Кластерный анализ в программе STATISTICA

ЛабРаб 2

Данные Росстат

Обновлено 29.08.2017

Продажа алкогольных напитков населению

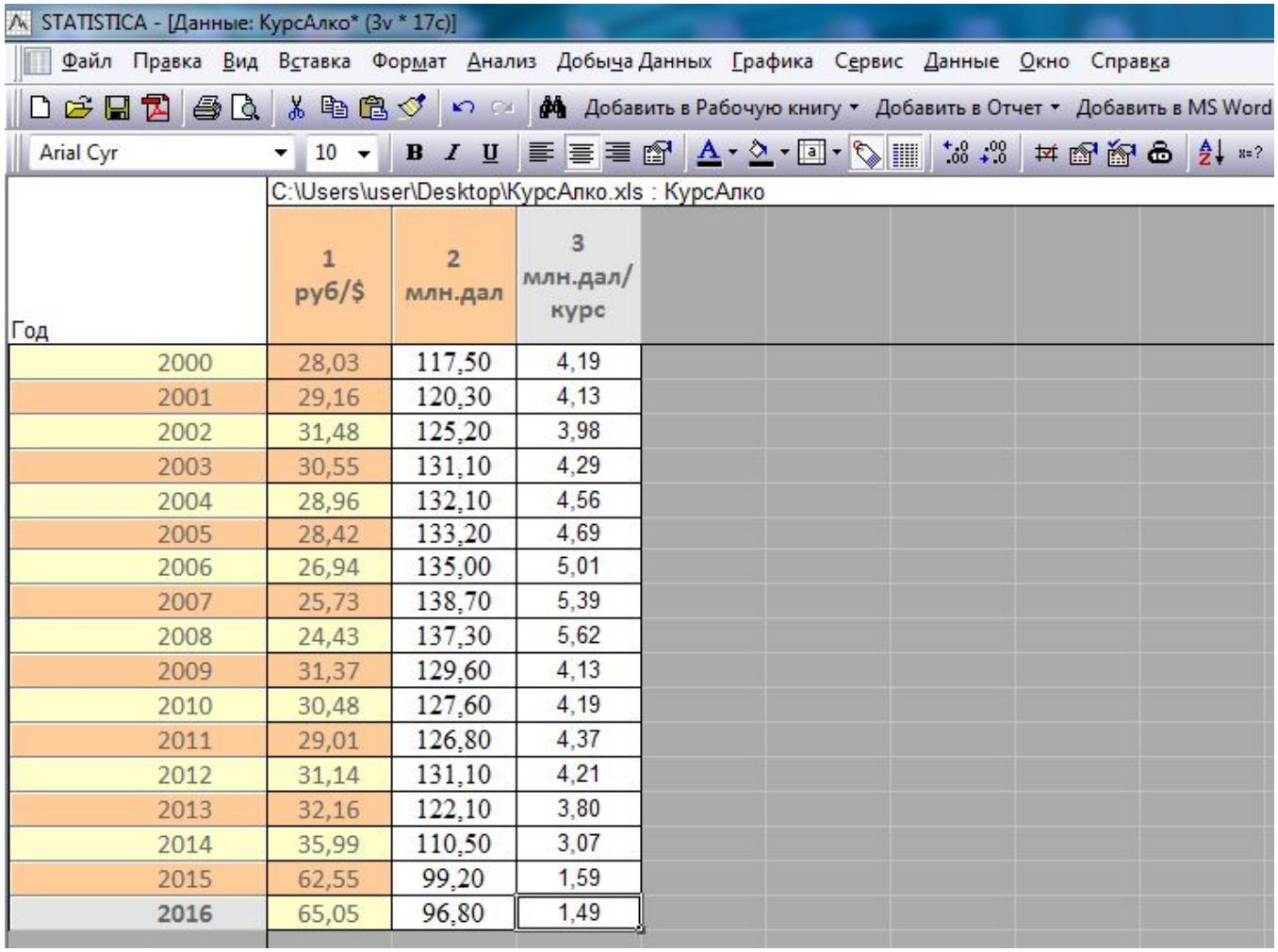
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
В абсолютном алкоголе:																	
всего, млн. дкл	118	120	125	131	132	133	135,0	139	137	130	128	127	131,1	122,1	110,5	99,2	96,8
на душу населения, л	8,0	8,2	8,6	9,1	9,2	9,3	9,4	9,7	9,6	9,1	8,9	8,9	9,2	8,5	7,6	6,8	6,6
В натуральном выражении, млн. дкл:																	
водка и ликероводочные изделия	**215**	**209**	**211**	**216**	**209**	**204**	**197**	**185**	**177**	**166**	**158**	**156**	**153,0**	**133,6**	**112,4**	**97,2**	**96,6**
коньяки, коньячные напитики (включая бренди, кальвадосы)	4,1	4,5	5,2	5,9	6,2	6,8	7,6	8,9	10,8	10,6	11,1	11,6	12,4	12,1	11,5	10,6	10,3
напитки слабоалкогольные (с содержанием этилового спирта не более 9%)	-	-	-	-	-	-	-	-	-	-	31,9	31,4	26,9	23,3	17,8	10,0	8,8
винодельческая продукция (без вин шампанских и игристых) [1)]	52,4	56,7	62,6	72,2	77,7	84,9	81,1	94,9	103	103	103	97,1	93,6	83,6	90,4	86,9	84,9
из нее вино	-	-	-	-	-	-	-	-	-	-	-	-	61,8	60,2	57,8	53,6	51,9
шампанские и игристые вина	18,3	18,5	17,7	18,7	18,5	19,4	20,6	24,1	26,0	25,5	27,3	28,5	28,3	27,7	26,4	23,6	22,0
пиво, кроме коктейлей пивных и напитка солодового	524,6	634,6	707,8	762,5	844,7	892,1	1003	1155	1138	1025	1004,0	1011,5	1017,5	984,2	895,9	810,2	780,6

Данные ЦБ РФ

Динамика курса доллара

№№	Год	руб/$
1	1998	6,206
2	1999	24,55
3	2000	28,03
4	2001	29,16
5	2002	31,4758
6	2003	30,5547
7	2004	28,9641
8	2005	28,4244
9	2006	26,9423
10	2007	25,7297
11	2008	24,4262
12	2009	31,3733
13	2010	30,4769
14	2011	29,0075
15	2012	31,1408
16	2013	32,1561
17	2014	35,989
18	2015	62,5472
19	2016	65,0539
20	2017	58,3152
21	2018	62,3431

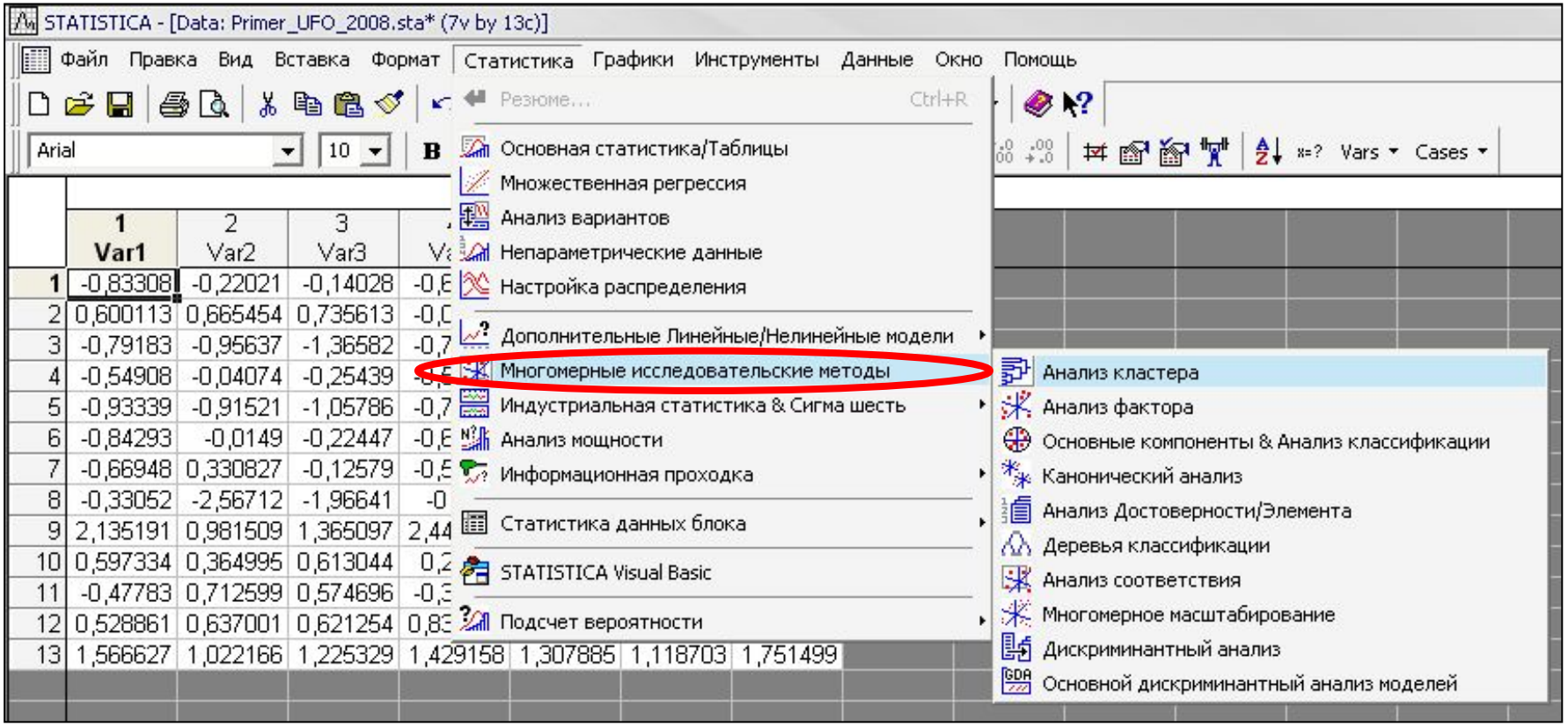
1. Запускаем программу Statistica
2. Открываем документ (Open Document):
   В меню **Файл** выбрать команду **Open**
3. Открываем **лист КурсАлко** в файле КурсАлко.xls



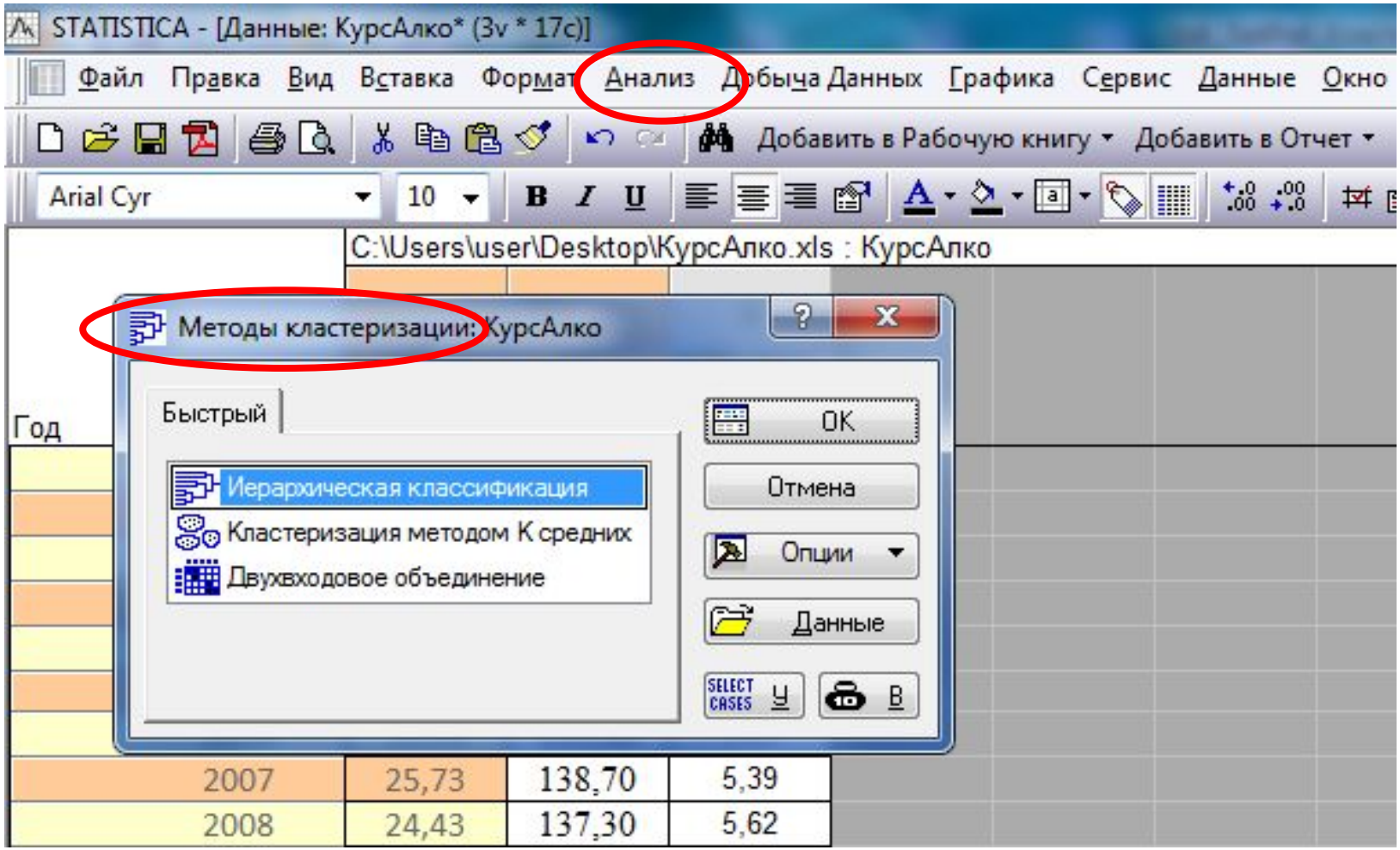
Год	1 руб/$	2 млн.дал	3 млн.дал/ курс
2000	28,03	117,50	4,19
2001	29,16	120,30	4,13
2002	31,48	125,20	3,98
2003	30,55	131,10	4,29
2004	28,96	132,10	4,56
2005	28,42	133,20	4,69
2006	26,94	135,00	5,01
2007	25,73	138,70	5,39
2008	24,43	137,30	5,62
2009	31,37	129,60	4,13
2010	30,48	127,60	4,19
2011	29,01	126,80	4,37
2012	31,14	131,10	4,21
2013	32,16	122,10	3,80
2014	35,99	110,50	3,07
2015	62,55	99,20	1,59
2016	65,05	96,80	1,49

В Statistica 6 вызываем диалог кластерного анализа

Меню **Статистика / Многомерные исследовательские методы / Анализ кластера**

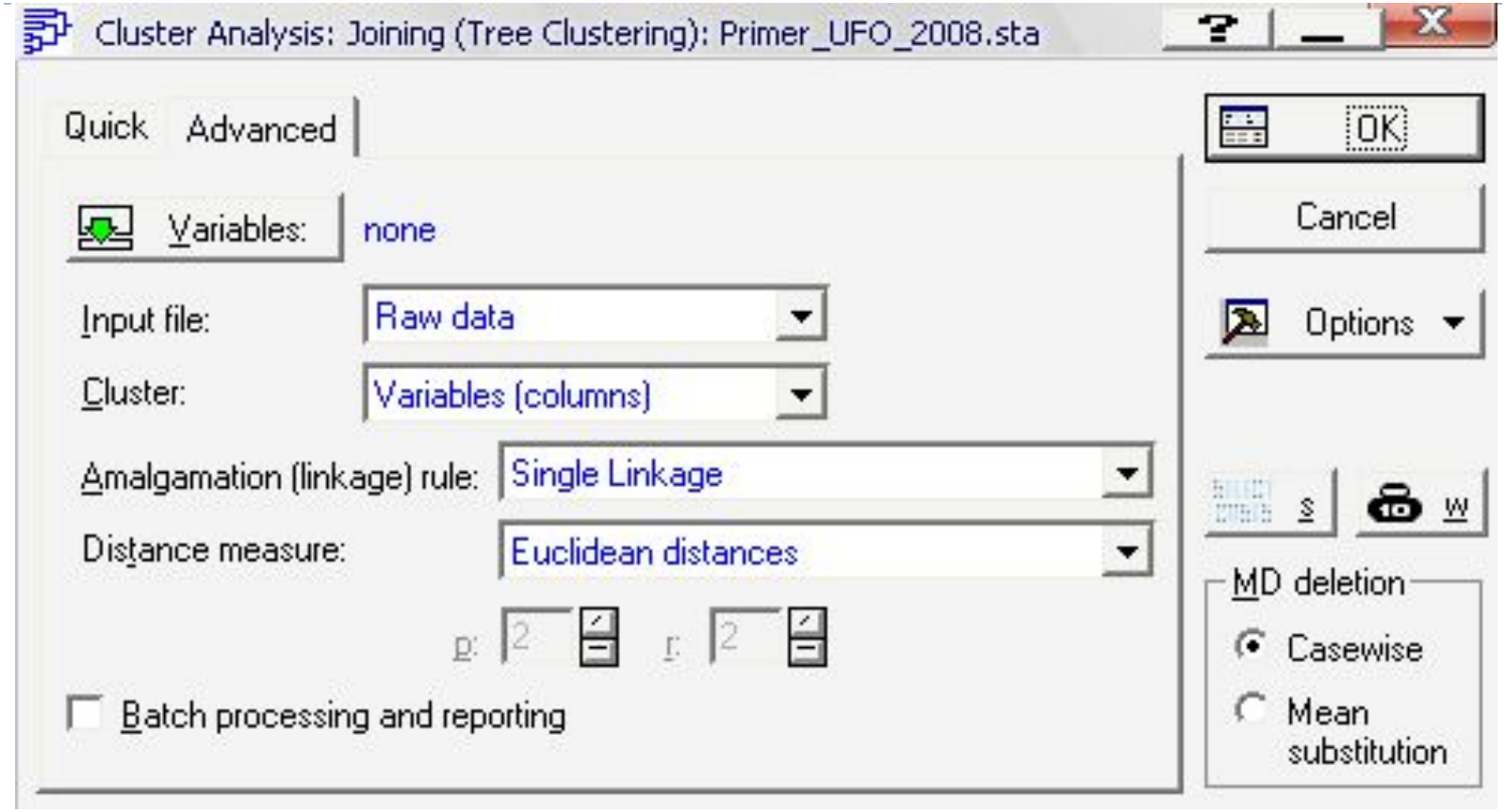


В Statistica 10 переходим на вкладку Анализ→Многомерный разведочный→Кластерный

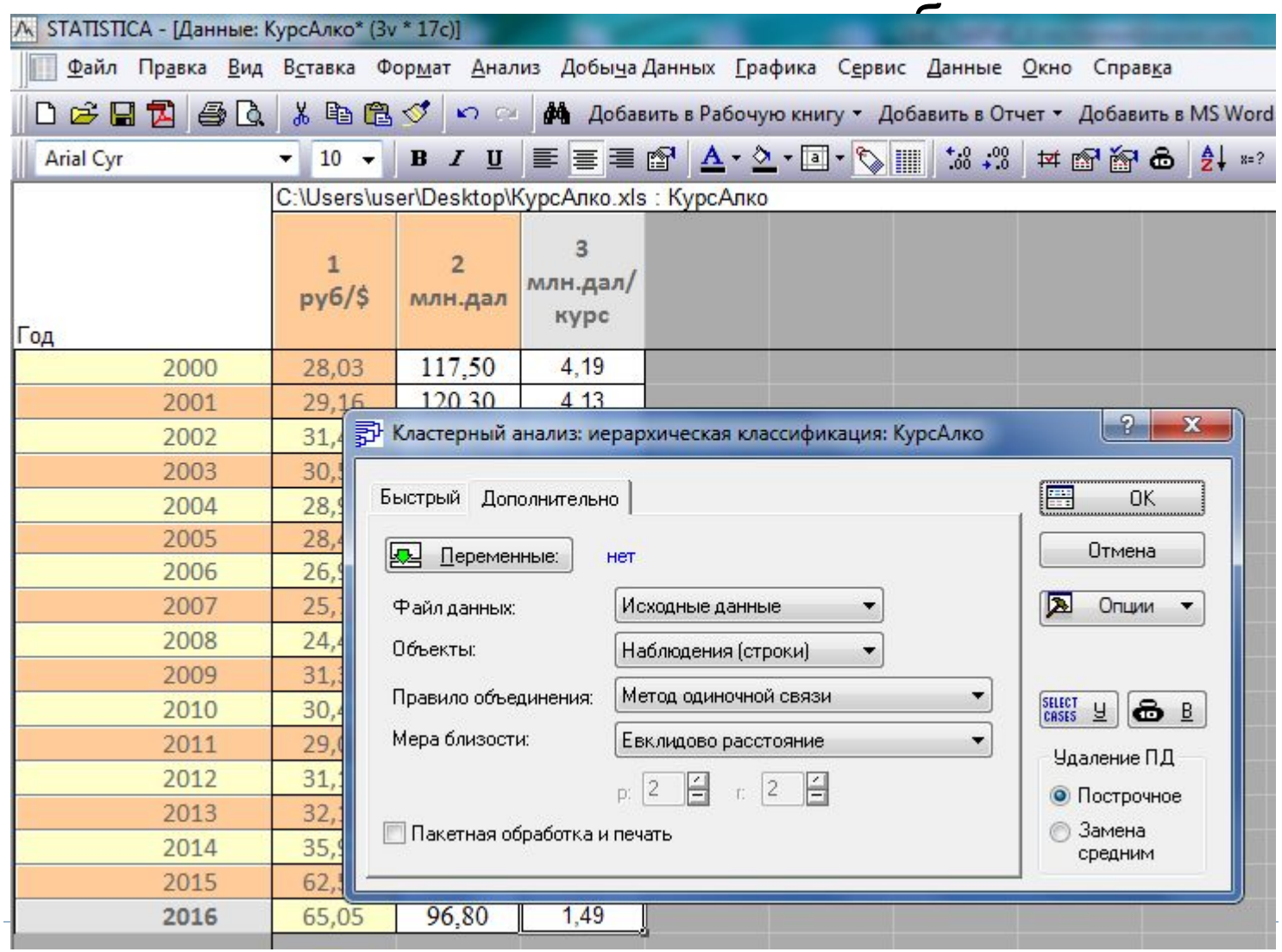


В появившемся окне диалога Statistica 6 выбираем иерархические агломеративные методы **Joining (tree clustering)**

Открываем вкладку **Advanced**

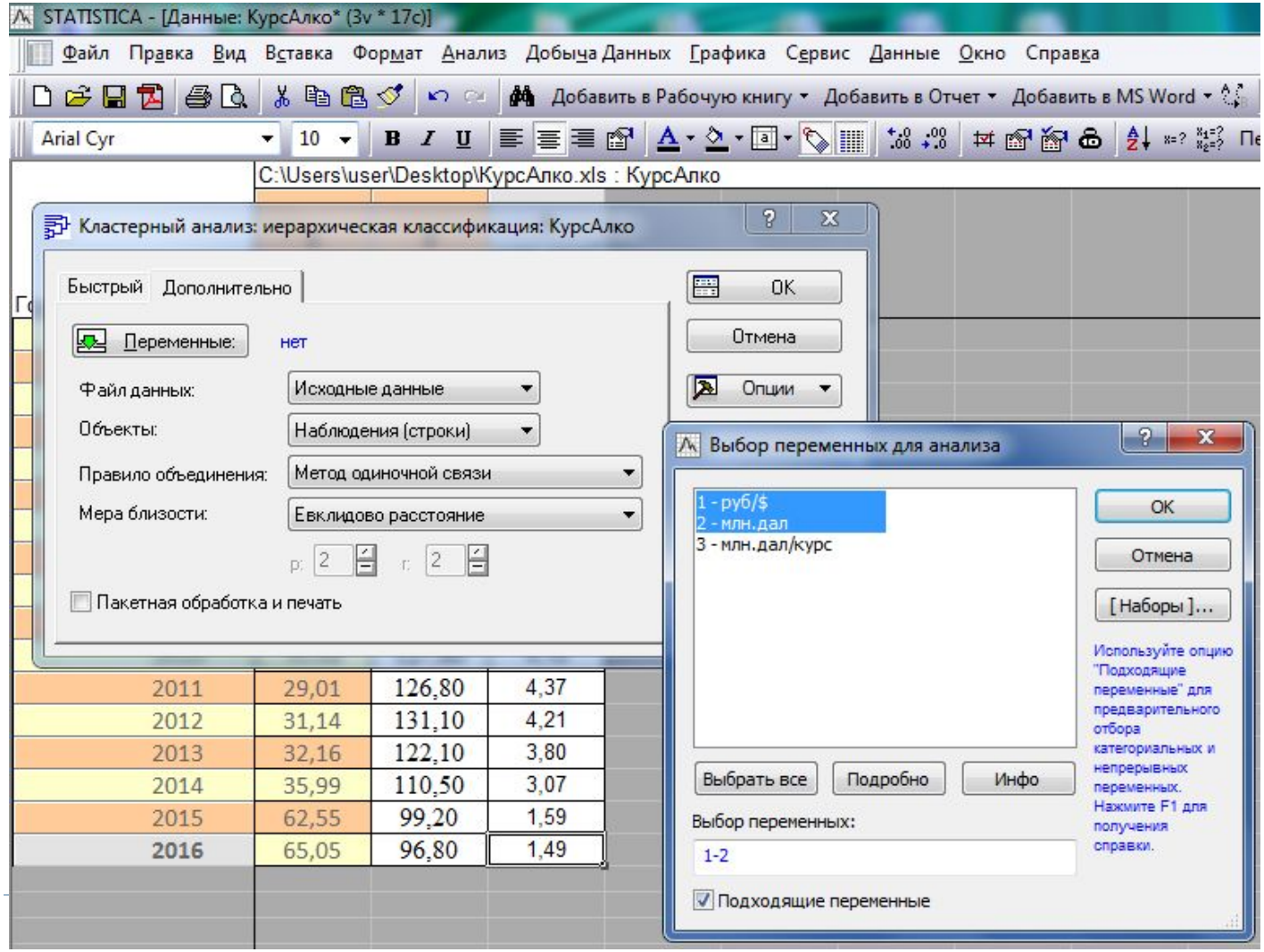


Выбираем переменные
- кнопка **Variables (Дополнительно)**



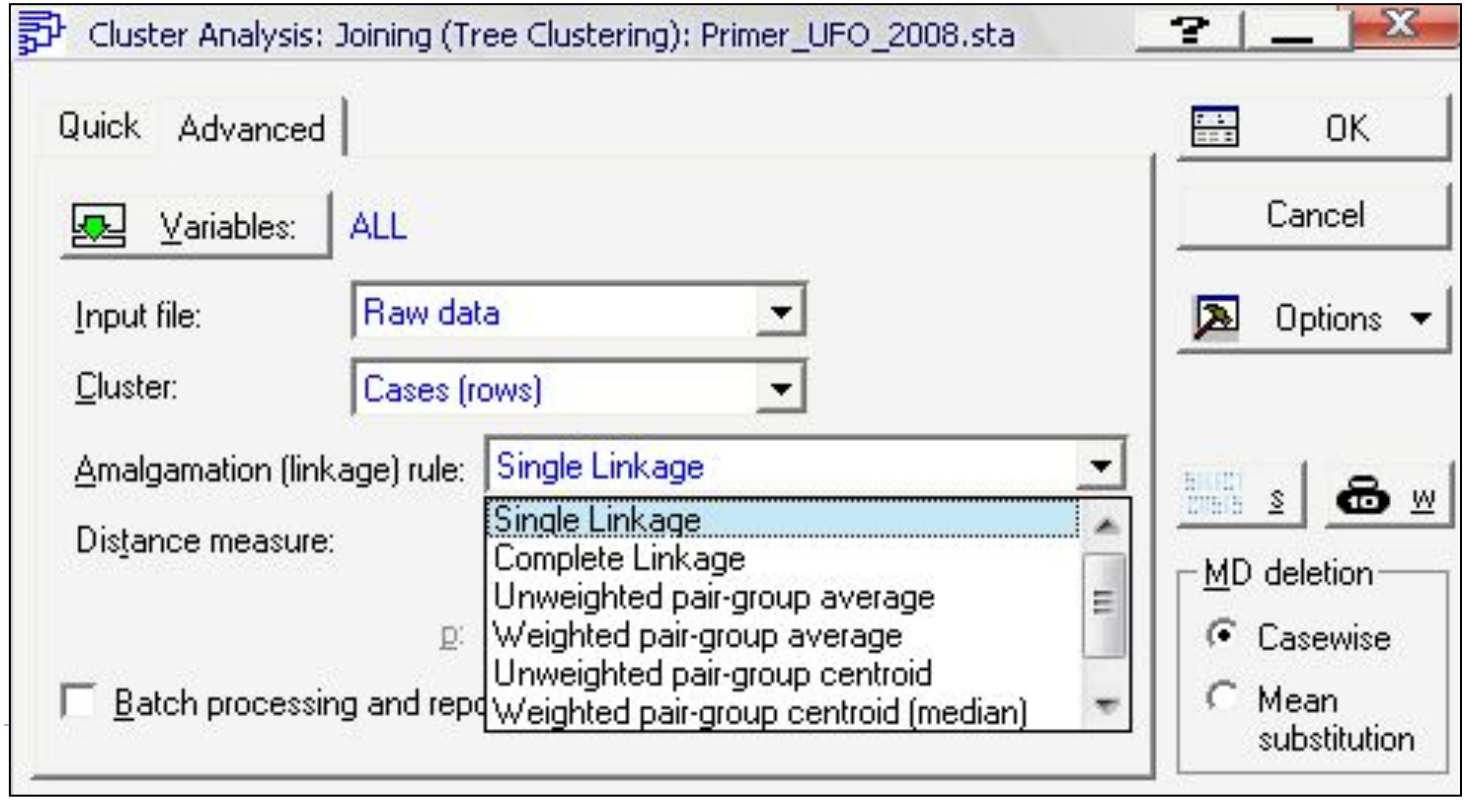
Выбираем переменные

- в открывшемся окне диалога выбрать первые две переменные и нажать кнопку **Ok**



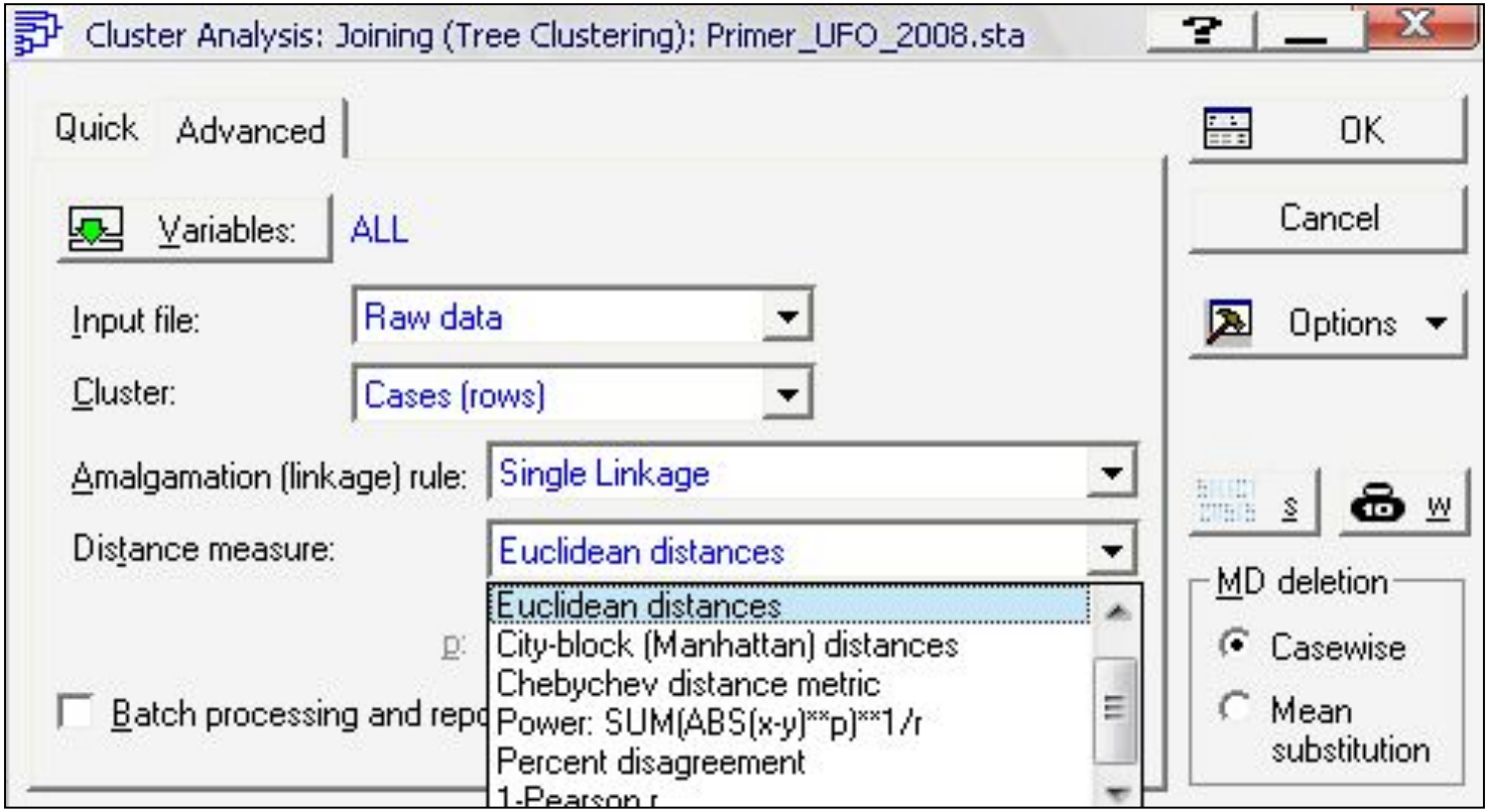
В открывающемся списке **Amalgamation (linkage) rule** (правило объединения в кластеры) можно указать:

- **Single Linkage** (метод одиночной связи)
- **Complete Linkage** (метод полной связи)
- **Unweighted pair-group average** (метод невзвешенного попарного среднего)
- **Ward's method** (метод Уорда) и другие



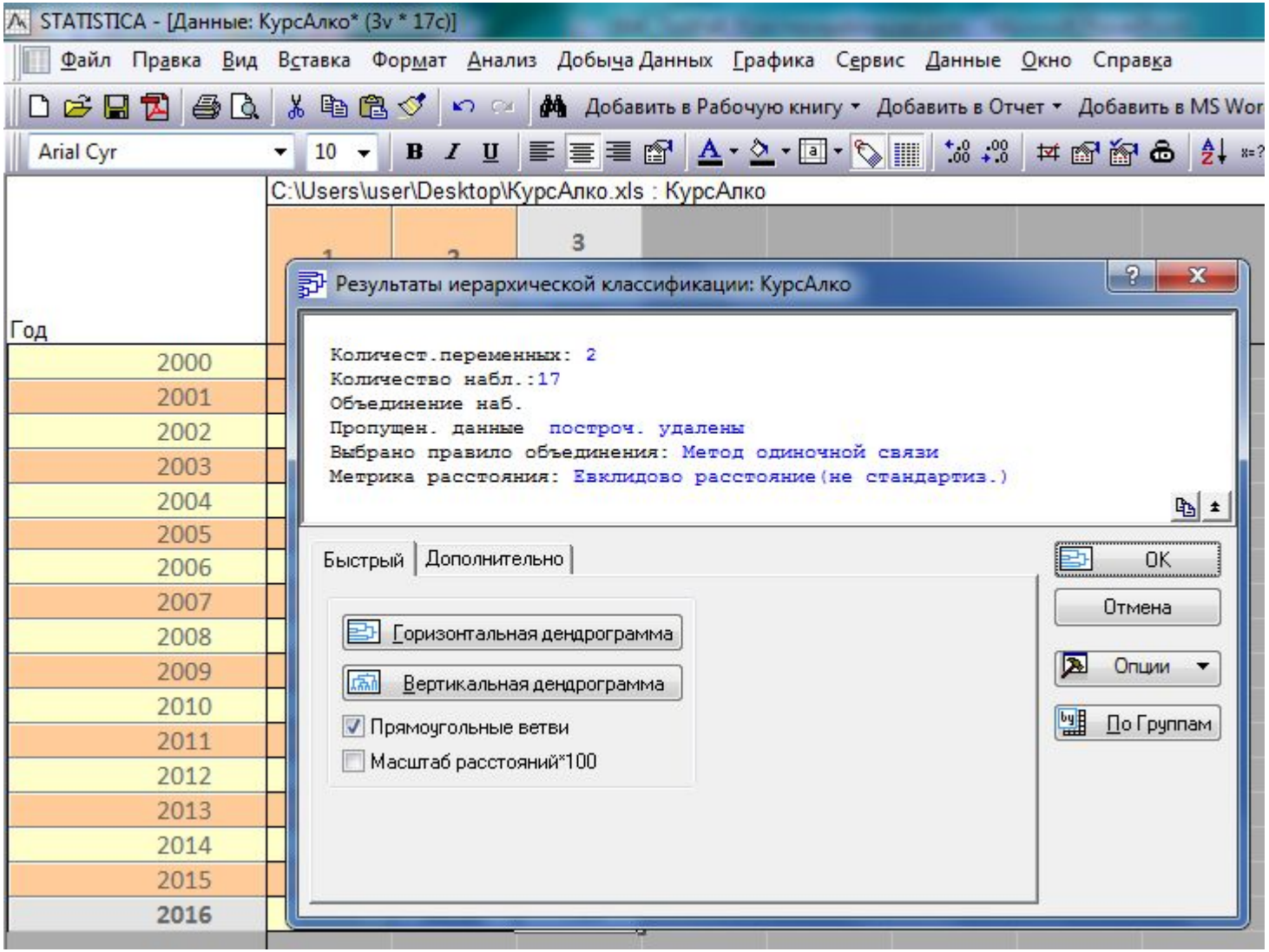
В открывающемся списке **Distance measure** (метрика – расстояние между объектами) можно указать:

- **Euclidean distance** (евклидова метрика)
- **City-block Manhattan distance** (манхеттенское расстояние городских кварталов)
- **Percent disagreement** (процент несогласия) и другие

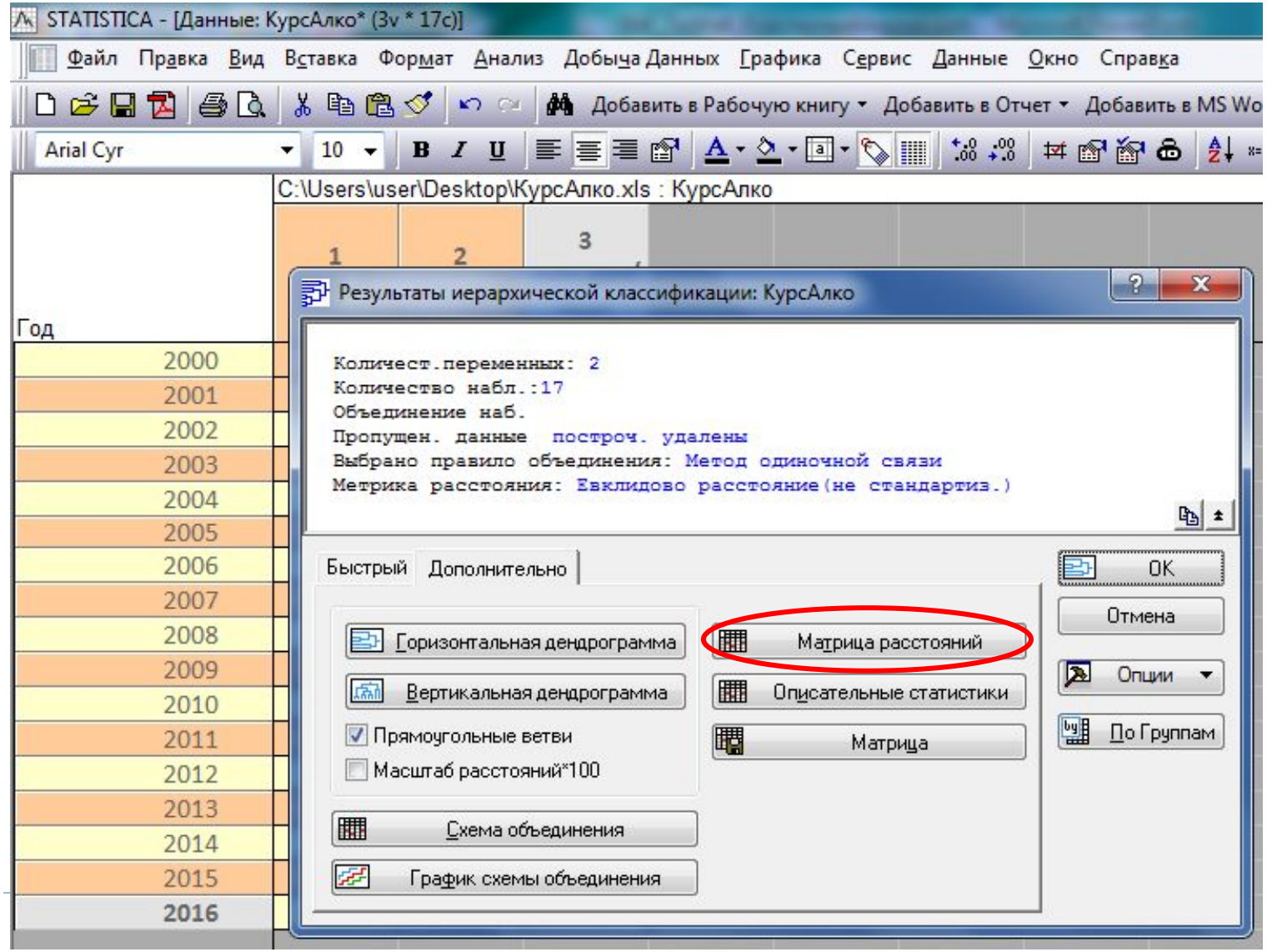


Выберем **метод одиночной связи** и **евклидову метрику**

После нажатия кнопки **Ok** появится окно, в котором по умолчанию откроется вкладка Быстрый



На вкладке **Advanced (Дополнительно) представлен** расширенный набор результатов



Кнопка **Distance matrix** выводит матрицу расстояний (с учетом выбранной метрики)

STATISTICA - [Рабочая книга3* - Евклидово расстояние (КурсАлко)]

Файл Правка Вид Вставка Формат Анализ Добыча Данных Графика Сервис Данные Рабочая книга Окно Справка

Добавить в Рабочую книгу Добавить в Отчет Добавить в MS Word

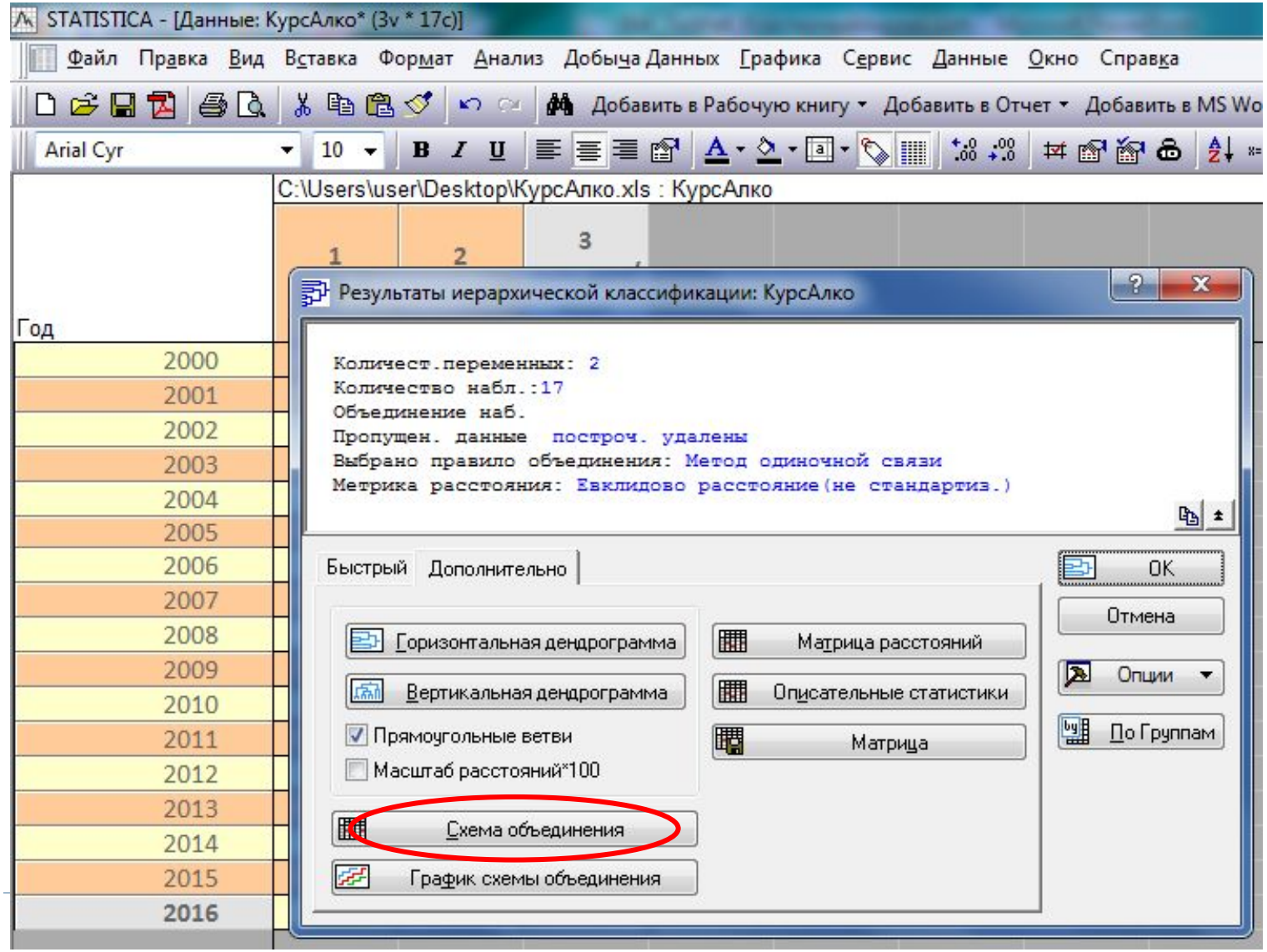
Arial 10 Перемен

Рабочая книга3*
Кластерный а
Результаты
Евклид

наблюд.	Евклидово расстояние (КурсАлко)																
	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
2000	0,0	3,0	8,4	13,8	14,6	15,7	17,5	21,3	20,1	12,6	10,4	9,4	14,0	6,2	10,6	39,1	42,4
2001	3,0	0,0	5,4	10,9	11,8	12,9	14,9	18,7	17,6	9,6	7,4	6,5	11,0	3,5	11,9	39,5	42,9
2002	8,4	5,4	0,0	6,0	7,3	8,6	10,8	14,7	14,0	4,4	2,6	2,9	5,9	3,2	15,4	40,5	44,0
2003	13,8	10,9	6,0	0,0	1,9	3,0	5,3	9,0	8,7	1,7	3,5	4,6	0,6	9,1	21,3	45,2	48,6
2004	14,6	11,8	7,3	1,9	0,0	1,2	3,5	7,3	6,9	3,5	4,7	5,3	2,4	10,5	22,7	47,0	50,5
2005	15,7	12,9	8,6	3,0	1,2	0,0	2,3	6,1	5,7	4,7	6,0	6,4	3,4	11,7	23,9	48,2	51,6
2006	17,5	14,9	10,8	5,3	3,5	2,3	0,0	3,9	3,4	7,0	8,2	8,5	5,7	13,9	26,1	50,5	54,0
2007	21,3	18,7	14,7	9,0	7,3	6,1	3,9	0,0	1,9	10,7	12,1	12,3	9,3	17,8	30,0	54,0	57,5
2008	20,1	17,6	14,0	8,7	6,9	5,7	3,4	1,9	0,0	10,4	11,4	11,5	9,1	17,1	29,2	53,9	57,4
2009	12,6	9,6	4,4	1,7	3,5	4,7	7,0	10,7	10,4	0,0	2,2	3,7	1,5	7,5	19,6	43,5	47,0
2010	10,4	7,4	2,6	3,5	4,7	6,0	8,2	12,1	11,4	2,2	0,0	1,7	3,6	5,8	18,0	42,8	46,3
2011	9,4	6,5	2,9	4,6	5,3	6,4	8,5	12,3	11,5	3,7	1,7	0,0	4,8	5,7	17,7	43,4	46,9
2012	14,0	11,0	5,9	0,6	2,4	3,4	5,7	9,3	9,1	1,5	3,6	4,8	0,0	9,1	21,2	44,8	48,2
2013	6,2	3,5	3,2	9,1	10,5	11,7	13,9	17,8	17,1	7,5	5,8	5,7	9,1	0,0	12,2	38,1	41,5
2014	10,6	11,9	15,4	21,3	22,7	23,9	26,1	30,0	29,2	19,6	18,0	17,7	21,2	12,2	0,0	28,9	32,1
2015	39,1	39,5	40,5	45,2	47,0	48,2	50,5	54,0	53,9	43,5	42,8	43,4	44,8	38,1	28,9	0,0	3,5
2016	42,4	42,9	44,0	48,6	50,5	51,6	54,0	57,5	57,4	47,0	46,3	46,9	48,2	41,5	32,1	3,5	0,0

Например, расстояние между 2000 и 2001 годом равно 3,0 по евклидовой метрике

На вкладке **Advanced (Дополнительно)** можно вывести **Схему объединения**



Кнопка **Amalgamation schedule (Схема объединения)** выводит последовательность объединения в кластеры в виде таблицы

STATISTICA - [Рабочая книга3* - Схема объединения (КурсАлко)]

Файл Правка Вид Вставка Формат Анализ Добыча Данных Графика Сервис Данные Рабочая книга Окно Справка

Добавить в Рабочую книгу Добавить в Отчет Добавить в MS Word

Arial 10 Переменные Наблюдения

Рабочая книга3*
Кластерный а
Результаты
Евклид
Дендро
Дендро
Схема

Схема объединения (КурсАлко)
Метод одиночной связи
Евклидово расстояние

расст. объедин.	Объект 1	Объект 2	Объект 3	Объект 4	Объект 5	Объект 6	Объект 7	Объект 8	Объект 9	Объект 10	Объект 11	Объект 12	Объект 13	Объект 14	Объект 15	Объект 16	Объект 17
,5861000	2003	2012															
1,225266	2004	2005															
1,517912	2003	2012	2009														
1,673062	2010	2011															
1,878832	2003	2012	2009	2004	2005												
1,912881	2007	2008															
2,191696	2003	2012	2009	2004	2005	2010	2011										
2,331656	2003	2012	2009	2004	2005	2010	2011	2006									
2,599577	2002	2003	2012	2009	2004	2005	2010	2011	2006								
3,019420	2000	2001															
3,173769	2002	2003	2012	2009	2004	2005	2010	2011	2006	2013							
3,408923	2002	2003	2012	2009	2004	2005	2010	2011	2006	2013	2007	2008					
3,470381	2015	2016															
3,495228	2000	2001	2002	2003	2012	2009	2004	2005	2010	2011	2006	2013	2007	2008			
10,59932	2000	2001	2002	2003	2012	2009	2004	2005	2010	2011	2006	2013	2007	2008	2014		
28,86222	2000	2001	2002	2003	2012	2009	2004	2005	2010	2011	2006	2013	2007	2008	2014	2015	2016

При уровне 1,91 (расстояние между кластерами не превышает два стандартных отклонения) образовано 5 кластеров:

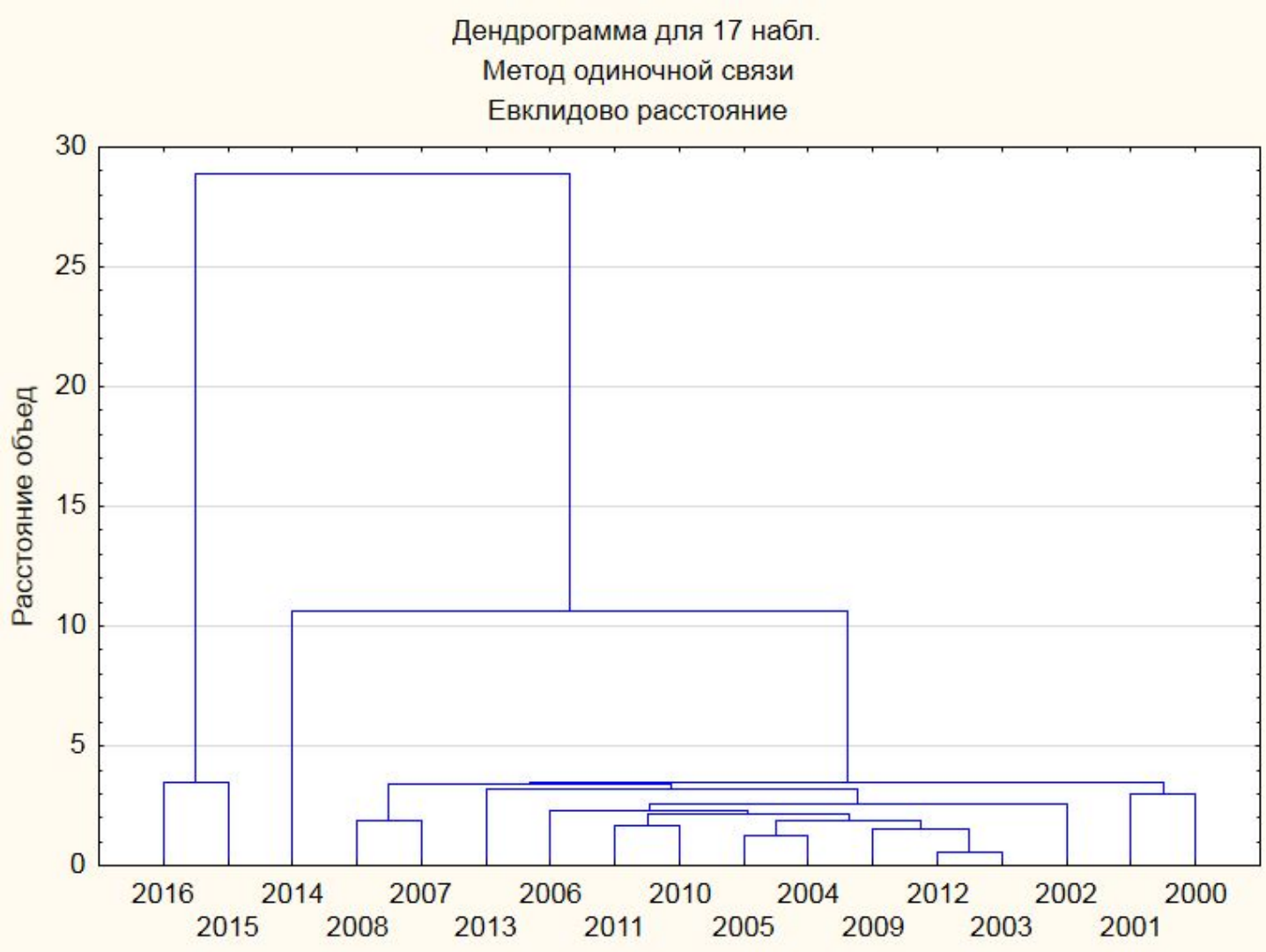
1 КЛАСТЕР: 2003,2012 3 КЛАСТЕР: 2009 4 КЛАСТЕР: 2010,2011

2 КЛАСТЕР: 2004,2005 5 КЛАСТЕР: 2007,2008

Содержательная интерпретация

- **Замечание**
  - Возрастание номера кластера не означает возрастание характеристик, входящих в состав кластера
  - В кластерах с меньшим номером объекты расположены ближе друг к другу (плотнее) и были раньше объединены в один кластер

Кнопка **hierarchical tree plot (вертикальная)** выводит дендрограмму в вертикальном виде



Содержательная интерпретация

- Наиболее удалены от основной группы 2000-2013гг. следующие годы:
  - 2015-2016
  - 2014
  - В эти годы наблюдается максимальный рост курса доллара и минимальное потребление алкоголя

- Таким образом, рост курса доллара содействует снижению потребления алкоголя населением.

Иллюстрация динамики

Иллюстрация динамики

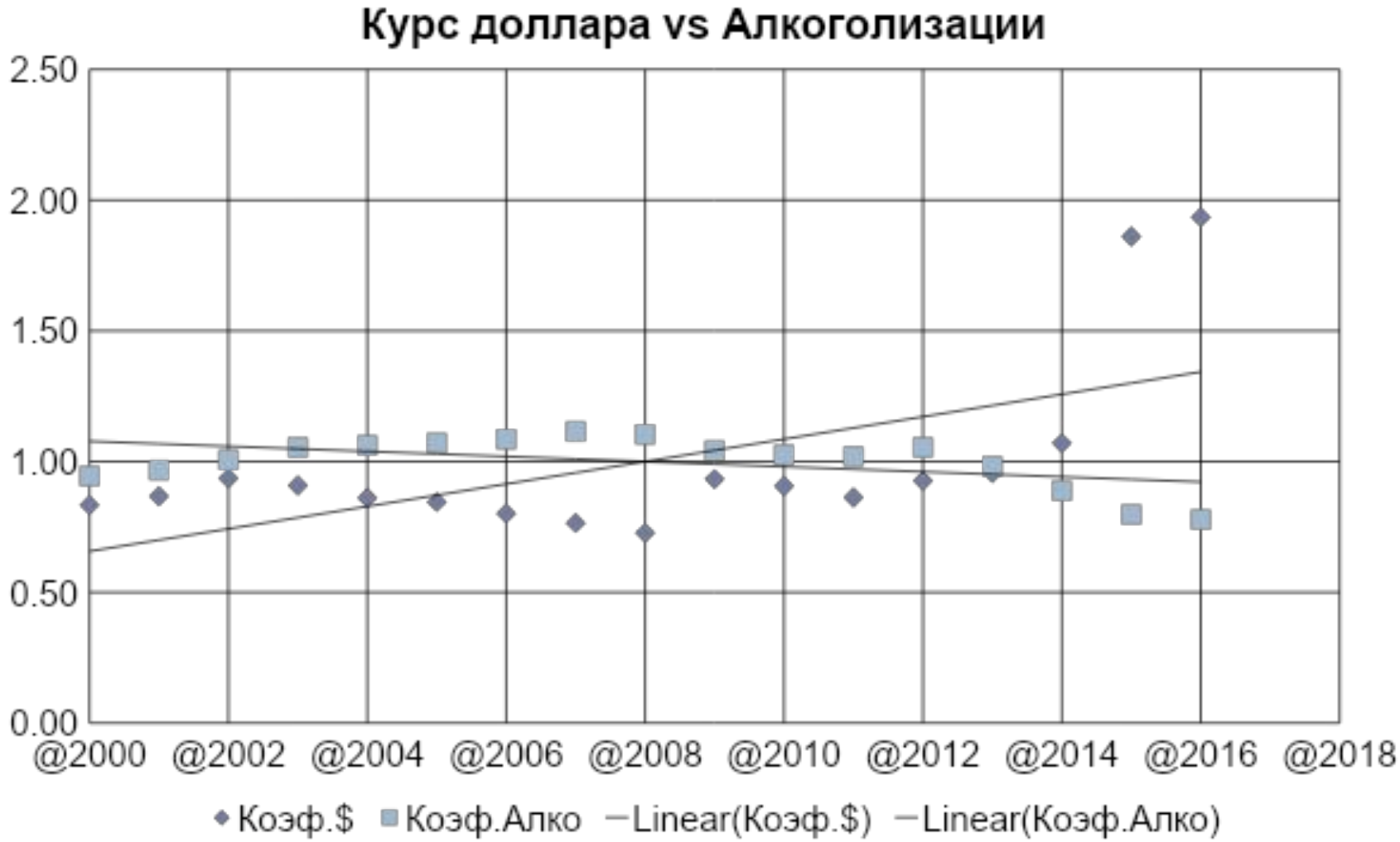
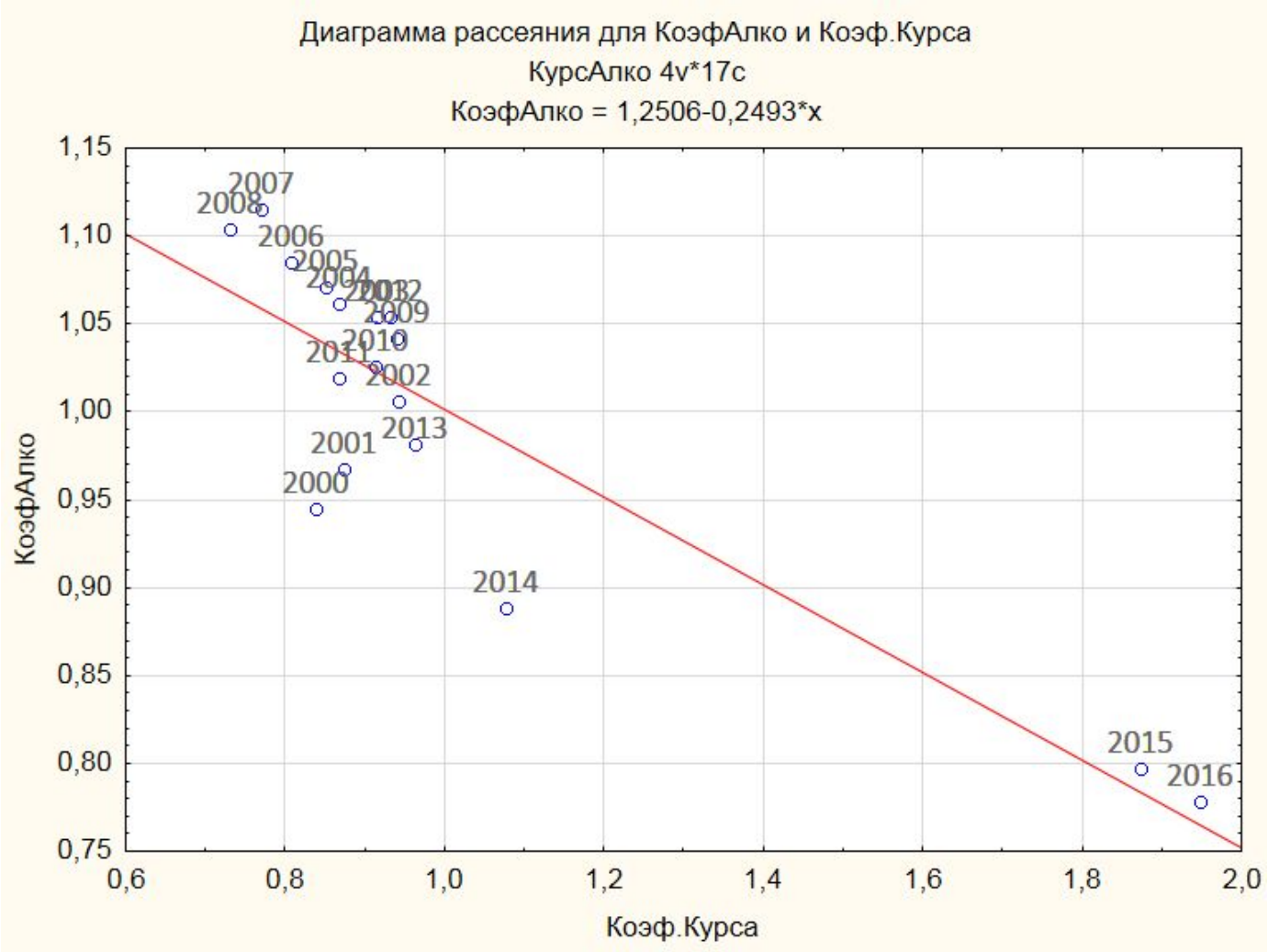


Иллюстрация динамики



Удвоение курса на четверть снижает потребление