Data Pre-Processing for Ecosystem Behavior Analysis

NATALYA REZOVA, LEV KAZAKOVTSEV, GUZEL SHKABERINA, DENIS DEMIDKO, ANDREY GOROSHKO

The most common data quality problems*

- incompleteness: the data does not contain attributes, or values are missing;
- noise: data contains erroneous records or outliers;
- inconsistency: data contains conflicting records or discrepancies.

*J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques. Third Edition, Morgan Kaufmann Publishers, 2011

The most common data preprocessing methods

- processing of missing values;
- data normalization;
- data discretization;
- dimensionality reduction;
- cleaning text fields.

Previous studies

1. J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques. Third Edition, Morgan Kaufmann Publishers, 2011.
2. A. Jain, R. Dubes, Algorithms for Clustering Data. Prentice-Hall: New Jersey, USA, 1988, P. 320.
3. D. Chicco, “Ten quick tips for machine learning in computational biology” in BioData Mining, vol. 10(35), 2017.
4. Sh Wu., “A review on coarse warranty data and analysis”, in Reliability Engineering & System Safety, vol. 114, pp. 1-11.
5. S. García, S. Ramírez-Gallego, J. Luengo, “Big data preprocessing: methods and prospects” in Big Data Analysis, vol. 1(9), 2013.
6. H.J. Jeong, K.S. Park, Y.G. Ha, “Image Preprocessing for Efficient Training of YOLO Deep Learning Networks”, IEEE International Conference on Big Data and Smart Computing (BigComp), 2018, pp. 635-637.
7. H.C. Lu, E.W. Loh, S.C. Huang, “The Classification of Mammogram Using Convolutional Neural Network with Specific Image Preprocessing for Breast Cancer Detection”, 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), 2019, pp. 9-12.
8. D. Tuia, B. Kellenberger, S. Beery et al. “Perspectives in machine learning for wildlife conservation” in Nature Communications, vol. 13(792), 2022.

The main problem

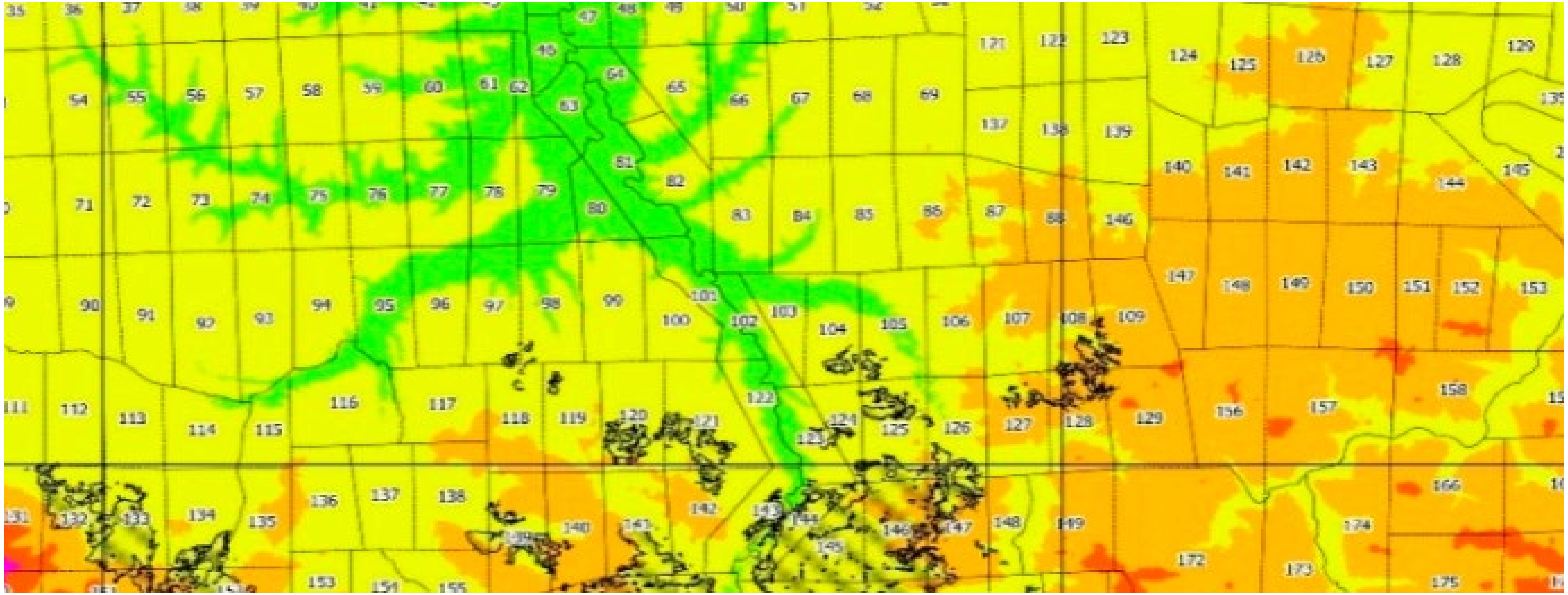
- Development an ecosystem of predictive analytics for the occurrence of outbreaks of mass reproduction of the Siberian silk moth

Data preprocessing problem

- Reduction in the number of input characteristics without loss of forecast accuracy

Data preprocessing problem

- n=15523 forest compartments $FS = \{FS_1, FS_2, ..., FS_i, ..., FS_n\}$



Data preprocessing problem

1996	8	13	69	0	0	5П1Е1К2П1Л	7	1	1	0	0	1	0	0	0	0	60	4	13	14	0,9	140	ХВЗМ	ВЛА)
1996	9	8	94	0	0	4П3Е3Л	4	3	3	0	0	0	0	0	0	0	150	3	24	26	0,7	228	ХВЗМ	ВЛА)
1996	9	9	84	0	0	4П2Е3Л1К	4	2	3	0	0	1	0	0	0	0	150	4	22	24	0,6	220	ХВЗМ	ВЛА)
1996	10	7	37	0	0	4Л3П2Е1К	3	2	4	0	0	1	0	0	0	0	220	4	27	40	0,6	240	БРЗМ	СВЕ)
1996	11	7	73	0	0	8П2Е	8	2	0	0	0	0	0	0	0	0	75	3	17	16	0,9	223	ППКТ	СВЕ)
1996	13	5	221	0	0	5П2Л2П1К	7	0	2	0	0	1	0	0	0	0	180	3	24	24	0,8	340	ХВЗМ	ВЛА)
1996	14	3	78	0	0	6П3Е1К	6	3	0	0	0	1	0	0	0	0	160	4	21	22	0,6	190	ХВКТ	ВЛА)
1996	14	4	56	0	0	5Е3П2К	3	5	0	0	0	2	0	0	0	0	150	3	23	24	0,9	350	ХВЗМ	ВЛА)
1996	14	5	128	0	0	5Е2Л2П1К	2	5	2	0	0	1	0	0	0	0	180	3	24	24	0,8	340	ХВЗМ	ВЛА)
1996	15	8	57	0	0	5Е3П2К	3	5	0	0	0	2	0	0	0	0	170	4	21	22	0,8	110	ХВЗМ	ВЛА)
1996	15	11	24	0	0	8Е1К1ОС	0	8	0	0	1	1	0	0	0	0	160	3	23	24	0,8	310	ХВЗМ	ВЛА)
1996	15	20	208	0	0	4П3Е2Л1К	4	3	2	0	0	1	0	0	0	0	160	3	23	24	0,8	230	ХВКТ	ВЛА)
1996	19	11	192	0	0	4Е3Л1К2Б	0	4	3	2	0	1	0	0	0	0	100	4	18	18	0,8	230	ХВЗМ	ВЛА)
1996	19	16	40	0	0	3Е2П2Л1К2Б	2	3	2	2	0	1	0	0	0	0	100	4	18	18	0,8	230	ХВЗМ	СВЕ)
1996	21	7	57	0	0	4Е2П2ОС1К1Л	2	4	1	0	2	1	0	0	0	0	150	4	22	24	0,7	290	ХВЗМ	СВЕ)
1996	21	11	9	0	0	5Е2П2ОС1Л	2	5	1	0	2	0	0	0	0	0	150	4	22	22	0,7	270	ХВЗМ	СЫР.
1996	22	1	91	0	0	3Е3П2ОС1Л1К	3	3	1	0	2	1	0	0	0	0	155	4	22	24	0,7	290	ОСРТ	ВЛА)
1996	22	2	110	0	0	3Е2П1Л4ОС	2	3	1	0	4	0	0	0	0	0	160	4	22	24	0,7	290	ХВЗМ	ВЛА)
1996	22	5	29	0	0	5Е2П2ОС1К	2	5	0	0	2	1	0	0	0	0	160	4	22	24	0,7	270	ХВЗМ	ВЛА)
1996	23	3	65	0	0	2Е2П2Л1С3ОС	2	2	2	0	3	0	1	0	0	0	160	4	22	24	0,7	290	ХВЗМ	ВЛА)
1996	24	4	294	0	0	4Е2П1К1Е1Л1ОС	2	5		0	1	1	0	0	0	0	140	4	22	26	0,7	290	ХВЗМ	СВЕ)
1996	24	19	171	0	0	4Л2К2Е2П	2	2	4	0	0	2	0	0	0	0	200	3	23	26	0,6	220	БРЗМ	СВЕ)
1996	24	21	58	0	0	3Е3П2Л1К1Б	3	3	2	1	0	1	0	0	0	0	170	4	22	24	0,7	270	ХВЗМ	ВЛА)
1996	37	8	27	0	0	5П2Е3Б	5	2	0	3	0	0	0	0	0	0	80	4	16	16	0,8	180	ХВЗМ	ВЛА)
1996	38	17	16	0	0	2Е2П2Л4Б	2	2	2	4	0	0	0	0	0	0	60	2	18	18	0,7	190	ОСРТ	ВЛА)
1996	39	18	11	0	0	3П2Е5Б	3	2	0	5	0	0	0	0	0	0	60	3	15	14	0,6	110	ППКТ	СВЕ)
1996	45	16	20	0	0	7Б1П1Е1К	1	1	0	7	0	1	0	0	0	0	60	3	17	16	0,9	180	ХВЗМ	СЫР.
1996	41	2	107	0	0	6Е2П1К1Л	2	6	1	0	0	1	0	0	0	0	60	3	23	28	0,8	300	ХВЗМ	ВЛА)
1996	42	2	121	0	0	3Е2П2Л1К2Б	2	3	2	2	0	1	0	0	0	0	150	3	23	24	0,7	270	ХВЗМ	ВЛА)
1996	43	1	133	0	0	3Е2П2Л1К2Б	2	3	2	2	0	1	0	0	0	0	150	3	23	24	0,7	270	ХВЗМ	ВЛА)
1996	44	1	524	0	0	6Е2К2Л	0	6	2	0	0	2	0	0	0	0	150	3	24	26	0,7	290	ХВЗМ	ВЛА)
1996	44	7	73	0	0	4П3Л2Е1К	4	2	3	0	0	1	0	0	0	0	160	3	23	24	0,8	280	ППКТ	СВЕ)
1996	45	1	65	0	0	8Е2К	0	8	0	0	0	2	0	0	0	0	150	3	23	24	0,6	230	ХВЗМ	ВЛА)
1996	45	4	120	0	0	4Е3П1К1Л1Б	3	4	1	1	0	1	0	0	0	0	140	4	22	24	0,7	250	ХВЗМ	ВЛА)
1996	45	11	136	0	0	7П2Е1К	7	2	0	0	0	1	0	0	0	0	160	4	21	24	0,4	120	ХВЗМ	ВЛА)
1996	45	12	41	0	0	5Е4П1Л	4	5	1	0	0	0	0	0	0	0	160	3	23	24	0,8	310	ХВЗМ	ВЛА)
1996	46	1	21	0	0	4Е3П1К1Л1Б	3	4	1	1	0	1	0	0	0	0	140	4	22	24	0,7	250	ХВЗМ	ВЛА)
1996	46	3	17	0	0	4Е1П1К4Б	1	4	0	4	0	1	0	0	0	0	100	4	19	18	0,9	260	ХВЗМ	ВЛА)
1996	46	8	30	0	0	4Е1П1К4Б	1	4	0	4	0	1	0	0	0	0	100	4	19	18	0,9	260	ХВЗМ	ВЛА)
1996	48	2	303	0	0	3Е2П1Л4Б	2	3	1	4	0	0	0	0	0	0	100	4	17	16	0,7	150	ХВЗМ	ВЛА)
1996	49	21	223	0	0	5Е2Л2П1К	2	5	2	0	0	1	0	0	0	0	160	4	22	28	0,8	310	ХВЗМ	СЫР.

63,5	1
58,6	1
3,2	0
3,9	0
480,7	1
7	0
1,9	0
5,2	0
11,3	0
4,6	0
3,5	0
4,4	0
3,6	0
2,5	0
8	0
20,3	0
4,8	0
9,3	0
18	0
0,3	0
0,3	0
3,3	0
3,4	0
65,6	1
19,9	0
21,6	0
15	0
3,7	0
4,9	0
6	0
1,4	0
1,9	0
0,3	0
3,3	0
4	0
4	0
3,3	0
0,6	0
2,9	0
1,4	0

Taxation characteristics

- T_1^i - area of the *i*th forest compartment, hectare;
- T_2^i - slope exposure of the *i*th forest compartment (qualitative characteristics were replaced with quantitative ones, eastern slope - 1, western - 2, northern - 3, southern - 4, northeastern - 5, northwestern - 6, southeastern - 7, southwestern - 8);
- stand composition – proportion of fir (T_3^i), spruce (T_4^i), larch (T_5^i), birch (T_6^i), aspen (T_7^i), Siberian pine (T_8^i) and pine (T_9^i) among stands of *i*th forest compartment $(\sum_{j=3}^{9} T_j^i = 1, i = 1, \ldots, n)$
- T_{10}^i - average age of stands of *i*th forest compartment, years;
- T_{11}^i – growth class of the *i*th forest compartment;
- T_{12}^i – average height of the *i*th forest compartment, m;
- T_{13}^i – average diameter of the *i*th forest compartment, cm;
- T_{14}^i – relative completeness of the forest stand of the *i*th forest compartment;
- T_{15}^i - stock of stands of the *i*th forest compartment, m^3/hectare;
- T_{16}^i - soil moisture of the *i*th forest compartment;
- T_{17}^i – mossiness of the *i*th forest compartment.

Bioclimatic characteristics*

- $C1_{tj}^{i}$, where $t = 1, 2, \ldots, T$, $j = 6, 7$ - soil temperature (0 - 10 cm underground) for June and July of the *i*th year, K;
- $C2_{t}^{i}$, where $t = 1, 2, \ldots, T$ - maximum soil temperature (0 - 10 cm underground) in the winter months of the *i*th year, K;
- $C3_{t}^{i}$, where $t = 1, 2, \ldots, T$ - maximum snow depth of the *i*th year, m;
- $C4_{tj}^{i}$, where $t = 1, 2, \ldots, T$, $j = 5, 6, \ldots, 10$ - near surface air temperature for the period May - October of the *i*th year, K;
- $C5_{tj}^{i}$, where $t = 1, 2, \ldots, T$, $j = 5, 6, \ldots, 9$ – total evaporation and transpiration for the period May-September of the *i*th year, kg/(m^3*s);
- $C6_{tj}^{i}$, where $t = 1, 2, \ldots, T$, $j = 5, 6, \ldots, 9$ – rainfall flux for the period May-September of the *i*th year, kg/(m^3*s);
- $C7_{tj}^{i}$, where $t = 1, 2, \ldots, T$, $j = 1, 2, \ldots, 12$ – soil moisture (0 - 10 cm underground) for the *j*th months of the *i*th year, %.

*The global climate database Land Data Assimilation System (FLDAS)

Data preprocessing problem

- n=15523 forest compartments $FS = \{FS_1, FS_2, \dots, FS_i, \dots, FS_n\}$
- $FS_i = \{T_1^i, \dots, T_{17}^i, C1_{16}^i, C1_{17}^i, \dots, C1_{76}^i, C1_{77}^i, C2_1^i, \dots, C2_7^i, C3_1^i, \dots, C3_7^i, C4_{15}^i, \dots, C4_{110}^i, \dots, C4_{75}^i, \dots, C4_{710}^i, C5_{15}^i, \dots, C5_{19}^i, \dots, C5_{75}^i, \dots, C5_{79}^i, C6_{15}^i, \dots, C6_{19}^i, \dots, C6_{75}^i, \dots, C6_{79}^i, C7_{11}^i, \dots, C7_{112}^i, \dots, C7_{71}^i, \dots, C7_{712}^i, Y_i\}$

17 input taxation characteristics, 217 bioclimatic characteristics

- $$Y_i = \begin{cases} 1, Y1_i \geq K \text{ or } Y2_i > 0 \\ 0, \text{otherwise} \end{cases}$$

- *K* is set by experts. For the considered ecosystem *K* = 25

Data classification problem for ecosystem behavior analysis

There is a dataset FS with dimension n = 15523 containing taxation and bioclimatic characteristics of forest compartments. One of the characteristics (Yi) determines the class of the object (presence or absence of an outbreak of the Siberian silk moth in a given area) and can take values from a fixed set {0, 1}. Based on the training sample, it is necessary to form a classification tree (decision tree) containing a set of logical conditions that allow for an arbitrary measurement FS_i from FS to indicate the quality class to which it may belong.

Set of forest compartments $FS = \{FS_1, FS_2, ..., FS_i, ..., FS_n\}$ → 2 groups → quality class 0 / quality class 1

Computational Experiments

Pre-classification accuracy

Method	LR	KNN	RF	DT
Accuracy	0.523	0.631	0.913	0.949

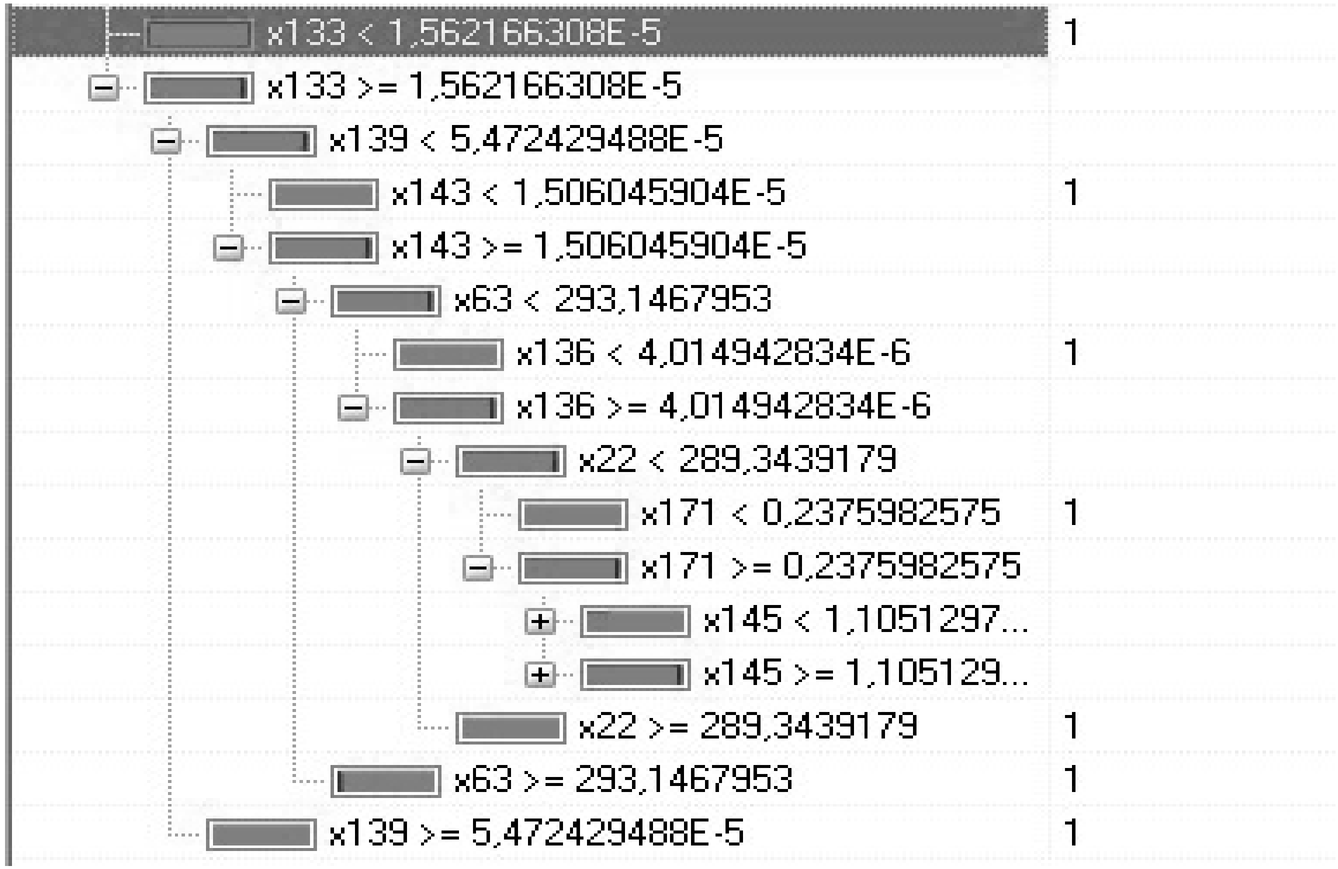
LR - logistic regression;

KNN - K-nearest neighbors;

RF - random forest;

DT - decision tree.

The result of building a decision tree in Deductor



Computational Experiments

№	The proportion of correctly classified outbreaks	№	The proportion of correctly classified outbreaks
1	0.97887	26	0.97917
2	0.97936	27	0.97862
3	0.97949	28	0.97836
4	0.97904	29	0.97942
…	…	30	0.97962

Input characteristics, the significance of which more than 1%

Feature	Significance, %		Feature	Significance, %	
	Max value	Min value		Max value	Min value
$C1_{16}$	4.244	1.706	$C6_{26}$	8.052	7.607
$C1_{17}$	3.17	3.069	$C6_{35}$	6.096	0.408
$C1_{76}$	2.175	0.14	$C6_{39}$	8.831	5.413
$C1_{77}$	3.703	1.401	$C6_{47}$	3.547	2.039
$C4_{57}$	10.413	3.205	$C6_{58}$	8.525	0.256
$C4_{67}$	6.129	6.129	$C6_{59}$	29.145	27.403
$C4_{75}$	2.305	2.274	$C6_{69}$	5.796	0.875
$C5_{35}$	2.478	0.663	$C6_{75}$	2.518	0.251
$C5_{38}$	5.994	5.994	$C7_{47}$	2.865	2.398
$C5_{47}$	3.975	1.802	$C7_{48}$	1.009	1.009
$C5_{56}$	9.688	1.146	$C7_{53}$	2.456	2.456
$C5_{58}$	7.062	6.341	$C7_{58}$	4.715	1.876
$C5_{65}$	1.083	0.266	$C7_{64}$	3.239	2.84
$C5_{66}$	2.019	2.019	$C7_{78}$	1.749	1.747
$C6_{16}$	7.325	0.38	T_8	1.048	0.087

Data preprocessing problem

- n=15523 forest compartments $FS = \{FS_1, FS_2, ..., FS_i, ..., FS_n\}$

$T_1, T_3, T_4, T_5, T_6, T_7, T_8, T_9, T_{10}, T_{12}, T_{13}, T_{14}, T_{15}, C1_{tj}$ (t = 1, 2, ..., T, j = 6, 7), $C4_{tj}$ (t = 5, 6, T, j = 5, 6, 7, 8), $C5_{tj}$ (t = 2, ..., 6, j = 5, 6, 7, 8), $C6_t$ (t = 1, 2, ..., T, j = 5, 6, ..., 9), $C7_{tj}^{i}$ (t = 1, 2, ..., T, j = 3, 4, ..., 10)

13 input taxation characteristics, 137 bioclimatic characteristics

- $$Y_i = \begin{cases} 1, Y1_i \geq K \; or \; Y2_i > 0 \\ 0, \text{otherwise} \end{cases}$$

- K is set by experts. For the considered ecosystem K = 25

Computational Experiments

№	The proportion of correctly classified outbreaks
1	0.99777
2	0.99736
3	0.99849
…	…
28	0.99836
29	0.99842
30	0.99762

Conclusions

The work presents a method for preliminary processing of taxation and climatic characteristics of ecosystems. The application of this method made it possible to identify significant factors in the development of these ecosystems and to remove from consideration a large number of characteristics that were identified by experts as significant, but during the experiment did not have any effect on the classification accuracy. In addition, the paper considers various types of classification models. The results showed that the decision tree method allows solving the classification problem with high accuracy (0.95-1.00). Based on the classification with the help of models trained on the existing taxation and climatic characteristics of ecosystems, it is possible to further analyze and predict the behavior of these ecosystems.