



Л е к ц и я № 5

Кластерный анализ

Разработал профессор П. М. Васильев
Кафедра фармакологии и биоинформатики

Для студентов, обучающихся по направлению 06.03.01 «Биология»
профили Биохимия, Генетика
при изучении дисциплины «Цифровые технологии в биологии»

П л а н л е к ц и и

- **Что такое кластерный анализ**
- **Методы кластерного анализа**
- **Метрики расстояния**
- **Метод k-средних**
- **Иерархическая кластеризация**

Что такое кластерный анализ

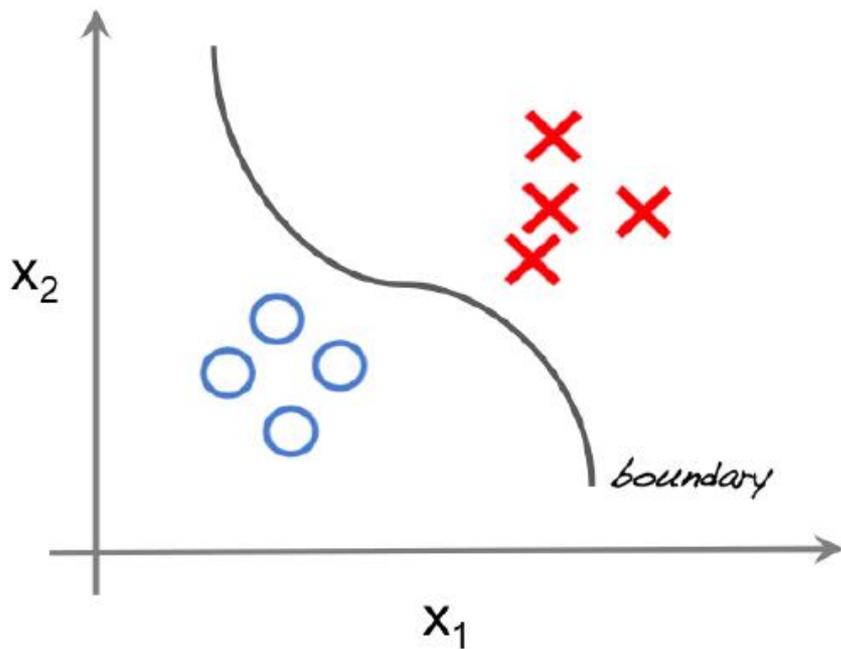
Общее название множества
вычислительных процедур,
используемых при создании
классификации
и направленных на внутреннюю
структуризацию данных.

Гипотеза компактности

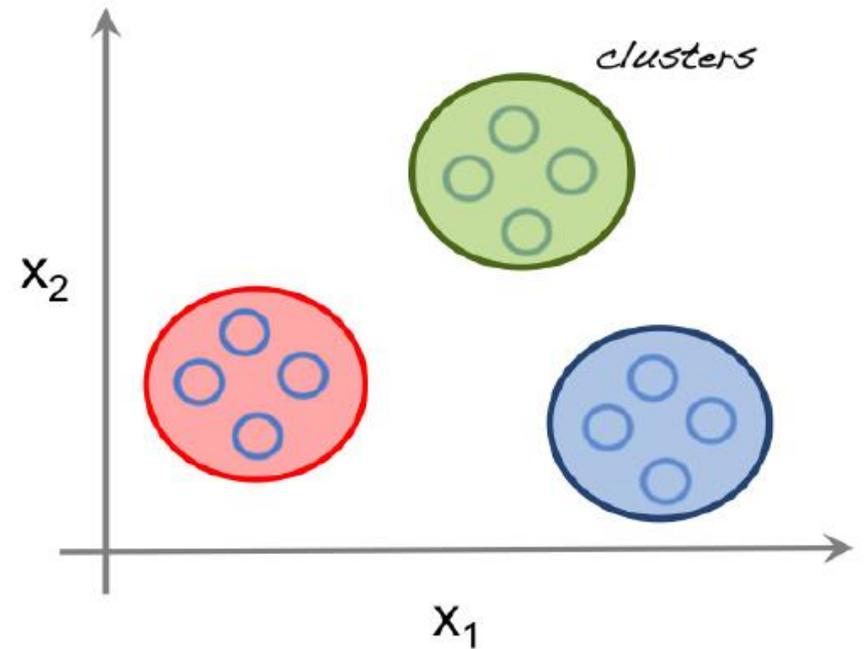
Расстояние между двумя объектами, принадлежащими к одному и тому же кластеру, всегда меньше расстояния между двумя объектами, принадлежащими к разным кластерам.

Виды обучения

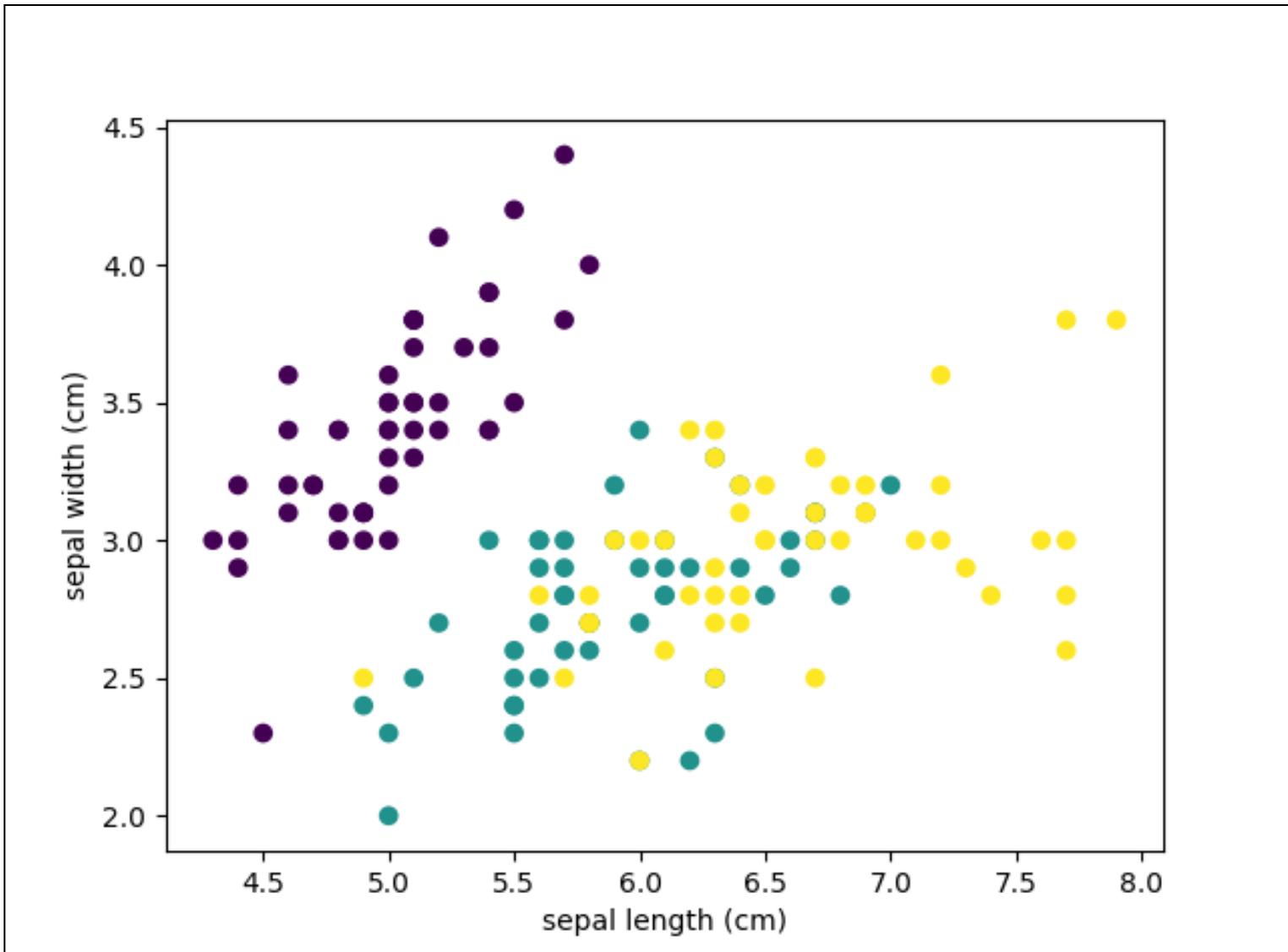
Supervised learning



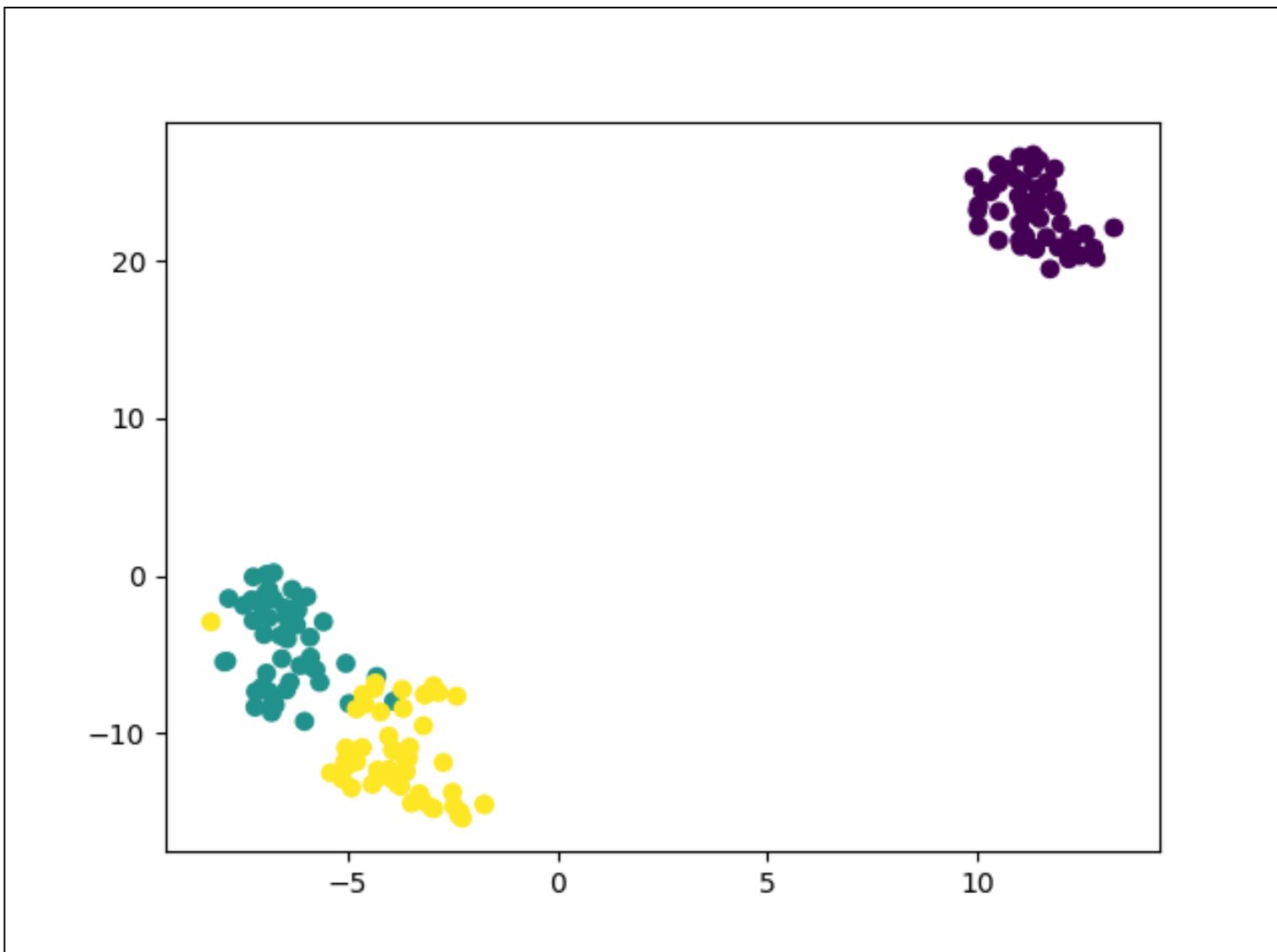
Unsupervised learning



Помеченное исходное множество



Результат кластеризации



Методы кластерного анализа

- Иерархические агломеративные методы
- Иерархические дивизимные методы
- Итеративные методы группировки
- Факторные методы
- Методы сгущений
- Методы теории графов
- Деревья решений
- Нейронные сети

Метрики расстояния

Эвклидово расстояние

$$d_{ij} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

Требует обязательного нормирования показателей

Метрики расстояния

Расстояние городских кварталов

$$d_{ij} = \sum_{k=1}^p |x_{ik} - x_{jk}|$$

Менее чувствительно к выбросам

Метрики расстояния

Расстояние Чебышева

$$d_{ij} = \max |x_{ik} - x_{jk}|$$

Хорошо учитывает различия объектов

Метрики расстояния

Расстояние Минковского

$$d_{ij} = \left(\sum_{k=1}^p |x_{ik} - x_{jk}|^p \right)^{1/r}$$

Взвешивает расстояние между переменными

Кластеризация методом k-средних

Пример исходных данных

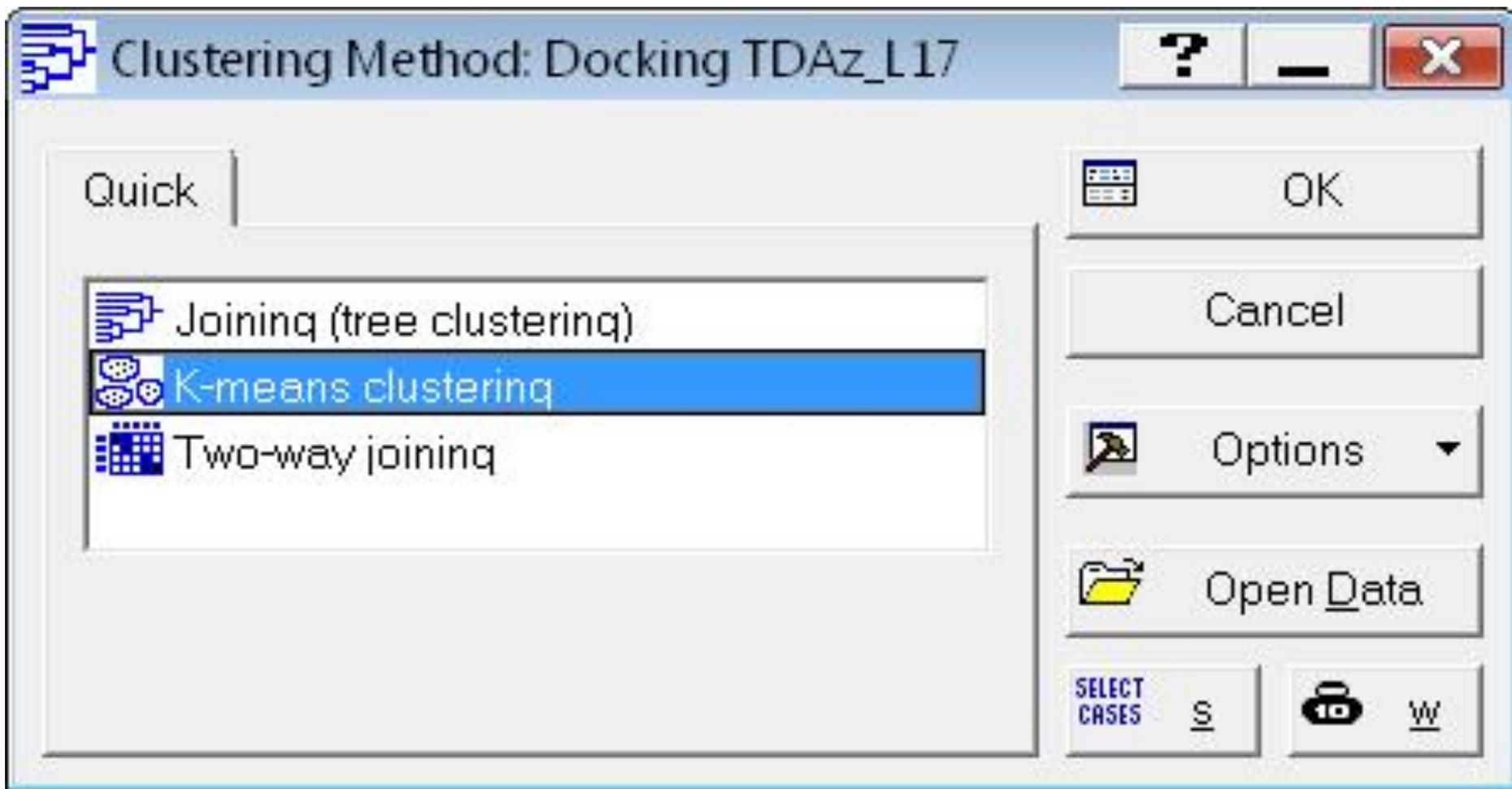
Code	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
LT-47	-8.9	-8.5	-6.5	-7.9	-7.1	-8.2	-8.7	-8.3	-6.2	-6.7	-7.5	-8.5	-6.4	-6.3
LT-51	-8.9	-8.4	-6.7	-7.4	-6.6	-8.1	-8.6	-8.3	-6.1	-6.5	-7.4	-8.4	-6.8	-6.1
LT-52	-8.9	-8.6	-7.1	-8.2	-7.0	-9.4	-9.2	-8.6	-6.4	-6.8	-8.2	-8.6	-7.5	-6.4
L-91	-8.7	-8.6	-6.9	-6.9	-8.3	-9.2	-8.7	-8.9	-6.1	-6.4	-7.5	-8.2	-7.0	-6.3
LT-13	-8.7	-8.6	-7.2	-8.2	-7.2	-9.0	-8.9	-8.7	-6.3	-6.5	-7.5	-8.2	-6.4	-5.8
LT-23	-8.7	-8.5	-6.8	-7.0	-7.7	-8.8	-8.2	-8.4	-6.1	-6.4	-7.6	-8.1	-7.0	-6.0
LT-48	-8.7	-8.5	-6.4	-7.4	-6.8	-8.4	-8.6	-8.2	-6.2	-6.6	-7.6	-8.0	-6.6	-6.1
LT-46	-8.6	-8.2	-6.2	-7.8	-7.8	-7.9	-8.3	-7.9	-6.0	-6.4	-7.4	-8.2	-6.2	-6.0
H-16	-8.5	-8.3	-6.5	-7.1	-6.6	-8.0	-8.1	-8.1	-5.7	-6.1	-7.3	-8.0	-6.6	-6.4
H-88	-8.5	-8.6	-7.0	-7.3	-8.0	-9.3	-8.4	-8.0	-6.1	-6.4	-7.2	-7.9	-6.8	-6.3
LT-22	-8.5	-8.5	-6.9	-7.4	-8.3	-9.0	-8.3	-8.4	-6.0	-6.3	-7.5	-8.2	-6.8	-6.5
LT-49	-8.5	-9.0	-6.6	-7.0	-6.8	-8.4	-8.5	-8.4	-6.0	-6.3	-7.6	-8.0	-6.3	-6.3
L-87	-8.4	-8.5	-7.0	-7.5	-7.8	-8.6	-8.5	-8.9	-6.2	-6.4	-7.3	-8.2	-6.9	-6.3
LT-55	-8.4	-8.6	-7.0	-7.5	-7.3	-8.5	-8.3	-9.1	-6.1	-6.4	-7.8	-8.1	-6.7	-6.0
LT-9	-8.4	-8.6	-6.8	-7.4	-8.0	-8.4	-8.2	-8.4	-6.1	-6.4	-6.9	-7.7	-6.3	-6.0
L-20	-8.3	-8.4	-6.6	-7.1	-7.3	-8.2	-8.3	-8.1	-6.0	-6.3	-7.4	-8.1	-6.6	-6.1
L-31	-8.3	-8.4	-6.9	-7.4	-7.8	-8.0	-8.4	-8.1	-6.2	-6.6	-7.4	-8.0	-6.4	-6.6
L-86	-8.3	-8.3	-6.8	-6.8	-8.5	-8.9	-8.2	-8.6	-6.1	-6.2	-7.2	-8.0	-6.9	-6.2
LT-21	-8.3	-8.6	-6.8	-7.1	-8.2	-8.8	-8.2	-8.1	-6.0	-6.2	-7.3	-8.0	-6.9	-6.1
LTG-2	-8.3	-7.8	-6.8	-6.9	-7.3	-7.9	-8.0	-7.4	-6.2	-6.5	-8.1	-8.4	-6.9	-6.8
L-19	-8.2	-8.3	-6.6	-7.1	-6.7	-8.3	-8.1	-7.9	-6.2	-6.4	-7.6	-7.9	-6.6	-6.1
LT-24	-8.2	-8.8	-7.1	-6.8	-7.9	-8.8	-8.5	-8.2	-5.9	-6.3	-7.4	-7.9	-6.5	-6.4
LT-33	-8.2	-8.0	-6.1	-7.2	-7.0	-7.9	-7.9	-7.7	-5.8	-5.9	-7.1	-8.0	-6.1	-5.5
L-17	-8.1	-8.0	-6.6	-7.4	-8.3	-8.1	-7.9	-7.8	-6.1	-6.3	-7.4	-8.0	-6.5	-6.0
LT-34	-8.1	-8.2	-6.3	-7.6	-6.5	-8.3	-8.1	-7.9	-6.0	-6.3	-7.1	-8.1	-6.3	-6.1
LT-36	-8.1	-8.2	-6.3	-7.1	-6.2	-7.8	-8.1	-7.7	-5.8	-6.0	-7.0	-8.1	-6.7	-5.9

Кластеризация методом k-средних

The screenshot displays the STATISTICA software interface. The 'Statistics' menu is open, and 'Cluster Analysis' is highlighted. The background data table is as follows:

	Clustering													
	1	2												
	Y	X1												
LT-47	-8.9	-8.5												
LT-51	-8.9	-8.4												
LT-52	-8.9	-8.6												
L-91	-8.7	-8.6												
LT-13	-8.7	-8.6												
LT-23	-8.7	-8.5												
LT-48	-8.7	-8.5												
LT-46	-8.6	-8.2												
H-16	-8.5	-8.3												
H-88	-8.5	-8.6												
LT-22	-8.5	-8.5												
LT-49	-8.5	-9.0												
L-87	-8.4	-8.5	-7.0	-7.5	-7.0	-8.0	-8.5	-8.9	-6.2	-6.4	-7.3	-8.2	-6.9	-6.3
LT-55	-8.4	-8.6	-7.0	-7.5	-7.3	-8.5	-8.3	-9.1	-6.0	-6.4	-7.8	-8.1	-6.7	-6.0
LT-9	-8.4	-8.6	-6.8	-7.4	-8.0	-8.4	-8.2	-8.4	-6.1	-6.4	-6.9	-7.7	-6.3	-6.0

Кластеризация методом k-средних



Кластеризация методом k-средних

Cluster Analysis: K-Means Clustering: Docking TDAz_L17

Quick | Advanced

 Variables:

Cluster:

Number of clusters:

Number of iterations:

Initial cluster centers

- Choose observations to maximize initial between-cluster distances
- Sort distances and take observations at constant intervals
- Choose the first N (Number of clusters) observations

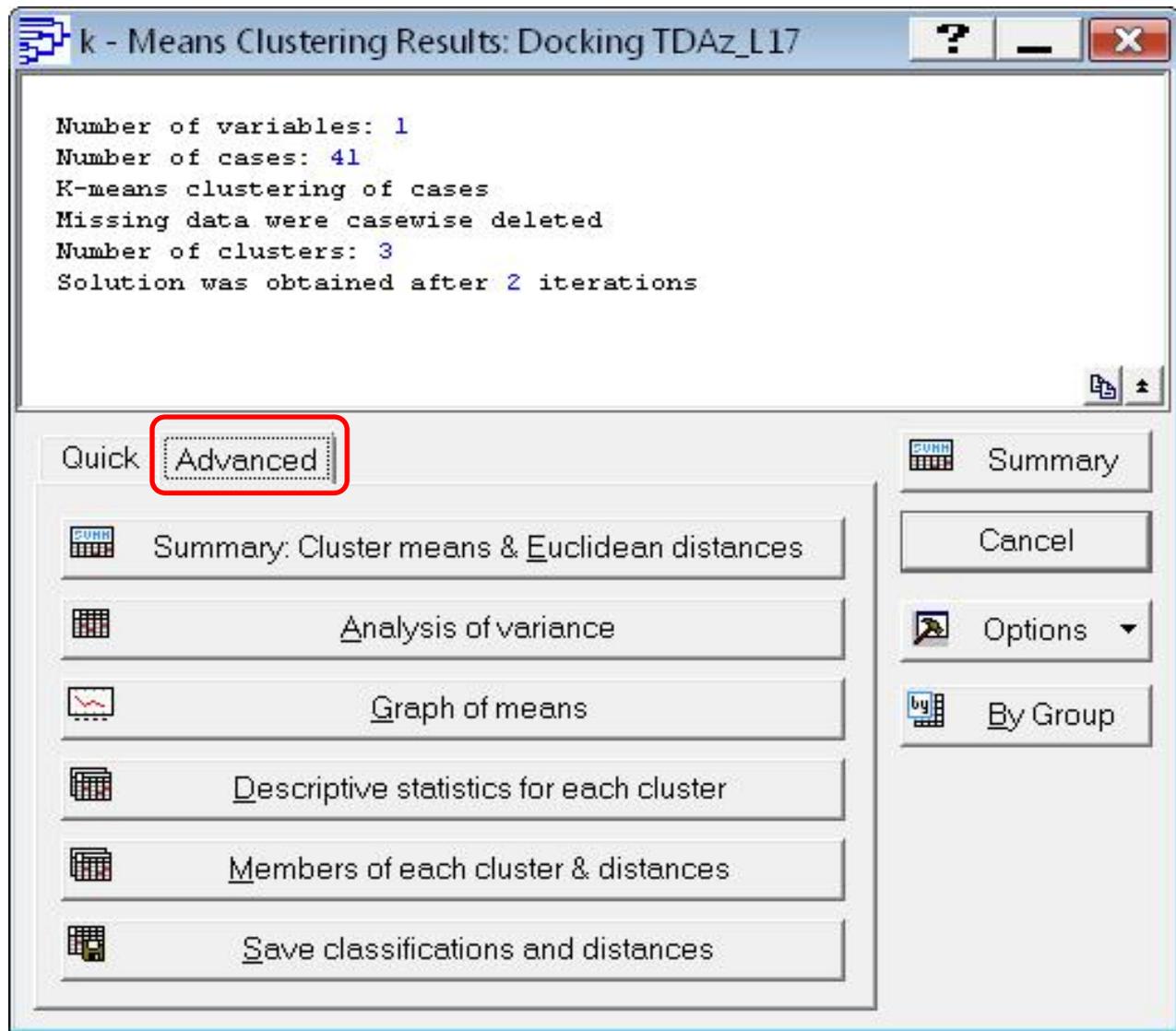
Batch processing and reporting

 Options ▾

MD deletion

- Casewise
- Mean substitution

Кластеризация методом k-средних



Кластеризация методом k-средних

Workbook1*

- Cluster Analysis (Docking TDAz_L17)
 - K-means clustering results dialog
 - Cluster Means (Docking TDAz_L17)**
 - Euclidean Distances between Clusters (Docking TDAz_L17)
 - Analysis of Variance (Docking TDAz_L17)
 - Descriptive Statistics for Cluster 1 (Docking TDAz_L17)
 - Descriptive Statistics for Cluster 2 (Docking TDAz_L17)
 - Descriptive Statistics for Cluster 3 (Docking TDAz_L17)
 - Members of Cluster Number 1 (Docking TDAz_L17)
 - Members of Cluster Number 2 (Docking TDAz_L17)
 - Members of Cluster Number 3 (Docking TDAz_L17)

Variable	Cluster Means (Docking TDAz_L17)		
	Cluster No. 1	Cluster No. 2	Cluster No. 3
Y	-8.22353	-7.62500	-8.67500
	↓ Средняя	↓ Низкая	↓ Высокая

Cluster Means (Docking TDAz_L17) | Euclidean

Кластеризация методом k-средних

The screenshot displays a software interface with a file explorer on the left and an ANOVA table on the right. The file explorer shows a hierarchy starting with 'Workbook1*', followed by 'Cluster Analysis (Docking TDAz_L17)', and then 'K-means clustering results dialog'. Under this dialog, several files are listed, with 'Analysis of Variance (Docking TDAz_L17)' highlighted by a red box. The ANOVA table on the right is titled 'Analysis of Variance (Docking TDAz_L17)' and contains the following data:

Variable	Between SS	df	Within SS	df	F	signif. p
Y	6.668802	2	1.175588	38	107.7820	0.000000

The 'F' and 'signif. p' values for variable Y are also highlighted with a red box. At the bottom of the interface, there are two tabs: 'Analysis of Variance (Docking TDAz_L17)' and 'Descriptive Statistics'.

Кластеризация методом k-средних

Members of Cluster Number 3 (Docking TDAz_L17)
and Distances from Respective Cluster Center
Cluster contains 12 cases

	Distance
LT-47	0.225000
LT-51	0.225000
LT-52	0.225000
L-91	0.025000
LT-13	0.025000
LT-23	0.025000
LT-48	0.025000
LT-46	0.075000
H-16	0.175000
H-88	0.175000
LT-22	0.175000
LT-49	0.175000

Members of Cluster Number 3 (Docking TDAz_L17)

Кластеризация методом k-средних

Members of Cluster Number 1 (Docking TDAz_L17) and Distances from Respective Cluster Center
Cluster contains 17 cases

	Distance
L-87	0.176471
LT-55	0.176471
LT-9	0.176471
L-20	0.076471
L-31	0.076471
L-86	0.076471
LT-21	0.076471
LTG-2	0.076471
L-19	0.023529
LT-24	0.023529
LT-33	0.023529
L-17	0.123529
LT-34	0.123529
LT-36	0.123529
LT-50	0.123529
LT-5	0.123529
H-18	0.223529

Members of Cluster Number 1 (Docking TDAz_L17) Members of Cluster Nun

Кластеризация методом k-средних

Members of Cluster Number 2 (Docking TDAz_L17) and Distances from Respective Cluster Center
Cluster contains 12 cases

	Distance				
H-69	0.275000				
LT-30	0.275000				
LT-31	0.175000				
LT-35	0.175000				
LT-39	0.175000				
LT-29	0.075000				
LT-32	0.025000				
LT-38	0.025000				
LT-37	0.125000				
H-42	0.225000				
LT-53	0.225000				
H-29	0.525000				

Members of Cluster Number 2 (Docking TDAz_L17) Members of Cluster Nun

Кластеризация методом k-средних

Границы кластеров

$$[Y(\text{LT-49}) + Y(\text{L-87})] / 2 = -8.45$$

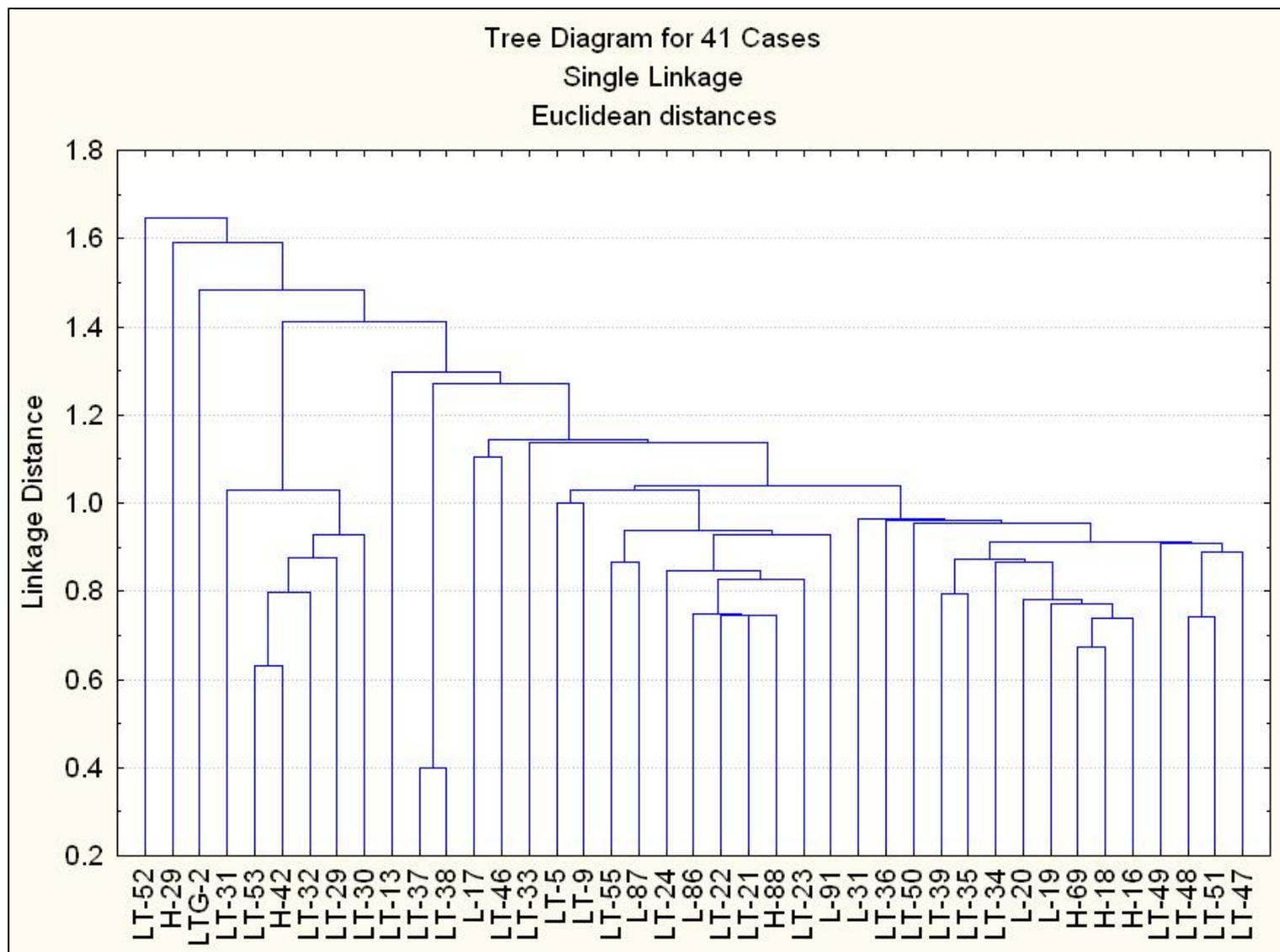
$$[Y(\text{L-18}) + Y(\text{L-69})] / 2 = -7.95$$

Высокая $Y < -8.45$

Умеренная $-8.45 \leq Y < -7.95$

Низкая $Y \geq -7.95$

Иерархическая кластеризация



Самоорганизующиеся сети Кохонена

STATISTICA - Docking TDAz_L17

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help

Automated Neural Networks

Machine Learning (Bayesian, Support Vectors, Nearest Neighbor)

General Classification/Regression Tree Models

General CHAID Models

Interactive Trees (C&RT, CHAID)

Boosted Tree Classifiers and Regression

Random Forests for Regression and Classification

Generalized Additive Models

MARSplines (Multivariate Adaptive Regression Splines)

Generalized EM & k-Means Cluster Analysis

Independent Components Analysis

Text & Document Mining, Web Crawling - Methods

Association Rules

Sequence, Association, and Link Analysis

Rapid Deployment of Predictive Models (PMML)

Goodness of Fit, Classification, Prediction

Feature Selection and Variable Screening

Combining Groups (Classes) for Predictive Data Mining

Data Mining - Workspaces

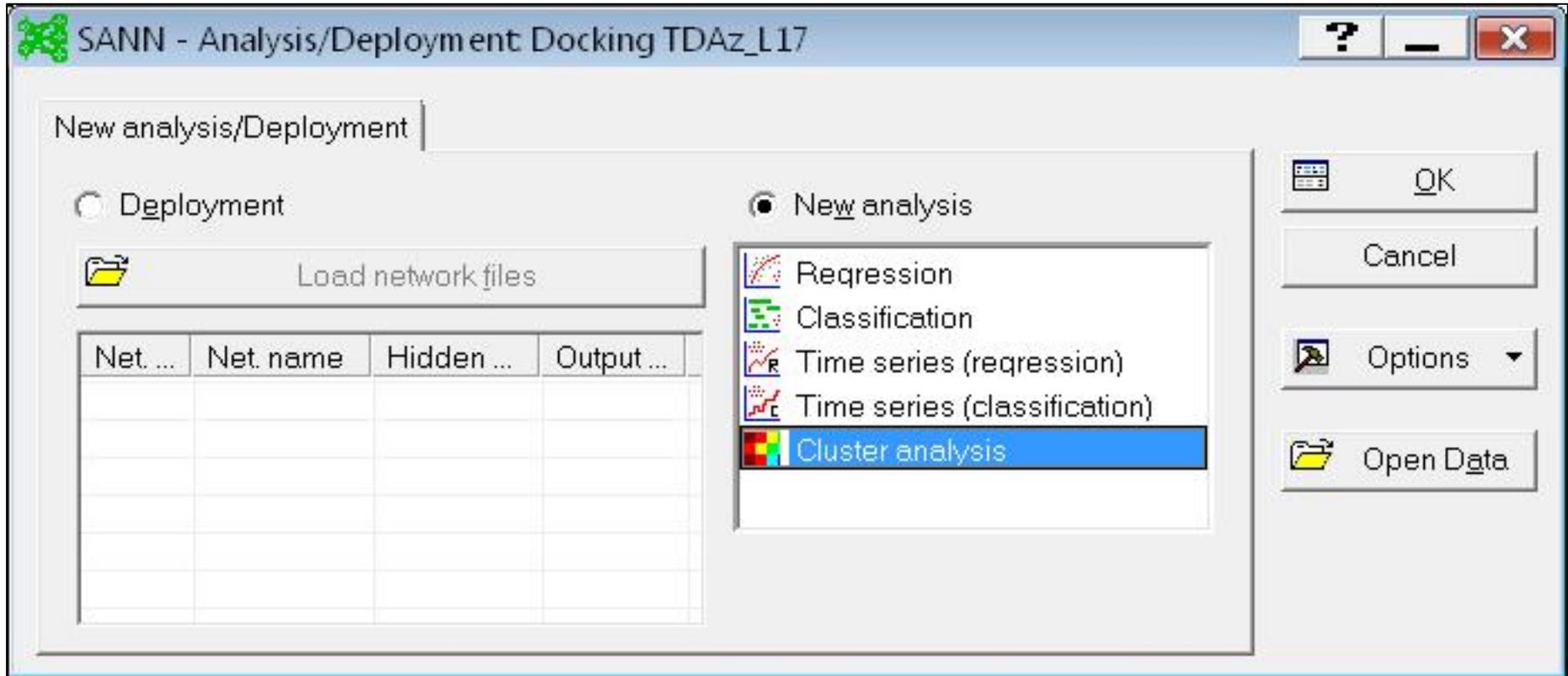
QC Data Mining & Root Cause Analysis - Methods

Data: Docking TDAz_L17* (14v by 43c)

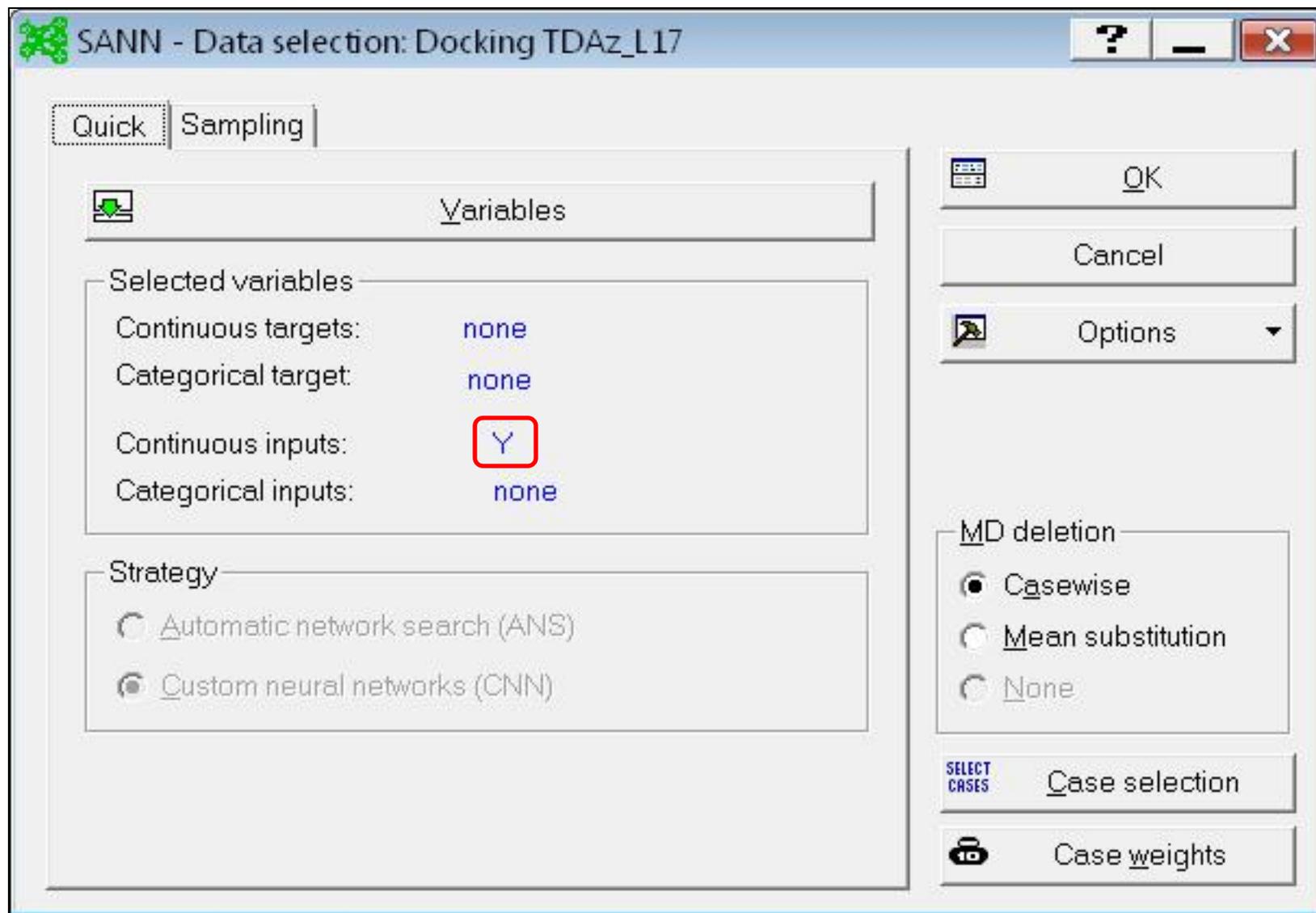
	Clustering		
	1 Y	2 X1	3 X2
LT-47	-8.9	-8.5	-6.5
LT-51	-8.9	-8.4	-6.7
LT-52	-8.9	-8.6	-7.1
L-91	-8.7	-8.6	-6.9
LT-13	-8.7	-8.6	-7.2
LT-23	-8.7	-8.5	-6.8
LT-48	-8.7	-8.5	-6.4
LT-46	-8.6	-8.2	-6.2
H-16	-8.5	-8.3	-6.5
H-88	-8.5	-8.6	-7.0
LT-22	-8.5	-8.5	-6.9
LT-49	-8.5	-9.0	-6.6
L-87	-8.4	-8.5	-7.0
LT-55	-8.4	-8.6	-7.0
LT-9	-8.4	-8.6	-6.8
L-20	-8.3	-8.4	-6.6
L-31	-8.3	-8.4	-6.9
L-86	-8.3	-8.3	-6.8

	12 X11	13 X12	14 X13
	-8.5	-6.4	-6.3
	-8.4	-6.8	-6.1
	-8.6	-7.5	-6.4
	-8.2	-7.0	-6.3
	-8.2	-6.4	-5.8
	-8.1	-7.0	-6.0
	-8.0	-6.6	-6.1
	-8.2	-6.2	-6.0
	-8.0	-6.6	-6.4
	-7.9	-6.8	-6.3
	-8.2	-6.8	-6.5
	-8.0	-6.3	-6.3
	-8.2	-6.9	-6.3
	-8.1	-6.7	-6.0
	-7.7	-6.3	-6.0
	-8.1	-6.6	-6.1
	-8.0	-6.4	-6.6
	-8.0	-6.9	-6.2

Самоорганизующиеся сети Кохонена



Самоорганизующиеся сети Кохонена



Самоорганизующиеся сети Кохонена

SANN - Custom Neural Network: Docking TDAz_L17

Active neural networks

Net ...	Net. name	Algorithm

Quick (Kohonen) Kohonen Training

Training

Training cycles: 1000

Learning rates

Start: .1

End: .02

Neighborhoods

Start: 3

End: 0

Stopping conditions

Enable stopping conditions

Improvement: .0000001

Cycles: 10

Network randomization

Normal randomization

Uniform randomization

Mean\Min: 0.

Variance\Max: .1

Train

Go to results

Save networks

Data statistics

Summary

Cancel

Options

Самоорганизующиеся сети Кохонена

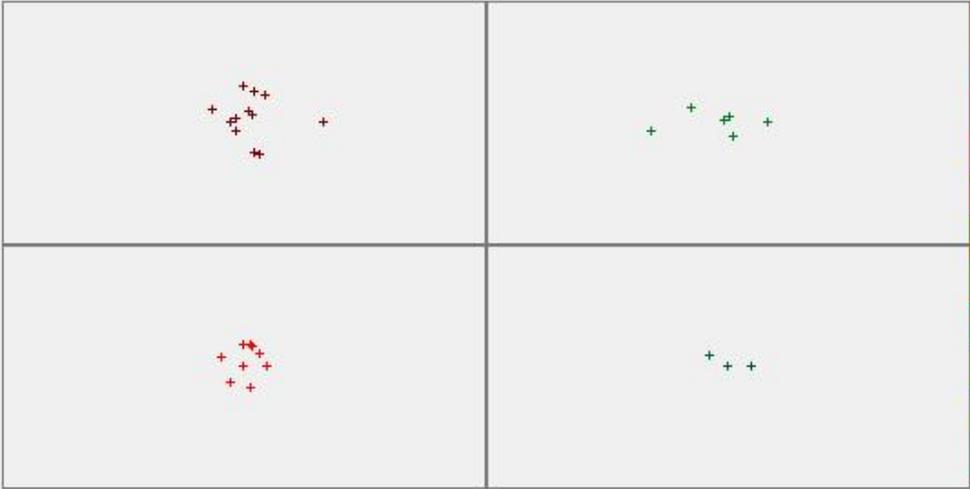
SANN - Results: Docking TDAz_L17

Active neural networks

Net ...	Net. name	Algorithm
1	SOFM 1-4	Kohonen 1...

Select/Deselect active networks Delete networks

Predictions (Kohonen) Graphs (Kohonen) Kohonen graph



Summary
Save
Cancel
Options

Sample
 Train
 Test
 Validation

0
0
none

Kohonen Kohonen Select All Clear

To be continued ...

