**Липецкий государственный технический университет**

Факультет Автоматизации и Информатики

Кафедра прикладной математики

ЛАБОРАТОРНАЯ РАБОТА №5

по дисциплине «Статистические методы в прикладных задачах»

Кластерный анализ

Вариант 8

Студенты Станиславчук Сергей

Группа АС-21-1 Коновалов Кирилл

Руководитель

доцент Рыжкова Д.В.

Липецк, 2023 г

# Задание кафедры

Задание 1. По данным Таблицы 1 и алгоритму кластерного анализа провести классификацию объектов иерархическим методом (древовидная кластеризация).

Таблица 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| № п/п. | Страны | Число врачей на 10000 населения | Смертность на 100000 населения | ВВП по паритету покупательной способности, в % к США | Расходы на здравоохранение, в % к США |
| X1 | X2 | X3 | X4 |
| 1 | Россия | 44.5 | 84.98 | 20.4 | 3.2 |
| 2 | Австралия | 32.5 | 30.58 | 71.4 | 8.5 |
| 3 | Австрия | 33.9 | 38.42 | 78.7 | 9.2 |
| 4 | Азербайджан | 38.8 | 60.34 | 12.1 | 3.3 |
| 5 | Армения | 34.4 | 60.22 | 10.9 | 3.2 |
| 6 | Беларусь | 43.6 | 60.79 | 20.4 | 5.4 |
| 7 | Бельгия | 41 | 29.82 | 79.7 | 8.3 |
| ~~8~~ | ~~Болгария~~ | ~~36.4~~ | ~~70.57~~ | ~~17.3~~ | ~~5.4~~ |
| 9 | Великобритания | 17.9 | 34.51 | 69.7 | 7.1 |
| 10 | Венгрия | 32.1 | 64.73 | 24.5 | 6 |
| 11 | Германия | 38.1 | 36.63 | 76.2 | 8.6 |
| 12 | Греция | 41.5 | 32.84 | 44.4 | 5.7 |
| 13 | Грузия | 55 | 62.64 | 11.3 | 3.5 |
| 14 | Дания | 36.7 | 34.07 | 79.2 | 6.7 |
| 15 | Ирландия | 15.8 | 39.27 | 57 | 6.7 |
| 16 | Испания | 40.9 | 28.46 | 54.8 | 7.3 |
| 17 | Италия | 49.4 | 30.27 | 72.1 | 8.5 |
| 18 | Казахстан | 38.1 | 69.04 | 13.4 | 3.3 |
| 19 | Канада | 27.6 | 25.42 | 79.9 | 10.2 |
| 20 | Киргизия | 33.2 | 53.13 | 11.2 | 3.4 |

Задание 2. Решить Задание 1, предварительно нормировав исходные данные.

Задание 3*.* Решить Задание 1 при условии, что расстояния между кластерами измеряются по принципу “дальнего соседа”, предварительно нормируя исходные данные.

Задание 4*.* Решить Задание 1, но в качестве расстояния между объектами принять “расстояние городских кварталов (Манхэттенское расстояние)”, а расстояния между кластерами измерять по методу Варда. Не нормируя предварительно исходные данные.

Задание 5*.* Решить Задание 1 методом К-средних. Предварительно нормируя исходные данные.

# Ход выполнения лабораторной работы

Задание 1. По данным таблицы 1, за исключением 8-ой строки, и алгоритму кластерного анализа проведем классификацию объектов иерархическим методом (древовидная кластеризация). В качестве расстояния между объектами принимаем “обычное евклидово расстояние”, а расстояния между кластерами измеряем по принципу “ближайшего соседа”.

Исходные данные не нормируем.

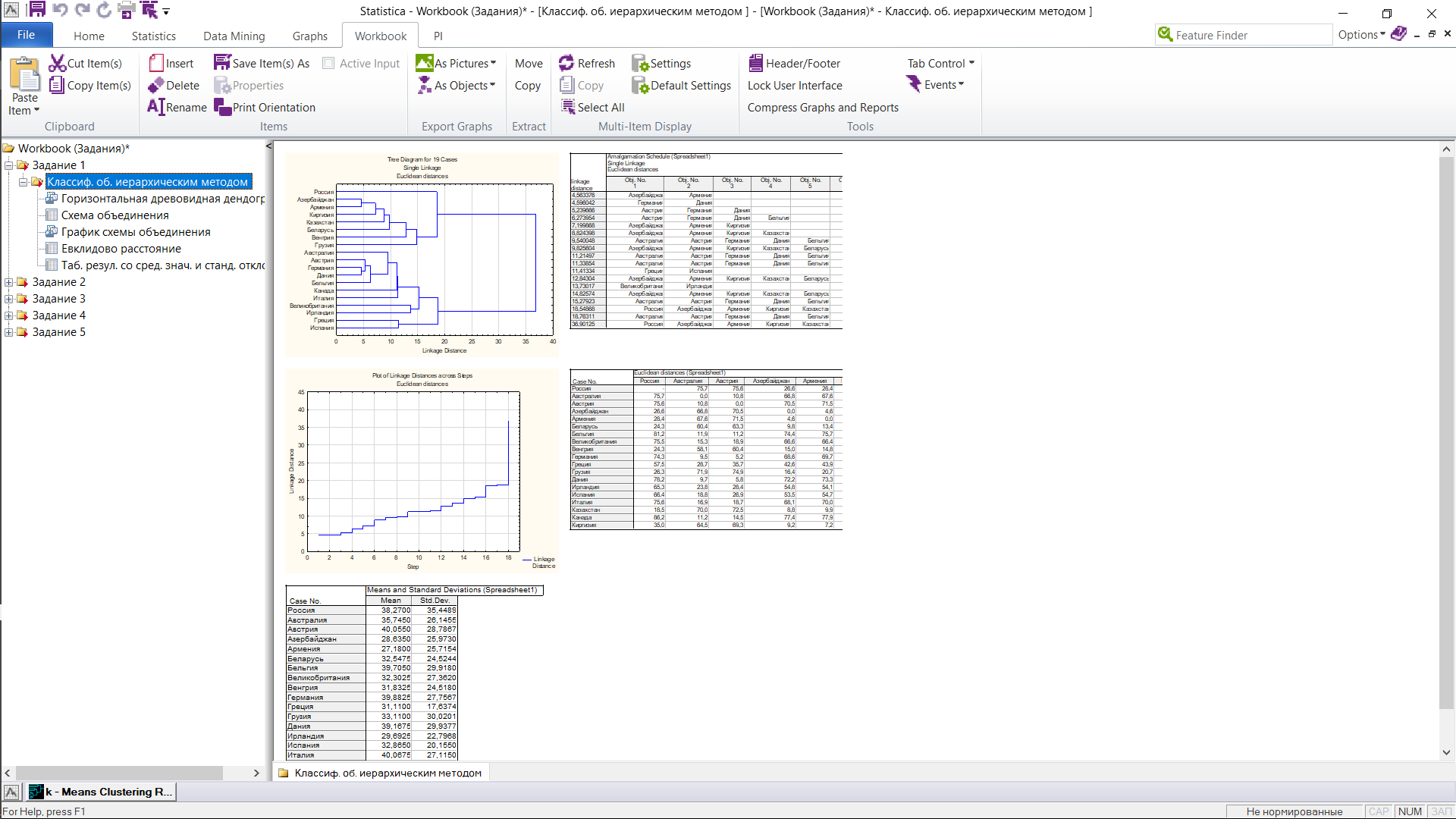
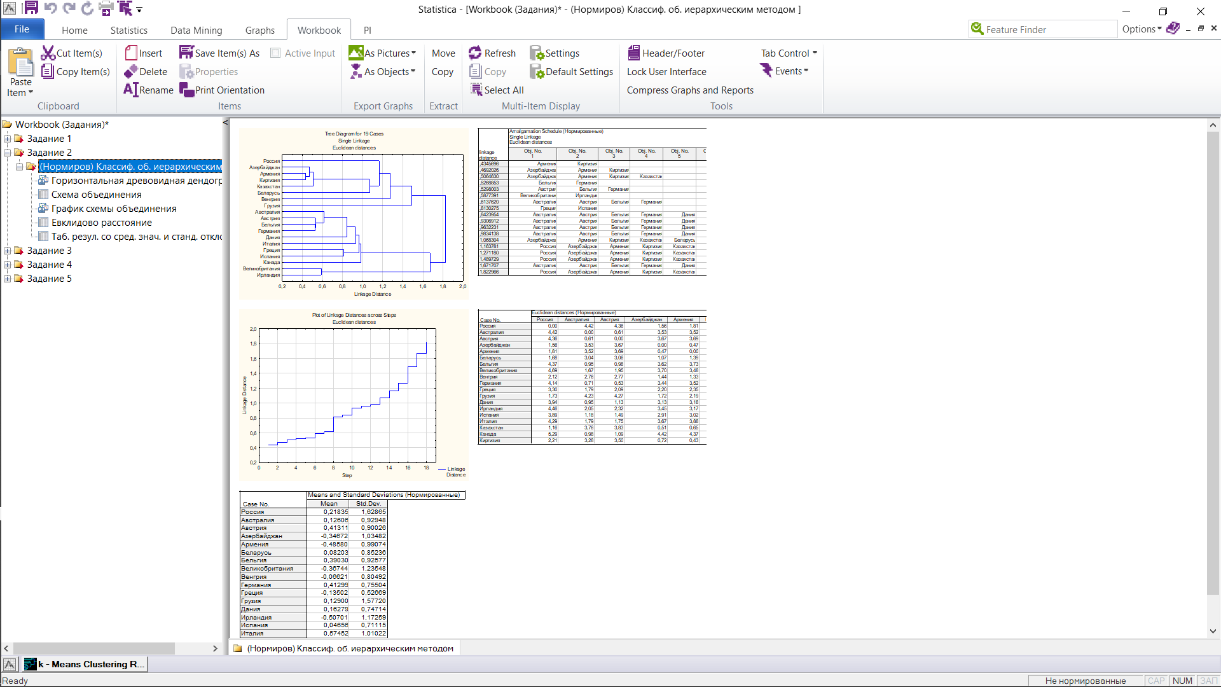


Рисунок 1 - Классификация объектов иерархическим методом (не нормированные данные)

Задание 2. По данным таблицы 1, за исключением 8-ой строки, и алгоритму кластерного анализа проведем классификацию объектов иерархическим методом (древовидная кластеризация). В качестве расстояния между объектами принимаем “обычное евклидово расстояние”, а расстояния между кластерами измеряем по принципу “ближайшего соседа”.

Исходные данные нормируем.

Рисунок 2 - Классификация объектов иерархическим методом (нормированные данные)

Задание 3. По данным таблицы 1, за исключением 8-ой строки, и алгоритму кластерного анализа проведем классификацию объектов иерархическим методом (древовидная кластеризация). В качестве расстояния между объектами принимаем “обычное евклидово расстояние”, а расстояния между кластерами измеряем по принципу “дальнего соседа”.

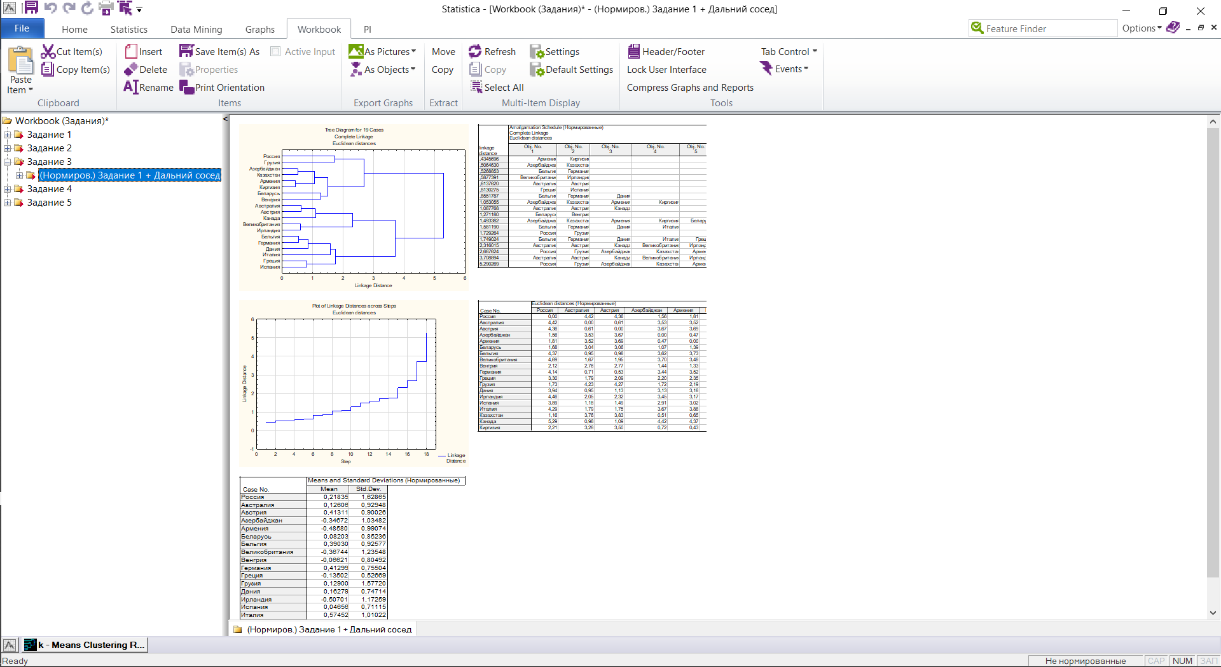
Исходные данные нормируем.

Рисунок 3 - Классификация объектов иерархическим методом с изменениями (нормированные данные)

Задание 4. По данным таблицы 1, за исключением 8-ой строки, и алгоритму кластерного анализа проведем классификацию объектов иерархическим методом (древовидная кластеризация). В качестве расстояния между объектами принимаем “расстояние городских кварталов (Манхэттенское расстояние)”, а расстояния между кластерами измеряем по методу Варда.

Исходные данные не нормируем.

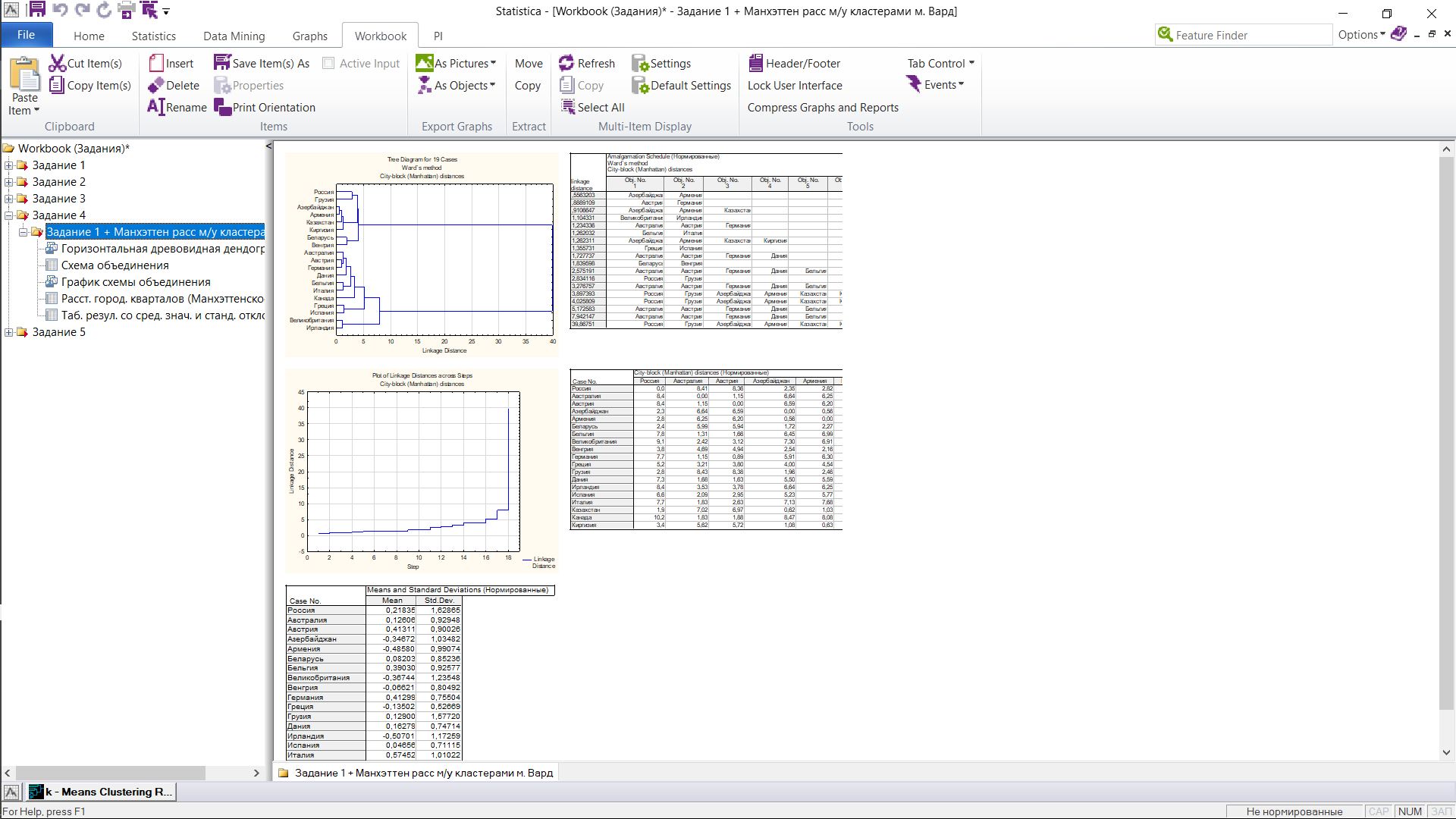


Рисунок 4 - Классификация объектов иерархическим методом с изменениями (не нормированные данные)

Задание 5. По данным таблицы 1, за исключением 8-ой строки, и алгоритму кластерного анализа проведем классификацию объектов методом K-средних.

Исходные данные нормируем.

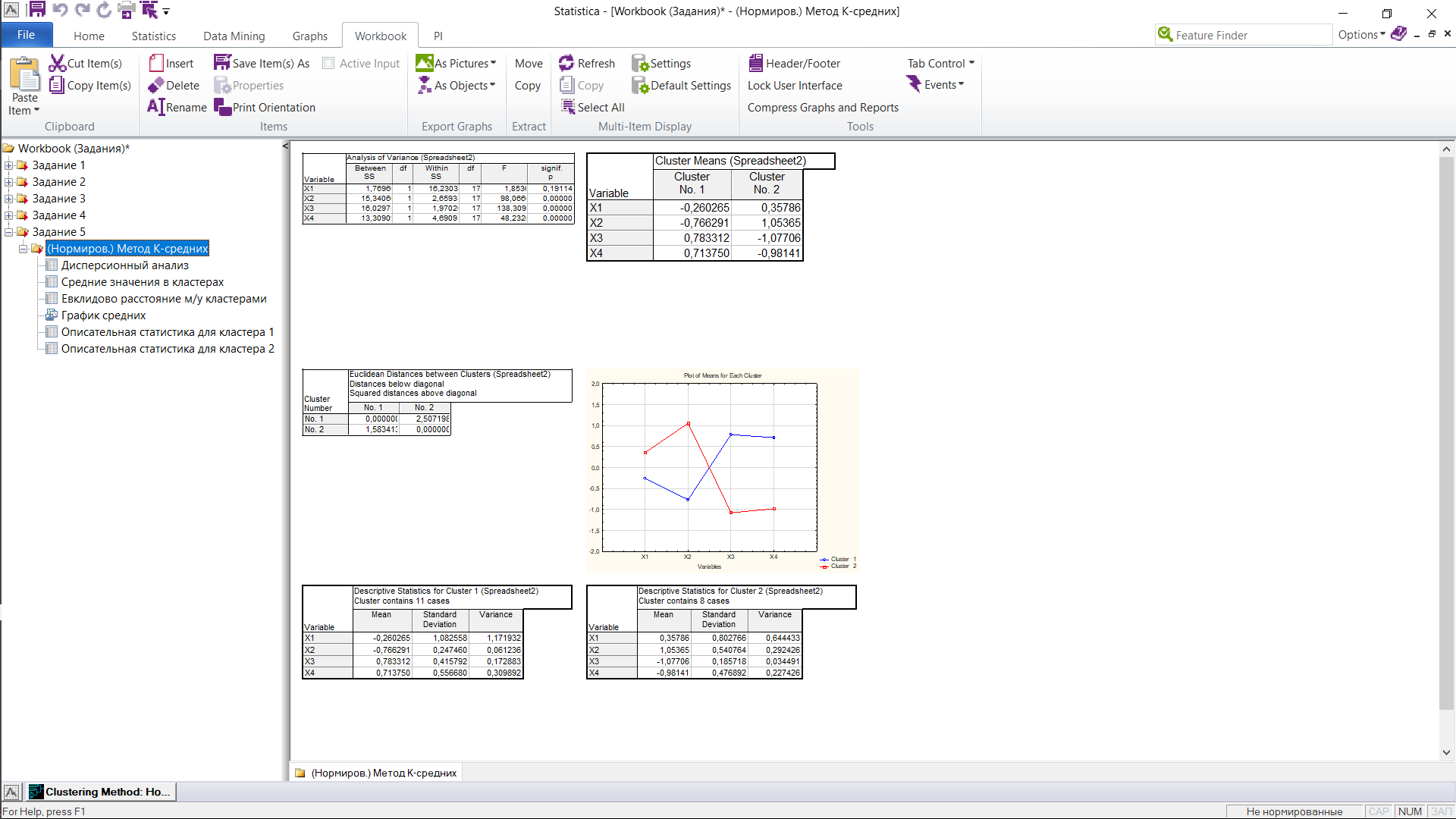


Рисунок 5 - Классификация объектов методом K-средних (нормированные данные)

Вывод:

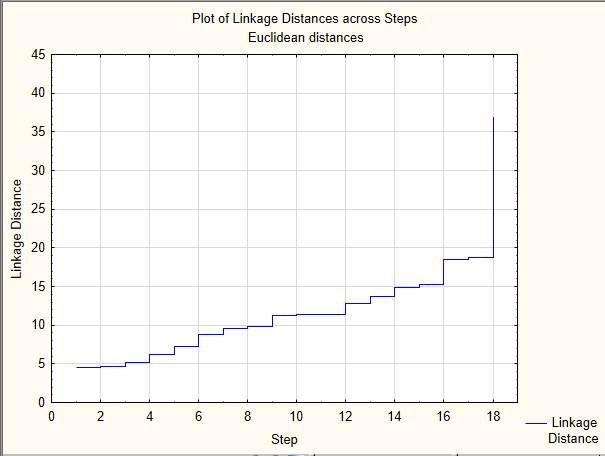
1. В результате выполненного задания 1 можно выделить 2 образованных больших кластера(по схемам объединения и дендограмме) – [Россия, Азербайджан, Армения, Киргизия, Казахстан, Беларусь, Венгрия, Грузия] и [Австралия, Австрия, Германия, Дания, Бельгия, Канада, Италия, Великобритания, Ирландия]. Это значит, что в каждом кластере данные страны имеют похожие показатели переменных X. Данные выводы можно сделать исходя из того, что нам предоставили схема объединения и дендограмма. Матрица евклидовых дистанций лишь описала это в конкретных значениях, взятых и посчитанных из переменных, принадлежащим этим странам (что для нас выглядит не очень и наглядно). А таблица средних значений и средних отклонений показала, что страны второго выделенного кластера имеют значения в среднем на 9 единиц выше, чем страны первого кластера!

**Таким образом**, Первый кластер стран [Россия, Азербайджан, Армения, Киргизия, Казахстан, Беларусь, Венгрия, Грузия] можно интерпретировать как кластер, включающий страны Евразийского региона. Эти страны действительно имеют схожие географические, исторические и культурные характеристики. Они расположены вблизи друг от друга и, вероятно, имеют схожие социально-экономические условия, политические связи или торговые отношения.

Второй кластер стран [Австралия, Австрия, Германия, Дания, Бельгия, Канада, Италия, Великобритания, Ирландия] можно интерпретировать как кластер развитых индустриальных стран. Эти страны обычно характеризуются высоким уровнем экономического развития, развитой инфраструктурой, высоким уровнем жизни и образования. Они являются членами организаций, таких как Европейский союз или Организация экономического сотрудничества и развития, и у них может быть схожий подход к экономическим, политическим и социальным вопросам.

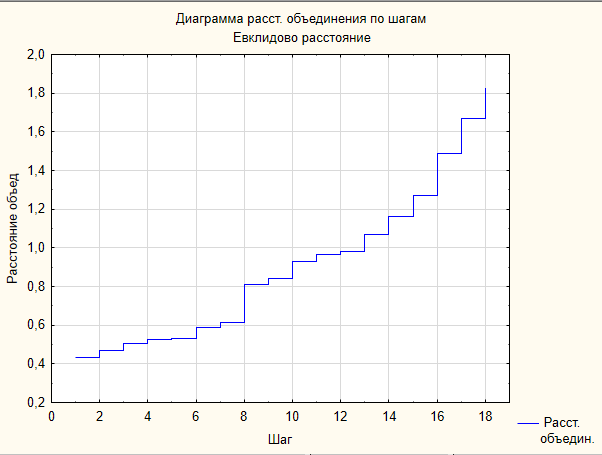
Оба кластера представляют собой группы стран с схожими характеристиками, но различаются в территориальном расположении и уровне развития. Кластерный анализ помогает выделить такие группы и понять, какие страны могут иметь схожие свойства или взаимодействия.

2. При нормированных наблюдениях ситуация выглядит немного иначе (лучше). Я специально умолчал о диаграмме расстояний объединения по шагам в первом задании, чтобы упомянуть об этом сейчас. Давайте теперь взглянем на неё:



Видим, что наблюдается сильный скачок в дистанции у 18-ого шага, а остальные шаги не выглядят действительно существенными (малюсенькие шажочки). Это может говорить о том, что данные (переменные), принадлежащие странам либо слишком сильно отличаются, либо не отличаются вообще (на коротких промежутках).

Давайте взглянем, что стало с этим графиком при использовании нормализованных данных:



Это уже больше напоминает нормальную лестницу (хоть и не везде идеально, но разница с первым заданием ощутима). Это говорит нам о том, что теперь расстояние объединений, в целом, будет шире постепенно (но уверенно), с каждым шагом. Следовательно, дальнейший анализ будет более объективным, так как евклидово расстояние приобрело более адекватную форму (без всяких резонансных всплесков или долгих топтаний на месте).

Нормализация никак не сказалась на тех самых двух кластерах, выделенных в первом задании. Зато, сказалась на самых маленьких кластерах, например:  
Армения теперь имеет намного больше общего с Киргизией, нежели с Азербайджаном или вот: Германия раньше была в одном маленьком кластере с Данией, теперь она с Бельгией. Можно долго перечислять изменения, но одно можно сказать точно: нормализация данных имеет значительное влияние на формирование мелких кластеров и может изменять относительные расстояния и сходства между странами. Однако, в случае с более крупными кластерами, влияние нормализации может быть менее заметным или незначительным!

3. Рассматривая полученную дендограмму, она мне показалась более наглядной, чем дендограмма ближнего соседа. При использовании метода полной связи (дальнего соседа), расстояние между кластерами определяется как наибольшее расстояние между любыми двумя точками в разных кластерах. Это означает, что в дендрограмме будут отображаться длинные вертикальные линии, которые представляют собой расстояния между кластерами. Таким образом, дендрограмма метода полной связи демонстрирует явное разделение кластеров и их удаление друг от друга.

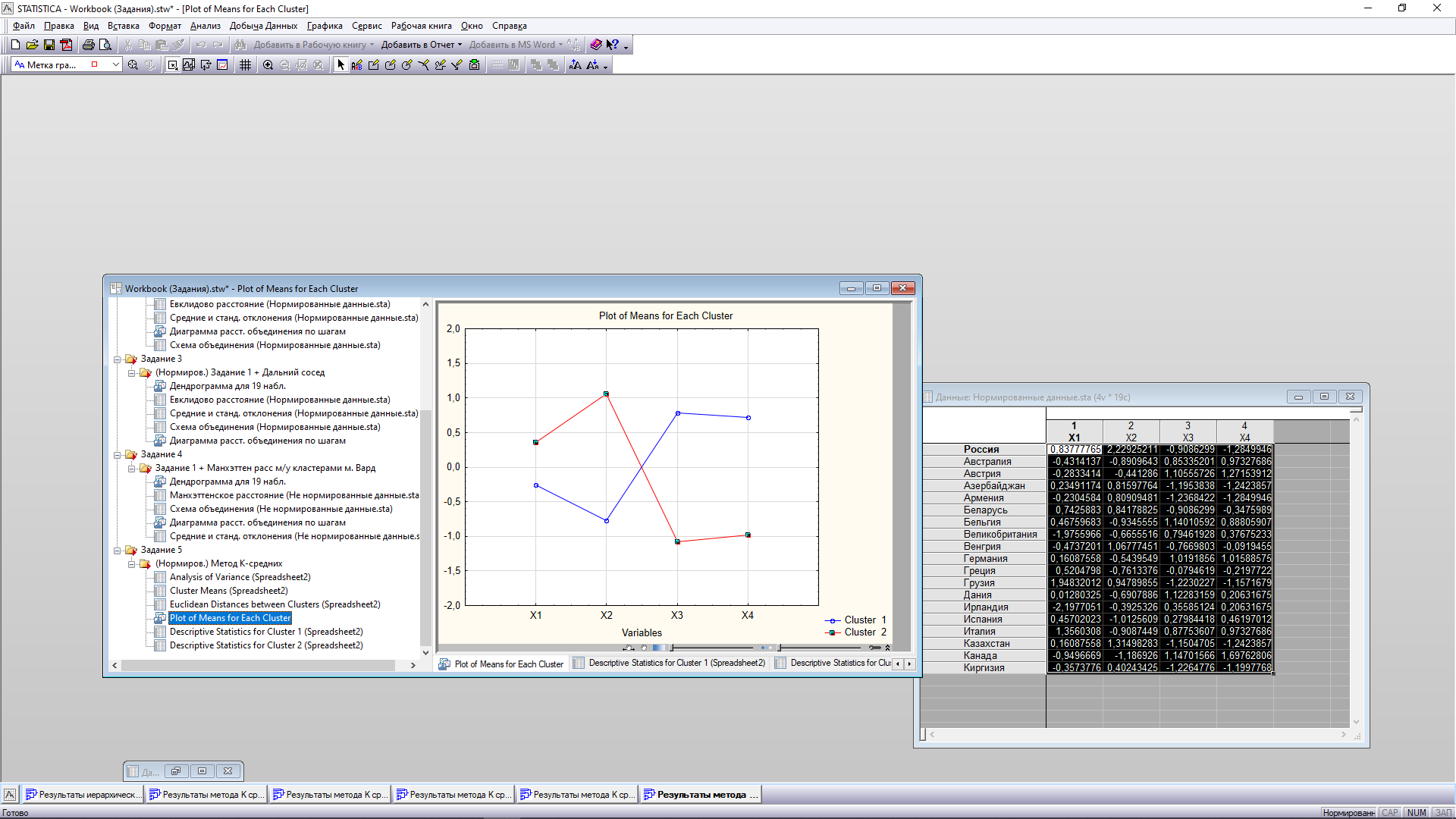
С другой стороны, при использовании метода одиночной связи (ближайшего соседа), расстояние между кластерами определяется как расстояние между самыми близкими точками в разных кластерах. Это может приводить к созданию более длинных и сложных вертикальных линий на дендрограмме, так как ближайшие соседи в разных кластерах могут иметь большое расстояние между ними. В результате, дендрограмма метода одиночной связи может выглядеть менее наглядно и более запутанно.

В остальном – кластеры практически такие же, как и во втором задании.

4. Данный метод отобрал кластеры точно также, как и в первом задании. Дендограмма достаточно визуально понятная. Тут также присутствует большой шаг в диаграмме расстояний (так как в конце оба огромных кластера объединяются, хотя они слишком разные, поэтому и возникает такой большой шаг). Таким образом, метод Варда склонен формировать более компактные и однородные кластеры, тогда как метод одиночной связи может приводить к более разреженным и разнообразным кластерам. Каждый метод имеет свои преимущества и ограничения, и выбор метода зависит от характеристик данных и целей анализа.

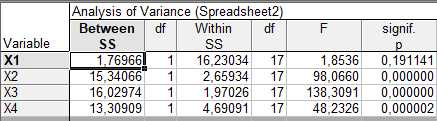
5. У этого метода уже нет дендограмм, заместо них используются графики, непрерывных прототипов (переменных)

На основе прошлых измерений можно уже предположить каким странам принадлежат кластер 1 и 2. Большое количество врачей и высокая смертность относятся к странам Евразийского региона, а высокое ВВП и расходы на здравоохранение к более западным странам.

\

Особенность метода в том, что мы можем указывать желаемое количество кластеров, вследствие чего, вероятно, наши новые кластеры будут занимать какие-нибудь промежуточные значения.

Анализ дисперсии:



"Число врачей на 10000 населения" равно 0,1911414, что больше уровня значимости 0,05. Это означает, что F-статистика для этой переменной не является статистически значимой, и мы не можем сделать вывод о наличии статистически значимых различий между группами по этой переменной.

В то же время, для переменных "Смертность на 100000 населения", "ВВП по паритету покупательной способности, в % к США" и "Расходы на здравоохранение, в % к США" значения p равны соответственно 0,00000001786194, 0,000000001368919 и 0,000002365992, что гораздо меньше уровня значимости 0,05. Это указывает на наличие статистически значимых различий между группами по этим переменным.

Было интересно!