http://statsoft.ru/products/STATISTICA\_Advanced/the-discriminant-analysis.php [[](http://statsoft.ru/)](http://statsoft.ru/)

Мы поставили и развили в России и мире   
современные технологии анализа данных!

 [Пресс-центр](http://statsoft.ru/coordination/news/)

 [Мероприятия](http://statsoft.ru/coordination/events/)

 [Контакты](http://statsoft.ru/coordination/contacts.php)

* [**Главная**](http://statsoft.ru/)
* [**О компании**](http://statsoft.ru/company/)
* [**Продукты**](http://statsoft.ru/products/)
* [**Решения**](http://statsoft.ru/solutions/)
* [**Обучение**](http://statsoft.ru/academy/)
* [**Консалтинг**](http://statsoft.ru/consulting/)
* [**Ресурсы**](http://statsoft.ru/resources/)
* [**Портал**](http://statistica.ru/)
* [**ТВ**](http://statsoft.ru/tv/)

Начало формы





Конец формы

Дискриминантный анализ

Содержание

[Введение](http://statsoft.ru/products/STATISTICA_Advanced/the-discriminant-analysis.php#vv)

[Оценка качества дискриминации](http://statsoft.ru/products/STATISTICA_Advanced/the-discriminant-analysis.php#oce)

[Описание данных и постановка задачи](http://statsoft.ru/products/STATISTICA_Advanced/the-discriminant-analysis.php#opi)

[Решение задачи по шагам](http://statsoft.ru/products/STATISTICA_Advanced/the-discriminant-analysis.php#res)

[Классификация](http://statsoft.ru/products/STATISTICA_Advanced/the-discriminant-analysis.php#cla)

[Литература](http://statsoft.ru/products/STATISTICA_Advanced/the-discriminant-analysis.php#lit)

Введение

В задачах построения объяснительных моделей часто встает вопрос о нахождении критерия, по которому можно было бы классифицировать данные (построение дискриминирующей функции) и определить переменные, которые различают две или более возникающие совокупности (группы).

Например, некий исследователь в области образования может захотеть исследовать, какие переменные относят выпускника средней школы к одной из трех категорий: (1) поступающий в колледж, (2) поступающий в профессиональную школу или (3) отказывающийся от дальнейшего образования или профессиональной подготовки. Для этой цели исследователь может собрать данные о различных переменных, связанных с учащимися школы.

После выпуска большинство учащихся естественно должно попасть в одну из названных категорий.

Затем можно использовать *Дискриминантный анализ* для определения того, какие переменные дают наилучшее предсказание выбора учащимися дальнейшего пути.

Пусть имеется *n*наблюдений, разбитых на *k* групп (классов).

Каждое наблюдение характеризуется набором из *m*значений (независимые переменные). Также для каждого наблюдения известно, к какой из *k* групп оно принадлежит. Принадлежность объектов к разным классам выражается в том, что для объектов данного класса имеет распределение Fj, j=1,…,k.

Задача состоит в том, чтобы для нового наблюдения определить группу (класс), к которой оно принадлежит.

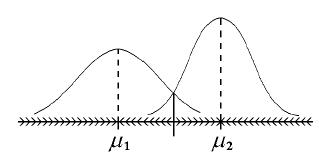
Дискриминантный анализ предполагает, что Fj являются *m*-мерными нормальными распределениями N(мюj,суммаj), j=1,…,k и имеющими плотности:

Формула для pj(x), j=1,…,k      (1)

Здесь мюj- m-мерный вектор средних значений, а суммаj- невырожденная ковариационная матрица (mxm).

В таком случае, исходя из принципа максимального правдоподобия, будем считать областью притяжения закона N(мюj,суммаj) множество таких наблюдений x принадлежит R^m, где плотность распределения N(мюj,суммаj) больше других. См. рис. 1.

В данном примере, где графики плотности пересекаются только в одной точке, получается, что вся прямая разбивается на 2 области притяжения.



*Рис. 1. Области притяжения для k=2, m=1*

Это равносильно тому, что линейно связанная с логарифмом плотности величина:

Формула для hj(x)  , j=1,…,k      (2)

имеет наименьшее значение среди h1,...hk. Таким образом, *n+1* наблюдение будет отнесено к*i-*группе, если hi(x)- имеет наименьшее значение.

[в начало](http://statsoft.ru/products/STATISTICA_Advanced/the-discriminant-analysis.php#toc)

Оценка качества дискриминации

Рассмотрим *модель Фишера*, которая является частным случаем нормальной дискриминантной модели при сумма1=сумма2=..=сумма.

При *k*=2 нетрудно проверить, что поверхность, задаваемая условием h1=h2, разделяет два класса уравнением:

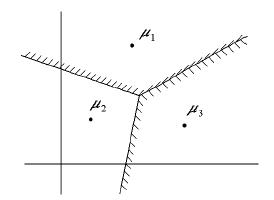
Формула, где Формула для a .

Линейную функцию фи(x)=(a^r)x часто называют дискриминантной функцией, как функцию, описывающую гиперплоскость, по которой разделяются две группы. Мы же будем под дискриминантной функцией понимать линейную часть функции  hi(x).

Обозначим через Формула для d^2 *расстояние Махаланобиса* между мю1и мю2. Чем более далекими в метрике Махаланобиса являются мю1и мю2, тем меньше вероятность ошибочной классификации.

В общем случае, image240.gif. *Расстояние Махаланобиса* является мерой расстояния между двумя точками x и y в пространстве, определяемым двумя или более коррелированными переменными. Можно заметить, что в случае, когда переменные не коррелированны (Формула), расстояние Махаланобиса совпадет с евклидовым.

При k>2 с помощью гиперплоскостей hi=hj m–мерное пространство разбивается на *k* частей. В каждой из них содержится только одна точка из мюj – та, к которой все точки данной части находятся ближе, чем к остальным в смысле расстояния Махаланобиса. См. рис. 2.



*Рис. 2. Области притяжения для k=3, m=2*

Для проверки гипотезы H0 о равенстве средних мю1=мю2=...=мюkв качестве статистик критерия используют статистики Уилкса (лямбда Уилкса):

L=detE/detT        (3)

Здесь T = Сумма – общая матрица рассеяния, матрица внутриклассового разброса: E=сумма от 1 до k Ej,

где Формула для Ej- матрица рассеяния*j*-го класса.

Очевидно, что ее значение меняется от 1.0 (нет дискриминации) до 0.0 (полная дискриминация).

Оказывается, что верно матричное тождество:

T = E + R,

где R = Сумма – матрица разброса между элементами класса, nj – число элементов в *j*-м классе.

При выполнении гипотезы H0:

Гипотеза     (4)

имеет распределение Фишера.

H0 отклоняется (т.е. дискриминация значима), если

Неравенство,

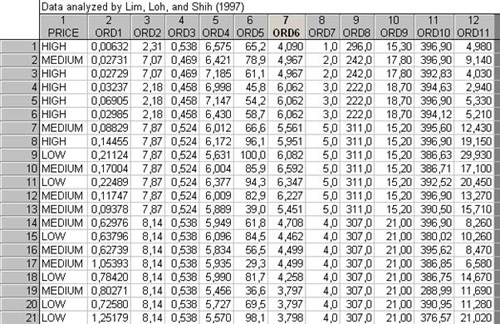
где Fa- квантиль уровня 1-альфа.

[в начало](http://statsoft.ru/products/STATISTICA_Advanced/the-discriminant-analysis.php#toc)

Описание данных и постановка задачи

Имеется файл с данными *boston.sta* с ценами земельных участков в Бостоне. Всего в файле содержится 1012 участков (наблюдений).

Участок характеризуется 11 параметрами ORD1,…, ORD11 – непрерывные предикторы, а также одной группирующей категориальной переменной PRICE – характеризующий ценовой класс, к которому относиться данный участок (HIGH, MEDIUM, LOW). См. рис. 3.



*Рис. 3. Таблица с исходными данными boston.sta*

**Цель:** определить критерий, по которому можно классифицировать наблюдения по категории PRICE в зависимости от параметров участка (ORD1-ORD11), и, c его помощью, определить категорию PRICE для нового наблюдения.

[в начало](http://statsoft.ru/products/STATISTICA_Advanced/the-discriminant-analysis.php#toc)

Решение задачи по шагам

Для решения задачи перейдем на вкладку *Анализ/Многомерный Разведочный анализ/Дискриминантный анализ*. См. рис. 4.

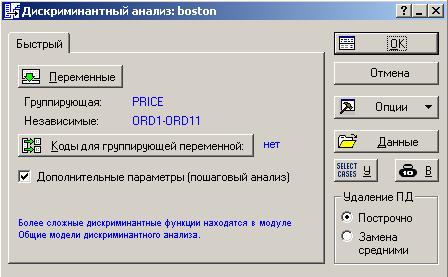
В качестве группирующей переменной укажем переменную PRICE, в качестве независимых – переменные ORD1-ORD11. Анализ будем проводить пошагово. Количество шагов соответствует числу переменных.

Пошаговый анализ с включением/исключением на каждой итерации при помощи статистики Фишера определяет, стоит ли включать в модель соответствующую переменную.

Обычно в пошаговом анализе дискриминантной функции, переменные включают в модель, если соответствующее им значение F больше, чем значение F-*включить*, переменные удаляют из модели, если соответствующее им значение F меньше, чем значение F-*исключить*.

Заметим, что значение F-*включить* всегда должно быть больше, чем значение F-*исключить*. Если при проведении пошагового анализа с включением, вы пожелаете включить все переменные, установите в поле F-*включить*значение, равное очень маленькому числу (например, 0.0001), а в поле F-*исключить* значение 0.0.

Если при проведении пошагового анализа с исключением, вы пожелаете исключить все переменные из модели, установите в поле F-*включить* значение, равное очень большому числу (например, 0.9999), а в поле F-*исключить*чуть-чуть меньшее значение того же порядка (например, 0.9998).

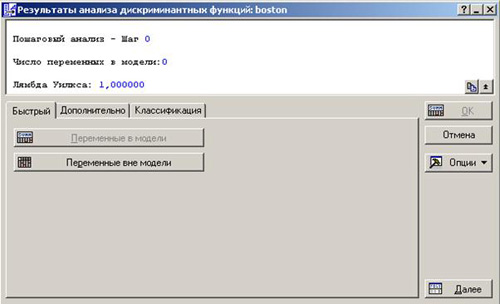


*Рис. 4. Пошаговый дискриминантный анализ*

Нажмем кнопку *ОК*.

В следующем меню на вкладке *Дополнительно* установим опцию: Пошаговый с включением с F-вкл = 10 и вывод результатов на каждом шаге.

Нажмем кнопку *ОК*. См. рис. 5.



*Рис. 5. Результаты анализа на 0-м шаге*

**Шаг 0.**

Лямбда Уилкса равна 1 на 0-м шаге, т.к. никакой дискриминационной модели еще нет.

Нажмем кнопку *Переменные вне модели*. См. рис. 6.

**Лямбда Уилкса.** Значение посчитано по формуле (3) и определяет значение L, если бы соответствующая переменная была включена в модель на этом шаге.

**Частная лямбда Уилкса.** Эта статистика для одиночного вклада соответствующей переменной в дискриминацию между совокупностями является аналогом частной корреляции. Так как в модель еще не введено ни одной переменной, частная лямбда Уилкса равна лямбда Уилкса.

**F-включить и p-значение.** Считается также как и F-статистика для всей модели (формула (4)), только вместо лямбды Уилкса подставляется Частная лямбда Уилкса.

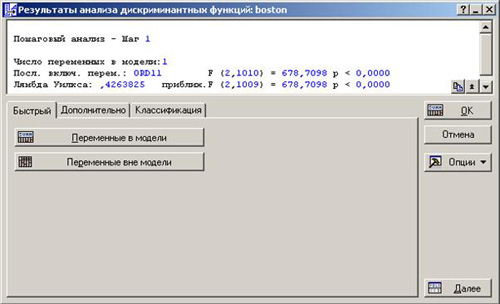
Взглянув на таблицу, вы видите, что наибольшие значения величины *F-включить* дает переменная ORD11 (последняя строка). Переменная с максимальным значением *F-включить* будет включена в модель на первом шаге (т.е. вносящая наибольший вклад в модель).



*Рис. 6. Переменные вне модели на 0-м шаге*

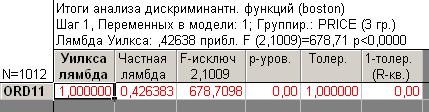
**Шаг 1.**

Анализ включил в модель переменную ORD11, т.к. она несет наибольший вклад среди прочих переменных в дискриминационную модель (наибольшее значение F-вкл). См. рис. 7.



*Рис. 7. Результаты анализа на 1-м шаге*

Нажав кнопку переменные в модели, получим следующую таблицу (рис. 8.)



*Рис. 8. Переменные в модели на 1-м шаге*

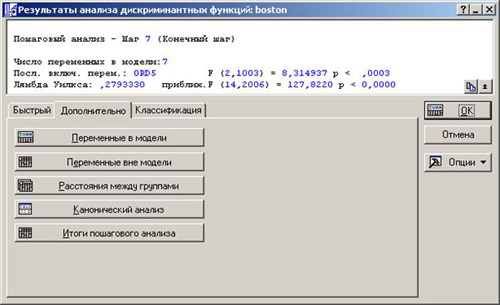
Далее, проводя аналогичные рассуждения, в модель будет включена переменная ORD4. См. рис. 9.



*Рис. 9. Переменные в модели на 2-м шаге*

Алгоритм дискриминантного анализа останавливается, если на очередном шаге F-вкл. в модель оказывается меньше заданного значения (в нашем примере F-вкл. = 10) или если на очередном шаге уже все переменные будут в модели.

В нашем случае анализ остановился на 7-м шаге (т.к. F (2, 1003) = 8,314937 < F-вкл. = 10 ). См. рис. 10.



*Рис. 10. Итоги дискриминантного анализа*

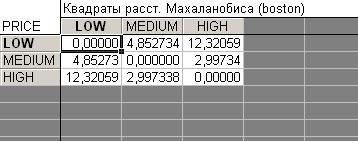
На вкладке *Дополнительно* можно вызвать пункт *Итоги пошагового анализа* (либо пункт *Переменные в модели*). См. рис. 11.



*Рис. 11. Переменные, включенные модель к концу анализа*

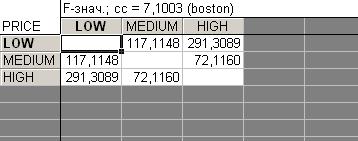
В итоге в модель было включено 7 переменных.

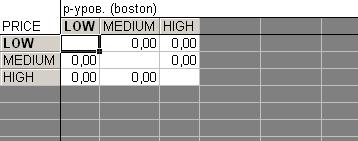
Кнопка *Расстояние между группами* выдаст таблицу с квадратами расстояний Махаланобиса между центрами групп. См. рис. 12.



*Рис. 12. Квадраты расстояний Махаланобиса*

Вместе с таблицей результатов расстояний Махаланобиса выводятся две другие таблицы результатов: одна с F-значениями, связанными с соответствующими расстояниями, а другая – с соответствующими p-уровнями. См. рис. 13.





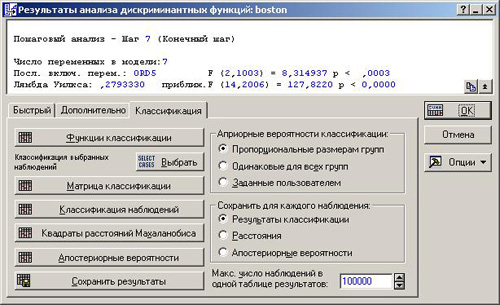
*Рис. 13. Значения F-статистики и p-уровней для расстояний Махаланобиса*

Эти p-уровни должны интерпретироваться с осторожностью, если только в анализ не привносится сильная априорная гипотеза относительно того, какие пары групп должны показывать особенно большие (и значимые) расстояния.

[в начало](http://statsoft.ru/products/STATISTICA_Advanced/the-discriminant-analysis.php#toc)

Классификация

Перейдем к подменю *Классификация*. См. рис. 14.



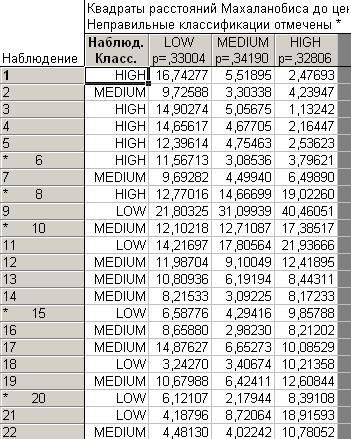
*Рис. 14. Подменю Классификация Дискриминантного анализа*

Здесь, кроме уже описанных выше расстояний Махаланобиса (таблица с расстояниями на рис. 16), можно вывести коэффициенты функции классификации для каждой группы. См. рис. 15.



*Рис. 15. Функции классификации (дискриминации)*

На рис. 15 в каждом столбце находятся коэффициенты дискриминирующей функции для соответствующего класса (стоит еще раз отметить, что подразумевается линейная функция).



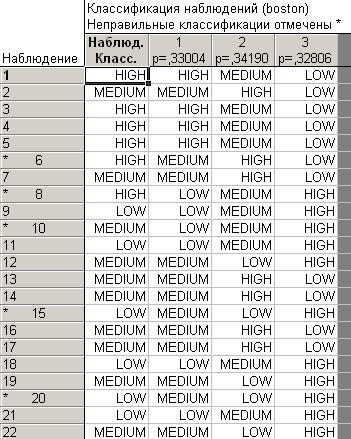
*Рис. 16. Квадраты расстояния Махаланобиса до центров соответствующих групп*

Также можно вывести матрицу классификации и классификацию наблюдений. См. рис. 17. и рис. 18.



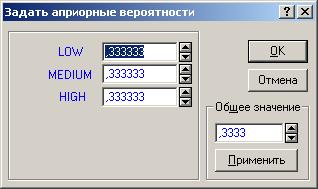
*Рис. 17. Матрица классификации*

Обе таблицы основываются на результатах таблицы с квадратами расстояний Махаланобиса (см. выше).



*Рис. 18. Классификация наблюдений*

Стоит обратить внимание, что в предыдущих таблицах каждая группа была помечена априорной вероятностью (см. в названии переменных таблиц). Их можно задать на панели справа (См. рис. 14 и рис. 19).



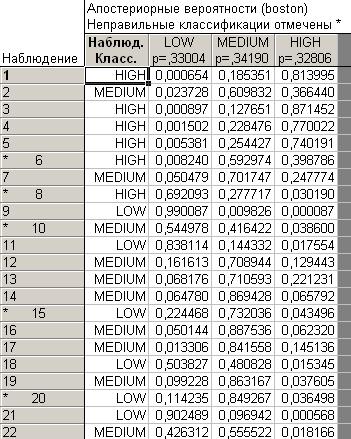
*Рис. 19. Априорные вероятности*

Априорные вероятности отражают наши знания о природе явления перед проведением эксперимента.

Например, если мы знаем, что в начальных данных преобладают элитные земельные участки (PRICE = HIGH), то этот факт, конечно, должен повлиять на анализ, увеличивая долю наблюдений, помеченных в результате дискриминации как HIGH.

По умолчанию, в системе *STATISTICA* априорные вероятности задаются пропорционально размеру групп.

Вероятности, полученные после эксперимента, называются апостериорными. Они приведены в таблице на рис. 20.



*Рис. 20. Апостериорные вероятности*

Апостериорные вероятности связаны с априорными по следующей формуле:

Формула,

где  P(wj/wi)=сумма pi(x).

Здесь wi- событие, «x принадлежит Si», где Si - множество элементов *i*-группы.

[в начало](http://statsoft.ru/products/STATISTICA_Advanced/the-discriminant-analysis.php#toc)

Литература

1. Рао С.Р. Линейные статистические методы и их применения, Наука 1968.
2. Розанов Ю.А. Теория вероятностей, случайные процессы и математическая статистика, Наука 1985.
3. Боровиков В.П. *STATISTICA*, искусство анализа данных на компьютере, Питер 2001.
4. Боровиков В.П. Нейронные сети. *STATISTICA Neural Networks*, Горячая линия – Телеком 2008.

[Есть вопросы?](http://statsoft.ru/coordination/feedback/)

[Специалисты StatSoft всегда на связи.](http://statsoft.ru/coordination/feedback/)

[Продукты](http://statsoft.ru/products)

 [Общий обзор](http://statsoft.ru/products/overview/)

 [Новая комплектация](http://statsoft.ru/products/configuration.php)

 [*STATISTICA Base*](http://statsoft.ru/products/STATISTICA_Base/)

 [*STATISTICA Advanced*](http://statsoft.ru/products/STATISTICA_Advanced/)

 [Промышленная *STATISTICA*](http://statsoft.ru/products/STATISTICA_QC/)

 [Нейронные сети](http://statsoft.ru/products/STATISTICA_Neural_Networks/)

 [Data Mining, Text Mining](http://statsoft.ru/products/STATISTICA_Data_Miner/)

 [Enterprise Solutions](http://statsoft.ru/products/Enterprise/)

 [Управление даннымиnew](http://statsoft.ru/products/data-management/)

 [Специализированные продукты](http://statsoft.ru/products/special-products/)

 [Интеграция](http://statsoft.ru/products/integration/)

 [Галерея](http://statsoft.ru/products/galery/)

 [Trial-версии](http://statsoft.ru/products/trial/)

 [Каталог продуктов](http://statsoft.ru/products/catalog.php)

[Календарь](http://statsoft.ru/academy/calendar.php)

| ‹‹ | Ноябрь 2018 | | | | | ›› |
| --- | --- | --- | --- | --- | --- | --- |
| Пн | Вт | Ср | Чт | Пт | Сб | Вс |
|  |  |  | 1 | 2 | 3 | 4 |
| 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 19 | 20 | 21 | 22 | 23 | 24 | 25 |
| 26 | 27 | 28 | 29 | 30 |  |  |
|  |  |  |  |  |  |  |

|  |
| --- |
|  |  |  |

[Разобраться в](http://statsoft.ru/academy/)*[STATISTICA](http://statsoft.ru/academy/)*[?](http://statsoft.ru/academy/)

[Легко!](http://statsoft.ru/academy/)

[Отраслевые решения на базе](http://statsoft.ru/solutions/ready_solutions.php)*[STATISTICA](http://statsoft.ru/solutions/ready_solutions.php)*[!](http://statsoft.ru/solutions/ready_solutions.php)

|  |  |
| --- | --- |
| [http://statsoft.ru/include/block-video-lessons.png](http://statsoft.ru/tv/screen-video/) | [Видеоролики](http://statsoft.ru/tv/screen-video/)*[STATISTICA](http://statsoft.ru/tv/screen-video/)* |

[Консалтинг StatSoft:](http://statsoft.ru/consulting/)

[помощь в принятии верных решений](http://statsoft.ru/consulting/)

[Big Data:](http://statsoft.ru/products/Enterprise/big-data.php)

[Как извлечь полезную информацию?](http://statsoft.ru/products/Enterprise/big-data.php)

Актуально:

[Открытая дискуссия по вопросам контроля качества и автоматизации анализа промышленных процессов](http://statsoft.ru/coordination/news/news_detail.php?ELEMENT_ID=1856)

[Новые медицинские кейсы на Портале Знаний StatSoft](http://statsoft.ru/coordination/news/news_detail.php?ELEMENT_ID=1699)

[Data Mining. Новые достижения: Обнаружение взаимодействий между лекарствами](http://statsoft.ru/coordination/news/news_detail.php?ELEMENT_ID=1632)

[Все новости](http://statsoft.ru/coordination/news/)

[База примеров](http://statsoft.ru/solutions/ExamplesBase/branches/)

[](http://statsoft.ru/solutions/ExamplesBase/branches/)

[*STATISTICA* Trial 13](http://statsoft.ru/products/trial/)

[](http://statsoft.ru/products/trial/)

[Академия Анализа Данных](http://statsoft.ru/academy/lections.php)

[](http://statsoft.ru/academy/lections.php)

|  |  |
| --- | --- |
| [info@statsoft.ru](http://statsoft.ru/coordination/contacts.php)       [(495) 787-77-33](http://statsoft.ru/coordination/contacts.php)       [(499) 674-06-15](http://statsoft.ru/coordination/contacts.php)      [*STATISTICA Data Miner* 13.3 Trial](http://statsoft.ru/products/trial/) | [http://statsoft.ru/upload/medialibrary/6f5/youtube.png](http://www.youtube.com/user/StatSoftRussia/) |
| Авторские права на дизайн и материалы сайта принадлежат компании StatSoft Russia.  Все права защищены. | © StatSoft Russia 1999-2018 | |
| StatSoft Russia – компания, зарегистрированная и действующая в соответствии с законами России, которые могут отличаться от законов других стран, имеющих офисы StatSoft. Каждый офис StatSoft является самостоятельным юридическим лицом, имеет право предлагать услуги и разрабатывать приложения, которые могут быть, а могут и не быть представлены в офисах StatSoft других стран.  [Лицензионное соглашение](http://statsoft.ru/license.php)      [Карта сайта](http://statsoft.ru/site_map.php) | | |