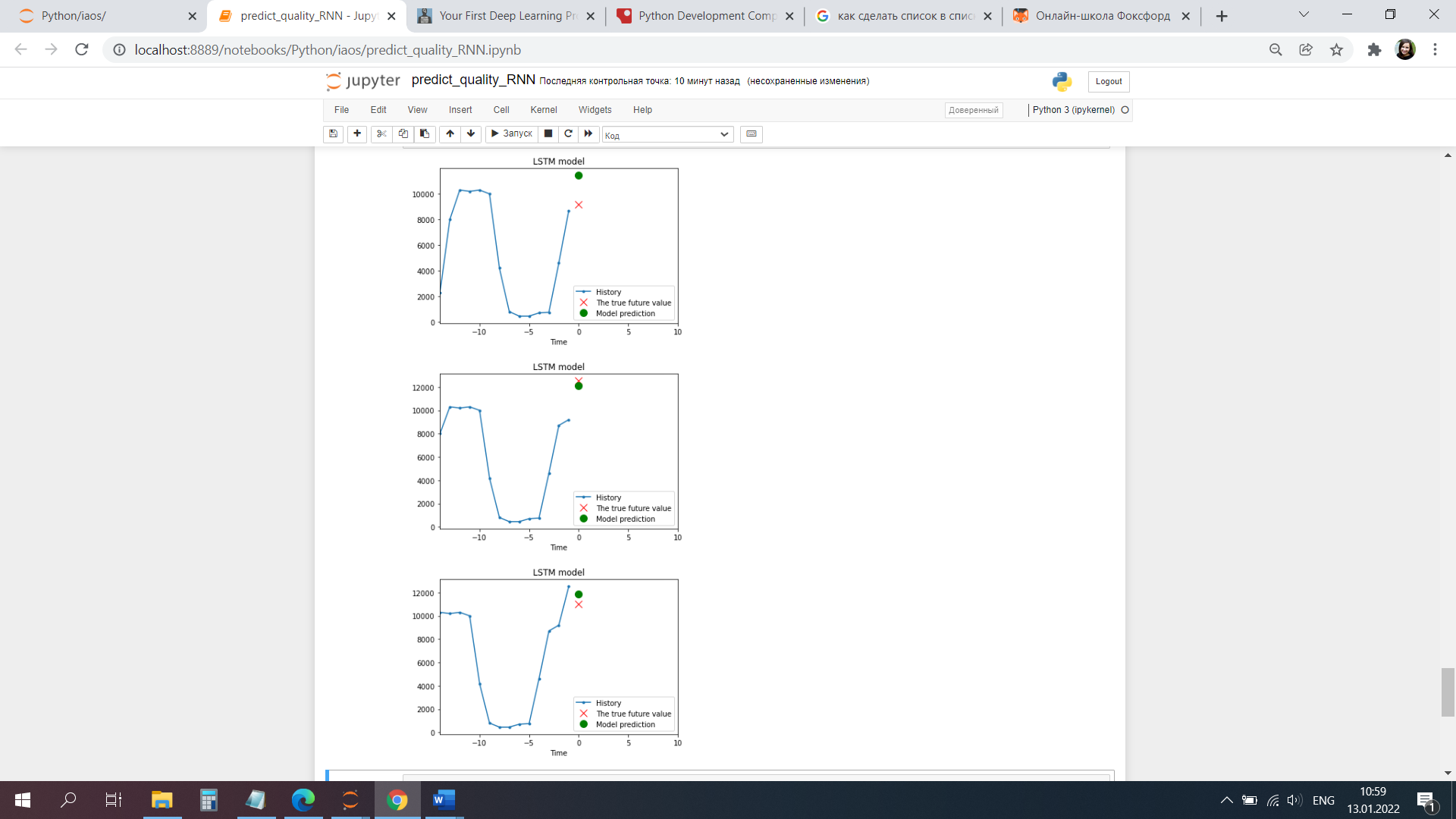


**Figure 8: Loss curves**

Then the prediction was created for each set from the test sample. The predictions for 3 sets from the test sample are shown in figure 9.

The forecast as of December 2021 is 10 854 Gcal.

The quality evaluation of the built model was also implemented based on a set from 12 last real values and those values which the model predicts. Thus, we got the values of errors MAE and MAPE equal to 957,6 Gcal and 18,7 % accordingly.



**Figure 9: Predictions for test sets**

Real and forecasting values of the model are listed in the table 3.

**Table 3: Real and forecasting values of the model**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 01-01-2021 | 01-02-2021 | 01-03-2021 | 01-04-2021 | 01-05-2021 | 01-06-2021 | 01-07-2021 | 01-08-2020 | 01-09-2021 | 01-10-2021 | 01-11-2021 | 01-12-2021 |
| Real values  (Gcal) | 9 200 | 12 524 | 11 000 | 10 200 | 9 085 | 800 | 680 | 954 | 954 | 827 | 4 090 | 8 700 |
| Forecasted values  (Gcal) | 11 442 | 12 083 | 11 872 | 10 000 | 3 508 | 863 | 653 | 812 | 553 | 1 038 | 2 517 | 7 628 |

A large difference between the predicted value and the real one is also on the date 01-05-2021, which notably degrades the quality of the model. After recalculation of errors without this predicted value, we got values MAE and MAPE equal to 650,1 Gcal and 15,9 % accordingly.

Quality evaluation for each of the built models is listed in table 4.

**Table 4: Quality evaluations for models**

|  |  |  |  |
| --- | --- | --- | --- |
| Models  Evaluation | ARIMA | Fbprophet | RNN |
| MAPE, % | 9,47 | 12,5 | 15,9 |
| MAE, Gcal | 402,9 | 481,7 | 650,1 |

That’s the way we conclude that for this time series the best quality shows predictive model ARIMA, but values of errors can differ for time series with other statistical indicators.

# **Conclusions**

Statistical information is one of the main components of the state information resource. Nowadays it is the subject of higher requirements. After analyzing the methods of processing statistical data, we found out that data validation of statistical data involves a great deal of work. In some cases, the staff is required to manually check the data again, that is why there is a necessity for additional checks. And it is a real problem. Also, the use of machine learning methods by statistical services was analyzed.

In this study, we offer to implement the methods of Machine learning during the official statistics production in order to improve the quality of statistical data. Different methods of ML are able to predict an indicator for the corresponding date on the basis of historical data. If we have a large deviation of a forecasted value from the real one the data should be returned to correction. The system also selects the best model according to its quality and also depending on time series, which include a certain statistical indicator, and depending on time series characteristics (presence of seasonality, trend, and abnormal values).

In this work, the time series of production of heat energy of one anonymized enterprise of the Brest region was chosen from the dataset. Three predictive models of Machine learning were built (ARIMA, regressive model, and recurrent neural network) on the basis of chosen time series, which predict the volume of heat energy production as of December 2021. The best quality was shown by ARIMA model from the three built models.

# **References**

1. Statistics Netherlands, 2021, Innovations of the Center for Big Data Statistics. Statistics Netherlands, 2021 – Mode of access: <https://www.cbs.nl/nl-nl/over-ons/innovatie>
2. Federal Statistical Office Switzerland, 2021, Experimental statistics. Federal Statistical Office Switzerland, 2021. – Mode of access: <https://www.experimental.bfs.admin.ch/expstat/en/home/innovation-data-science/overview.html>.
3. United Nations Committee of Experts on Big Data and Data Science for Official Statistics, 2021, Data science and official Statistics. Part of the “Road to Bern” Webinar Series, 2021.
4. Eurostat, 2022, Experimental Statistics. Eurostat, 2022. – Mode of access: <https://ec.europa.eu/eurostat/web/experimental-statistics>.
5. Moscow Institute of Physics and Technology, Yandex, 2021, Applied data analysis tasks. The global online learning platform, Coursera, 2021. – Mode of access: <https://www.coursera.org/learn/data-analysis-applications>.
6. Open Data science community, 2017, Predicting the Future with the Facebook Prophet Library. Community of IT-specialists, Habr, 2017. – Mode of access: [https://habr.com/ru/company/ods/blog/323730/.](https://habr.com/ru/company/ods/blog/323730/)
7. Sebastian Thrun, CEO Udacity, 2019, Introduction to Machine Learning. Community of IT-specialists, Habr, 2017. – Mode of access: [https://habr.com/ru/post/453558/.](https://habr.com/ru/post/453558/)
8. Wes McKinney, 2015, Python for Data Analysis. O’Reilly Media, pp. 93.
9. Samarth Agrawal, 2019, Hyperparameters in Deep Learning. Corporation, Towards Data Science, 2019. – Mode of access: <https://towardsdatascience.com/hyperparameters-in-deep-learning-927f7b2084dd>.