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»

План лекции

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

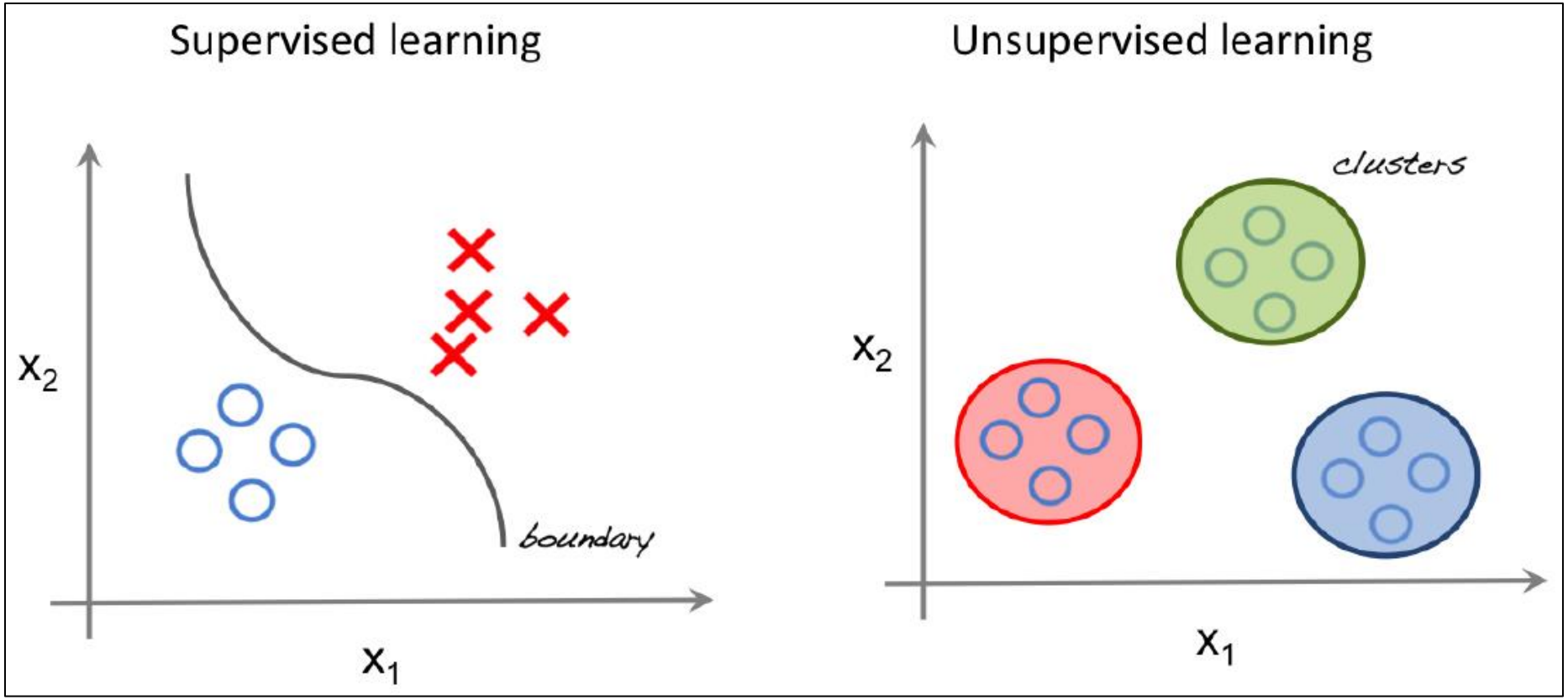
Что такое кластерный анализ

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

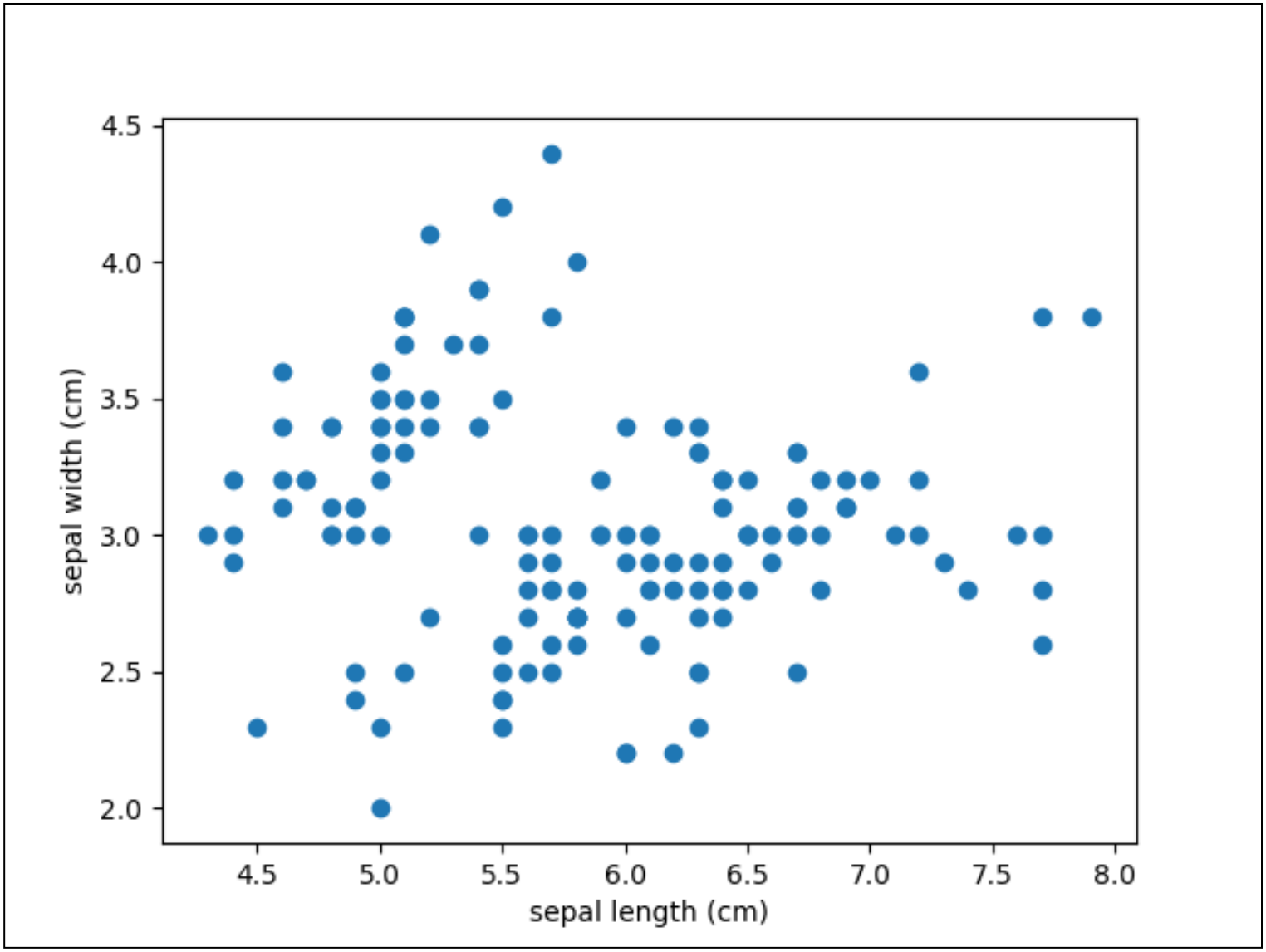
Гипотеза компактности

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

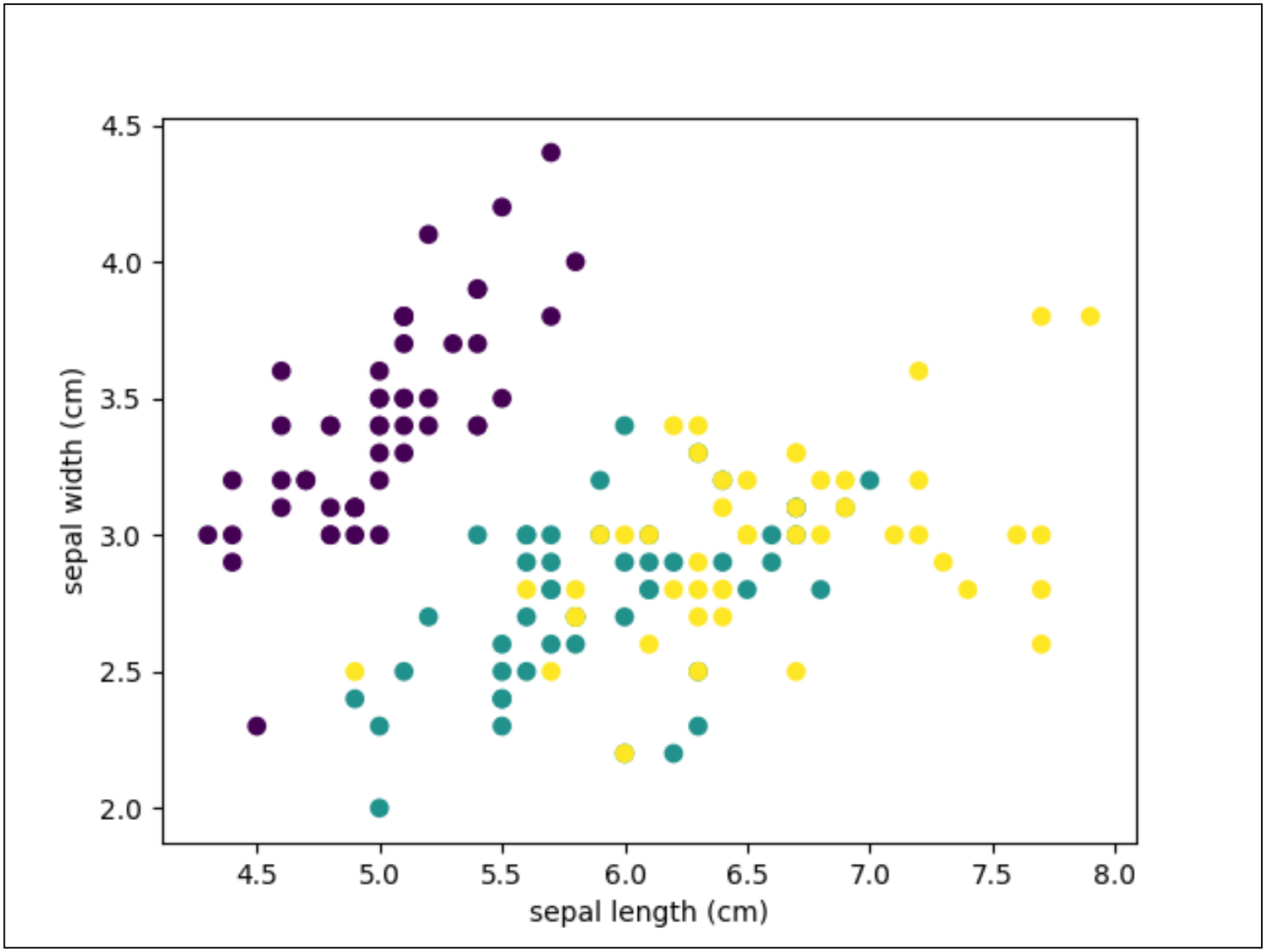
Виды обучения



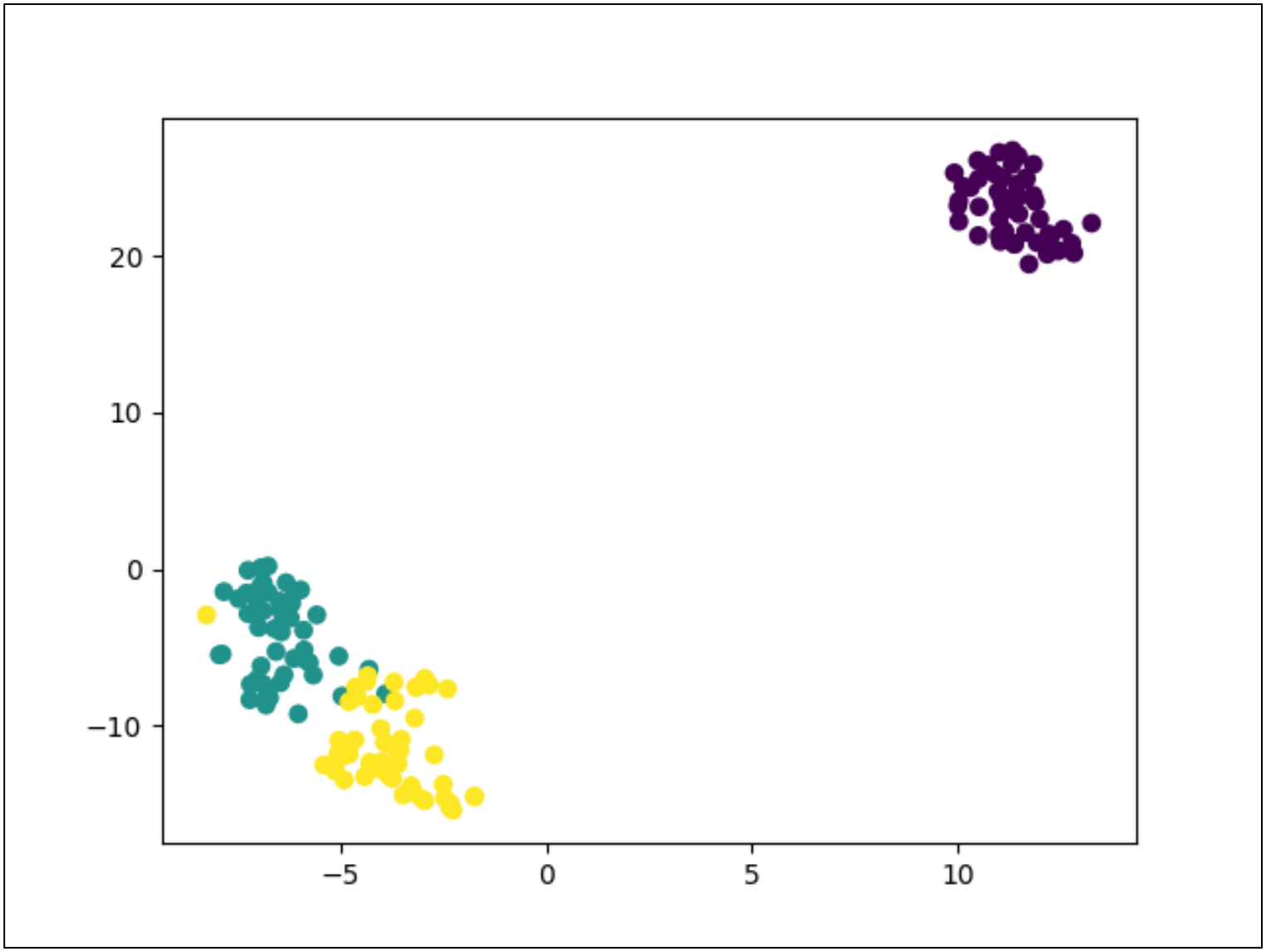
Исходное множество



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

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

Менее чувствительно к выбросам

Метрики расстояния

Расстояние Чебышева

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

Хорошо учитывает различия объектов

Метрики расстояния

Расстояние Минковского

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

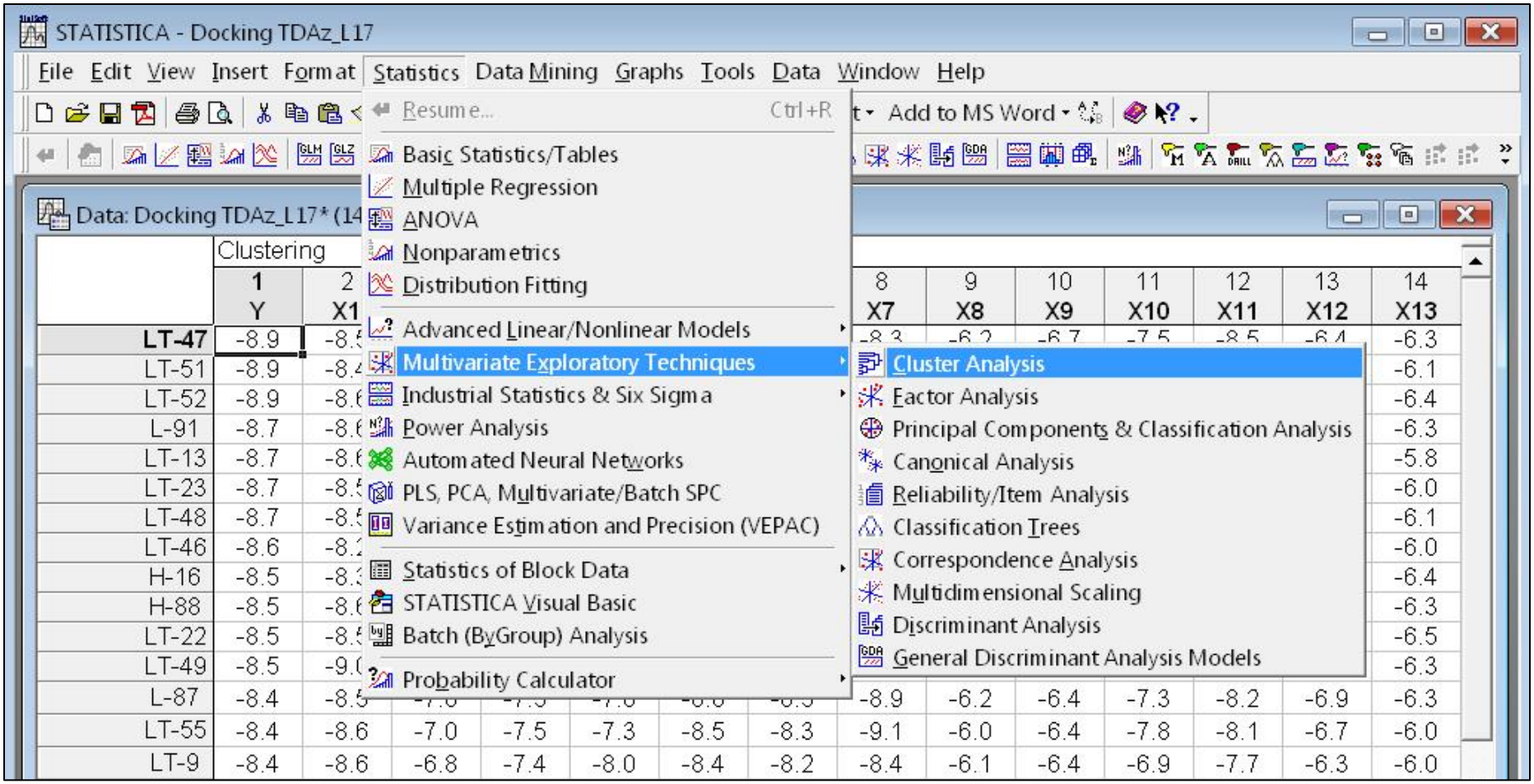
Взвешивает расстояние между переменными

Кластеризация методом k-средних

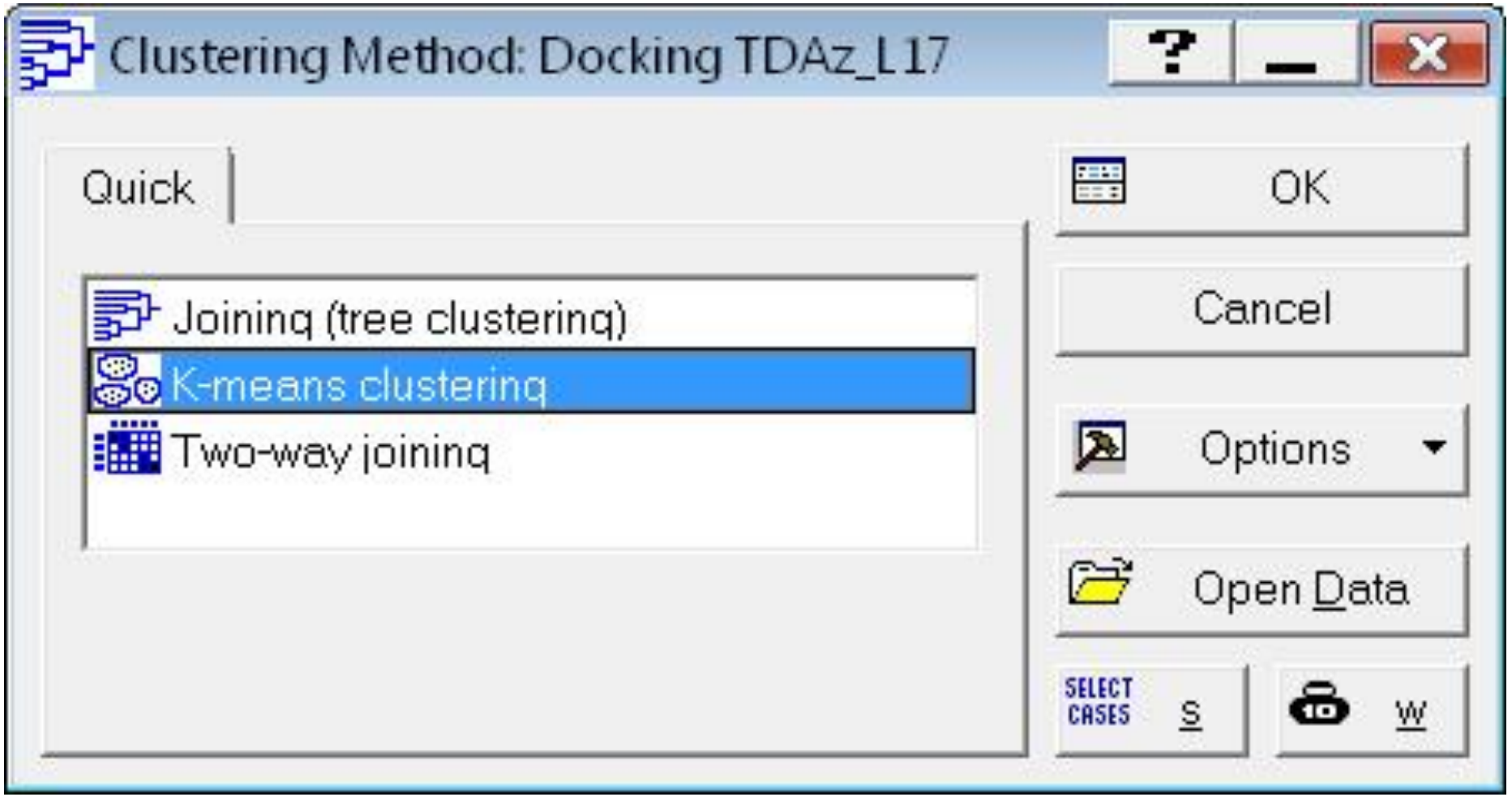
Пример исходных данных

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

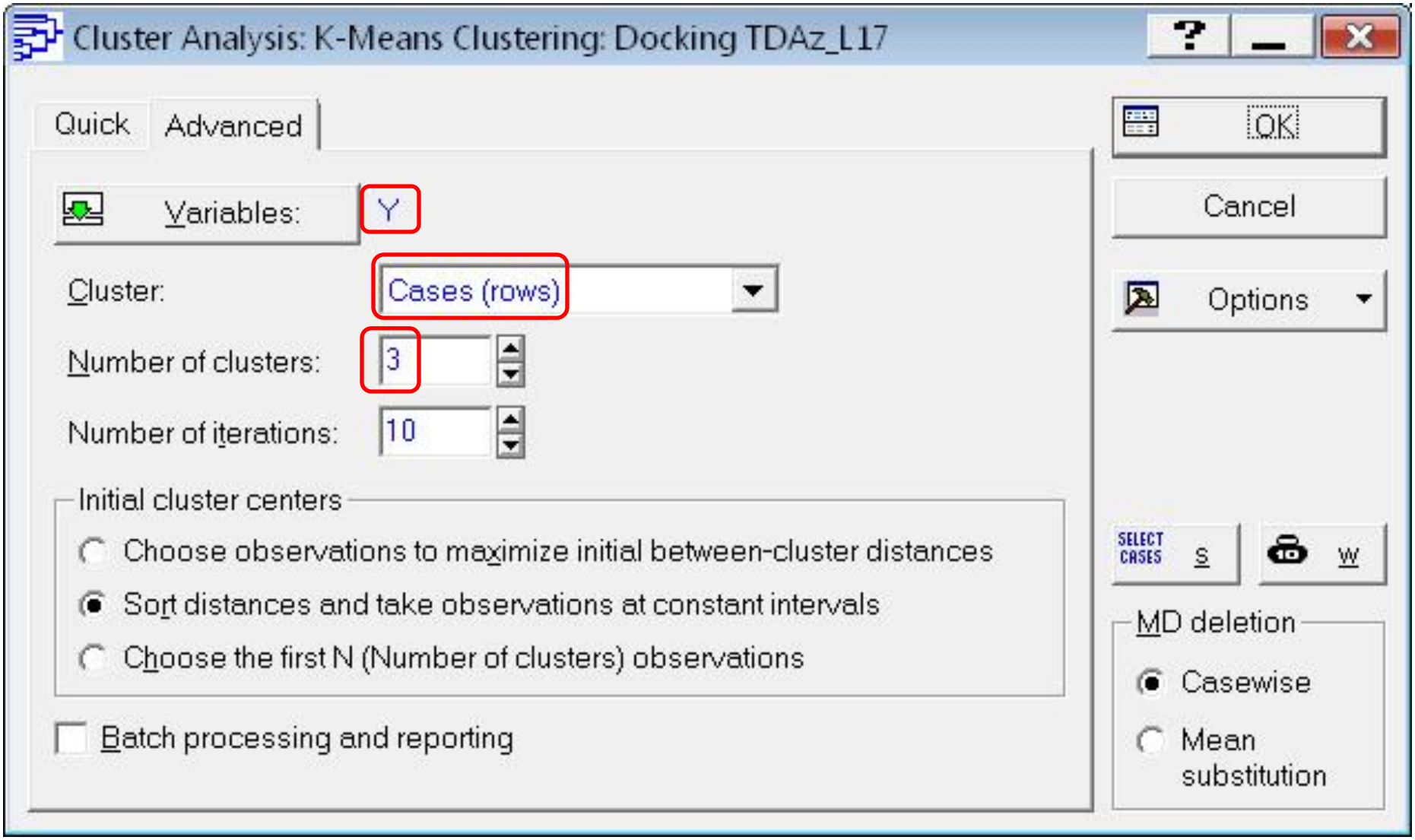
Кластеризация методом k-средних



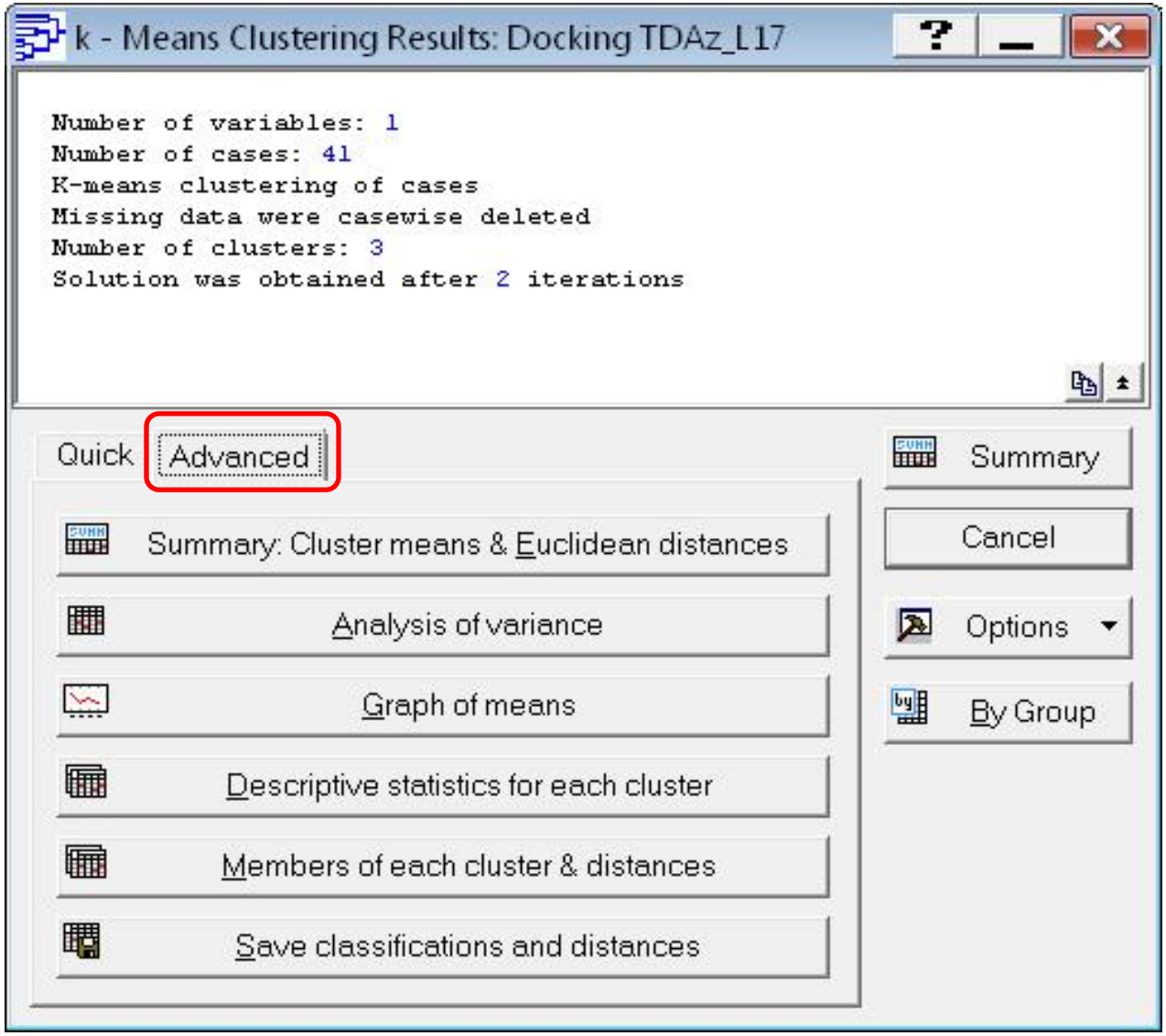
Кластеризация методом k-средних



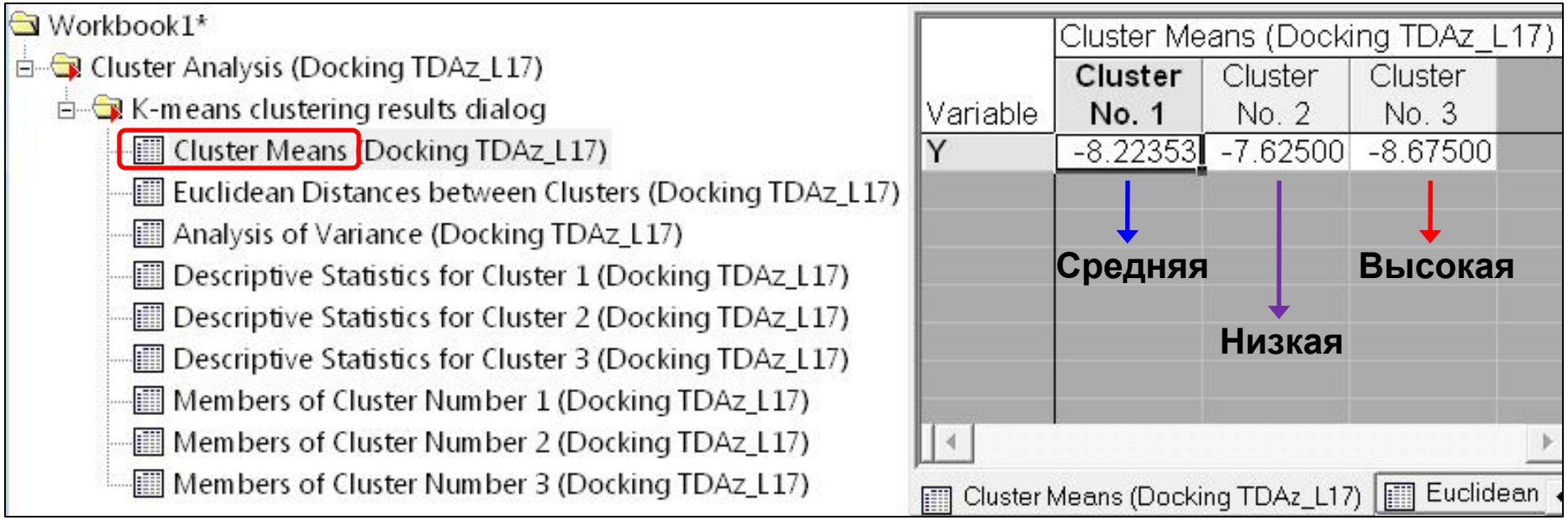
Кластеризация методом k-средних



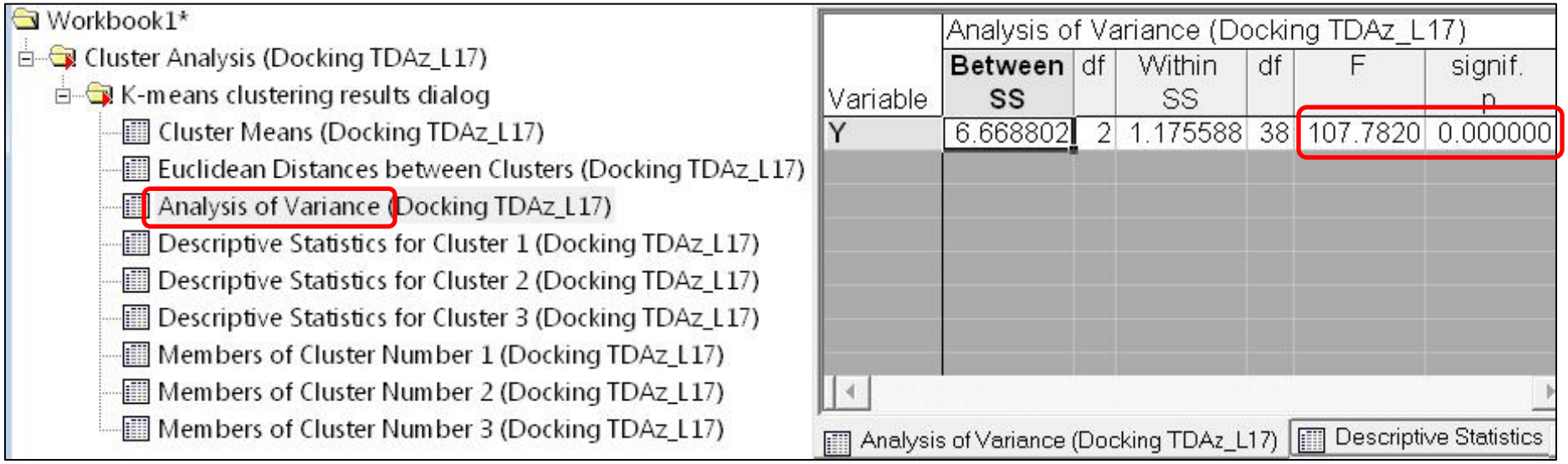
Кластеризация методом k-средних



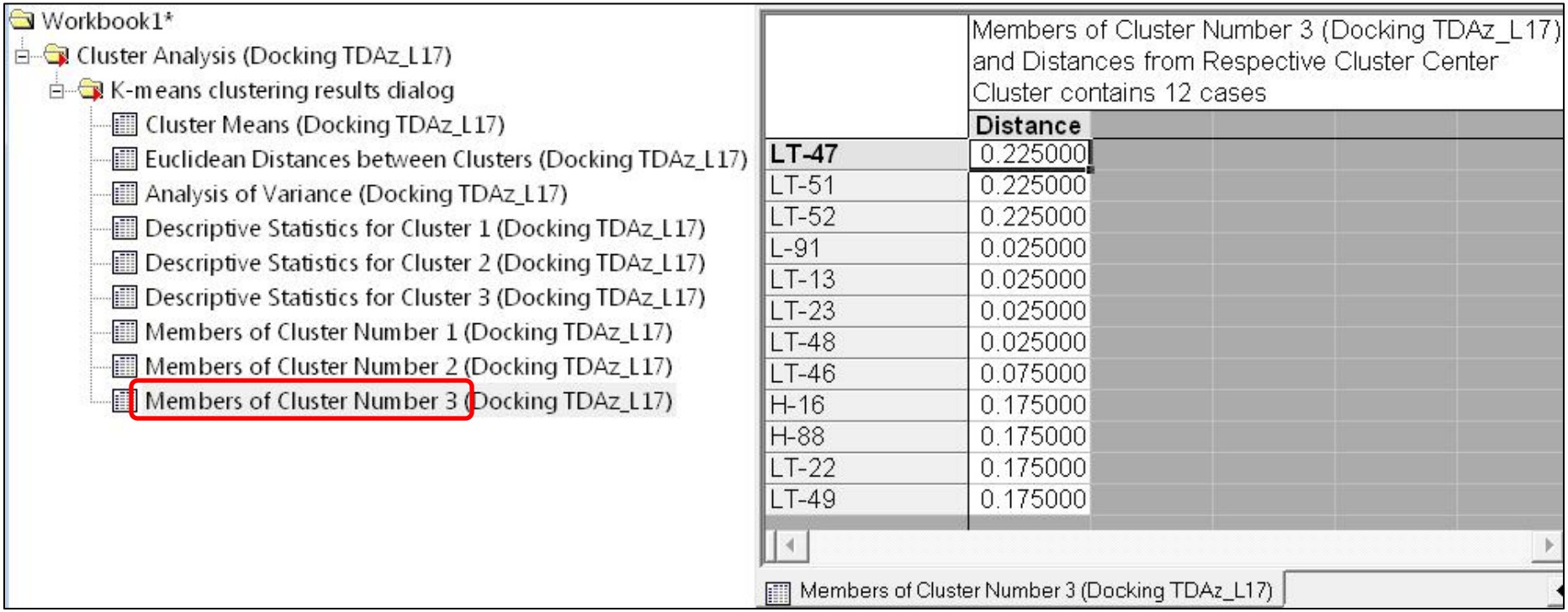
Кластеризация методом k-средних



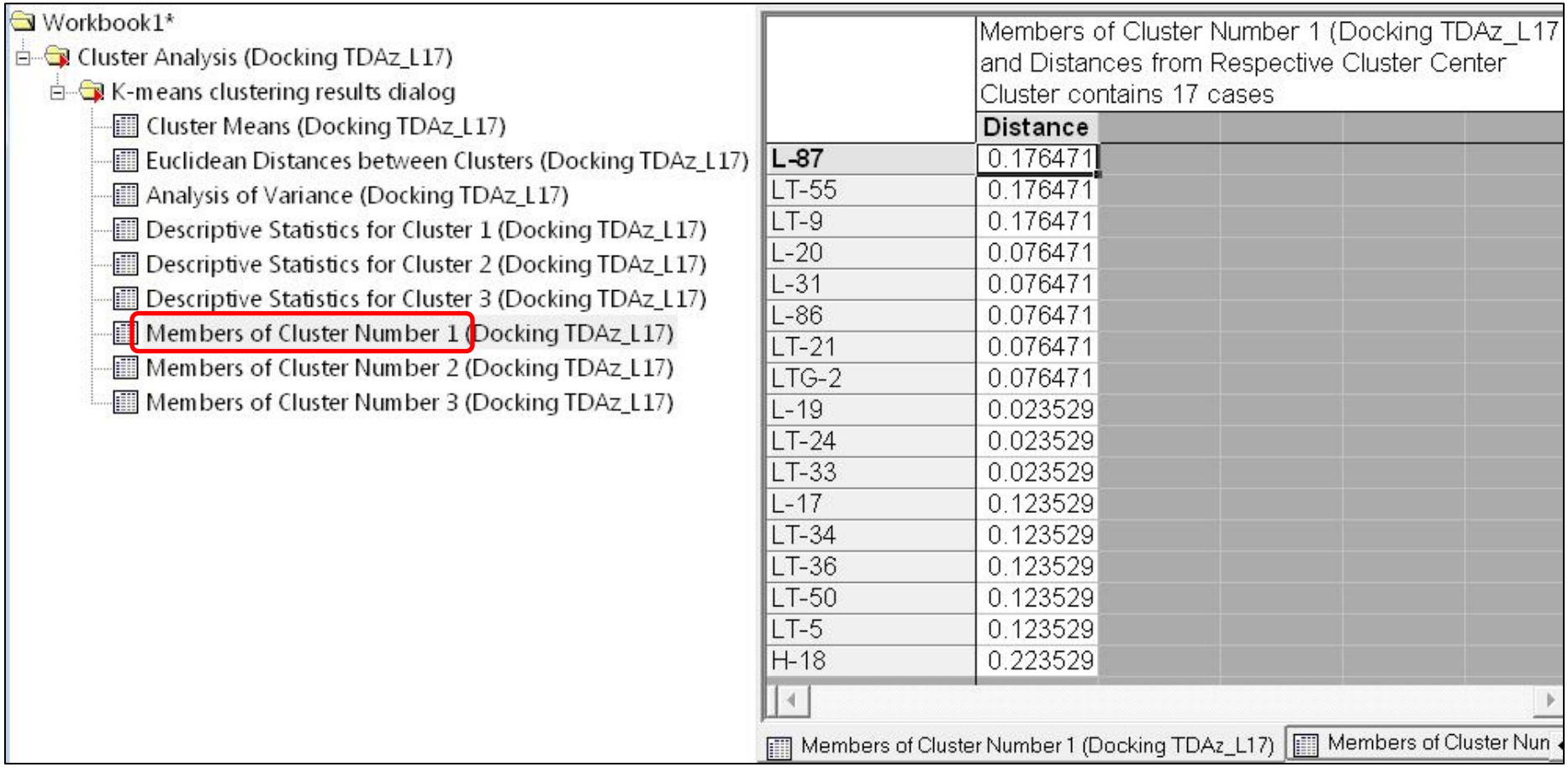
Кластеризация методом k-средних



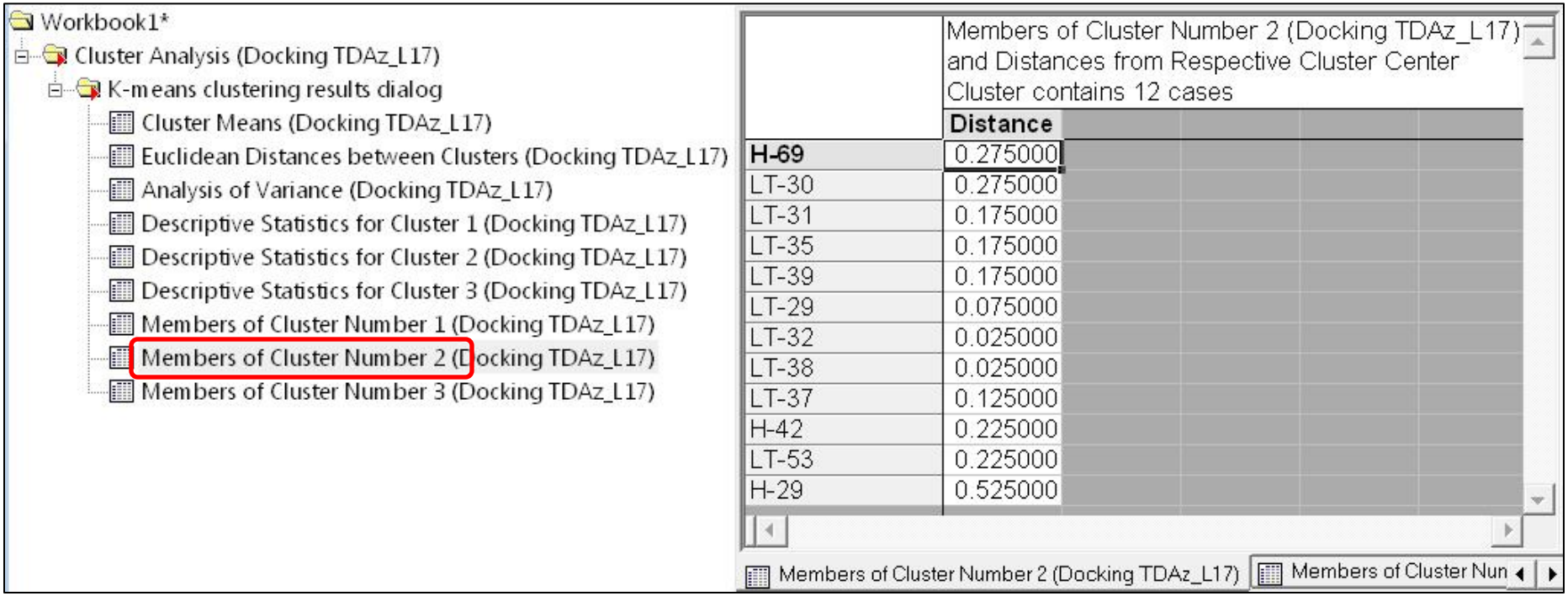
Кластеризация методом k-средних



Кластеризация методом k-средних



Кластеризация методом k-средних



Кластеризация методом k-средних

Границы кластеров

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

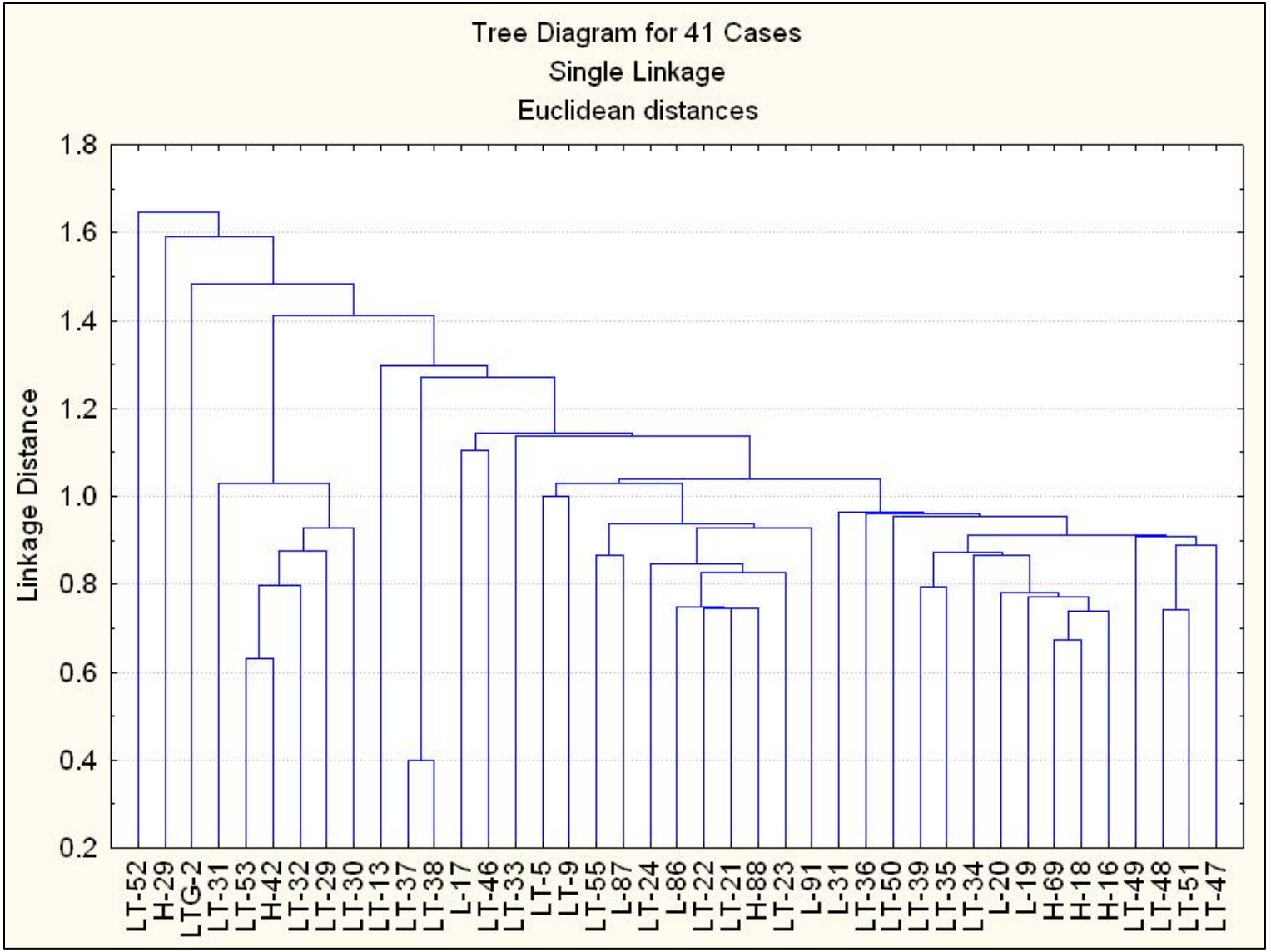
$[Y(L\text{-}18) + Y(L\text{-}69)] / 2 = -7.95$

Высокая $Y < -8.45$

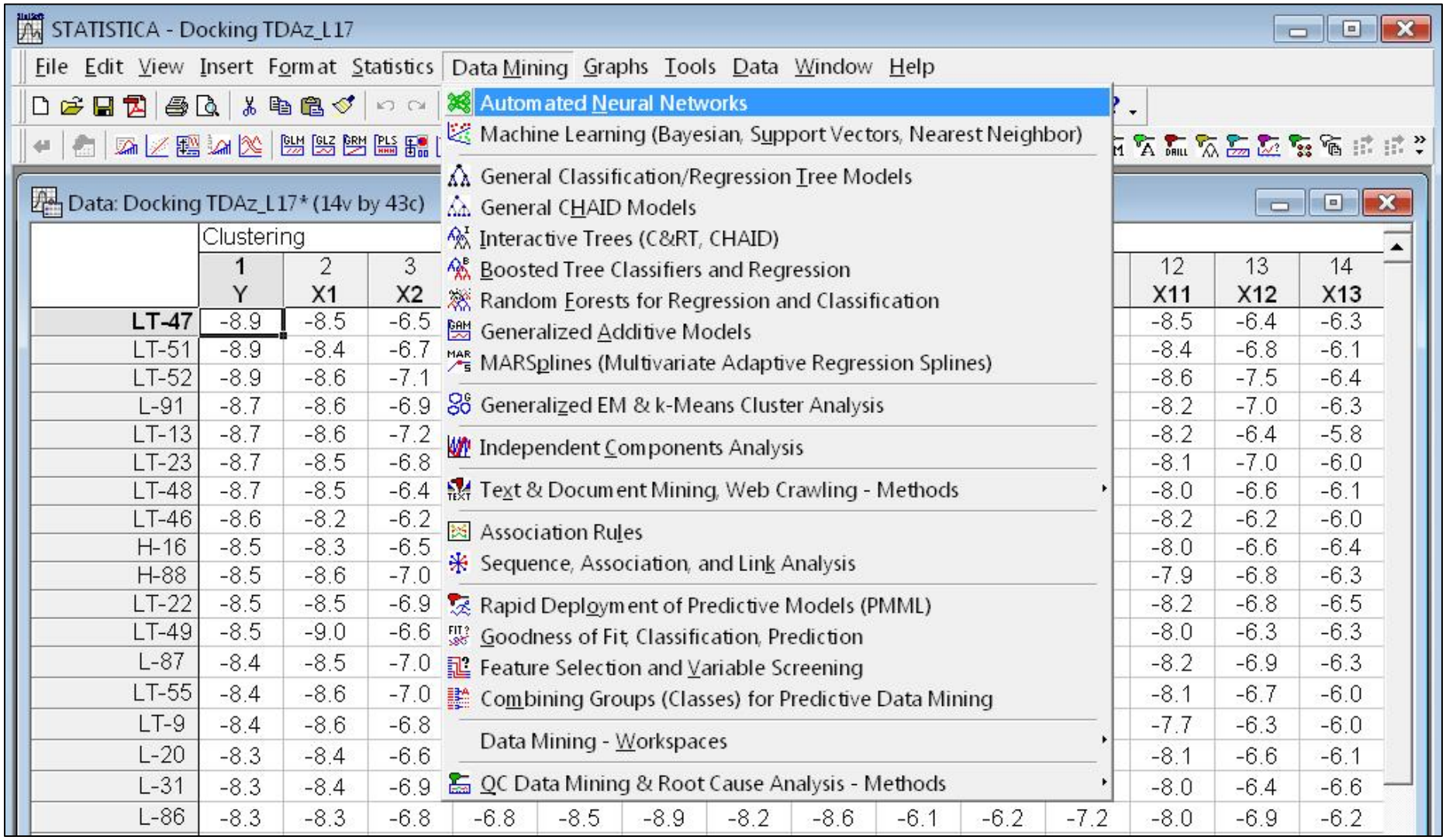
Умеренная $-8.45 \leq Y < -7.95$

Низкая $Y \geq -7.95$

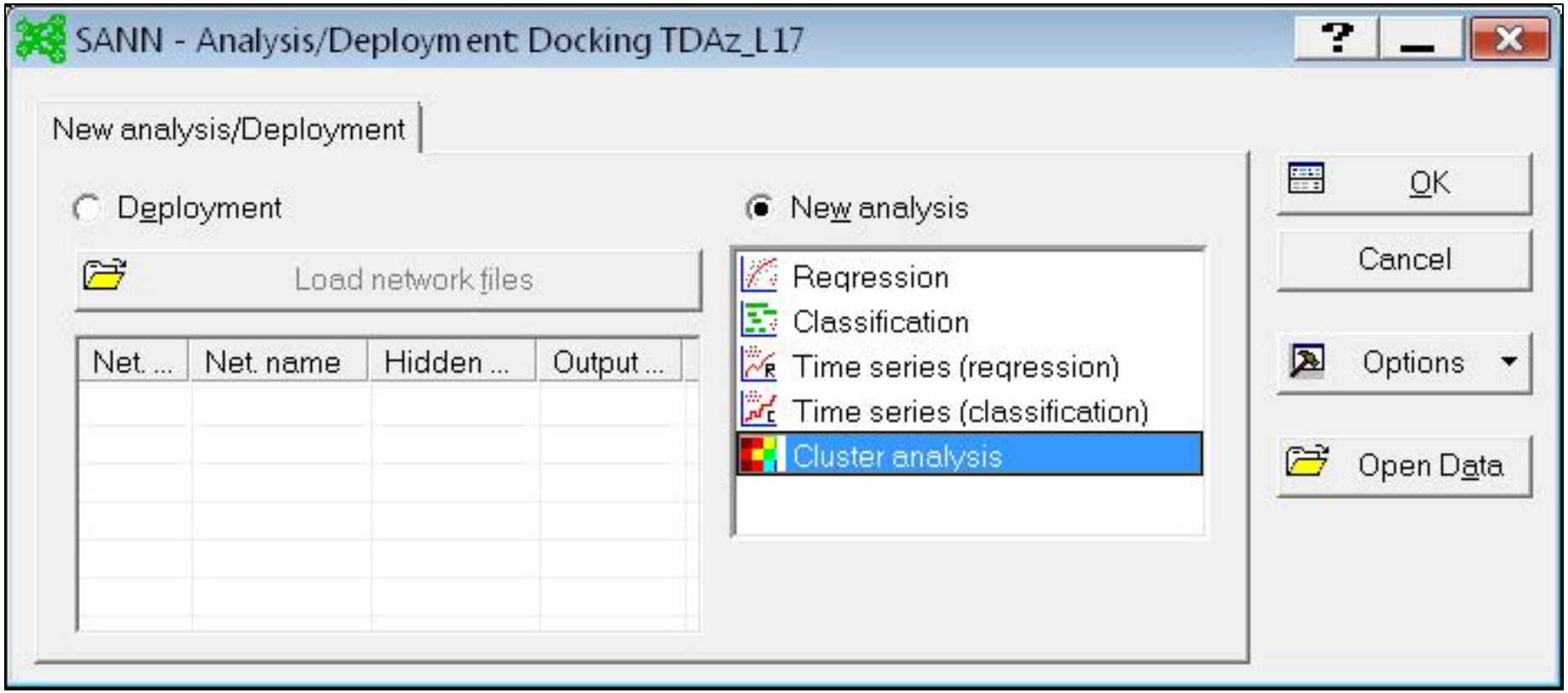
Иерархическая кластеризация



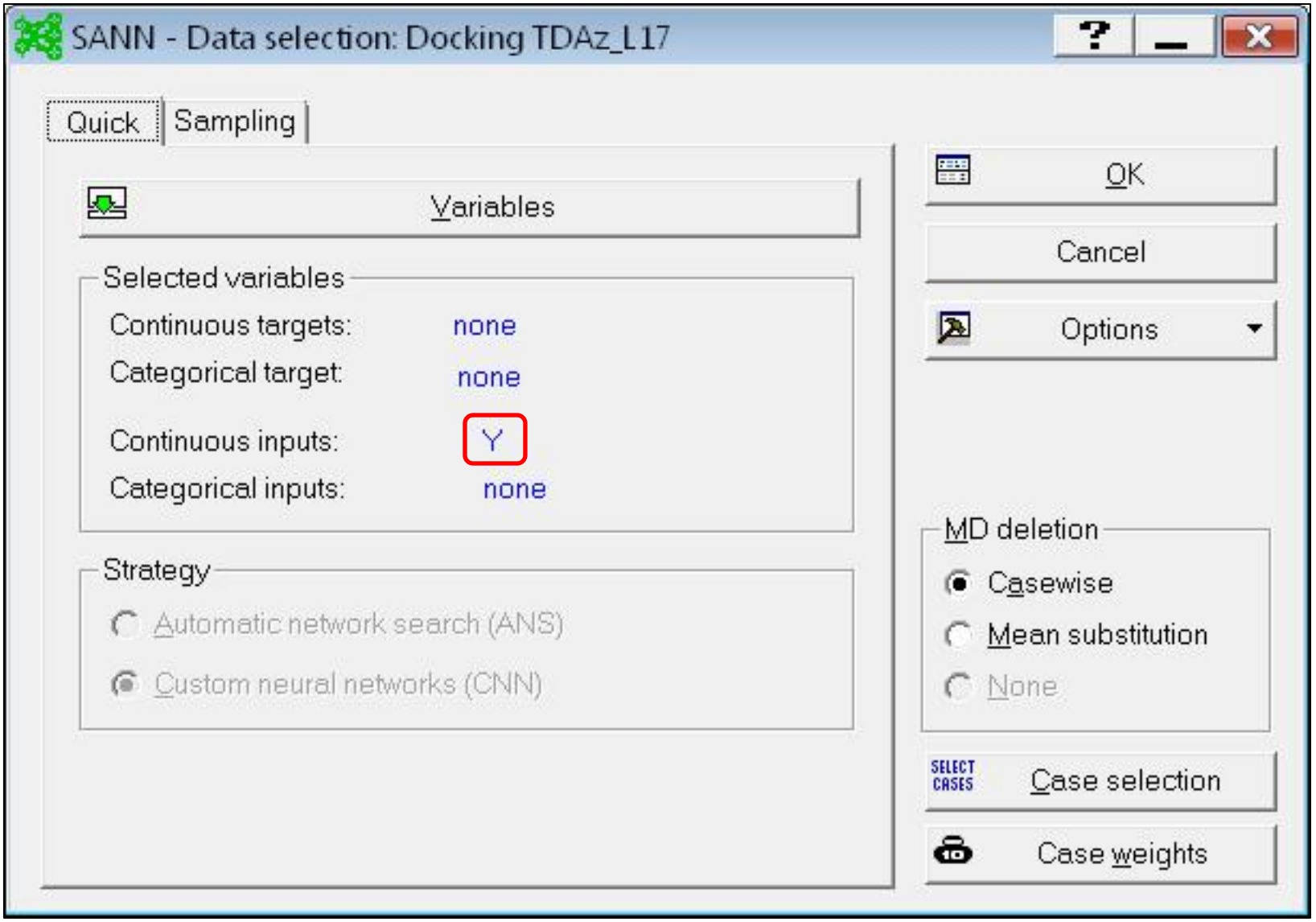
Самоорганизующиеся сети Кохонена



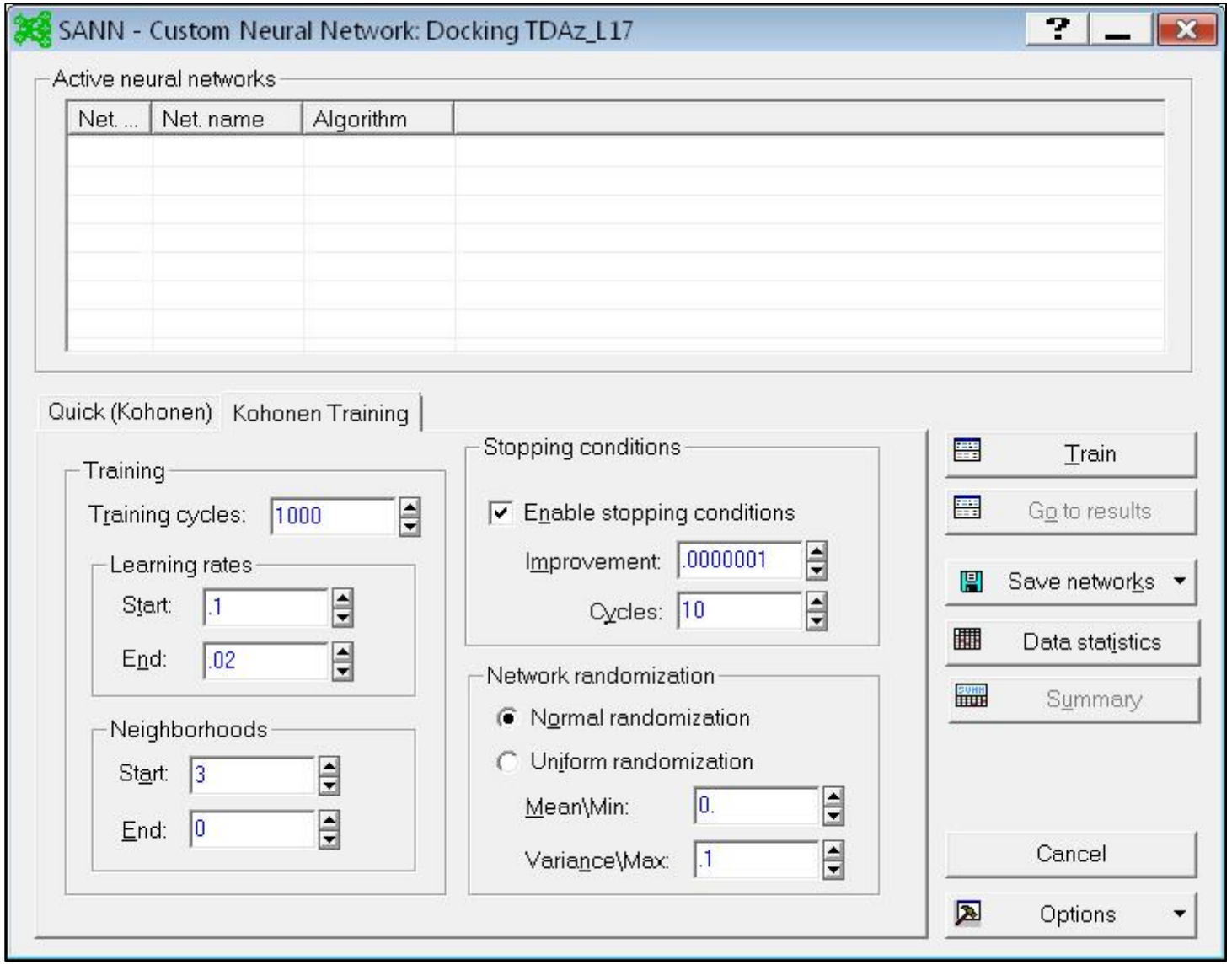
Самоорганизующиеся сети Кохонена



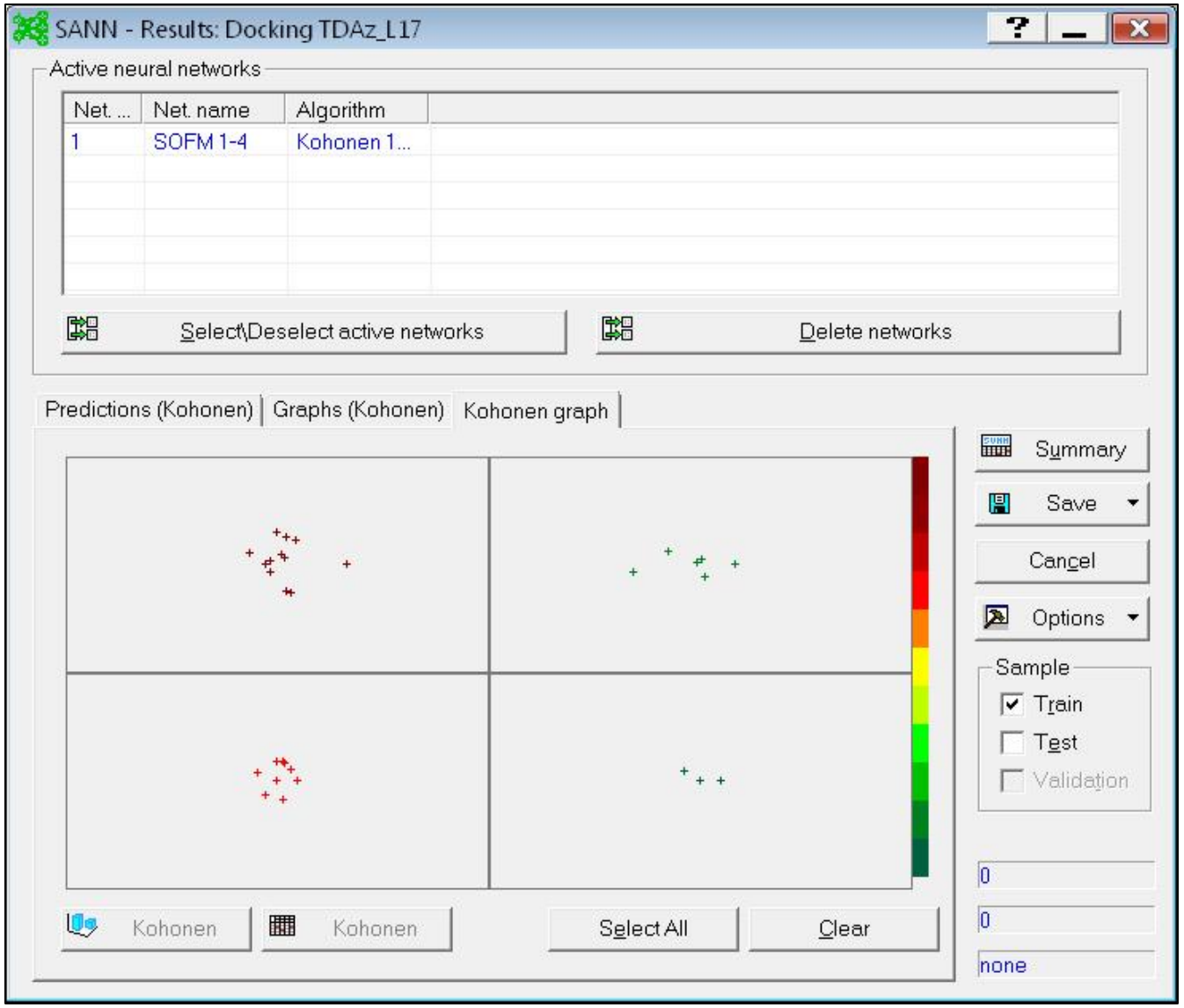
Самоорганизующиеся сети Кохонена



Самоорганизующиеся сети Кохонена



Самоорганизующиеся сети Кохонена




To be continued ...