Modelling the production process of Roman snail using RIDGE and LASSO regression

S N Tkachenko¹, I A Tkachenko², S G Shpilevaya¹ and V P Dedkov²

¹ Institute of Physics, Mathematics and Information Technology, Immanuel Kant Baltic Federal University, 14, A. Nevskogo str., Kaliningrad, 236016, Russia
² Institute of Living Systems, Immanuel Kant Baltic Federal University, 2a, Universitetskaya str., Kaliningrad, 236040, Russia

E-mail: tkasergey@yandex.ru

Abstract. This article is a continuation of a series of works on the study and modelling of the production process of the Roman snail (Helix pomatia L.). The paper considers the applicability of three models for simulating the average weight of groups of Roman snails based on the values of environmental parameters, snail’s ration and specific factors characteristic of molluscs. The main focus of the simulation is on two models: RIDGE and LASSO regression. It is proved that RIDGE regression gives the best results in modelling. The combination of qualitative and quantitative parameters (air temperature, relative humidity, nutrition and some specific factors characteristic of molluscs) is sufficient for modelling the production process of Roman snail.

1. Introduction

In the previous work [1], the question of the applicability of regression models for modelling the production process of biological objects was considered. The Roman snail (Helix pomatia L.) was taken as a simulated species.

A detailed description of the experiment with snails, the features of their keeping, feeding and measurement of average weight is presented in [2].

The Roman snail is widely distributed in Europe. The Kaliningrad region of the Russian Federation is one of the borders of its areal. Human actively uses this type of mollusc in various spheres from gastronomy to cosmetology.

As well know that demands on Roman snail’s meat are outstripping the natural capacity. Therefore, its artificial breeding is necessary, which in turn leads to the need to develop a model of the production process for this type of mollusc.

The production process is understood as weight gain in conditions of artificial breeding, that is, keeping in controlled conditions of the farm. In this condition, it is possible to control various parameters that affect the weight of the snail. Usually these are environmental parameters (air temperature and humidity), as well as factors related to the vital activity of the snail itself (conditions of keeping, feeding, etc.).

Modelling of the production process is reduced to the mathematical modelling based on a number of parameters that are used to calculate the average weight of a group of snails. A six groups of snails are considered. There are ten snails in each group. Groups are designated using a combination of an
alphabetic character and a number. Three groups are marked with the letters "Ryb", which corresponded to the place of collection of molluscs (the village of Rybachy in the Curonian Spit national park in the Kaliningrad region), and three more groups are marked with the letters "Lad" (Ladushkin city in the Bagrationovsk district of the Kaliningrad region). The character is followed by a number corresponding to a specific ration: 1 – ‘herbs’ diet (the feed consisted of a mix of the plant leaves), 2 – vegetable diet (a mixture of vegetables grown in the Kaliningrad region), 3 – a mix of plant leaves and vegetable (50/50).

Each of the six groups was monitored for 68 days. According to the results of observations, environmental parameters were recorded, as well as the average weight of each group, feeding features and behaviour of snails. The initial data for the simulation is described as set of 68 points corresponding to the days of the experiment [2]. For each group of snails, a different data set are used.

In the previous paper [1], we showed that regression models have a number of limitations as a tool for modelling the production process of the Roman snail. To overcome these limitations, we need to add independent variables. In this paper, a modernized model is created taking into account additional parameters and appropriate mathematical apparatus of contemporary machine learning.

### 2. Data preparation

In the previous work [1], good results on modelling the average weight of snails were obtained only when adding "noise". Additional variables that do not have any physics or biology interpretation have been added to the data set.

In [1], the number of the day, week, month, and converted UNIX time (time in seconds since January 1, 1970) were added to the data set, and shifts of the main variables: air temperature and humidity – were performed.

This approach is not correct. Therefore, we upgraded the original data as follows: we removed all "noise" variables, deleted variable shifts, and replaced the parameter associated with the snail’s diet with the food weight.

The qualitative feature that denoted one of the types of diet, namely: herbs or vegetable or mix was removed from the original data set.

Instead, two attributes were added, one of which is the weight of food in grams that was given to a group of snails on a specific date, the second is the weight of food residues that were removed from the snail habitat the next day after the food portion was given (variable $OUT$). Removing food residues is an important process, as snails should not eat stale food, it is also important that the place where the snails live have no mould.

The weight of the introduced food was divided into two independent variables ($IN1$, $IN2$) depending on the diet: $IN1$ – the weight of the vegetables given out, $IN2$ – the weight of the plant leaves if available. If the weight of vegetables on a particular date is zero, it means the fact that they were not issued on that day.

Thus, the data set is an eight-dimensional time series with the following variables:

- **Date** has the next format "dd.mm.yyyy", is a time stamp of the series, covers the period from 03.07.2017 to 08.09.2017 with a discreteness of one day.
- The average weight of a group of snails is the target variable (denoted by the letter W, followed by the designation of the corresponding group of snails Lad1..3, or Ryb1..3), the value of which is required to predict, as well as to identify dependencies on other variables and parameters of the model.
- $T$ – air temperature.
- $\varphi$ – relative air humidity.
- $A$ – parameter that characterizes the conditions of adaptation of snails to changed environmental conditions, takes values from -1 to 1.
- $R$ is a discrete variable that denotes the number of snails participating in the breeding process as of a certain date, taking values from 0 to 1 in increments of 0.1.
• IN1, IN2, OUT – values of the weight of food that was given to snails and the weight of residues.

Python version 3.7 with additional open libraries: NumPy, Pandas, etc. [3] was used for working with data and creating models.

The data is a time series and the number of points is small, in that case the process of training the models should include cross-validation and splitting the set of points into 5 folds. To evaluate the models, as before in [1], the prediction horizon consists of the last 12 points in time. This is due to the fact that the test set should have included a section without a sharp rise in weight and with its presence. Since the data set is time series, the prediction horizon can only consist of the rightmost points (the last in time).

3. Modelling of the production process
To model the production process, it is decided to use regression models. Currently, regression is widely used, including forecasting tasks. The purpose of regression analysis is to determine the relationship between the target variable and a set of external factors (regressors) [4], which is required in this task. The goal is to create a model that predicts the value of an average weight based on several input variables.

Linear regression is most often used. Mainly because of its ease of interpretation, trainability, and good accuracy. When constructing a linear regression model, the input is a time series or a set of values of some functional dependence, the behaviour of which needs to be predicted. A certain averaged line is built on the set of input values, based on which further forecasting is made:

\[
y = w_1 \cdot x + w_2
\]  

There are also other methods based on the linear regression. For example, the LASSO, RIDGE regression [5].

The LASSO algorithm uses the so-called L1 regularization. The method of RIDGE regression – L2 regularization. They are quite similar [6], both represent an additional term in the formulas for calculating the values of loss functions:

\[
J_{\text{LASSO}} = \sum_{i=1}^{n} (y_n - \hat{y}_n)^2 + \lambda \|w\|_1
\]

\[
J_{\text{RIDGE}} = \sum_{i=1}^{n} (y_n - \hat{y}_n)^2 + \lambda \|w\|_2^2
\]

RIDGE regression is a method used to potential problems with multicollinearity when independent variables are correlated with each other.

Predictive power can decrease with an increased correlation between predictor variables.

The method consists in introducing an additional regularizing term into the minimized functional (strictly speaking, this method is not a feature selection method, since it does not specify which features should be excluded from the model).

LASSO (least absolute shrinkage and selection operator) method for estimating parameters of a linear model. When evaluating, a limit is imposed on the sum of the absolute values of the model parameters.

MAPE (Mean Absolute Percentage Error) and ME (Max Error) are used as metrics to evaluate the accuracy of the results, which are suitable for evaluating all models used in the work. MAPE is calculated as the average of the modulus of the ratio of the difference between the actual and predicted values to the actual values, multiplied by 100%. In other words, the MAPE value can be used to estimate how much the model is wrong on average. ME – the maximum value of the modulus of the difference between the actual and predicted values. The ME value reflects the maximum deviation of
the predicted value from the actual value, which allows to estimate the accuracy of the model in the worst case.

Modelling the average weight values for groups Lad2 and Ryb2 is the most difficult, since for these groups all values of the variable IN2 are zero (the ration consists only of vegetables). Further, we consider the data and simulation results for the average weight of the Lad2 group of snails. For the other five groups, the conclusions drawn for the Lad2 group are identical.

First, we analysed the dependencies of variables on each other. The most interesting is the graph of the relationship between the variables $T$, $\varphi$ and $WLad2$ (figure 1). The figure 1 shows that with a decrease in air temperature and a simultaneous increase in air humidity, the values of the average weight of snails increase.

![Figure 1](image)

**Figure 1.** Three-dimensional relationship between air temperature, relative air humidity and average weight of a group Lad2.

A correlation matrix of various variables (figure 2) was constructed in the whole data set, as well as separately in its training part. Almost all parameters have approximately the same level of correlation with the target variable $WLad2$. This fact confirms the conclusion that it is difficult to distinguish a direct relationship between the values of individual variables, since they all affect each other in the aggregate.

Based on the results of the correlation matrix analysis, the air temperature has a fairly high level of correlation with the target variable, and the relative humidity of the air correlates, in turn, with $T$ and $WLad2$. This triple of interrelated variables was also identified by experiments with the construction of
points in three-dimensional space, as one of the most obvious (figure 1), which once again confirms the dependence of the average weight of snails on environmental parameters.

As can be seen from figure 2, a strong negative correlation is observed for variables $A$ and $WLad2$ (the correlation coefficient is -0.82). The values of variable $a$ are monotonically increasing (in absolute value monotonically decreasing) during the first 12 days. All the remaining time value of this variable is equal to zero. Variable $A$ characterizes the adaptation of snails to changed environmental conditions. Negative values mean that conditions have worsened, as was the case for the Lad2 group. Positive values, on the contrary, mean that the living conditions of a group of snails have improved compared to their natural habitat. The duration of the adaptation period (12 days) was determined experimentally.

The presence of a strong correlation (positive or negative) between the variable associated with adaptation and the average weight of a group of snails confirms the need to introduce variable $A$ into the dataset. Without this variable, it will be impossible to take into account the monotonous decrease in average weight for groups Lad1...3 in the first days of observation. Also, without presence variable $A$, it is impossible to explain the weight gain for groups Ryb1...3 in the same time period.

Therefore, the inclusion of a variable in the data set that is associated with adaptation mechanisms is a necessary condition for the validity of the developed model.

Before starting the simulation, the data (except for the target variable) were normalized using a standard algorithm [7].

We used the complementary method plotModelResults for model testing and validation. This method calculates the predicted values of the target variable, the values of the two selected metrics, and plots the real and predicted values with confidence intervals. Each of the models was examined by this method.

The metric values for each of the three models are shown in table 1.
The most accurate model, both in terms of metrics and based on the graph of predicted values, is the RIDGE model. Classical linear regression and LASSO regression give higher MAPE error values.

The highest error values are typical for the LASSO regression model. This can be explained by the fact that the LASSO algorithm, instead of charging penalties for each feature in the data, charged penalties only for features with a large coefficient value, which is not entirely useful in this case. L1 regularization shrinks many regression coefficients to zero.

Linear regression gives error values close to LASSO method. This is probably due to that the linear model could not correctly take into account all the initial data.

The results of the model based on RIDGE regression are shown in figure 3. The confidence interval is highlighted with red dotted lines. The confidence interval is calculated for each of the constructed models. The values are called anomalous if the model curve does not match the experimental one even within the confidence interval. This is observed when the predicted values differ too much from the experimental ones. The presence of anomalies leads to an increase in ME values.

The width of the confidence interval was defined as the average value of the standard deviation obtained during cross-validation. This width is different for various models. With a smaller error on the training set, we get a smaller spread, but at the same time we can get an increase in the error on the test set.

As can be seen from figure 3, the curve predicted by the RIDGE regression method coincides with the experimental curve within the confidence interval. There are no anomalies in this graph. This indicates the good quality of this model.

![Figure 3](image_url)  
**Figure 3.** Experimental and predicted values (for the last twelve days) of the average weight for the Lad2 group calculated by the RIDGE regression.

The results of the model based on LASSO regression are shown in figure 4.

As can be seen from figure 4, the curve predicted by the LASSO regression method coincides with the experimental curve within the confidence interval at only two points. All other points are anomalies. At the same time, as follows from table 1, the values of MAPE and ME for the RIDGE and

### Table 1. Values of model metrics.

| Model                  | Mean Absolute Percentage Error, % | Max Error (without normalizing the target variable) |
|------------------------|----------------------------------|-----------------------------------------------------|
| Linear Regression      | 4.50                             | 1.49                                                |
| LASSO Regression       | 4.72                             | 1.55                                                |
| RIDGE Regression       | 3.69                             | 1.31                                                |
LASSO methods do not differ very significantly. However, the graph shows these differences more clearly.

![Image](image_url)

**Figure 4.** Experimental and predicted values (for the last twelve days) of the average weight for the Lad2 group calculated by the LASSO regression.

The presence of such a large number of anomalies is influenced by the absolute values of the predicted average weight, the mathematical model, and the results of cross-validation. In this regard, it is difficult to single out a main factor that would be responsible for the unsatisfactory results of the LASSO algorithm.

Linear regression shows results similar to the LASSO method. Only two points fit into the confidence interval, so it is also concluded that this model is not applicable for this data set. This method can suffer greatly from the effects of multicollinearity.

For a more complete comparison of the simulation results using the LASSO and RIDGE algorithms, these values are shown in one figure. Figure 5 shows three graphs: experimental data and two graphs of values predicted from the LASSO and RIDGE models.

![Image](image_url)

**Figure 5.** Experimental and predicted values (for the last twelve days) of the average weight for the Lad2 group calculated by the LASSO and RIDGE regression.
As can be seen from figure 5, the curve for LASSO regression lies under the curve for RIDGE regression over the entire test time interval. The curve for linear regression is almost the same as the curve for LASSO regression, so it is not shown in figure 5.

The observed placement of the curves in figure 5 visualizes the difference in the MAPE and ME metrics. It can also be seen that the RIDGE regression curve is closest to the experimental curve, so this model is the best one considered.

The main difference between LASSO and RIDGE regressions is that the LASSO algorithm can cause some independent variables to go to zero, while the RIDGE reduces them to values close to zero.

Consider for simplicity a two-dimensional space of independent variables. In the case of LASSO regression, the constraint on the coefficients is a rhombus $|w_1| + |w_2| \leq t$, in the case of RIDGE regression, a circle $w_1^2 + w_2^2 \leq t^2$, where $t$ is the regularization parameter. It is necessary to minimize the error function, but at the same time observe the restrictions on the coefficients. From a geometric point of view, the task is to find the point of contact of the line reflecting the error function with the figure reflecting the constraints on $w_i$.

In the case of LASSO regression, this point is very likely to be in the corners of the rhombus, that is, to lie on the axis, whereas in the case of RIDGE regression, such a position of the point is very rare. If the intersection point lies on the axis, one of the coefficients will be zero, which means that the value of the corresponding independent variable will not be taken into account.

The main difference between RIDGE and LASSO regression from the point of view of model construction is that the RIDGE algorithm takes into account the maximum possible number of independent variables, namely six (since for Lad2 group all values of $IN2$ are zero), while the LASSO algorithm uses only three independent variables: $A$, $T$ and $IN1$.

The values of the coefficients for the variables included in the dataset for all three constructed models are presented in table 2. A large value of the coefficient in absolute value means that the target variable strongly depends on this parameter. If the coefficient value is zero, then the target variable does not depend on this parameter.

|                  | Linear Regression | LASSO Regression | RIDGE Regression |
|------------------|-------------------|------------------|------------------|
| $T$              | -1.60             | -1.03            | -1.77            |
| $\phi$           | -0.23             | 0.00             | 0.60             |
| $A$              | -10.63            | -10.57           | -5.50            |
| $R$              | -0.44             | 0.00             | -0.04            |
| $IN1$            | 0.61              | 0.49             | 0.49             |
| $OUT$            | 0.04              | 0.00             | 0.68             |

As can be seen from table 2, all three models almost equally assess the contribution of the variable associated with the amount of food that was given to snails. Similar values include the contribution of air temperature, although this contribution is lower for LASSO regression compared to other models.

For variables such as relative humidity and the weight of food residues, there is a large difference in the contribution to the models studied. These variables only make a noticeable contribution to the RIDGE regression model. In the other two, their contribution is either minimal or zero. From a biological point of view, these parameters should affect the average weight values, since the humidity of the air determines the activity of snails. By the weight of the remains, you can judge how much food the snails actually ate, that is, again, how active they were at one time or another. Therefore, the RIDGE algorithm is most correct when dealing with these two parameters.
The contribution of such a parameter as adaptation is very large for linear and LASSO regression models and half as much for RIDGE regression. It can be assumed that excessive dependence on the variable associated with the adaptation of snails to changing environmental conditions negatively affects the quality of the model. As mentioned earlier (figure 2), the target variable has a high correlation with variable $A$. Therefore, the model must reduce the effect of multicollinearity otherwise the maximum error values increase.

Unexpectedly low is the contribution of a variable that is associated with snail reproduction. It makes a significant contribution only to the values of linear regression, and this does not lead to any qualitative improvements in this model. The other two models practically do not take into account the effect of reproduction on the production process. This is the strangest fact, because if a snail is involved in the breeding process, it can lose up to 20% of its weight in a short period of time. Perhaps there is a problem with how the values of the variable $R$ are set and they do not reflect the essence of this process.

It should be noted that all three models studied give underestimated values compared to the experimental ones. This may indicate that some factors are not taken into account in the mathematical model or are not present in the original dataset. It can be expected that if all factors are taken into account, the predicted values will be located symmetrically relative to the experimental data (the predicted points will be higher and lower than the corresponding experimental values). This fact may be due to the fact that the models do not take into account the influence of the breeding process on the production process.

As for the factors affecting the target variable, the graphs presented in the work showed that there is no specific relationship between any variables from the data set. All variables affect the target variable, which, in turn, does not have a pronounced periodicity and predictable behaviour in advance.

4. Conclusion

Thus, the model based on RIDGE regression most accurately (in comparison with other models) reflects the influence of environmental parameters on the production process of a group of Roman snails. Indeed, with an increase in relative humidity (up to 60-70%) and a decrease in its temperature (up to 22-24 °C), favourable conditions are created for weight gain and reproduction of snails.

The model takes into account that the average weight of a group of snails reacts quickly to changes in air temperature, but with a regular delay to changes in air humidity. This behaviour can be explained by the influence of endogenous factors – snails themselves moistened the soil with their mucus. As a result, microclimate becomes a more suitable.

At the moment to take into account in the mathematical model the influence of endogenous factors is difficult, as it is hard to determine the exact amount of mucus secreted by molluscs per unit of time, for example, for one day or to get the value of the net weight of mucus in each time interval. However, the formalization of this attribute could improve and refine the constructed model.

It is also worth noting the need for parameter $A$, which characterizes the conditions of adaptation of snails to changed environmental conditions. It is the key parameter associated with the average weight of a group of molluscs in the first days (when the weight changes monotonously). Once in new conditions for themselves, snails must adapt to the environment and the new ration, which takes some time. During this period, many molluscs in the group refuse to eat, which leads to a decrease in average weight. Thus, the necessity of including this parameter in the model was objectively proved.

During the breeding (mating) and egg laying period, snails can also refuse to eat. At these moments, it is quite difficult to measure the weight of specific individuals. Therefore, a discrete variable $R$ was introduced, denoting the number of snails participating in the breeding process as of a specific date, taking values from 0 to 1 in increments of 0.1.

Accounting for all independent variables in the RIDGE model allowed it to predict the values of the target variable most accurately. Other models showed a higher error value, since they did not take into account all variables (LASSO, due to its apparatus, took into account only three indicators - $A$, $T$ and $IN1$) or due to the limitations of its apparatus (as a linear regression). Nevertheless, it is proved
that taking into account all quantitative and qualitative factors affects more significantly than stochastic error reduction.

Thus, we can conclude that RIDGE regression has shown better results for modelling the data of the Roman snail’s production process.

References

[1] Tkachenko I A, Tkachenko S N and Kopytov G V 2020 Modeling of the Roman snail production process using regression models Proc. Int. Conf. Actual problems of applied mathematics, computer science and mechanics (Voronezh 11 to 13 November 2019) pp 1121-1125 (in Russian)

[2] Tkachenko I A, Dedkov V P and Tkachenko S N 2018 Influence of ration and environmental parameters on weight of Helix pomatia terrestrial snail (Linnaeus 1758) under the conditions of artificial breeding Vestnik of North-Eastern Federal University 6 pp 34-42 (in Russian)

[3] McKinney W 2017 Python for Data Analysis Data Wrangling with Pandas, NumPy, and IPython (O’Reilly) p 960

[4] Gheyas I A and Smith L S 2009 A Neural Network Approach to Time Series Forecasting Proceedings of the World Congress on Engineering (London) pp 1292-1296

[5] Garreta R, Moncecchi G, Hauck T and Hackeling G 2017 scikit-learn: Machine Learning Simplified (Packt Publishing) p 530

[6] Demir-Kavuk O, Kamada M, Akutsu T and Knapp E. 2011 Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features BMC Bioinformatics 12 pp 412-422

[7] Nikulin V N, Kanishchev I S and Bagaev I V 2016 Methods of balancing and normalizing the data to improve the classification quality Computer tools in education 3 pp 16-24 (in Russian)