

# Principal component analysis in the epidemiology of diarrhoea in calves

Ablaikhan Kadyrov<sup>1</sup>, Altay Ussenbayev<sup>2</sup>, Dariyash Kurenkeyeva<sup>3</sup>, Berdaly Kurenkey<sup>4</sup>,  
Sarsenbay Abdrakhmanov<sup>2</sup>, Nurlan Tashatov<sup>1</sup>

<sup>1</sup>Department of Computer and Software Engineering, Faculty of Engineering Science, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

<sup>2</sup>Department of Veterinary Medicine, Faculty of Veterinary Medicine and Animal Husbandry Technology, Seifullin Kazakh Agrotechnical University, Astana, Kazakhstan

<sup>3</sup>Department of Information Systems, International Information Technology University, Almaty, Kazakhstan

<sup>4</sup>Faculty of Physics and Technology, Al Farabi Kazakh National University, Almaty, Kazakhstan

## Article Info

### Article history:

Received Sep 25, 2022

Revised Feb 6, 2023

Accepted Feb 12, 2023

### Keywords:

Calve

Diarrhoea

Eigenvalue

Eigenvector

Epidemiology

Kazakhstan

Principal component analysis

## ABSTRACT

The paper is the first to use principal component analysis (PCA) in veterinary epidemiology and aims to identify the most significant agents of diarrhoea in calves. Data were collected from 245 calves on 32 farms within 13 districts of Northern Kazakhstan. Epidemiological data were simulated in R, using standard PCA modules. The data clustering was carried out based on the prevalence of seven enteropathogens for an age group dataset (represented by four animal groups) and a farm type dataset (by three farm types depending on herd size). The simulation revealed that the components identified here for one and two-week calves explained 91.31% of the variance. The first two components of large and middle-sized farms covered 90.3% of the variance. The coordinates corresponding to pathogens were approximately visualised in directions of eigenvectors for each age group and farm type. The coordinate for *Cryptosporidium parvum* was in directions of eigenvectors for one-to three-week calves and large farms. The coordinate for rotaviruses was in the direction of eigenvectors for four-week calves and medium-sized and small farms. So, PCA can potentially be useful in the clustering of epidemiological datasets and making decisions on control of infectious diseases with a multi-pathogenic nature.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Dariyash Kurenkeyeva

Department of Information Systems, International Information Technology University

Manas Street 34/1, 050040 Almaty, Kazakhstan

Email: dariyash.kurenkeyeva@gmail.com

## 1. INTRODUCTION

The Singular value decomposition (SVD) is the most frequently used tool in computational linear algebra [1], and is considered the foundation of modern computational science and technology. The numerical implementation of SVD is very useful from a mathematical point of view, and computations based on SVD have generated many important results. The main idea of this method is a generalisation of the Fourier transform, in which the primary data are mapped to a new coordinate system where the task becomes simpler [2]–[4].

Principal component analysis (PCA) is a statistical interpretation of SVD and is used to reduce the dimensionality of a task. This technique applies an orthogonal transformation to several correlating variables to create a smaller number of linearly uncorrelated principal components. PCA was first described at the beginning of last century [5], and has since proven to be valuable in statistical theory and practice. The method

is widely used in data science and machine learning applications, as it allows the original problem to be reduced to a data-driven hierarchical coordinate system, based on directions that limit the maximum number of statistical variations of the dataset. PCA is often used in practice thanks to its capability to interpret different scientific problems. However, although this method is popular and has a wide range of applications in bioinformatics, ecology, and other biomedical sciences [6]–[9], it has rarely been used in the epidemiology of contagious diseases in the field of public and animal health research.

It is known that diarrhoea in neonatal calves is a common condition that has a negative impact on welfare and remains the main cause of death for new-born animals in many countries [10]–[12]. The economic importance of diarrhoea is associated with mortality and stunting in calves, as well as the costs of diagnosis, treatment and control [13]. Bovine rotaviruses (BRV) and coronaviruses (BCV), the enterotoxigenic strain of *escherichia coli* K99, and a protozoa species *cryptosporidium parvum* are widespread endemic pathogens causing diarrhoea in new-born calves. These pathogens are found in mono and mixed infections in diarrhoeal calves, and the severity of the disease depends on the interactions between these pathogens, environmental factors, and factors associated with carrier animals [14], [15]. Etiological diagnosis of diarrhoea is carried out by identifying pathogens in faeces [16]. Identification of the significance of these causative agents is therefore necessary in order to control diarrhoea in neonatal calves on farms in Kazakhstan. In addition, *c. parvum* is a zoonotic pathogen, and an assessment of its epidemiological importance is relevant for public health. In this study, we provide a theoretical framework based on a matrix-dimensionality reduction method, namely PCA, and explain its inference and application in the epidemiological context for understanding the epidemiology of diarrhoeal disease in new-born calves, with the aim of defining the most important infectious pathogens on cattle farms in Northern Kazakhstan.

## 2. MATHEMATICAL FRAMEWORK AND NOTATION

Let us consider a dataset of measurements from independent experiments, written in the form of the row vectors of a matrix  $X$ . Each row vector  $x_i$ , in this case, represents measurements from one individual experiment. Since the PCA method is a statistical interpretation of the SVD, it consists of the following steps.

### 2.1. Step 1: standardisation

The PCA method is sensitive to deviations in the original variables. The initial values of the variables in the dataset have large ranges, and to avoid variables with large ranges dominating the other variables, the source data must be centred. As a result, the original data are transformed on a comparable scale. Mathematically, standardisation of the original data can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

### 2.2. Step 2: covariance matrix

The purpose of this step is to define the relationship between the dataset variables. It is known that where there is strong correlation between variables, redundant information may be present. To determine the correlations between variables, a correlation matrix is calculated, the  $(i, j)$ -th element of which represents the correlation of the features  $(X_i, X_j)$ . According to the covariance formula:

$$\text{Cov}(X_i, X_j) = E \left[ (X_i - E(X_i)) (X_j - E(X_j)) \right] = E(X_i, X_j) - E(X_i)E(X_j)$$

since  $E(X_i) = E(X_j) = 0$ , then  $\text{Cov}(X_i, X_j) = E(X_i, X_j)$ . The covariance of a variable with itself is its variance, i.e.,  $\text{Cov}(X_i, X_i) = \text{Var}(X_i)$  when  $X_i = X_j$ . On the main diagonal, we have the variance of each original variable. The covariance matrix is symmetric and is a generalisation of the variance to the case of multidimensional random variables; it also describes the shape (scatter) of the random variable, similarly to the variance.

### 2.3. Step 3: eigenvalues and eigenvectors

The covariance matrix is a generalisation of the variance to higher dimensions. Let us consider a unit vector onto which we will project our random vector  $X$ . Then, the projection will be equal to  $v^T X$ . The variance projection onto the vector will be correspondingly equal to  $\text{Var}(v^T X)$ . In general, in the vector form (for centred quantities), the variance is expressed as, and  $\Sigma$  is a matrix singular value.

$$\text{Var}(X) = \Sigma = E(X \cdot X^T)$$

$$\text{Var}(v^T X) = \Sigma^* = E((v^T X) \cdot (v^T X)^T) = E(v^T X \cdot X^T v) = v^T E(X \cdot X^T) v = v^T \Sigma v$$

The variance is maximised at a value of  $v^T \Sigma v$ . According to the Rayleigh distribution, which is a special case for covariance matrices, when  $x$  is an eigenvector, and  $\lambda$  an eigenvalue. Thus, the direction of the maximum variance in the projection always coincides with the eigenvector that has the maximum eigenvalue equal to the magnitude of this variance. This also holds true for projections on a larger number of dimensions: the variance (covariance matrix) of the projection on an  $m$ -dimensional space will be maximal in the direction of  $m$  eigenvectors with maximum eigenvalues.

$$R(M, x) = \frac{x^T M x}{x^T x} = \lambda \frac{x^T x}{x^T x} = \lambda$$

#### 2.4. Step 4: derive the new dataset

In this step, we need to calculate the vector  $v^T X$ . If we have a hyperplane rather than a vector, then instead of the vector  $v^T$  we take the matrix of basic vectors  $V^T$ . The resulting vector (or matrix) will be an array of projections of our observations. To interpret experimental data, it is often necessary to estimate the amount of lost information, and the most convenient way to represent this is as a percentage. We take the variances along each of the axes and divide them by the total sum of the variance along the axes (i.e., the sum of all eigenvalues of the covariance matrix). If the relationship between the signs is very strong, then the loss of information will be minimal. The principal components are the new variables, which are constructed as linear combinations, constructed in such a way that the new variables (i.e., the principal components) are uncorrelated, and most of the information in the original variables is compressed into the first components. The aim of PCA is to put the maximum possible information into the first component, then the maximum remaining information into the second, and so on.

### 3. MATERIALS AND METHODS

#### 3.1. Data collection

Epidemiological data were collected between January and August of 2019 by a cross-sectional investigation of randomly selected 32 farms, including dairy, fattening and small householding entities, in 13 districts of Northern Kazakhstan. In total, faeces were sampled from 245 neonatal calves under one month of age with clinical signs of diarrhoea. Material for the study was collected individually during one-off trips to farms via per rectum sampling of faeces. The samples were examined microscopically for the presence of *Cryptosporidium* oocysts after Heine (1982) [17] and tested for *C. parvum*, *E. coli* K99, *BRV*, and *BCV* with a commercial FassisiBoDia immune chromatographic test (Fassisi GmbH, Germany).

#### 3.2. Simulation

Simulations of epidemiological data were performed in R using the standard PCA modules. Classification of the data on the prevalence of pathogens was carried out separately: the  $x_i$  line vector for the age group dataset was represented by four groups of new-born animals (age 1: 35 calves under the age of one week; age 2: 86 calves that were two weeks old; age 3: 62 calves that were three weeks old; age 4: 86 calves that were four weeks old). The vector for the farm type dataset was sampled from three types of farms with varying sizes of dairy herds (large farms: 10 enterprises with more than 500 milked cows; medium-sized farms: eight peasant farms with more than 150 productive animals; small farms: 14 households with fewer than 20 dairy cows). Data clustering was carried out considering the influence of the age groups of the calves and the types of farms on the infection level of animals with seven combinations of enteropathogens, as shown in Table 1. The survey was approved by the research ethics committee of the faculty of veterinary and animal husbandry technology, Seifullin Kazakh Agrotechnical University.

### 4. RESULTS AND DISCUSSION

#### 4.1. Epidemiological data on the prevalence of enteropathogens in calves in the north of Kazakhstan

The most common pathogens in the calves' intestines were *C. parvum* and *BRV*, these were found in 70.6% of the studied farms. There was also *BCV* on 47.05% of the farms and *E. coli* K99 on 23.5% Table 1. Intestinal pathogens were found mainly in the form of mixed infections: in nine cases (52.9% of farms), a combination of two pathogens (*C. parvum* + *BRV*, or *C. parvum* + *BCV*) was noted, and this combination was observed most often in two-week-old calves. Two-component mixed infections were mainly found in enterprises with a large number of cows Tables 1 and 2.

Table 1. Combinations of intestinal pathogens in calves up to one month of age

Pathogens and their combinations	Number of infected farms (%), n=17	Number of infected calves (%) by age			
		1-7 days (n=35)	8-14 days (n=86)	15-21 days (n=62)	22-31 days (n=62)
<i>C. parvum</i> (total positive)	12 (70.6)	7 (20)	15 (17.4)	14 (22.6)	1 (1.6)
<i>BRV</i> (total positive)	12 (70.6)	4 (11.4)	15 (17.4)	5 (8.1)	3 (4.8)
<i>BCV</i> (total positive)	8 (47.05)	2 (5.7)	1 (1.7)	4 (6.4)	4 (6.4)
<i>E. coli</i> K99 (total positive)	4 (23.5)	1 (2.8)	2 (2.3)	3 (4.8)	0
<i>C. parvum</i> + <i>BRV</i>	5 (29.4)	3 (8.6)	4 (4.6)	1 (1.6)	0
<i>C. parvum</i> + <i>BCV</i>	4 (23.5)	2 (5.7)	3 (3.5)	1 (1.6)	0
<i>BRV</i> + <i>BCV</i>	3 (17.6)	0	2 (2.3)	1 (1.6)	0

Table 2. Infections with intestinal enteropathogens by size of farm

Pathogens and their combinations	Type of farm based on numbers of dairy cows		
	>500 cows (n=171)	>150 cows (n=43)	<20 cows (n=31)
<i>C. parvum</i> (total positive)	32 (18.7)	4 (9.3)	1 (3.2)
<i>BRV</i> (total positive)	12 (7.01)	7 (16.3)	8 (25.8)
<i>BCV</i> (total positive)	10 (5.8)	0	1 (3.2)
<i>E. coli</i> K99 (total positive)	1 (0.6)	5 (11.6)	0
<i>C. parvum</i> + <i>BRV</i>	7 (4.09)	1 (2.3)	0
<i>C. parvum</i> + <i>BCV</i>	6 (3.5)	0	0
<i>BRV</i> + <i>BCV</i>	3 (1.7)	0	0
	71 (41.5)	17 (39.5)	10 (32.2)

4.2. PC analysis of epidemiological data

Our datasets showed some statistical distribution and there was some statistical variability in this information. Statistical data are presented in Table 3. The infection rates varied or fluctuated more strongly among the older calves and the highest variations in prevalence rate were observed on small farms. This was due to the small amounts of data used Table 3.

The sizes of the variance for both datasets suggest that the first two principal components account for most of variations in Table 4. The age group PC1 explains 68.9% of the variation. In the farm type dataset, the corresponding group PC1 explains 61.5% of the variation. These proportions were obtained by dividing each eigenvalue by the total of all eigenvalues. When dealing with these linear combinations, we can ignore extremely small values. The first two components for the age groups explain 91.31% of the changes in the dataset, while the first two components for the farm types explain 90.3% of the changes.

Table 3. Preliminary statistical analysis of the total numbers of infections of calves with enteropathogens

	Mean	Standard deviation	Coefficient of variation (%)	n
Age group dataset				
Age 1	2.7142	2.2886	84.3	19
Age 2	6.0000	6.2182	103.6	42
Age 3	4.1428	4.6342	111.8	29
Age 4	1.1428	1.6761	146.6	8
Farm type dataset				
Large farms	10.1428	10.3509	102.0	71
Medium-sized farms	2.4285	2.8784	118.5	17
Small farms	1.4285	2.9358	205.52	10

Table 4. Significance of principal components

	Standard deviation	Proportion of variance	Cumulative proportion
Principal components for the age group dataset			
PC1	1.6599	0.6888	0.6888
PC2	0.9473	0.2243	0.9131
PC3	0.49851	0.06213	0.97525
PC4	0.31466	0.02475	1.00000
Principal components for the farm type dataset			
PC1	1.3587	0.6154	0.6154
PC2	0.9297	0.2881	0.09649
PC3	0.53802	0.09649	1.00000

The eigenvalues of the covariance matrix represent the variance of the principal components. To visually compare the eigenvalues, we constructed scree plots. Figure 1 shows a scree plot of the eigenvalues for the four age groups. The curve is asymptotic regarding the horizontal axis, as can be seen from the last two

values; this means that the last two components are not very different between the four age groups and can therefore be neglected. A scree plot of eigenvalues for the three types of farms is shown in Figure 2. The sharp angle on the third component shows that it is possible to discard it. Corresponding eigenvector loadings for both datasets are given in Table 5. From the PCA graph in Figure 3, we can see clusters of pathogens that reflect their similarity.

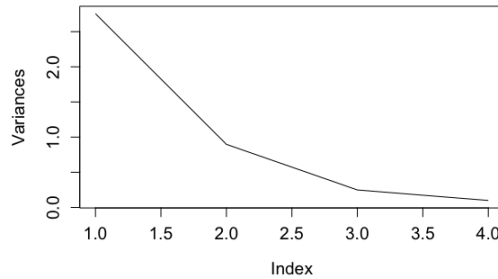


Figure 1. Scree plot of eigenvalues for the age group dataset

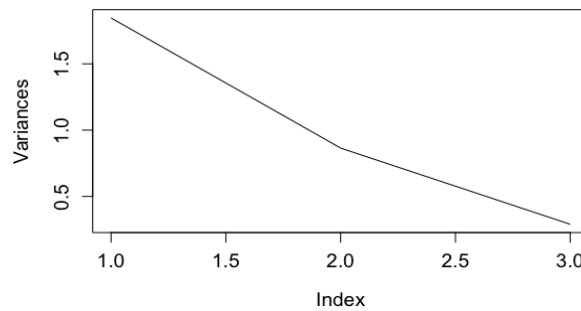


Figure 2. Scree plot of eigenvalues for the farm type dataset

Table 5. Eigenvector loadings for the principal components

	PC1	PC2
Age group dataset		
Age 1	-0.5749045	0.1691971
Age 2	-0.5524039	0.1163859
Age 3	-0.5554445	0.1221487
Age 4	-0.2362544	-0.9710337
Farm type dataset		
Large farms	-0.3792410	-0.9200110
Medium-sized farms	-0.6649462	0.1967453
Small farms	-0.6434460	0.3389264

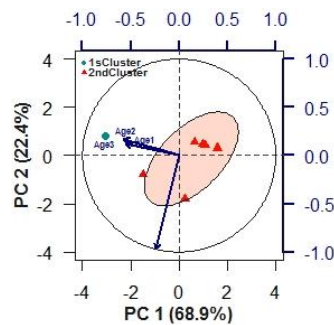


Figure 3. Clustering of the indicators of enteropathogen infection for the age group dataset

The loading plots reflect how strongly each characteristic affects a particular principal component. The projections of the eigenvectors onto the components indicate that the variances in the age 1, age 2, age 3 groups are explained by PC1, while that of the age 4 group is almost completely explained by PC2. For the farm type dataset, the variance in the infection level for small and medium-sized farms is explained by PC1. Table 6 shows the coordinates on the new plane, based on the first two principal components of the age group dataset. The new plane based on the first two principal components for the farm type dataset is represented in Table 7. The coordinates on the new plane for the farm type dataset are shown in Figure 4.

Table 6. New coordinates for the age group dataset

Pathogens and their combinations	PC 1	PC2
<i>C. parvum</i>	-3.0373762	0.8278567
<i>BRV</i>	-1.4869840	-0.7897851
<i>BCV</i>	0.2380130	-1.8053526
<i>E. coli</i> K99	1.0840257	0.4303557
<i>C. parvum</i> + <i>BRV</i>	0.6436795	0.5629286
<i>C.parvum</i> + <i>BCV</i>	0.9837092	0.4702843
<i>BRV</i> + <i>BCV</i>	1.5749328	0.3037124

Table 7. New coordinates for the farm type dataset

Pathogens and their combinations	PC 1	PC 2
<i>C. parvum</i>	-1.06988314	-1.88476492
<i>BRV</i>	-2.56433015	0.90603164
<i>BCV</i>	0.66017670	-0.20277231
<i>E. coli</i> K99	0.05406532	0.82346674
<i>C. parvum</i> + <i>BRV</i>	0.75825677	0.01677719
<i>C.parvum</i> + <i>BCV</i>	1.02590002	0.03730857
<i>BRV</i> + <i>BCV</i>	1.13581449	0.30395308

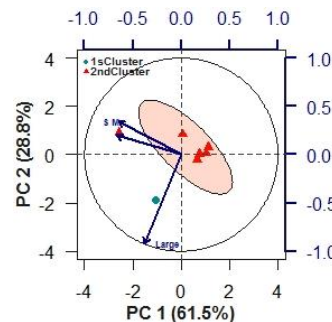


Figure 4. Clustering of the indicators of enteropathogen infection for the farm type dataset

SVD has been applied as part of numerous related methods for dimensionality reduction, to obtain reduced order models that can be used to describe and visualise the key characteristics of large amounts of data for a huge range of scientific and practical problems. These techniques include PCA in the area of statistics [18], the Karhunen-Loève transform [19], empirical orthogonal functions for climatic investigations [20], proper orthogonal decomposition in the dynamics of fluids [21], and canonical correlation analysis [22]. Although these methods were developed independently in different sciences, many of them vary only in terms of the type of data collection and how the data are processed [23]. PCA is the most widely applied technique and has been used in many branches of science and types of analytical procedure. It reduces the dimensionality of the data by extracting the eigenvalues and eigenvectors of the correlation matrix and generating linear combinations of vectors for the variables, from which one can select the component that represents significant or low volatility according to the studied problem.

This study presents a pioneering PCA analysis of epidemiological data on the distribution of infectious diarrhoea pathogens (*C. parvum*, *BRV*, *BCV*, and *E. coli* K99) among newborn calves with diarrhoea. The aim was to identify the most significant infectious agents of the disease, which in recent years has been a significant global economic problem in dairy herds, including those in Northern Kazakhstan [24]–[26].

Field studies showed that under the conditions in the north of the country, these pathogens could be found in 70.6% of dairy farms. Infection with cryptosporidia was the greatest in calves that were eight to 14

days old. Among the other enteropathogens, rotaviruses were also more common in two-week-old animals, in the same way as for cryptosporidia, whereas coronaviruses were observed mainly among calves aged up to seven days. A significant increase in the prevalence of *E. coli* K99 was established in the second and third weeks of a calf's life. It has been shown that there is a strong and very significant relationship between *C. parvum* infection and the occurrence of diarrhoea [27]. When the epidemiological data for our age group dataset were used in a PCA, it was revealed that the first eigenvalue for one-week-old calves (PC1) had a relatively high proportion of 0.6888, thus explaining 68.88% of the variability, while the next highest eigenvalue for two-week-old calves (PC2) was 22.43%. Together, these PCs explain 91.31% of the information in the dataset.

Our epidemiological research also demonstrates that distribution of diarrhoea pathogens and their combinations among neonatal calves in the north of Kazakhstan depend on farm size. *C. parvum* and *BCV* were the most common infections in large industrialised dairy entities. Among new-born animals on the medium-sized and householders' farms, the prevalence of *BRV* was higher than for large enterprises and *E. coli* was found mainly in medium-sized entities. These results are aligned with information from similar surveys in other countries [28], [29]. A PCA simulation of our farm type dataset showed that the first two components for large and medium-sized farms explained 90.3% of the variance in the data. For both datasets, it is sufficient to choose the first and second components for further analysis. This means that we can define the main epidemiological factors and retain the number of PCs whose total corresponding eigenvalues match with the risk of diarrhoea.

In the next step of our study, we carried out a cluster analysis over PCA. The main advantage of this simulation is the creation of mutually exclusive groups that can easily be used in analysis. For the age group dataset, two clusters were formed on the new plane. The coordinate corresponding to *C. parvum* was in the direction of the age 1, age 2, and age 3 eigenvectors, while the eigenvector corresponding to the Age 4 group had a negative correlation. The coordinates of *BRV* and *BCV* were located almost on the border of the second cluster and were approximately in the direction of the Age 4 eigenvector. The coordinates for the rest (*E. coli* K99, *C. parvum* + *BRV*, *C. parvum* + *BCV*, and *BRV* + *BCV*) were close to each other, and had a negative correlation with all age groups. The clustering obtained in this way allows us to suggest that *C. parvum* can be considered the main contagious pathogen of diarrhoea among calves between one and three weeks of age. On the new plane for the farm type dataset, which showed two clusters, the *C. parvum* coordinate was located almost in the direction of the large farm eigenvector. The *BRV* coordinate was in the direction of the eigenvectors for the medium-sized and small farms. The remaining coordinates of *BCV*, *E. coli* K99, *C. parvum* + *BRV*, *C. parvum* + *BCV*, and *BRV* + *BCV* were in the second cluster, and were not correlated with the types of farms. These results for the directions of the eigenvectors supported the hypothesis that the predominant infectious agent for diarrhoea in calves in large dairy enterprises is *C. parvum*, while in medium-sized and small farms it is *BRV*. The other agents and their combinations as etiologic factors in diarrhoea we assess as being concomitant secondary pathogenic microorganisms.

## 5. CONCLUSION

PCA of diarrhoeal infection in calves on cattle farms in Northern Kazakhstan has been presented that demonstrates the power of this tool for the preliminary analysis of epidemiological data. Although the analysis of the correlation matrix for a multivariate dataset of this sort is quite complex, the use of PCA makes it possible to divide the data into simple components that can be easily interpreted. In addition, deep internal structures can be discovered in terms of the data and the relationships between them. Our results indicate that PCA can be potentially useful in the clustering of epidemiological datasets, and further research in public health and veterinary sciences is needed to apply this technique to the analysis of other infectious diseases with a multi-pathogenic nature.

## ACKNOWLEDGEMENTS

The study formed part of a research project funded by the Ministry of Education and Science of the Republic of Kazakhstan (AP05135550) in 2019–2020.




## REFERENCES

- [1] J. N. Kutz, *Data-driven modeling & scientific computation: Methods for complex systems & big data*, 1 st., no. Book Review 1. Oxford University Press, 2015.
- [2] G. H. Golub and C. F. V. Loan, *Matrix Computations*. JHU Press, 2013.
- [3] M. T. Heath, A. J. Laub, C. C. Paige, and R. C. Ward, "Computing the singular value decomposition of a product of two matrices," *SIAM Journal on Scientific and Statistical Computing*, vol. 7, no. 4, pp. 1147–1159, Oct. 1986, doi: 10.1137/0907078.
- [4] V. C. Klema and A. J. Laub, "The singular value decomposition: its computation and some applications," *IEEE Transactions on Automatic Control*, vol. 25, no. 2, pp. 164–176, Apr. 1980, doi: 10.1109/TAC.1980.1102314.

- [5] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, Nov. 1901, doi: 10.1080/14786440109462720.
- [6] W. Liebermeister, "Linear modes of gene expression determined by independent component analysis," *Bioinformatics*, vol. 18, no. 1, pp. 51–60, Jan. 2002, doi: 10.1093/bioinformatics/18.1.51.
- [7] S. Hussain, N. Z. Quazilbash, S. Bai, and S. Khoja, "Reduction of variables for predicting breast cancer survivability using principal component analysis," in *Proceedings-IEEE Symposium on Computer-Based Medical Systems*, Jun. 2015, vol. 2015-July, pp. 131–134, doi: 10.1109/CBMS.2015.62.
- [8] S. Salvatore, J. Roislien, J. A. B. Lomba, and J. G. Bramness, "Assessing prescription drug abuse using functional principal component analysis (FPCA) of wastewater data," *Pharmacoepidemiology and Drug Safety*, vol. 26, no. 3, pp. 320–326, Mar. 2017, doi: 10.1002/pds.4127.
- [9] I. Bakolis, P. Burney, and R. Hooper, "Principal components analysis of diet and alternatives for identifying the combination of foods that are associated with the risk of disease: A simulation study," *British Journal of Nutrition*, vol. 112, no. 1, pp. 61–69, Jul. 2014, doi: 10.1017/S0007114514000221.
- [10] A. M. K. Virtala, G. D. Mechor, Y. T. Gröhn, and H. N. Erb, "Morbidity from nonrespiratory diseases and mortality in dairy heifers during the first three months of life," *Journal of the American Veterinary Medical Association*, vol. 208, no. 12, pp. 2043–2046, 1996.
- [11] C. Svensson, K. Lundborg, U. Emanuelson, and S. O. Olsson, "Morbidity in Swedish dairy calves from birth to 90 days of age and individual calf-level risk factors for infectious diseases," *Preventive Veterinary Medicine*, vol. 58, no. 3–4, pp. 179–197, May 2003, doi: 10.1016/S0167-5877(03)00046-1.
- [12] I. Lorenz, J. Fagan, and S. J. More, "Calf health from birth to weaning. II. Management of diarrhoea in pre-weaned calves," *Irish Veterinary Journal*, vol. 64, no. 1, p. 9, Dec. 2011, doi: 10.1186/2046-0481-64-9.
- [13] J. A. Mawly, A. Grinberg, D. Prattley, J. Moffat, and N. French, "Prevalence of endemic enteropathogens of calves in New Zealand dairy farms," *New Zealand Veterinary Journal*, vol. 63, no. 3, pp. 147–152, May 2015, doi: 10.1080/00480169.2014.966168.
- [14] C. J. M. Bartels, M. Holzhauer, R. Jorritsma, W. A. J. M. Swart, and T. J. G. M. Lam, "Prevalence, prediction and risk factors of enteropathogens in normal and non-normal faeces of young Dutch dairy calves," *Preventive Veterinary Medicine*, vol. 93, no. 2–3, pp. 162–169, Feb. 2010, doi: 10.1016/j.prevetmed.2009.09.020.
- [15] M. M. Izzo, P. D. Kirkland, V. L. Mohler, N. R. Perkins, A. A. Gunn, and J. K. House, "Prevalence of major enteric pathogens in Australian dairy calves with diarrhoea," *Australian Veterinary Journal*, vol. 89, no. 5, pp. 167–173, May 2011, doi: 10.1111/j.1751-0813.2011.00692.x.
- [16] Y. Millemann, "Diagnosis of neonatal calf diarrhoea," *Revue de Medecine Veterinaire*, vol. 160, no. 8–9, pp. 404–409, 2009.
- [17] J. Heine, "Eine einfache Nachweismethode für Kryptosporidien im Kot," *Zentralblatt für Veterinärmedizin Reihe B*, vol. 29, no. 4, pp. 324–327, May 1982, doi: 10.1111/j.1439-0450.1982.tb01233.x.
- [18] I. Jolliffe, "Principal component analysis," in *Encyclopedia of statistics in behavioral science*, 2005.
- [19] M. Loeve, *Probability theory*. London: Courier Dover Publications, 1995.
- [20] E. N. Lorenz, "Empirical orthogonal functions and statistical weather prediction," in *Technical report Statistical Forecast Project Report 1 Department of Meteorology MIT 49*, vol. 1, 1956.
- [21] P. Holmes, J. L. Lumley, G. Berkooz, and C. W. Rowley, *Turbulence, coherent structures, dynamical systems and symmetry (Cambridge monographs on mechanics)*. England, 2012.
- [22] S. Chery, "Singular value decomposition analysis and canonical correlation analysis," *Journal of Climate*, vol. 9, no. 9, pp. 2003–2009, Sep. 1996, doi: 10.1175/1520-0442(1996)009<2003:SVDAAC>2.0.CO;2.
- [23] Z. Bai, A. Knyazev, and H. A. V. D. Vorst, *Linear algebra and its applications: Preface*, vol. 415, no. 1. New Delhi: Brooks/Cole, 2006.
- [24] A. D. Kruijff, R. Mansfeld, and M. Hoedemaker, *Tierärztliche Bestandesbetreuung beim Milchrind*, 2nd ed., vol. 149, no. 9. Stuttgart, 2007.
- [25] D. R. Snodgrass, H. R. Terzolo, D. Sherwood, I. Campbell, J. D. Menzies, and B. A. Syngé, "Aetiology of diarrhoea in young calves," *The Veterinary record*, vol. 119, no. 2, pp. 31–34, Jul. 1986, doi: 10.1136/vr.119.2.31.
- [26] A. Ussenbayev, D. Kurenkeyeva, C. Bauer, and A. Kadyrov, "Prevalence of Calves' Cryptosporidiosis in Northern Kazakhstan," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12253 LNCS, 2020, pp. 718–726.
- [27] L. P. Garber, M. D. Salman, H. S. Hurd, T. Keefe, and J. L. Schlater, "Potential risk factors for cryptosporidium infection in dairy calves," *Journal of the American Veterinary Medical Association*, vol. 205, no. 1, pp. 86–91, 1994.
- [28] R. D. L. Fuente *et al.*, "Cryptosporidium and concurrent infections with other major enteropathogens in 1 to 30-day-old diarrheic dairy calves in central Spain," *Veterinary Parasitology*, vol. 80, no. 3, pp. 179–185, Jan. 1999, doi: 10.1016/S0304-4017(98)00218-0.
- [29] J. Al Mawly, A. Grinberg, D. Prattley, J. Moffat, J. Marshall, and N. French, "Risk factors for neonatal calf diarrhoea and enteropathogen shedding in New Zealand dairy farms," *Veterinary Journal*, vol. 203, no. 2, pp. 155–160, Feb. 2015, doi: 10.1016/j.tvjl.2015.01.010.




## BIOGRAPHIES OF AUTHORS






**Abilaikhan Kadyrov**    he graduated from Seifullin Kazakh Agrotechnical University, Nur-Sultan, Kazakhstan, with Bachelor's Degree in Electrical Power Engineering in 2014, received MS degree in Technical Sciences at Bauman State Technical University, Moscow, Russia, in 2016, he is Ph.D. student at the Department of Computer and Software in L. N. Gumilyov Eurasian National University, Kazakhstan. He has published more than 15 journal papers in the fields of computer sciences, modelling of epidemiological processes. He can be contacted at email: kadyrov.abilaikhan@gmail.com.








**Altay Ussenbayev**    he is a Ph.D. in Epidemiology, Associate Professor at the Veterinary Medicine Department, Faculty of Veterinary Medicine and Animal Husbandry, Seifullin Kazakh Agrotechnical University, Kazakhstan. His current research interests include modelling the epidemiology of infectious diseases in humans and animals, implementation of IT in veterinary sciences and education. He has published more than 180 journal papers in the fields of modelling the dynamics of infections in veterinary parasitology, epidemiology, food safety. He can be contacted at email: altay\_us@mail.ru.






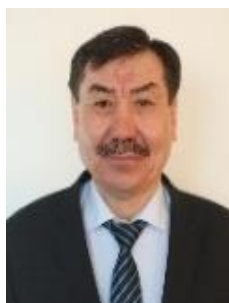
**Dariyash Kurenkeyeva**    she graduated from Al-Farabi Kazakh National University, Almaty, Applied Mathematics, in 1987, received Ph.D. degree at Computer Sciences Department of Dmitry Mendeleev University of Chemical Technology of Russia, Moscow, in 1993. She is an Associate Professor in Mathematics. She can be contacted at email: dariyash.kurenkeyeva@gmail.com.






**Berdaly Kurenkey**    he was graduated from Al-Farabi Kazakh National University, Almaty, Physics, in 1985, received Ph.D. in Physics degree at Ioffe Physical-Technical Institute of the Russian Academy of Science, Saint-Petersburg, Russia. Author of the book “Quantum Mechanics”, 2014. He can be contacted at email: berdaly.kurenkey@gmail.com.



**Sarsenbay Abdrakhmanov**    he is a Doctor of Veterinary Sciences, Professor. He is currently Dean of the Faculty of Veterinary Medicine and Animal Husbandry, Seifullin Kazakh Agrotechnical University, Kazakhstan. He has authored or coauthored more than 200 refereed journal and conference papers, 10 books edited with Ministry of Education and Science of Kazakhstan. His research interests include the modelling the epidemiology and control of zoonotic diseases, application machine learning for analysis the big epidemiological data. He can be contacted at email: s.abdrakhmanov@mail.ru.



**Nurlan Tashatov**    he is a Candidate of Physical and Mathematical Sciences, Associate Professor, L. N. Gumilyov Eurasian National University, Faculty of Information Technology, department of computer and software engineering. He teaches courses on information theory and coding. He can be contacted at email: tash.nur@mail.ru.