DETERMINANTS OF REGIONAL RAW MILK PRICES IN RUSSIA

Svetlana Kresova[a], Sebastian Hess[b]

skresova@gmail.com

[a]Institute of Agricultural Economics, Christian-Albrechts-University of Kiel,
Kiel, Germany

[b]Institute of Agricultural Policy and Markets, University of Hohenheim,
Stuttgart, Germany

2021

***Paper prepared for presentation at the 61th annual conference of the GEWISOLA** (German Association of Agricultural Economics)*

***„The Transformation of Agricultural and Food Systems:
Challenges for Economics and Social Sciences“
Berlin, Germany, September 22th – 24th, 2021***

DETERMINANTS OF REGIONAL RAW MILK PRICES IN RUSSIA

Abstract

Drivers of regional milk price differences across Russian regions are difficult to determine due to limited data availability and restrictions on data collection. In this study, official data from Russian regions for the period from 2013 to 2018 was analysed based on 18 predictor variables in order to explain the regional raw milk price. Due to various data-based restrictions, the use of conventional panel regression models was limited and the analysis was therefore performed based on a Random Forest (RF) machine learning algorithm. Model training and hyperparameter optimization was performed on the training data set with time folds cross-validation. The findings of the study showed that the RF algorithm has a good predictive performance in the test data set even with the default RF values. Finally, the RF variable importance showed that income, gross regional product, livestock density, and milk yield are the four most important variables for explaining the variation in regional milk prices.

Keywords: milk price, Russia, machine learning, random forest

1 Introduction

Self-sufficiency in milk and dairy products is a priority goal within Russian agricultural policy (PETRICK and GÖTZ, 2017). The problem of achieving self-sufficiency with milk and dairy products in Russia is considered by Russian policy makers to be related to other various social and economic problems (SOLODUCHA et al., 2019). Within the new Doctrine of food security, which is in place since January 2020, defines the minimal necessary level of food independence as the level of self-sufficiency in percent. For the milk sector, this level corresponds to the ratio of domestic milk and dairy production relative to the volume of internal consumption. This ratio should be at least at 90% (UKAZ, 2020). However, by the end of 2019, the self-sufficiency rate of milk and dairy products (in terms of milk) was around 82,4% (NOSOV et al., 2020). Milk and dairy product consumption continue to decrease slowly from 248 kg per capita in 2013 to 231 kg per capita in 2017 (KULIKOV and MINAKOV, 2019). In 2013 the domestic milk and dairy production accounted for 76,6% of domestic consumption with the remaining 23,4% being imported from various countries (WEGREN, 2014).

However, in 2014, the global political conflict regarding Ukraine (Crimea), resulted in sanctions to the Russian Federation that responded with the introduction of an embargo on the imports of agricultural and food products (including milk and dairy products) from the European Union (EU), Norway, Australia, the United States of America (USA) and Canada (FAO, 2014). However, it seems that the embargo had no major negative effect on the consumption level in Russia, because dwellers have adjusted their buying behavior to the new situation and started consuming a higher share of Russian products (WEGREN et al., 2016). In this direction, studies have shown that milk price changes in Russia are affected by innovations and investment in the Russian dairy sector, in a way that is independent from international dairy markets (CARVALHO et al., 2015).

However, for Russia, in recent years regional milk price variations have been observed from 115 Euro/t to 1720 Euro/t, and this variation remains unexplained so far but affects the whole dairy supply chain. As long as regional markets would be integrated, a homogenous milk price for the entire country should be expected. Instead, regional autocorrelation of milk prices is relatively low, and distances are huge, population density and income per capita as well as consumer preferences may vary widely from western Russia to the Far East, or between the European and Asian parts. Official data are scarce and surveys across Russia are nearly impossible given political and logistic restrictions. Therefore, the question arises: Why do regional milk prices differ, and what may explain their variation?

In the present paper, the determinants of regional raw milk prices in Russia are studied using machine learning algorithms (ML), in order to determine the factors which can explain the observed regional variation in milk prices.

The paper is organized as follows: Firstly, the methodological approach is presented (conceptual framework, data collection, imputation, features selection, random forest, and model training). Afterward, the results are presented and discussed. The paper concludes with a summary of the findings and implications for our understanding of the factors that affect the milk price in the Russian dairy sector.

2 Methodology

2.1 Data collection

The data was collected from two sources: Russian Federal State Statistics Service (RFSSS) and Dairy Intelligence Agency (DIA, former Russian Dairy Research Center). The data was collected for 78 regions of the Russian Federation over 6 years (2013 – 2018). At the time of data collection, the data for the year 2019 was not publicly available. The data are limited to 6 years because DIA provides data only since 2013. DIA is dedicated to collect statistical time series data regarding dairy science, containing information which is currently not offered from RFSSS. The represented regions are 78 instead of the officially existing 85 regions in the Russian Federation, as DIA considers the following 7 regions as sub-regions of their larger neighbor regions (Moscow city to Moscow region, Saint-Petersburg to Leningrad region, Sevastopol city to the Republic of Crimea, Republic Adygea to Krasnodar region, Khanti-Manssiskiy autonomous region and Jamalo-Nenetskij autonomous region to Tyumen region and Nenetsk autonomous region to Archangelsk region). However, RFSSS provides data for these 7 sub-regions independently. To be in accordance with the data from DIA, RFSSS data from these 7 regions were included into their larger neighbor regions (as above), using the weighted average. The weights were calculated based on the population difference between the pairs of regions, giving higher weight to the region with a higher population. In total 18 potential explanatory variables were collected from the above-mentioned sources (Table 1). Milk price (Euro/t), is provided from RFSSS for the period 2002 – 2018. However, the data before 2013 were not included in the analysis but presented in the Figure 1, because DIA does not provide the data before 2013.

Table 1: Explanatory variables used in the data set (variable in each region/year)

Variable (unit)	Abbreviation	Source
Population density (people/km^2)	popden	https://www.dairynews.ru/company/country/russia/stat/
Total number of milk producers (companies)	mprodr	https://www.dairynews.ru/company/country/russia/stat/
Milk production per capita (kg)	mprodn	https://www.dairynews.ru/company/country/russia/stat/
Milk consumption per capita (kg)	mconsn	https://www.dairynews.ru/company/country/russia/stat/
Total need for milk (t)	needm	https://www.dairynews.ru/company/country/russia/stat/
The total amount of processed milk (t)	aprocm	https://www.dairynews.ru/company/country/russia/stat/
Total drink milk production (t)	drinkm	https://www.dairynews.ru/company/country/russia/stat/
Total cheese production (t)	cheesp	https://www.dairynews.ru/company/country/russia/stat/
Total number of dairy companies (companies)	daircm	https://www.dairynews.ru/company/country/russia/stat/
The average amount of processed milk per company (t)	prmcom	https://www.dairynews.ru/company/country/russia/stat/
Gross regional product per capita (Euro)	grprod	https://www.fedstat.ru/indicator/43890
Population surplus (people)	popsur	https://www.fedstat.ru/indicator/31325
Livestock cattle (thous. heads)	livest	https://showdata.gks.ru/report/278934/
Investments in stock capital per capita (Euro)	invcap	https://www.gks.ru/investment_nonfinancial
Investments in stock capital (M Euro)	invtot	https://www.gks.ru/investment_nonfinancial
Income per month per capita (Euro)	income	https://www.gks.ru/folder/13397
Total population (people)	poptot	https://www.fedstat.ru/indicator/31556
Milk yield per cow (kg)	miyiel	https://www.fedstat.ru/indicator/31223

2.2 Data imputation

The resulting balanced panel data set contained 0,48% and 2,14% missing values, with respect to the explanatory and response variables, respectively. Imputation of missing values was conducted separately for each region, based only on available data from this region. Since the vast majority of explanatory variables do not show specific structural breaks, the median imputation was selected. Opposite to this, the milk price has an increasing linear trend over the years for the vast majority of the regions. Traditional approaches such as overall mean imputation and missing-indicator method ignore such structure and bring biased results (DONDERS et al., 2006). To maintain this trend, a linear interpolation was performed with the use of imuteTS package in R (MORITZ and BARTZ-BEIELSTEIN, 2017). The Explanatory Data Analysis (EDA) of the final data set was performed with the R programming language (R Core Team (2019)) and QGIS v.3.14 (QGIS Development Team (2019)).

2.3 Data transformation

The EDA showed that the milk price is highly right-skewed and thus a log-transformation was applied. Normality is not a required assumption for the Random Forest (RF) (BREIMAN, 2001), however, recent studies have shown that in case of severe skewness the appropriate transformation can improve the prediction performance (JIANG et al., 2008; STEVENS et al., 2015; LÜTKEPOHL and XU, 2012; CURRAN-EVERETT, 2018). After the model training, the predictions were back-transformed to express the prediction errors in its initial scale (Euro).

2.4 Random forest

Machine learning is a powerful data-driven method, which develops very rapidly, and many new approaches for classification and regression problems are used parallel to traditional methods (TRAWINSKI et al., 2012). The success of the machine learning approach is determined through many factors, such as data quality and quantity (e.g. well-designed sampling schemes with enough and representative data for all examined sub-cases). Moreover, the exclusion of irrelevant, redundant, noisy, or generally unreliable information that is used as predictors increases the model performance (HALL, 2000). Machine learning data-driven models are empirically optimized, looking for the optimal solution (GOLDSTEIN et al., 2017). Such models have been applied in various studies with agricultural data (MCQUEEN et al., 1995; BALDUCCI et al., 2018; STORM et al., 2020), as well as in the dairy science (BORCHERS et al., 2017; SHAHINFAR et al., 2014; MA et al., 2018). To our knowledge, the determinants of the producers' milk prices in Russia have not been studied with the use of machine learning algorithms until now. The motivation for using random forest model is that the tree-based methods are good in capture non-linear relationships in data and provide the variable importance (STORM et al., 2020). Thus, this study is the first effort to examine the potential of machine learning as a quantitative tool in dairy science.

RF is an ensemble machine learning method consisting of classification and regression trees (BREIMAN, 2001). In RF the input training data is randomly and with replacement divided into several samples. Each of these samples is again sub-divided into 2 sub-samples (in-bag and out-of-bag sub-samples). Then the in-bag sub-sample is used to grow the tree. Thus, each tree is grown independently from the other trees and it is not correlated with them, as it uses a different random sample of observations and predictors. The different predictions from the trees are averaged to a final prediction in a regression problem. In the classification problem, the majority vote is used. Parallel to this, the predictions performance and variable importance are calculated based on the out-of-bag sample. RF is resistant to over-fitting (CUTLER et al., 2007) and has high predictive accuracy, and even when a large number of explanatory variables are used (SVETNIK et al., 2003).

2.5 All-relevant feature selection

All-relevant feature (variables) selection aims to identify the relevance of explanatory variables to our response variable. This is an important step before machine learning modeling because it ensures that only the relevant explanatory variables are used for the model predictions. Thus, irrelevant and noisy predictors are not used, maximizing the model performance (KURSA and RUDNICKI, 2010). Among the different relevant feature selection methods we selected the Boruta algorithm (KURSA and RUDNICKI, 2010). Recent studies showed that Boruta has increased sensitivity and selective power in synthetic and real-world data sets, outperforming other common algorithms (DEGENHARDT et al., 2019). In short, the algorithm creates randomized copies of the explanatory variables (the so-called shadow variable) by shuffling them. After that, a relatively fast version of the random forest algorithm ("ranger random forest") algorithm (WRIGHT et al., 2018) is performed several times and the importance of each variable is calculated. Then the variables, which scored better than the shadow variables (shuffled explanatory variables), are considered as important, while the variables that scored lower are considered as unimportant for the analysis (KURSA and RUDNICKI, 2010).

2.6 Model training and hyperparameter optimization

The data was divided into training (80% of the observations) and test (20% of observations) data sets with stratified data sampling of the response variable to ensure better sample representativeness and modeling performance. The test data set was kept out of the training and cross-validation and used only for the final model assessment. RF is a relatively simple method regarding its tuning with the most influential hyperparameter to be the number of predictors available for splitting in each split when growing a tree, the so-called *mtry* parameter (PROBST et al., 2019). In total 17 different *mtry* values were tested and compared, based on the resulted Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The number of growing trees (*ntree* parameter) was kept constant in 500 trees, which is the default recommended value (LIAW and WIENER, 2002). Increasing the number of trees does not automatically improve the performance of the model (OSHIRO et al., 2012). The model training was conducted with the use of the *caret* R package (KUHN, 2008), which allows for hyperparameter optimization through time folds cross-validation. Time fold cross-validation is a more appropriate method for time series data as the traditional random cross-validation does not account for the temporal structure of the data, leading to over-optimistic prediction errors (ROBERTS et al., 2017). The training data was split into 6 folds, according to their time component (6 years from 2013 to 2018) with the use of CAST R package (MEYER et al., 2018). The final model was applied in the test data set for an independent assessment of the model performance in a new data set, in which RMSE, MAE, and residual characteristics were calculated.

2.7 RF variable importance

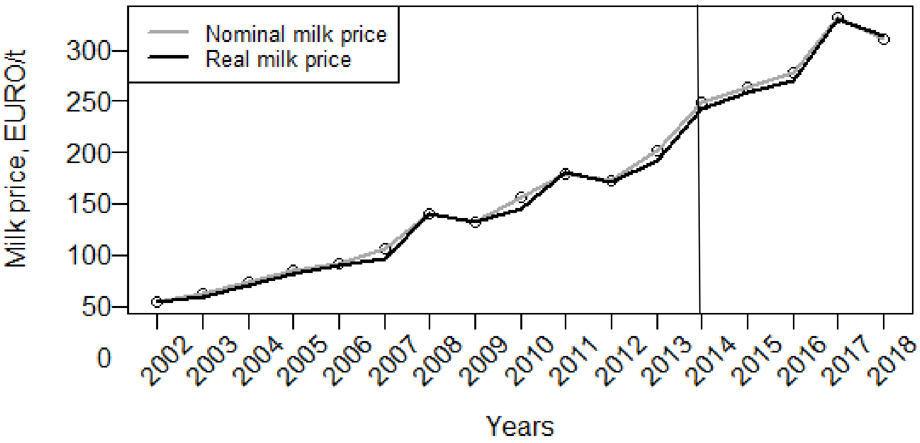
The RF calculates the variable importance of each predictor, either based on node impurity or permutation error. Based on recent studies, node impurity (the number of times that each variable is being chosen in each node in each tree of the ensemble) can be biased when the predictors vary in their scale or are correlated and the sampling of them is performed with replacement (STROBL et al., 2007). Opposite, permutation importance (which is calculated by estimating the increase in the prediction error when shuffling the variable while at the same time keep all other variables unchanged) is more stable and unbiased (STROBL et al., 2007; STROBL et al., 2008). Based on the above mentioned the permutation importance was calculated. Finally, the partial dependence plots (pdp) for the four most important variables were presented with the use of the pdp package (GREENWELL, 2017b). These plots visualize the partial dependence of the response variable on a single explanatory variable aiming to show the type of their relationship (GREENWELL, 2017a).

3 Results

3.1 Exploratory Data Analysis

The analysis of the mean milk price (mean value over all regions for a certain year) shows a constant increase from 2002 until 2017, with a small drop in 2018 (Figure 1). The mean milk price in the period 2002 – 2014 in Russia was around 132 Euro/t, while the mean milk price in the period 2015 – 2018 was 296 Euro/t, showing an increase after the establishment of the embargo (straight line on the graph). Recent studies have concluded similar results (MURTUZALIEVA et al., 2017).

Figure 1: Fluctuations of the mean milk price in Russia from 2002 until 2018



The analysis of the period 2013 – 2018 showed that the milk price grew the period 2013 – 2017, and then slightly declined in 2018. A Kruskal-Wallis test confirmed that there was a statistically significant difference in milk prices between different years (p-value < 2.2e-16). It is worth to be mentioned that the range of the milk price in the last two years (2017 and 2018) is greater, showing higher variance among different regions. Also both median and mean prices are higher for these years.The inflation rate in Russia varies from 6,45% in 2013 with increasing until 11,36% and 12,91% in 2014 and 2015, respectively and to 5,38% in 2016 with decreasing until 2,52% and 2,4% in 2017 and 2018, respectively(BS, 2017). Summary statistics of deflated and non-deflated milk prices are presented in the Table 2.

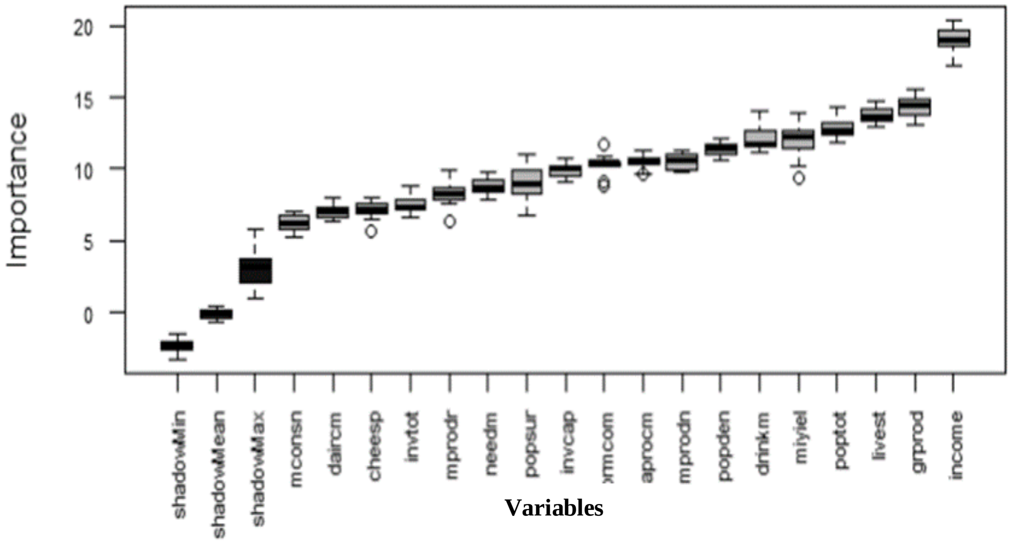
Table 2: Summary statistics of milk price (Euro/t) for the years 2013 – 2018

Milk price	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Real	114.7	231.9	269.6	306.6	318.7	1720.1
Nominal	115.0	239.2	273.9	312.4	320.7	1720.1

3.2 Results of the features selection

Feature selection was conducted with the training data set, which represents 80% of the whole data set. The Boruta analysis showed that all 18 predictor variables are relevant for the explanation of the milk price (Figure 2). We see that the most of variables have scored higher than the maximum shadow Z-score value among all shadow variables, indicating potentially strong predictors. In Figure 2, “Importance” is defined as Z-score of each predictor variable and Z-score is calculated by dividing the avarage accuracy loss by its standard deviation (KURSA and RUDNICKI, 2010). Income has the highest predictive ability, followed by gross regional product, livestock output, total population, and milk yield in the first five positions. However, some of the variables (milk consumption and cheese production) have a relatively low importance, with their minimum importance value about equal to the maximum shadow value.

Figure 2: Feature selection with the Boruta algorithm using log-transformed response variable in the training data set

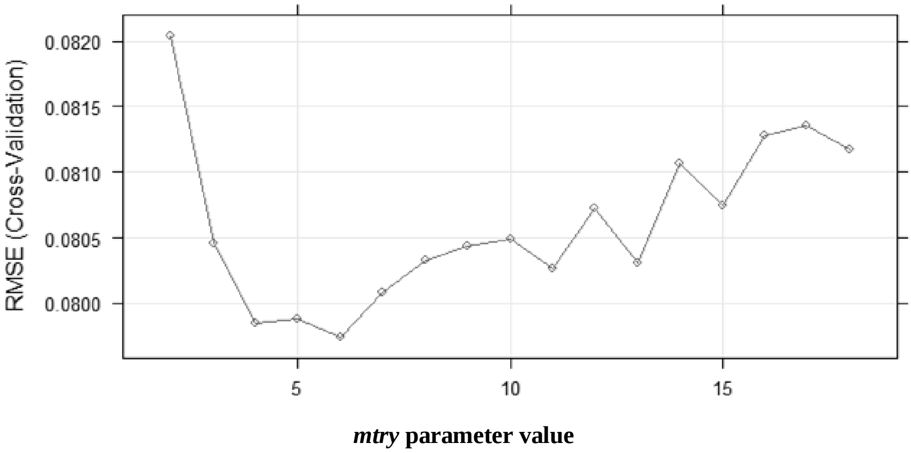


Nevertheless, the overall performance of these variables is higher than the maximum shadow Z-score, so that they can be included in the final model. The Boruta algorithm showed that all variables are relevant, that is why all variables were used in the subsequent analysis.

3.3 Results of the RF modeling

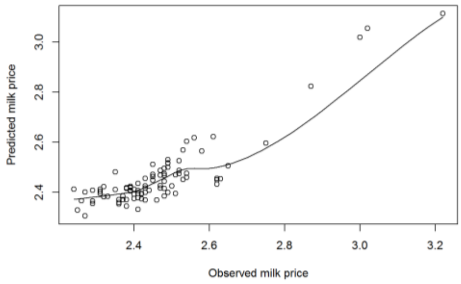
During the optimization process, the algorithm has examined 17 different *mtry* values (see sec tion 2.6), which are presented on the horizontal axis in the Figure 3. The RMSE is minimal wh en *mtry* is 6. This value corresponds to the default value of *mtry* for regression (number of pre dictor variables divided by 3) in the *random forest* package (LIAW and WIENER, 2002). The us e of *mtry* values that are much higher or much lower than 6 increases the error with the highes t error to be observed when *mtry* = 2 (Figure 3).

Figure 3: RF performance (in log-transformed training data) with different *mtry* values



After the optimal *mtry* value was found, the final model was applied to the independent test data set to evaluate its predictive ability. The analysis of the results showed good overall performance (Figure 4). This fact is also confirmed by good RMSE and MAE (Figure 4). However, the range of predicted values of the milk price is slightly smaller than the range of the observed values. The maximum predicted value is underestimated and the minimum predicted value is overestimated. This limitation occurs because in the RF regression the final result is the average value of all predictions (BREIMAN, 2001).

Figure 4: Scatter plot between observed and predicted milk price in the test data set (left panel), comparison of RF model performance in training and test data set (right panel)



Type of the data set	RMSE	MAE	R-squared
Training data set (log)	0.0797	0.0625	0.8032
Test data set (log)	0.0708	0.0535	0.7989
Test data set (anti-log)	63.3252	39.9690	0.9050

The good overall performance of the model is further depicted by the residual characteristics. The range of the residuals is relatively small and the majority of the residuals are distributed around the zero and approximately normally distributed (Table 3, Figure 5). The scatter plot between residuals and predicted values shows a random pattern with the residuals having constant variance (homoscedasticity).

Figure 5: Histogram of residuals distribution and distribution the residuals around zero

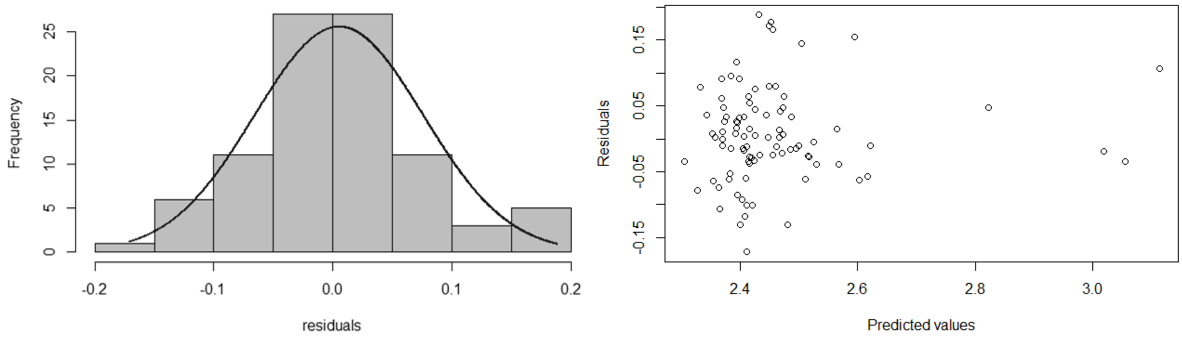


Table 3: Summary statistics of residuals of the RF model

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.171722	-0.034822	0.001687	0.005220	0.042783	0.188113

3.4 RF model interpretation

3.4.1 Variable importance

Based on RF variable importance, income is the most important predictor, followed by gross regional product, livestock, milk yield, population surplus, total population, and drink milk production (Figure 6, Table 4).

Figure 6: Variable importance for the milk price using machine learning algorithm and RF model

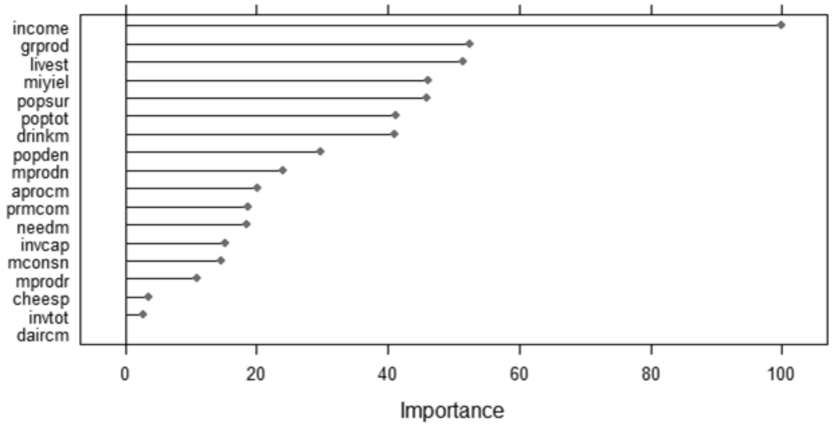


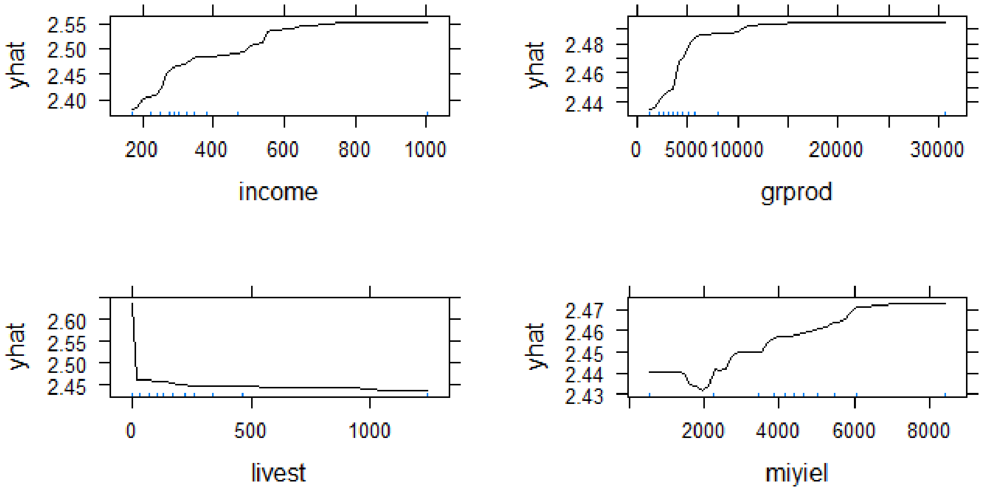
Table 4: Variable importance of predictor variables for the explaining milk price, %

Nr.	Variable	Importance, %	Nr.	Variable	Importance, %	Nr.	Variable	Importance, %
1.	income	100	7.	drinkm	40.904	13.	invcap	15.23
2.	grprod	52.471	8.	popden	29.797	14.	mconsn	14.665
3.	livest	51.526	9.	mprodn	24.043	15.	mprodr	10.864
4.	miyiel	46.071	10.	aprocm	20.108	16.	cheesp	3.586
5.	popsur	45.823	11.	prmcom	18.647	17.	invtot	2.741
6.	poptot	41.275	12.	needm	18.476	18.	daircm	0

3.4.2 Partial dependence plots

Based on the analysis of partial dependence plots (Figure 7) of the 4 most important variables we see that the income and gross regional product have a clear monotonic relationship with milk price. More specifically the increase of the income and gross regional product increases the milk price in these regions. Nevertheless, the maximum increase occurs in the specific range of the values and then we see the stabilization of milk price. Opposite to this, the increase in livestock decreases the milk price. Here we should mention that in the areas with extremely low livestock (Chukotka autonomous region, Magadan region, etc.) we can see an abrupt increase in the milk price. The milk yield shows more complex patterns but with similar to income and gross regional product monotonic trends.

Figure 7: Partial dependence plots between predicted milk price (yhat) and the four most important predictor variables: income, livestock, gross regional product, and milk yield



4 Discussion

In this study, we investigated the milk price variability in Russian regions over the years from 2013 to 2018. The data analysis showed that between 2013 and 2017 the milk price was continuously increasing with a small drop in 2018. This increase is probably the effect of the embargo which forced the Russian dairy sector to substitute imported dairy products through domestic production. Insufficient dairy production increases the competition on the dairy market and consequently enforced the increase of milk prices in Russia (Nosov et al., 2020). In general, the increase of the milk price might be considered a positive development, because farmers and milk producers receive higher payout prices for raw milk. However, providing high-quality raw milk constitutes a serious problem in Russia, due to the lack and low quality of feeding resources and hygiene, as well as lacking medicines for cattle (antibiotics).

The highest mean milk prices were observed in Chukotka autonomous district, Magadan region, and Kamchatka territory. All these regions are located in the Far East of Russia, where scarce feeding resources, extremely cold climate, low number of milk producers, livestock, and dairy companies take place. Moreover, in these regions, there are a lack of water and energy resources, low-developed infrastructure, lack of transport roads, low population density, and low purchasing power.

The RF variable importance shows that *income* is the most important factor, which determines the changes and fluctuations of milk prices in the Russian regions. Increasing average per capita income contributes to the improvement of livelihood, development of the region, growth of demand, and willingness to pay of consumers, additional purchasing power, etc. This all attracts new investments in the region and increase the prices for consumer goods, in particular for dairy products. The dairy companies achieve a higher profit and become able to pay more for the raw milk to the milk producers.

Similarly, *the gross regional product* is found to be an important variable for explaining the milk price in the Russian regions. The gross regional product shows the level of the development of the region. The more investments and cash flows are attracted to the region to support its development, the more developed will be the region. Thus, increasing production and a developing industry sector in the region seems to be associated with a growing number of dairy companies and milk producers. The increasing purchasing power of the population will build an attractive climate for developing the milk industry in the region that will constantly affect the raw milk price. Higher consumers' prices for dairy products could be a reason for growing the producers' prices for raw milk in the region. Thus, the milk price is influenced by gross regional product, so that the changes in gross regional product lead to differences in the milk price.

Besides, *livestock* is also considered as one of the most important variables for the explanation of the milk price in Russia. If in the Russian region the number of cattle is high, so the raw milk supply will be lower. Milk price will decrease because the provision of raw milk will grow. Therefore, livestock influences the milk price directly. The more livestock the region has, the higher is the milk supply and the bigger the amount of raw milk in the region. Then the dairy companies have more raw milk supply and more options to procure raw milk. For example, in most of the northern and far-eastern Russian regions the number of milk producers is low, combined with high raw milk prices. However, this is not necessarily a general regional pattern, as the analysis of Moran's I had revealed that the regional autocorrelation of milk price in Russia is statistically significant with p-value 0,03, but relatively low and corresponds to 0,142. A higher number of milk producers or increasing the size of dairy farms can lead to increasing livestock in the region. Higher milk yields from growing livestock numbers force milk producers and dairy companies to grow and to develop. If the provision with raw milk in the region is scarce, so the raw milk prices are higher in this region, but if the amount of raw milk provision is high, the raw milk prices are lower.

In the same direction, the *milk yield* in the region has also an important influence on the raw milk price in Russia. It is obvious that the higher is the milk supply in the region, the more milk is available, which leads to changes in the milk price. Increasing the milk yield in the specific region is the result of systematic breeding selection, scientific research, and favourable climate conditions for milk production, using high-quality feeding resources and applying technically efficient livestock farming methods.

In the present study, the milk price fluctuation in Russia was analyzed with the machine learning algorithm random forest (RF). RF has high predictive power and can detect which determinants are the most important for the explanation of the milk price. The feature selection algorithm Boruta was used to exclude any irrelevant predictors from the model training. The results of the Boruta algorithm confirmed the relevance of the collected data. The log transformation of the data could further improve the predictive performance of the model as it reduces the skewness of the milk price. Machine learning modeling is a data-driven approach and thus the presence of very few extreme observations reduces the ability of the model to learn these extreme cases. Time folds cross-validation was applied to consider the temporal time-series structure of the data. Here we should mention that the aim of the study was not the forecast of the milk price but understanding the main drivers that influence milk price.

Nowadays, the potential of the machine learning approaches in agriculture are continuously growing, as they provide an efficient way to process data with complex (non-linear) relationships. Future studies could include more potential predictors (e.g. feeding resources prices, climate data, etc.) and distinguish between highly collinear predictors such as regional income and gross regional product.

5 Conclusions

Milk prices increased during the years 2013 – 2018 in almost all Russian regions. Regarding the determinants of the regional milk price, the machine learning algorithm RF was used to estimate their influence on the milk price in the data set. Results show that income, gross regional product, livestock numbers and milk yield are the four most important determinants (from these used in this study) for explaining the raw milk price in Russia.

References

BALDUCCI, F., D. IMPEDOVO AND G. PIRLO (2018): Machine learning applications on agricultural datasets for smart farm enhancement. Machines 6(3): 38. https://doi.org/10.3390/machines6030038.

Banki Segodnja (2017): [online] Available: https://bankstoday.net/last-articles/inflyatsiya-v-rossii-po-godam

BORCHERS, M.R., Y.M. CHANG, K.L. PROUDFOOT, B.A. WADSWORTH, A.E. STONE AND J.M. BEWLEY (2017): Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. Journal of dairy science 100(7): 5664-5674. https://doi.org/10.3168/jds.2016-11526.

BREIMAN, L. (2001): Random forests. Machine learning 45(1): 5-32.

CARVALHO, G.R., D. BESSLER, T. HEMME AND E. SCHRÖER-MERKER (2015): Understanding International Milk Price Relationships. In Embrapa Gado de Leite-Artigoemanais de congresso (ALICE). In: Southern Agricultural Economics Association Annual Meeting, 2015, Atlanta, Georgia. [Proceedings...] Atlanta: SAEA, 2015.

CURRAN-EVERETT, D. (2018): Explorations in statistics: the log transformation. Advances in physiology education 42(2): 343-347. https://doi.org/10.1152/advan.00018.2018.

CUTLER, D.R., T.C. EDWARDS JR, K.H. BEARD, A. CUTLER, K.T. HESS, J. GIBSON AND J.J. LAWLER (2007): Random forests for classification in ecology. Ecology 88(11): 2783-2792. https://doi.org/10.1890/07-0539.1.

DEGENHARDT, F., S. SEIFERT AND S. SZYMCZAK (2019): Evaluation of variable selection methods for random forests and omics data sets. Briefings in bioinformatics 20(2): 492-503. https://doi.org/10.1093/bib/bbx124.

DONDERS, A.R., G.J. VAN DER HEIJDEN, T. STIJNEN AND K.G. MOONS (2006): Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology 59(10): 1087-1091. https://doi.org/10.1016/j.jclinepi.2006.01.014.

FAO (2014): Russia's restrictions on imports of agricultural and food products: An initial assessment. Rome: Food and Agriculture Organization of the United Nations.

GOLDSTEIN, B.A., A.M. NAVAR AND R.E. CARTER (2017): Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. European heart journal 38(23): 1805-1814. https://doi.org/10.1093/eurheartj/ehw302.

GREENWELL, B. (2017a): pdp: An R Package for Constructing Partial Dependence Plots. R Journal 9(1): 421-436.

GREENWELL, B. (2017b): Package 'pdp'. [online] Available: https://mran.microsoft.com/snapshot/2018-06-07/web/packages/pdp/pdp.pdf

HALL, M.A. (2000): Correlation-based feature selection of discrete and numeric class machine learning. Working paper 00/08. Hamilton, New Zealand: University of Waikato. Department of Computer Science.

JIANG, Y., B. CUKIC AND T. MENZIES (2008, July): Can data transformation help in the detection of fault-prone modules?. In Proceedings of the 2008 workshop on Defects in large software systems (pp. 16-20). https://doi.org/10.1145/1390817.1390822.

KUHN, M. (2008): Building predictive models in R using the caret package. Journal of statistical software 28(5): 1-26.

KULIKOV, I.M. AND I.A. MINAKOV (2019): Food security: problems and prospects in Russia. Scientific Papers: Management, Economic Engineering in Agriculture & Rural Development 19(4): 141-147.

KURSA, M.B. AND W.R. RUDNICKI (2010): Feature selection with the Boruta package. Journal of Statistical Software 36(11): 1-13.

LIAW, A. AND M. WIENER (2002): Classification and regression by randomForest. R news 2(3): 18-22.

LÜTKEPOHL, H. AND F. XU (2012): The role of the log transformation in forecasting economic variables. Empirical Economics 42(3): 619-638. https://doi.org/10.1007/s00181-010-0440-1.

MA, W., J. FAN, Q. LI AND Y. TANG (2018): A raw milk service platform using BP Neural Network and Fuzzy Inference. Information Processing in Agriculture 5(3): 308-319. https://doi.org/10.1016/j.inpa.2018.04.001.

MCQUEEN, R.J., S.R. GARNER, C.G. NEVILL-MANNING AND I.H. WITTEN (1995): Applying machine learning to agricultural data. Computers and electronics in agriculture 12(4): 275-293. https://doi.org/10.1016/0168-1699(95)98601-9.

MEYER, H., C. REUDENBACH, M. LUDWIG AND T. NAUSS (2018): CAST: "caret" Applications for Spatial-Temporal Models; R Package Version, 2018. [online] Available: https://cran.r-project.org/web/packages/CAST/CAST.pdf.

MORITZ, S. AND T. BARTZ-BEIELSTEIN, (2017): ImputeTS: Time Series Missing Value Imputation in R. The R Journal 9(1): 207-218.

MURTUZALIEVA, T.V., S.V. PANASENKO, E.V. SLEPENKOVA, T.A. TULTAEV AND B.I. POGORILYAK (2017): Import substitution strategy and ways of marketing, its implementation using the dairy industry of the Russian Federation as an example. Academy of Strategic Management Journal 16(2): 1-14.

NOSOV, V.V., N.M. SURAY, O.A. MAMAEV, O.V. CHEMISENKO, P.A. PANOV AND M.G. POKIDOV (2020, August): Milk production dynamics in the Russian Federation: causes and consequences. In IOP Conference Series: Earth and Environmental Science 548(2): 022091. IOP Publishing.

OSHIRO, T.M., P.S. PEREZ AND J.A. BARANAUSKAS (2012): How many trees in a random forest?. In International workshop on machine learning and data mining in pattern recognition (pp. 154-168). Springer, Berlin, Heidelberg.

PETRICK, M. AND L. GÖTZ (2017): The expansion of dairy herds in Russia and Kazakhstan after the import ban on Western food products. German Association of Agricultural Economists (GEWISOLA), 57th Annual Conference, Weihenstephan, Germany, September 13-15, 2017.

PROBST, P., M.N. WRIGHT AND A.L. BOULESTEIX (2019): Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9(3): e1301. https://doi.org/10.1002/widm.1301.

ROBERTS, D.R., V. BAHN, S. CIUTI, M.S. BOYCE, J. ELITH, G. GUILLERA-ARROITA, S. HAUENSTEIN, J.J. LAHOZ-MONFORT, B. SCHRÖDER, W. THUILLER, D.I. WARTON, B.A. WINTLE, F. HARTIG AND C.F. DORMANN (2017): Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8): 913-929. https://doi.org/10.1111/ecog.02881.

RUDNICKI, W.R., M. WRZESIEŃ AND W. PAJA (2015): All relevant feature selection methods and applications. In Feature Selection for Data and Pattern Recognition. Springer, Berlin, Heidelberg. 11-28.

SHAHINFAR, S., D. PAGE, J. GUENTHER, V. CABRERA, P. FRICKE AND K. WEIGEL (2014): Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. Journal of dairy science 97(2): 731-742. https://doi.org/10.3168/jds.2013-6693.

SOLODUKHA, P.V., E.A. MAIOROVA AND O.V. SHINKAREVA (2019): Social and economic consequences of influence of food embargo on production of milk and dairy products in Russia. Ecological Agriculture and Sustainable Development 2019(1): 297-305.

STEVENS, F.R., A.E. GAUGHAN, C. LINARD AND A.J. TATEM (2015): Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. PloS one 10(2): e0107042. https://doi.org/10.1371/journal.pone.0107042.

STORM, H., K. BAYLIS AND T. HECKELEI (2020): Machine learning in agricultural and applied economics. European Review of Agricultural Economics 47(3): 849-892. https://doi.org/10.1093/erae/jbz033.

STROBL, C., A.L. BOULESTEIX, A. ZEILEIS AND T. HOTHORN (2007): Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics 8(1): 25. https://doi.org/10.1186/1471-2105-8-25.

STROBL, C., A.L. BOULESTEIX, T. KNEIB, T. AUGUSTIN AND A. ZEILEIS (2008): Conditional variable importance for random forests. BMC bioinformatics 9(1): 307. https://doi.org/10.1186/1471-2105-9-307.

SVETNIK, V., A. LIAW, C. TONG, J.C. CULBERSON, R.P. SHERIDAN AND B.P. FEUSTON (2003): Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences 43(6): 1947-1958. https://doi.org/10.1021/ci034160g.

TRAWINSKI, B., M. SMĘTEK, Z. TELEC AND T. LASOTA (2012): Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. International Journal of Applied Mathematics and Computer Science 22(4): 867-881. https://doi.org/10.2478/v10006-012-0064-z.

UKAZ PRESIDENTA RF ot 21 Janvarya 2020 Nr. 20 “Ob utvergdenii Doktrini prodovolstvennoi besopasnosti Rossiyskoy Federacii. 22nd January 2020. www.garant.ru.

WEGREN, S.K. (2014): The Russian food embargo and food security: can household production fill the void?. Eurasian geography and economics 55(5): 491-513. https://doi.org/10.1080/15387216.2014.992449.

WEGREN, S.K., F. NILSSEN AND C. ELVESTAD (2016): The impact of Russian food security policy on the performance of the food system. Eurasian Geography and Economics 57(6): 671-699. https://doi.org/10.1080/15387216.2016.1222299.

WRIGHT, M.N., S. WAGER, P. PROBST AND M.M.N. WRIGHT (2018): Package ‘ranger’.[online] Available: https://cran.r-project.org/web/packages/ranger/ranger.pdf.

ZHOU, J., E. LI, H. WEI, C. LI, Q. QIAO AND D.J. ARMAGHANI (2019): Random forests and cubist algorithms for predicting shear strengths of rockfill materials. Applied Sciences 9(8): 1621. https://doi.org/10.3390/app9081621.