International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 12 (2020), pp. 5252-5261

Prediction the Turnover of Retail Trade Using Analysis of Time Series

Nadezhda Anatolevna Opokina[1]

Associate Professor Associate Professor (Department of Computer Mathematics and Informatics)

Institute of Mathematics & Mechanics named after N.I. Lobachevsky, Orcid 0000-0002-9753-0205, ID Scopus 55977505400

[1]Kazan Federal University,

Abstract

Many economic indicators presented by time series can be considered as the processes that depend on their previous values. In this case, modeling of such characteristics occurs via autoregressive models. The retail turnover of food products is one of the indicators affecting the economic development of the region. This characteristic is one of the assessments of quality level concerning the economic development of the region, since it represents the turnover of funds for the purchase of food products, both in cash and paid by credit cards, received from the sale of goods to the population. In this study, the retail turnover of food products was considered as a time series in order to predict this economic indicator using the example of the Republic of Tatarstan. An autoregressive model of the retail trade of food products was built and the assessment of the quality and forecasting accuracy of the constructed model was given in this paper. Based on the constructed model, conclusions were drawn about prediction the turnover of retail trade in food products in the conditions of self-isolation of the Republic of Tatarstan, as well as in terms of restriction absence. The proposed method of an autoregressive model development for the retail trade of food products can be applied to the retail trade of non-food products, as well as for other economic indicators.

Keywords: retail trade turnover, autoregression, food products, time series, stationarity, consumer price index.

I. INTRODUCTION

Analysis and forecasting of economic processes requires the use of a variety of mathematical methods and models. When they study time series, both econometric methods and the methods of fractal analysis are used [1, 2, 3].

This work is devoted to forecasting of such an economic indicator as the turnover of food retail trade on the example of the Republic of Tatarstan. This indicator, which is one of the components of the region economic stability description, is a time series and its study and forecasting is carried out using the apparatus of time series analysis, namely autoregressive models.

The constructed model gives a high indicator of its adequacy and an accurate forecasting result in the absence of restrictions such as self-isolation and an emergency situation that affect the economy of the region, as well as the country as a whole. In conditions associated with the measures limiting the spread of a pandemic, this model is not reliable and cannot be used for forecasting.

II. METHODS

The website of the Federal State Statistics Service (www.gks.ru) served as the information base of statistical data for the series under study. The time series of the retail trade in food products was taken as the data under study. The levels of this time series, presented in the form of monthly accumulated amounts from 01.2009 to 04.2020, were first reduced to absolute values without accumulations, expressed in million rubles. To bring the time series to the prices of 01.2009, it was necessary to take into account the inflation rate for the current period. The consumer price index was taken as the indicator of inflation. The source of the consumer price index data was the website "EMISS (state statistics)" (fedstat.ru). Before an autoregressive model development, the time series must be stationary. In this work, the following technique was used to achieve this goal. Let X_t be the levels of retail trade in food products, expressed in million rubles, where t is the time; I_t is the consumer price index.

Then

1. Let's define the consumer price indices relative to the basic index. Let us take the consumer price index 01.2009 as the base one and assign it a value equal to 1, i.e. $I_1^{base} = 1$. The rest of the indices are calculated using the formula $I_t^{base} = I_{t}\text{-}1^{base} \times I_t$.

2. Let's calculate the absolute values of the retail trade of food products taking into account the consumer price index using the formula $X_t^{base} = X_t / I_t^{base}$. In this case, we will get the retail turnover relative to the prices of January 2009.

3. Let's find the indices of the change in the retail trade turnover X_t base in relation to the previous month as follows:

$Y_1=1$, $Y_t= X_t^{base}/ X_{t-1}^{base}$ (t=2,3,…)

This sample was divided into two-time series:

- the training sample represented by data for the period 01.2009-12.2018. Based on this sample, a model was built;

- the sample for testing the adequacy of the constructed model. It includes the data for the period of 01.2019-04.2020.

Retail trade turnover is an endogenous variable; therefore, we will analyze this time series using autoregressive models. The volume of the training sample turned out to be equal to 120 values, which is sufficient for an autoregressive model development.

We build an autoregressive model for the resulting time series Y_t. First, to build a model for forecasting a given time series, it is necessary to examine it for stationarity. During the study for stationarity, the Dickey-Fuller test was carried out [4, 5]. To determine the type of autoregressive model, the analysis of autocorrelation coefficients and private autocorrelation was carried out. Based on the analysis, hypotheses are put forward about the type of autoregressive model, as well as time lags [6, 7]. The following autoregressive models were considered as the main models for the studied time series Y_t:

- autoregressive model of p - AR (p), which has the following form:

$$Y_t = \mu + \sum_{i=1}^{p} \beta_i Y_{t-i} + u_t,$$

where β_i (i=1,2,.., p)- are autoregression coefficients, u_t is "white noise", μ-const;

- moving average model. This model contains a combination of white noise as explanatory variables, that is, the series Y_t is described by the MA(q) process:

$$Y_t = \mu + u_t + \sum_{i=1}^{q} \alpha_i u_{t-i}$$

where $\{u_t\}$ – «white noise», μ-const, α_i(i=1,2,…q) - moving average coefficient;

- ARMA (p, q) model. It looks like this:

$$Y_t = \mu + \sum_{i=1}^{p} \beta_i Y_{t-i} + u_t + \sum_{i=1}^{q} \alpha_i u_{t-i},$$

where $\{u_t\}$ – «white noise», μ-const, α_i (i=1,2,…q) - moving average coefficients, β_i (i=1,2,.., p)- autoregression coefficients [8, 9, 10].

The resulting model must be checked for adequacy using the criteria such as the coefficient of determination, the value of the F-distribution, the Durbin-Watson coefficient, and the residual norm check.

After checking the quality of the constructed model, it is also necessary to compare the data of the test sample with the data obtained by forecasting using the constructed model.

To carry out all the analysis and processing of the statistical and econometric research, the Excel computer package and the Gretl statistical package were chosen. [11].

III. RESULTS AND DISCUSSION

Let's consider the original time series without accumulations (Fig. 1). This series has a trend and seasonality; therefore, it is not stationary.

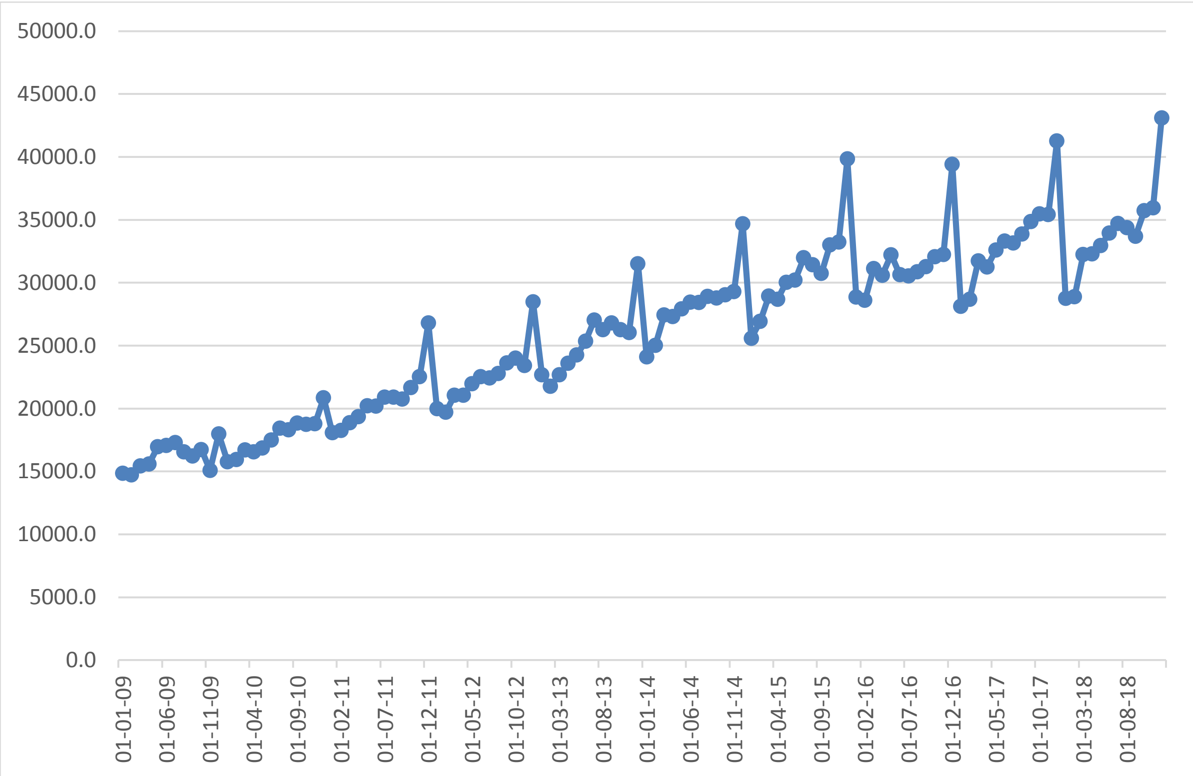

Fig. 1. Retail trade turnover of food products for the period 01.2009-12.2018

The time series of retail trade turnover of food products using the method described above was reduced via the consumer price index to the indices of changes in retail trade turnover relative to the previous month (Fig. 2).

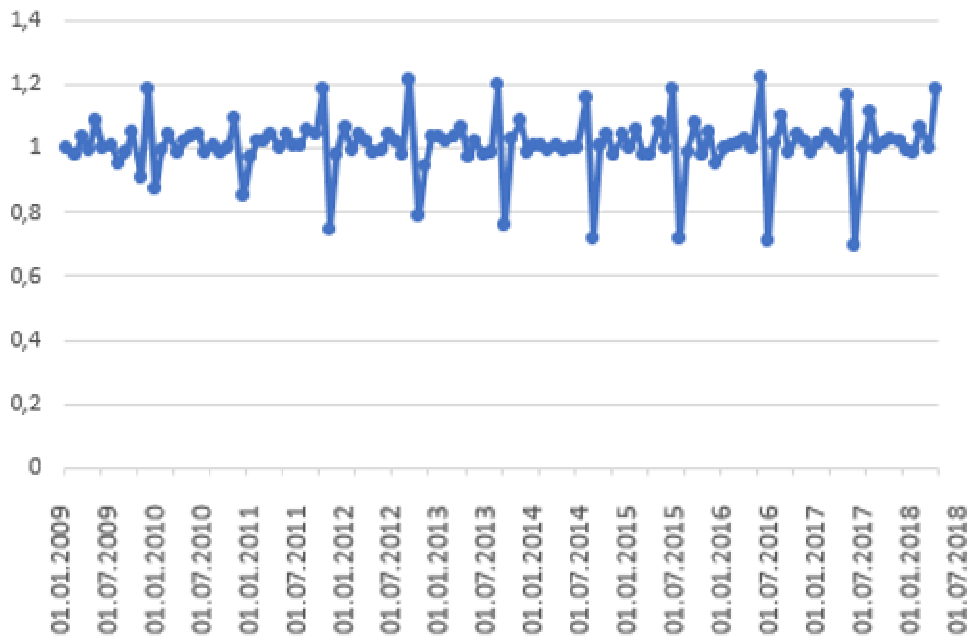

Fig. 2. Indices of changes in the turnover of retail trade in food products relative to the previous month

It can be seen on Fig. 2 that the trend has disappeared for the resulting series Y_t, but there is the periodicity with the lag of 12. Let's check the time series Y_t for stationarity using the criteria. First, let's check the Dickey-Fuller test (Table 1).

Table 1. Dickey-Fuller test results

The test without constant	model: (1-L)y=(a-1)*y(-1)+…+e first order autocorrelation coefficient for e: -0.058 F(11,96)=127.583 Test statistic: -0.0981152 p-value 0.65
The test with constant	model: (1-L)y=b0+(a-1)*y(-1)+…+e first order autocorrelation coefficient for e: 0.05 F(11,95)=37.616 Test statistic: -3.7928 p-value 0.002971
The test with constant and trend	model: (1-L)y=b0+b1t+(a-1)*y(-1)+…+e first order autocorrelation coefficient for e: 0.049 F(11,94)=37.138 Test statistic: -3.74328 p-value 0.01957

Based on the results for the stationarity of the time series, the test with a constant turned out to be the best, because in this case, the time series will be stationary at a significance level of less than 1%. Let's construct a correlogram for the time series Y_t.

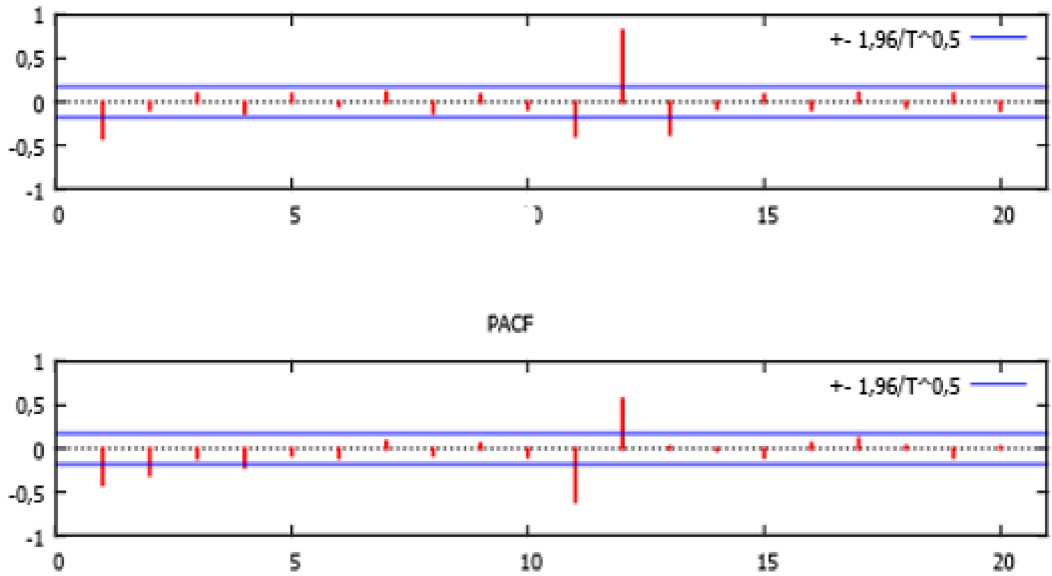

Fig. 3. Correlogram for food retail trade turnover

On Figure 3, we can see that the autocorrelation (ACF) and partial autocorrelation functions (PACF) of autoregression are an infinitely decreasing sinusoid with disturbances. This process is typical for the ARMA (p, q) autoregressive model. The greatest disturbance on both graphs of the autocorrelation and partial autocorrelation functions is achieved with the lag of 12, which is confirmed by the graph of the time series levels shown on Fig. 2. Hence it follows that there is an annual seasonality in the studied time series. Based on this, we assume that the autoregressive model for the time series Y_t is described by the ARMA model (12, 12). The results of the model construction are shown on Table 2.

Table 2. The model 1: ARMA(12, 12) for the time series Y_t

Variable	Coefficient	Std.error	T stat	P-value	
const	1,90466	0,729142	2,6122	0,01068	**
ort__1	-0,204279	0,0666705	-3,0640	0,00295	***
ort__2	-0,188897	0,0684607	-2,7592	0,00713	***
ort__3	-0,146962	0,070047	-2,0981	0,03894	**
ort__4	-0,132454	0,070017	-1,8917	0,06201	*
ort__5	-0,133497	0,0698354	-1,9116	0,05938	*
ort__6	-0,129668	0,0701425	-1,8486	0,06807	*
ort__7	-0,143439	0,070318	-2,0399	0,04454	**
ort__8	-0,147097	0,0708814	-2,0753	0,04106	**
ort__9	-0,177395	0,0710937	-2,4952	0,01457	**
ort__10	-0,16903	0,072297	-2,3380	0,02179	**
ort__11	-0,171049	0,0719946	-2,3759	0,01981	**
ort__12	0,850823	0,0714084	11,9149	<0,00001	***
u(-12)	-0,37653	0,0830625	-4,5331	0,00002	***

Mean of dependent variable	1,386789	Standard deviation of dep.var.	0,130435
Sum of squared residuals	0,098721	Adjusted R-squared	0,876072
Unadjusted R-squared	0,891726	P-value	4,39e-45
F-statistic (12, 83)	106,3235	Durbin-Watson statistic	1,730179

In this model, all the criteria indicate its good quality, namely the coefficient of determination $R^2 = 0.891726$ has a high value. The value of the F-distribution for this model is 106.3235, which is more than the tabular value of 1.83, therefore, the constructed model is adequate. The Durbin-Watson coefficient is close to 2, which indicates that there is no autocorrelation of residuals. The hypothesis of the normal distribution of the residuals of the series is confirmed, which is clearly demonstrated on the graph (Fig. 5).

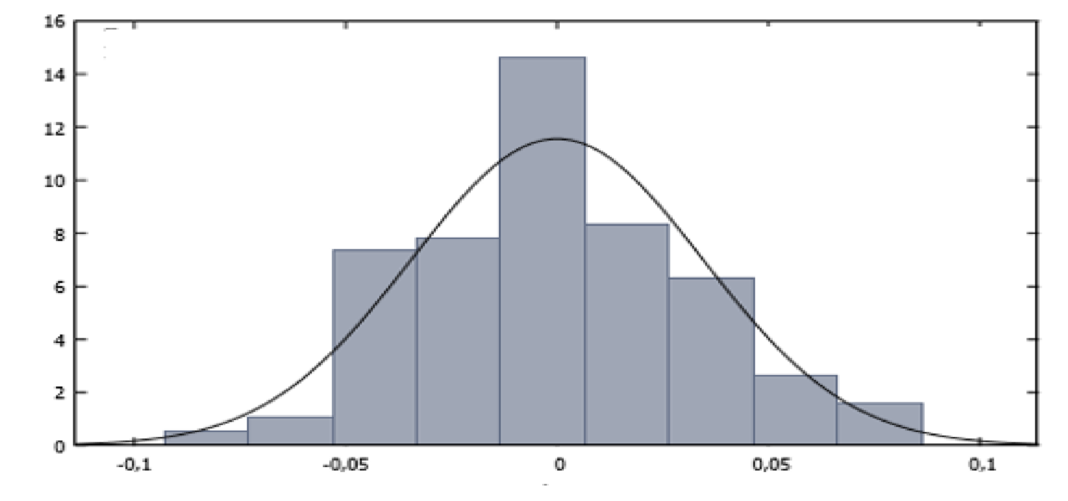

Fig. 5. Test for normal distribution of residuals of the ARMA model (12, 12).

Let's construct a forecast for the resulting model for the testing sample, namely, retail turnover data for the period 01.2019-05.2020 (Fig. 6).

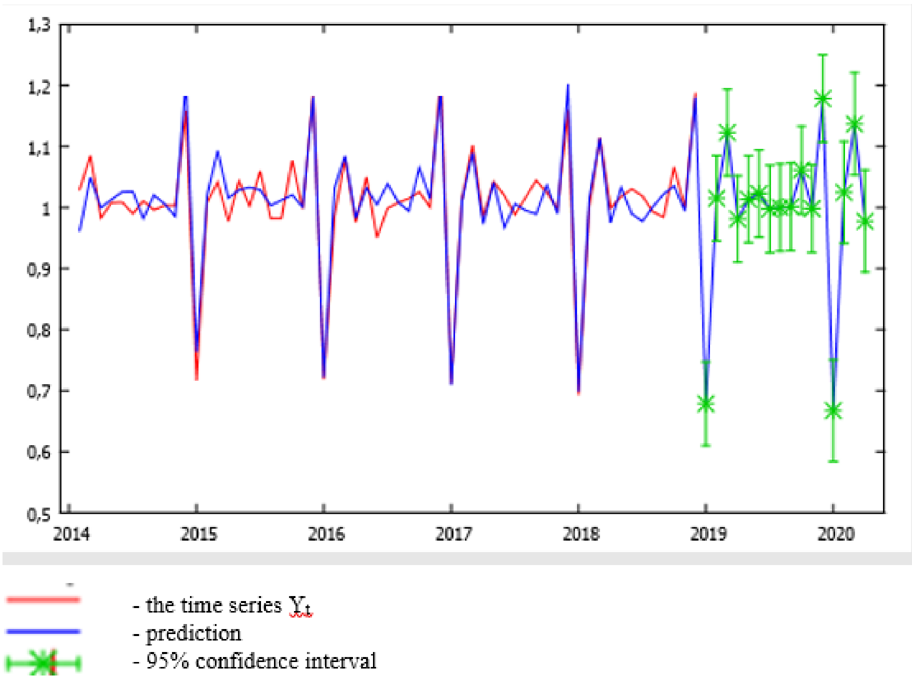

Fig. 6. Forecast of the time series Y_t for the period 01.2019-04.2020, built according to model 1.

The predicted values and confidence intervals for the retail trade turnover value prediction with the probability of 95% are shown in Table 2. Let's compare the values obtained using the model with the real values of the testing sample. The results are shown on the Table 3.

Table 3. Comparison of prediction results of model 1with the levels of the tested sample of the time series Y_t (01.2019-04.2020)

Observation	Time series Y_t	Prediction	std.error	95% confidence interval
01.19	0,685826	0,678578	0,0344878	(0,609984, 0,747173)
02.19	1,001087	1,01500	0,0352000	(0,944991, 1,08501)
03.19	1,105449	1,12289	0,0355640	(1,05215, 1,19362)
04.19	0,995912	0,981175	0,0356664	(0,910236, 1,05211)
05.19	1,023473	1,01369	0,0357237	(0,942641, 1,08475)
06.19	1,013279	1,02267	0,0357797	(0,951505, 1,09383)
07.19	1,023366	0,997342	0,0358188	(0,926100, 1,06858)
08.19	1,005032	0,999926	0,0358735	(0,928575, 1,07128)
09.19	1,009238	1,00152	0,0359167	(0,930086, 1,07296)
10.19	1,040979	1,06158	0,0360028	(0,989970, 1,13319)
11.19	1,008829	0,998247	0,0360367	(0,926571, 1,06992)
12.19	1,161585	1,17842	0,0360612	(1,10670, 1,25015)
01.20	0,717302	0,667250	0,0417960	(0,584119, 0,750380)
02.20	1,008377	1,02455	0,0420004	(0,941009, 1,10808)
03.20	1,088914	1,13694	0,0421005	(1,05320, 1,22067)
04.20	**0,830126**	**0,977554**	**0,0421077**	**(0,893803, 1,06130)**

The table shows that for this model all values of the tested sample (except 04.20) fell within the confidence interval with the probability of 95%. This is due to the fact that at that time there was a regime of self-isolation in the Republic of Tatarstan. In this regard, it can be concluded that this model will predict the situations only in the usual mode, and it is not suitable for forecasting in the mode of restrictions imposed on the territory of the Republic of Tatarstan related to the pandemic.

Note that not all lag variables are significant at a high level in the constructed model, some of them will be significant at a level less than 10%. Therefore, we will try to improve the quality of the model by excluding such variables from the model. For this purpose, we will use the analysis of the correlogram, while choosing the lags 1,2,11,12 (Fig. 3). The results of the proposed model are presented in Table 4.

Table 4. The model 2 for the time series Y_t

Variable	Coefficient	Std.error	T stat	P-value	
const	0,136849	0,106985	1,2791	0,20410	
ort__1	-0,0681873	0,0365769	-1,8642	0,06552	*
ort__2	-0,0533149	0,0318578	-1,6735	0,09766	*
ort__11	-0,0102746	0,0326532	-0,3147	0,75374	
ort__12	0,995922	0,0393957	25,2800	<0,00001	***
u(-12)	-0,382999	0,0820032	-4,6705	0,00001	***

Mean of dependent variable	1,393305	Standard deviation of dep.var.	0,131002
Sum of squared residuals	0,106936	Adjusted R-squared	0,877560
Unadjusted R-squared	0,882715	P-value	6,38e-53
F-statistic (4, 91)	324,0948	Durbin-Watson statistic	1,841294

Analyzing the resulting model, the significant variable at a significance level less than 1% is only the variable with the lag 12. Let's leave only lag equal to 12. The simulation result is presented in Table 5.

Table 5. The model 3 for the time series Y_t

Variable	Coefficient	Std.error	T stat	P-value	
const	-0,0345479	0,0293981	-1,1752	0,24289	
ort__12	1,03439	0,0290729	35,5792	<0,00001	***
u(-12)	-0,378096	0,0816993	-4,6279	0,00001	***

Mean of dependent variable	1,388367	Standard deviation of dep.var.	0,130572
Sum of squared residuals	0,111957	Adjusted R-squared	0,875901
Unadjusted R-squared	0,877208	P-value	2,47e-56
F-statistic (1, 94)	1265,879	Durbin-Watson statictic	1,930685

This model is identified by all criteria: determination coefficient $R^2 = 0.877208$; the value of the F-distribution for this model is 1265.879, which is more than the tabular value of 3.95, therefore, the constructed model is adequate; the Durbin-Watson coefficient is close to 2, which indicates the absence of autocorrelation of residuals. The hypothesis of the normal distribution of the residuals of the series is confirmed, which is clearly demonstrated in the graph (Fig. 7).

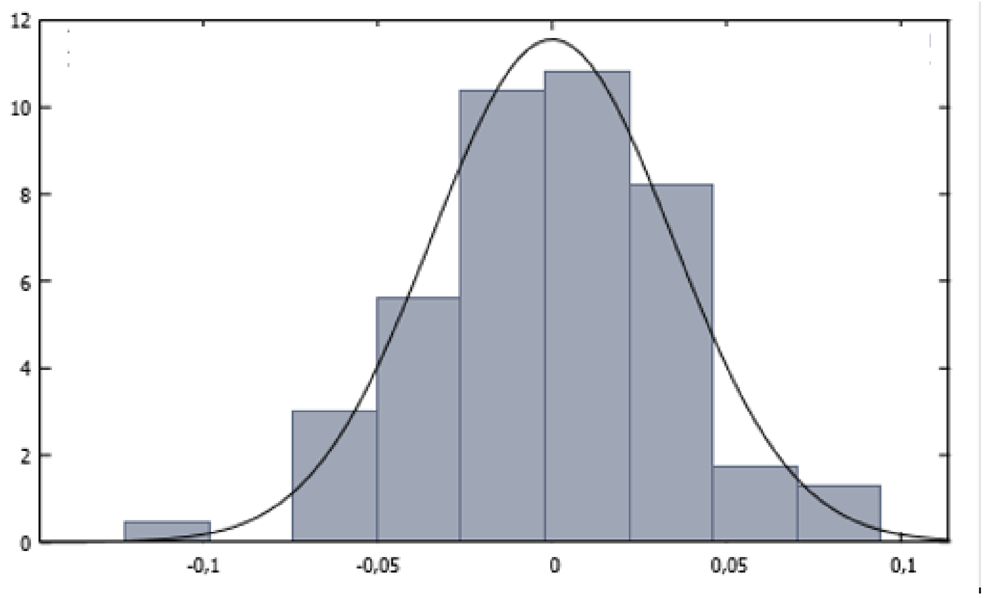

Figure 7. Test for normal distribution of residuals of the model 3.

The resulting autoregressive model can be written as follows:

$$Y_t = -0{,}0345479 + 1{,}03439Y_{t-12} - 0{,}378096u_{t-12} + u_t.$$

Let's evaluate the quality of this model. For this we use a testing sample. The model forecast for the period 01.2018-04.2020 is shown on Fig. 8.

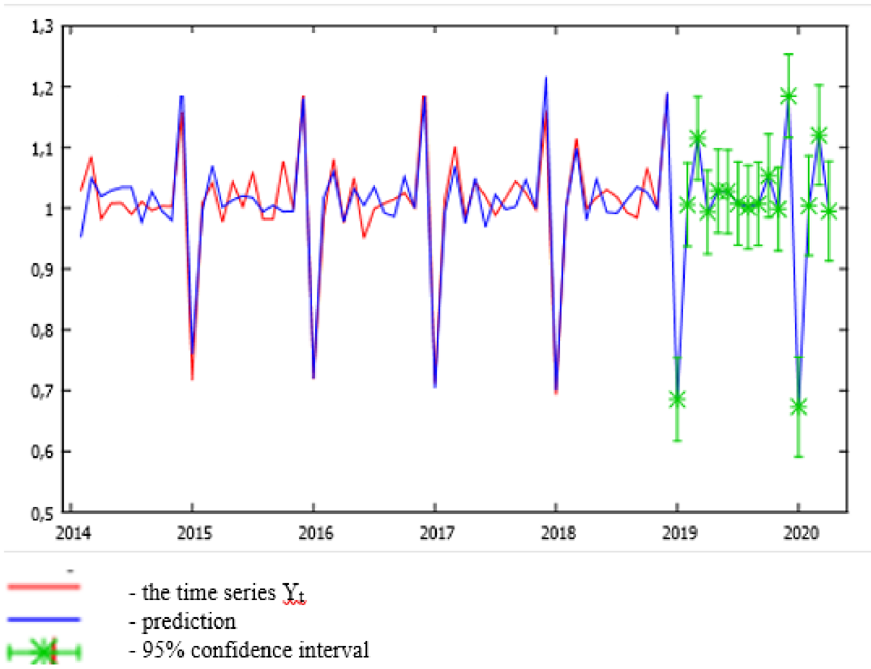

Fig. 8. Forecast for the time series Y_t, built using the model 3.

Let us compare the values obtained using the forecast with the real values of the time series Y_t (Table 6).

Table 6. Comparison of prediction results of model 3 with the levels of the tested sample of the time series Y_t (01.2019-04.2020)

Observation	Time series Y_t	Prediction	std.error	95% confidence interval
01.19	0,685826	0,685338	0,0345114	(0,616815, 0,753861)
02.19	1,001087	1,00578	0,0345114	(0,937254, 1,07430)
03.19	1,105449	1,11493	0,0345114	(1,04641, 1,18345)
04.19	0,995912	0,993387	0,0345114	(0,924864, 1,06191)
05.19	1,023473	1,02806	0,0345114	(0,959538, 1,09658)
06.19	1,013279	1,02746	0,0345114	(0,958932, 1,09598)
07.19	1,023366	1,00775	0,0345114	(0,939226, 1,07627)
08.19	1,005032	1,00186	0,0345114	(0,933340, 1,07039)
09.19	1,009238	1,00688	0,0345114	(0,938357, 1,07540)
10.19	1,040979	1,05312	0,0345114	(0,984599, 1,12165)
11.19	1,008829	0,998640	0,0345114	(0,930117, 1,06716)
12.19	1,161585	1,18471	0,0345114	(1,11619, 1,25323)
01.20	0,717302	0,673091	0,0412800	(0,591129, 0,755054)
02.20	1,008377	1,00391	0,0412800	(0,921949, 1,08587)
03.20	1,088914	1,12008	0,0412800	(1,03812, 1,20204)
04.20	**0,830126**	**0,995114**	**0,0412800**	**(0,913152, 1,07708)**

Let's compare the forecasting results of the first and third models. The comparison results are shown in Table 7.

Table 7. Comparative analysis of prediction of the model 1 and model 3.

Observation	Time series Y_t	Prediction of the model 1	Prediction of the model 3
01.19	0,685826	0,678578	*0,685338*
02.19	1,001087	1,01500	*1,00578*
03.19	1,105449	1,12289	*1,11493*
04.19	0,995912	0,981175	*0,993387*
05.19	1,023473	1,01369	*1,02806*
06.19	1,013279	*1,02267*	1,02746
07.19	1,023366	0,997342	*1,00775*
08.19	1,005032	0,999926	*1,00186*
09.19	1,009238	1,00152	*1,00688*
10.19	1,040979	1,06158	*1,05312*
11.19	1,008829	0,998247	*0,998640*
12.19	1,161585	*1,17842*	1,18471
01.20	0,717302	0,667250	*0,673091*
02.20	1,008377	1,02455	*1,00391*
03.20	1,088914	1,13694	*1,12008*
04.20	**0,830126**	**0,977554**	**0,995114**

From a comparative analysis of two autoregressive models built for the time series Y_t, the second is the best model, since it is the most accurate predictor of the time series values.

IV. CONCLUSIONS

In this paper, an autoregressive model was built for the coefficients showing the change in the retail turnover of food products as compared to the previous one, calculated taking into account the consumer price index, using the example of the Republic of Tatarstan. As the result of the study, it turned out that this time series has an annual lag. The constructed model has good predictive characteristics under normal conditions, but the model does not work within the introduced restrictions associated with the pandemic.

V. SUMMARY

In this work, autoregressive analysis of time series was applied to study the retail turnover of food products using the example of the Republic of Tatarstan. For this, the time series of retail price turnover was recalculated taking into account the consumer price index, which is a characteristic of inflation. The constructed autoregressive model for the retail trade in food products has a seasonal annual component.

This technique for autoregressive model development can be applied to other economic indicators described by time series.

VI. ACKNOWLEDGEMENTS

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

REFERENCES

[1] Makletsov SV, Opokina NA, Shafigullin IK. Application of fractal analysis method for studying stock market. International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies. 2019 Sep;11(1).

[2] Katz DB, Makletsov SV, Opokina NA. Using of Fractal Analysis and Artificial Neural Networks to Build a Mathematical Model for Determining Trends. Journal of Computational and Theoretical Nanoscience. 2019 Nov 1;16(11):4540-5.

[3] Saprykina EA. Forecasting diesel fuel prices using the autoregressive model. Modern high technologies. 2014;7(part 3):30-35.

[4] Afanasyev VN, Yuzbishev MM. Time Series Analysis and Forecasting. Moscow: Finance and Statistics. 2010:319 p.

[5] Greene WH. Econometric analysis. Stern School of Business, New York University.2018:1176 p.

[6] Wooldridge JM. Introductory econometrics: A modern approach. Nelson Education; 2016.

[7] Angrist JD, Pischke JS. Mastering'metrics: The path from cause to effect. Princeton University Press; 2014 Dec 21.

[8] Magnus YaR, Katyshev PK, Peresetskiy AA. Econometrics. M.: Delo. 2004:576 p.

[9] Kantorovich GG. Time Series Analysis. HSE Economic Journal. 2002;6(1):85–116.

[10] Krupko AM, Galaktionov ON, Shegelman IR, Vasilev AS, Sukhanov YV, Alkin RV. Justification of the mathematical model for describing the cross-cutting processes of functional food production with their enrichment. Caspian Journal of Environmental Sciences. 2020 Dec 1;18(5):421-35.

[11] Marneau V. Models Based on Panel Data. Applied Econometrics. 2006;1:94-135.

Nadezhda Anatolevna Opokina, Associate Professor at Faculty of Mechanics and Mathematics, Kazan Federal University, Russia. She is a Candidate of Physics and Mathematics Sciences. Fields of scientific interests: economic and mathematical modeling, econometrics, fractal analysis. The author of more than 30 scientific articles, including 6 articles indexed in the Scopus and WoS database.