TOMSK POLYTECHNIC UNIVERSITY

APPLYING DATA MINING TECHNIQUES IN DECISIONS MAKING TASKS

Olga V. Marukhina, Elena E. Mokina
Olga G. Berestneva, Maria D. Shagarova

Barcelona,
October, 10-14, 2016

- On the problem of data visualization reduces the problem of representation in the visual form of the experimental data or the results of theoretical research.
- Possibility of visualization tools are defined areas of its application.
- The main objective of data visualization is the problem of obtaining a visual image, uniquely corresponding data set.

- General format for data is a vector in finite-dimensional space *R*

$$A = (a_0, a_1, a_2, ..., a_{n-1}) \in R_n \qquad \textbf{(1)}$$

- To move from this vector to the visual image will be used a basis of orthonormal functions

$$\{\varphi_i(\tau)\}_{i=0}^{\infty}$$

- As such as this basis can be used are wellknown functions, in particular the orthonormal Legendre polynomials on the interval [0,1]

$$\{l_i(\tau)\}_{i=0}^{\infty}$$

- The point A can be associated with the function

$$F_A(\tau) = \sum_{i=0}^{n-1} a_i l_i(\tau) \qquad \textbf{(2)}$$

- Between (1) and (2) establishes a one-to-one relationship.

- If we introduce the second vector

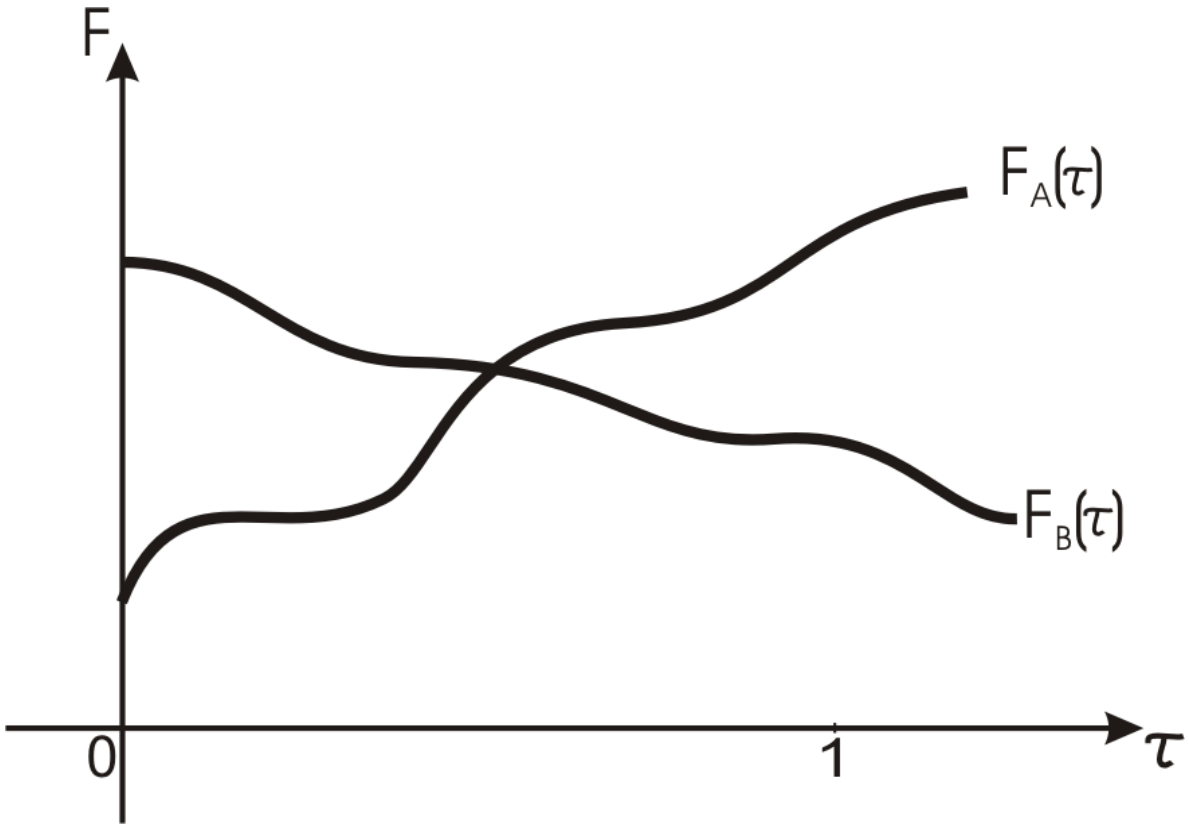
$$B = (b_0, b_1, b_2, \ldots, b_{n-1}) \in R_n$$

then it is assigned to the function

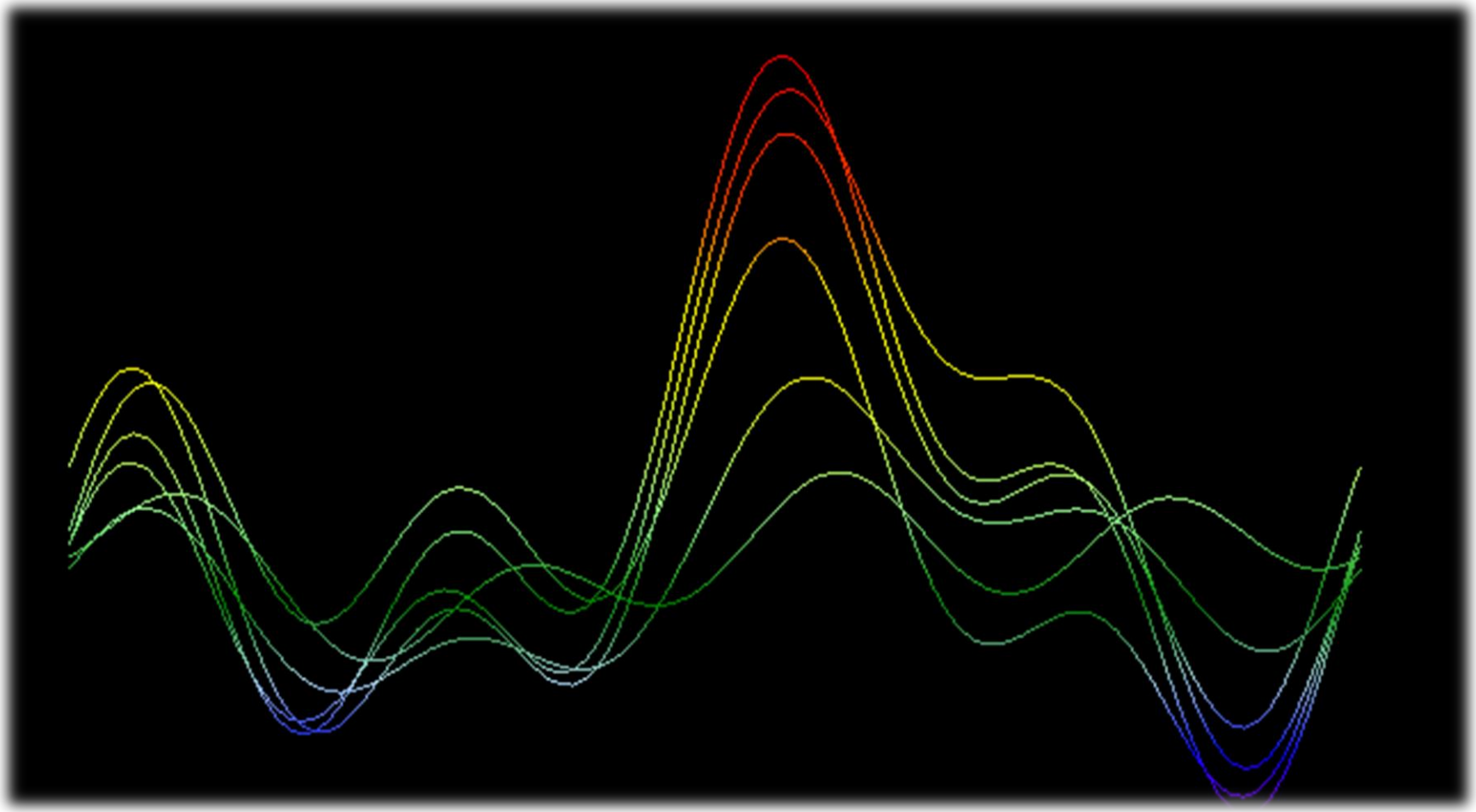
$$F_B(\tau) = \sum_{i=0}^{n-1} b_i l_i(\tau)$$

- Functions $F_A(\tau)$ and $F_B(\tau)$ are visual images of the points A and B, which belong to the space R_n.

Visual images of the points A and B, belonging to the space R_n



- This approach is implemented in a NovoSpark Visualizer (*www.novospark.com*).
- Thus, the basis of imaging (in NovoSpark) is to present the multivariate observations in the form of a two-dimensional image (as curve).
- For close values of the observations (A and B) will meet the similar images (curves). For different values of the observations, their images (curves) will differ markedly.
- *NovoSpark Visualizer* is a modern visualization toolkit; it combines the visual analysis with statistical methods and provides the most complete information about the multi-dimensional data.



Visualization of experimental data in the problem of estimating the parallel version of the test tasks

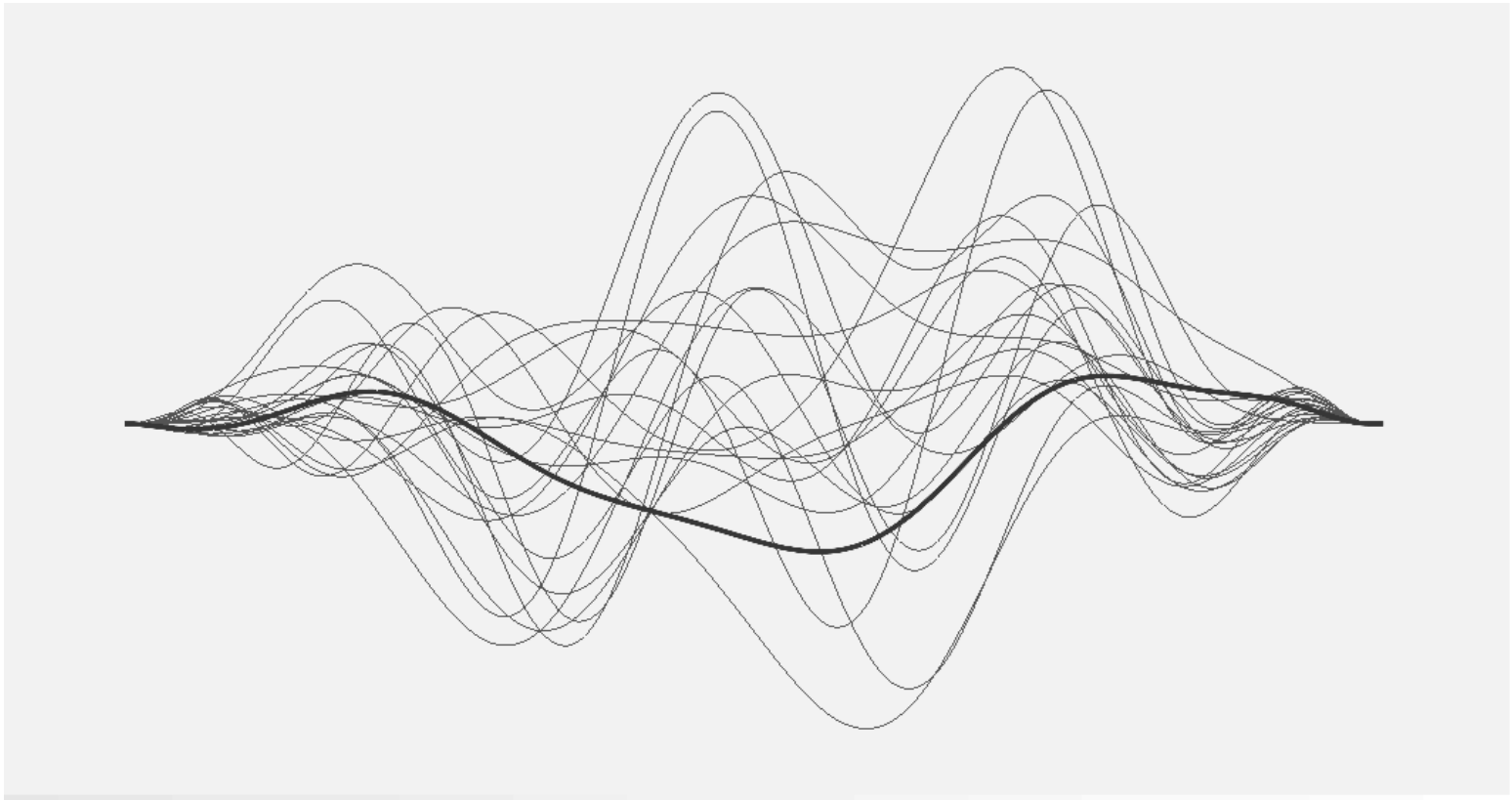
- Independent certification of students is the most objective assessment of their knowledge, the potential of their mental abilities.
- In this regard, in the Tomsk Polytechnic University has developed a system of independent evaluation of the quality of students' knowledge on general subjects.
- There are several versions of one test for each general subjects. For example - mathematics - there are twenty one version of the test tasks.
- As a consequence, there is a problem of similarities ("sameness") between these variants of tests.

- The experimental data is the results of the monitoring in mathematics (for first-year students of Tomsk Polytechnic University).

Таблица 1. Стандартизированные значения трудности заданий теста A_i по вариантам B_j [5]

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
B1	-1,41	-1,53	0,18	0,38	-0,4	-0,21	1,55	0,18	0,38	0,71	0,6
B2	-1,32	0,97	-1,57	0,97	0,62	-0,89	0	-0,69	0,73	0,73	0,73
B3	-0,94	-1,18	-0,31	-0,31	-0,23	-0,39	-0,55	1,54	1,38	2,14	0,45

The initial data set

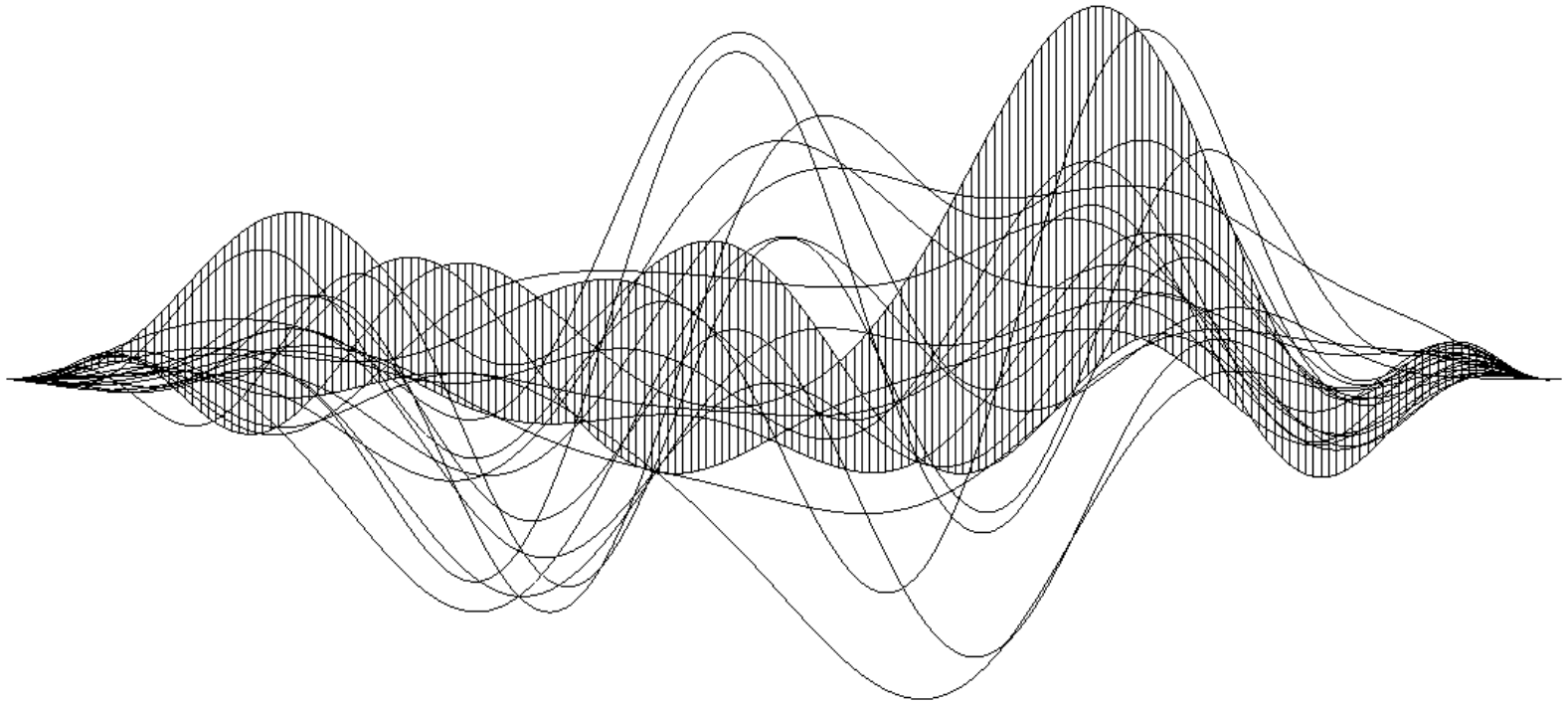


Data.pvp (Data*)

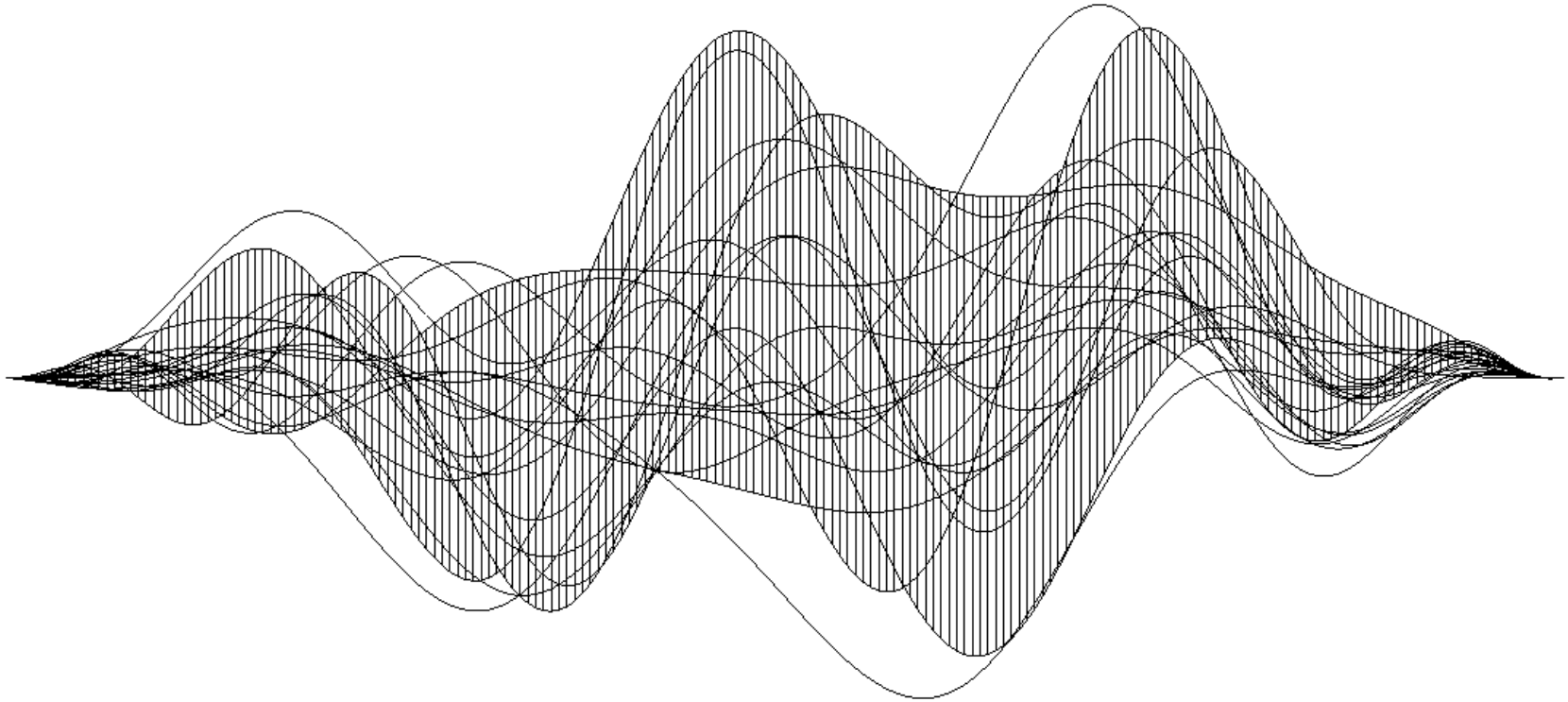
50%

	Строки	Z-Порядок	(z) A1	(z) A2	(z) A3	(z) A4	(z) A5	(z) A6	(z) A7	(z) A8	(z) A9	(z) A10	(z) A11	(i) Name [т]
Стб	Видимый на образе		☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	
iMin			-4,2064	-4,6473	-2,8359	-2,0971	-2,0261	-3,5692	-1,2539	-2,0379	-2,6568	-1,7726	-0,8513	
iMax			2,5083	3,2454	1,9787	2,0018	2,8004	2,2063	2,3977	1,9579	4,3873	3,6221	3,2265	
1	☑	2,4742	0,12	-1,28	-0,64	-0,84	1,63	-0,07	0,52	0,22	0,32	0,41	0,32	B1
2	☑	2,7892	-0,19	0,58	-0,06	0,44	0,18	-1,31	0,31	-1,04	1,8	0,06	1,02	B2
3	☑	2,8242	0,43	-1,34	0,02	1,05	0,12	-0,64	1,73	-1,02	-0,36	0,12	0,54	B3
4	☑	2,8554	-1,41	-1,53	0,18	0,38	-0,4	-0,21	1,55	0,18	0,38	0,71	0,6	B4
5	☑	3,0023	-0,88	0,61	0,03	-0,38	-0,38	-1,4	-0,18	-0,18	1,81	1,38	0,61	B5
6	☑	3,0562	-1,32	0,97	-1,57	0,97	0,62	-0,89	0	-0,69	0,73	0,73	0,73	B6
7	☑	3,1808	0,86	-0,55	0,16	-0,82	0,75	-2,18	0,16	-0,37	0,25	0,98	1,37	B7
8	☑	3,1986	-0,4	0,2	-0,4	-1,15	0,34	-1,4	0,23	0,02	2,07	-0,3	1,43	B8
9	☑	3,2835	-2,18	-0,5	0,14	0,06	-0,42	0,23	0,81	0,14	-0,02	1,64	1,47	B9

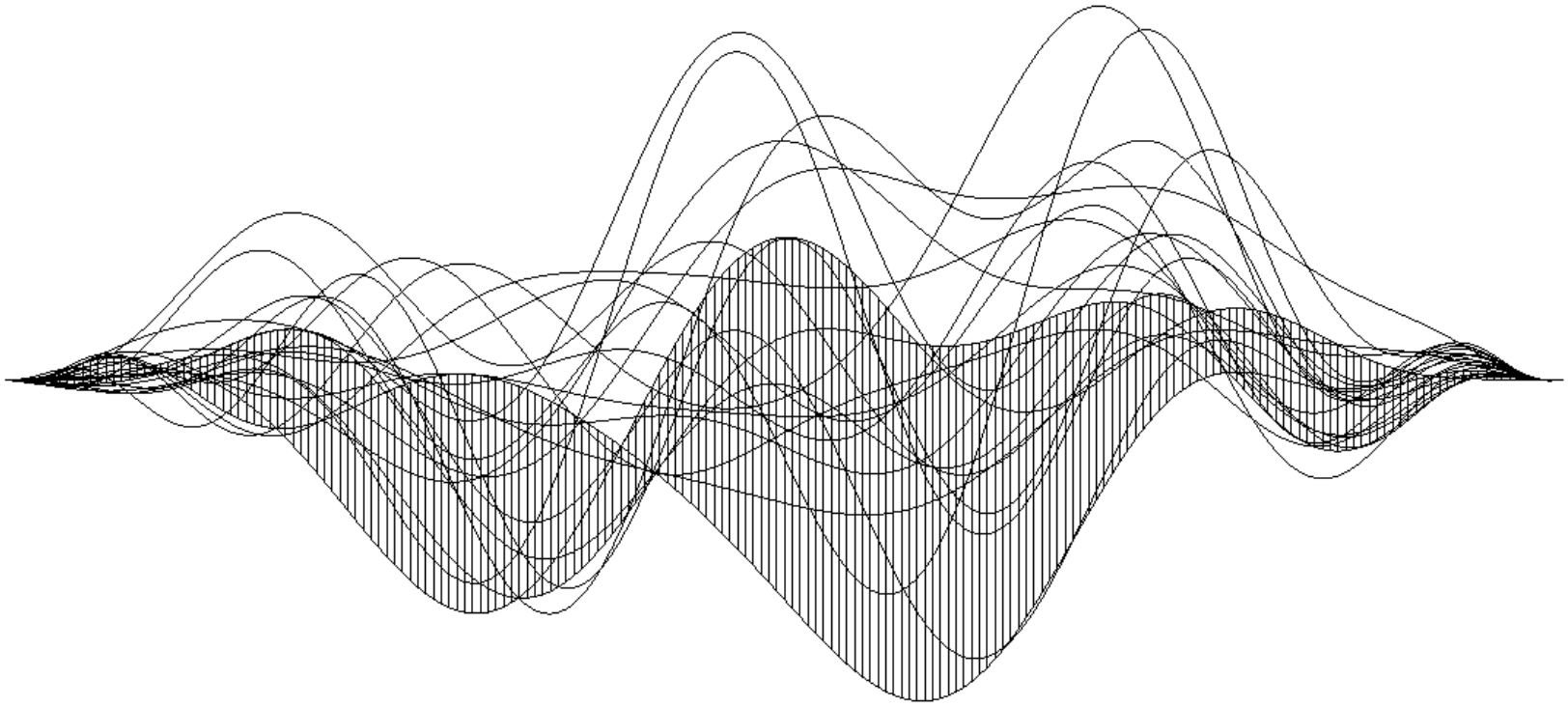
Cluster № 1 – variants 2, 5, 6, 8, 15, 21



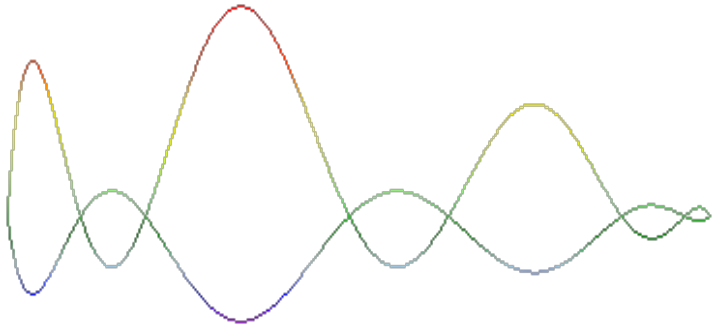
Cluster № 2 – variants 1, 4, 9, 10, 12, 14, 16, 17, 18, 19, 20

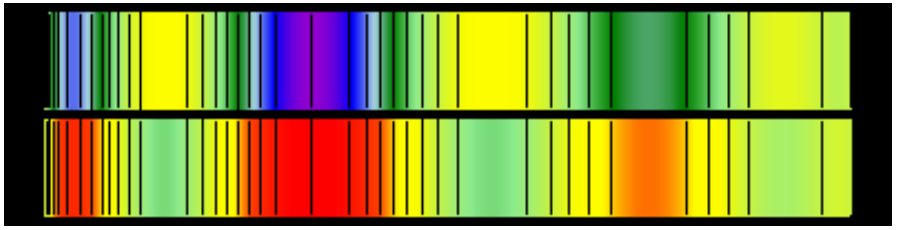


Cluster № 3 – variants 3, 7, 11, 13



- For comparison of individual observations can use their "spectral" representation. It emphasizes the distinctive characteristics of each curve and helps a more detailed study of their visual properties.





"Spectrum" of multivariate observations

- We have used this approach dealing with the problems of identifying hidden regularity in medical data, particularly analyzing the characteristics of various bronchopulmonary diseases.
- Background information is data of patients with four types of bronchopulmonary diseases:
  - Bronchial asthma non-psychogenic (BANP);
  - Bronchial asthma somatic psychogenic (BASP);
  - Bronchial asthma psychogenic-induced (BAPI);
  - Psychogenic dyspnea (PD).

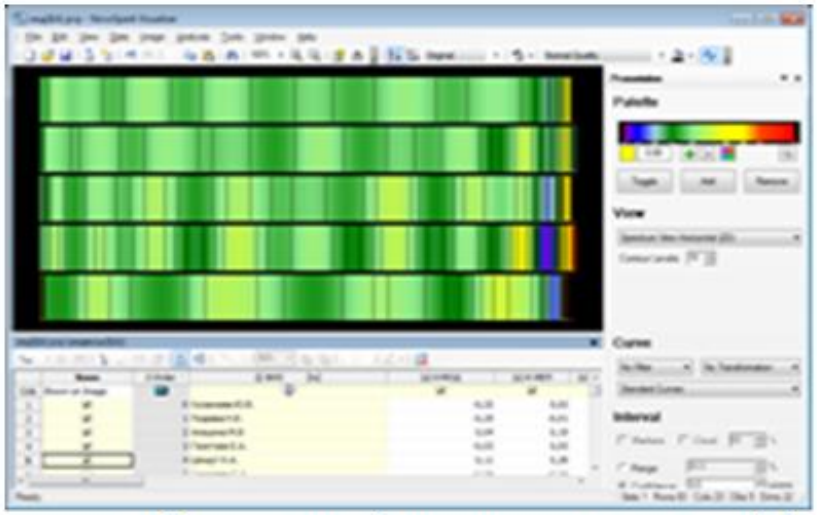


Fig. 2. The spectral representation of the data on patients diagnosed with BAPI

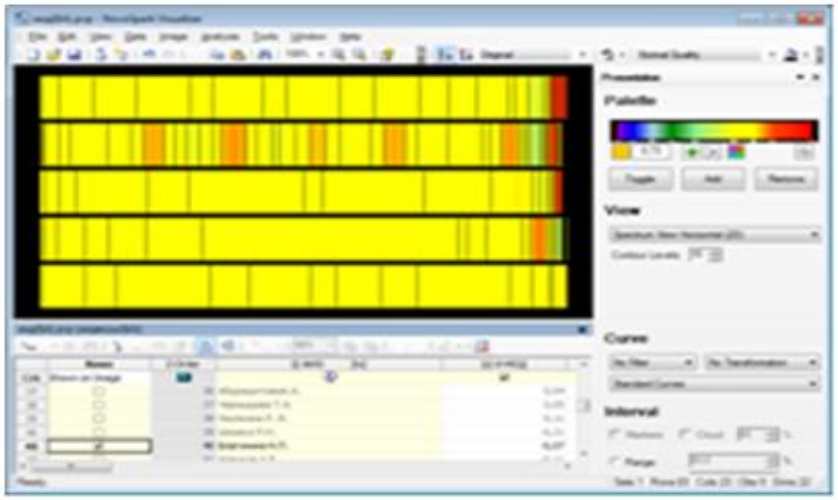


Fig. 3. The spectral representation of the data on patients with a diagnosis of BASP

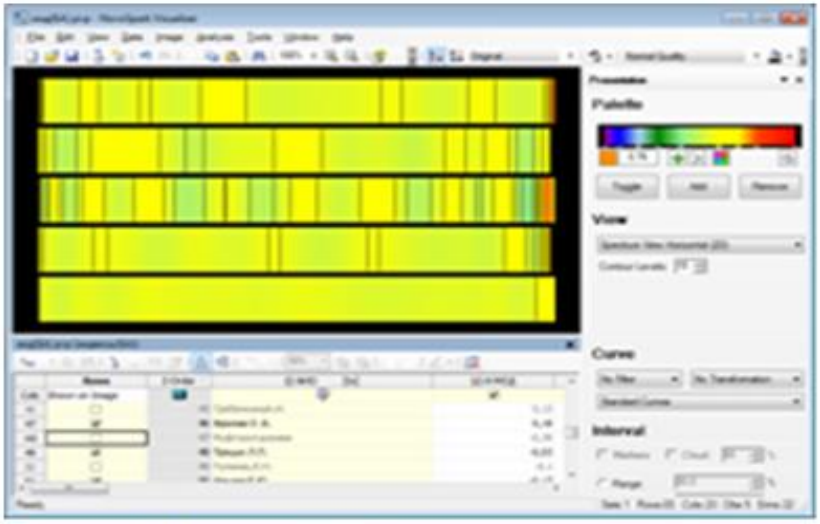


Fig. 4. The spectral representation of the data on patients diagnosed with BANP

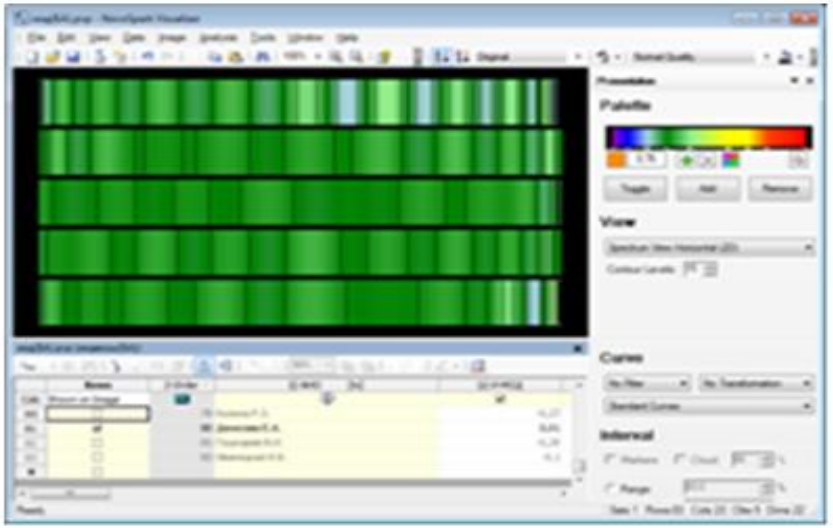


Fig. 5. The spectral representation of the data on patients diagnosed with PD

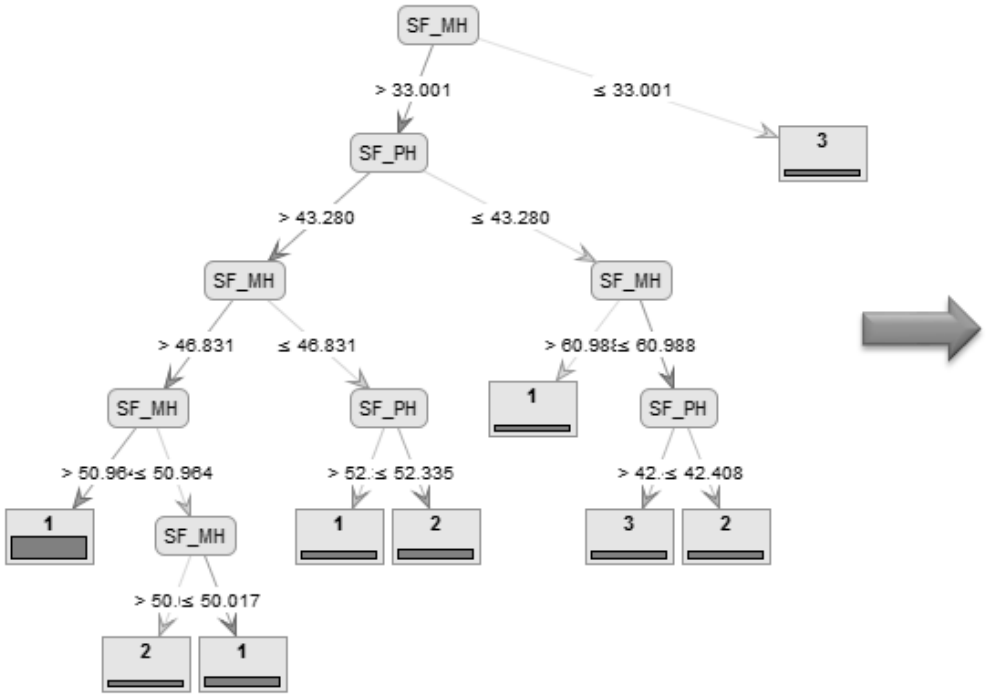
- On the basis of visualization it’s possible to solve a number of tasks, which then require finding a similar method of tasks statement. These are such tasks as segmentation, clusterization, prognostication.
- Users of such tasks will mainly be various decision takers, who will be offered a less complicated process of results achieving, which can have different forms. In the first place these are numerical results which have linguistic description outlining qualitative aspects of various subject areas.

Identifying quality life features for patients with neurological diseases

3

- Well-known questioner was used for researching of patients life quality - "Short Form Health Assessment - MOS SF-36» (Medical Outcomes Study-Short Form).
- This technique allows to register and quantify changes in the quality of patients life with a certain type of disease during a specific period of hospital treatment, as well as identify components that make the most significant contribution to the treatment caused changes in the quality of life.
- As a tool for building decision trees was used RapidMiner. It is an integrated system that implements the methods of Data Mining and Statistical Analysis.

An example of the graphical representation of the decision tree for Anxiety definition (HADS_T) depending MH indicator values ("mental health component") and PH («physical health component") which was generated by program RapidMiner, is presented as a graphic as well as text.




```
SF_MH > 33.001
|   SF_PH > 43.280
|   |   SF_MH > 46.831
|   |   |   SF_MH > 50.964: 1 {3=0, 1=51, 2=0}
|   |   |   SF_MH ≤ 50.964
|   |   |   |   SF_MH > 50.017: 2 {3=0, 1=0, 2=2}
|   |   |   |   SF_MH ≤ 50.017: 1 {3=0, 1=11, 2=0}
|   |   SF_MH ≤ 46.831
|   |   |   SF_PH > 52.335: 1 {3=0, 1=6, 2=0}
|   |   |   SF_PH ≤ 52.335: 2 {3=0, 1=0, 2=11}
|   SF_PH ≤ 43.280
|   |   SF_MH > 60.988: 1 {3=0, 1=2, 2=0}
|   |   SF_MH ≤ 60.988
|   |   |   SF_PH > 42.408: 3 {3=6, 1=0, 2=0}
|   |   |   SF_PH ≤ 42.408: 2 {3=0, 1=0, 2=6}
SF_MH ≤ 33.001: 3 {3=5, 1=0, 2=0}
```

The values 1, 2, 3, indicator HADS_T:
1 - Norm; 2 -Ssubclinical expressed anxiety / depression;
3 - Symptomatic anxiety / depression.

TOMSK POLYTECHNIC UNIVERSITY

OUR TEAM:

Olga Marukhina, Associate Professor, marukhina@tpu.ru

Elena Mokina, Researcher, Alisandra@tpu.ru

Olga Berestneva, Professor, ogb6@yandex.ru

Maria Shagarova, Student