***Internal clustering criteria***

SPSS macros by Kirill Orlov

kior@akado.ru, ttnphns@gmail.com

<https://www.spsstools.net/en/KO-spssmacros>

All rights reserved

*Внутренние кластерные критерии.* Вычисление индексов, таких как Calinski–Harabasz, Davies–Bouldin, Cubic clustering criterion, Ratkowsky–Lance, C-Index, корреляция, гамма-статистика, Dunn (несколько типов), силуэт-статистика (несколько типов), AIC, BIC и других индексов, помогающих выбрать лучшее кластерное разбиение, в частности решить, сколько кластеров следует выделить в кластерном анализе.

*Прочтите «*[*О SPSS макросах*](https://www.spsstools.net/ru/KO-aboutmacros)*» что они такое и как их запускать.*

*Ошибка “Protected directory”.* Некоторые из макросов, описанных в текущем документе, пишут временные файлы на жесткий диск. Если вы не обладаете полными правами Администратора вашего компьютера, это может вызвать ошибку, сообщающую среди прочего: *“SPSS Statistics cannot access a file... specifies a protected directory...”* и значащую, что дефолтная директория, какую макрос хочет использовать, защищена на вашем ПК. Чтобы решить эту проблему, в окне синтаксиса скомандуйте: CD 'myfolder'., где 'myfolder' есть путь/имя некоторой папки, куда вам разрешено сохранять файлы.

Предлагаются следующие внутренние кластерные критерии: Calinski–Harabasz, SSw “elbow”, Davies–Bouldin, Cubic Clustering Criterion, Log SS Ratio, Log Det Ratio, PBM, Ratkowsky–Lance, AIC, BIC, Point-biserial correlation, Goodman-Kruskal Gamma, C-Index, Dunn, Generalized Dunn, McClain–Rao, Silhouette index, Simplified (deviation) Silhouette index.

* Кластерные критерии на основе идеологии дисперсионного анализа в евклидовом пространстве. Основаны на отношениях сумм квадратов отклонений внутри и вне кластеров: *B/W*, *B/T* или *W/T*.
  + - **Calinski–Harabasz** является многомерным аналогом критерия Фишера. Он хорошо распознает всякие выпуклые кластеры. [!KO\_CALHARV](#_МАКРОС_!KO_CALHARV:_КРИТЕРИИ) берет к качестве входящих переменные, а [!KO\_CALHARM](#_МАКРОС_!KO_CALHARM:_КРИТЕРИИ) – матрицу расстояний. Эти макросы выдают также **SSw критерий «перегиба»** (“elbow”).
    - **Davies–Bouldin** подобен предыдущему, но без его тенденции к примерно равновеликим по числу объектов внутри кластерам; онскорее предпочитает равноудаленные друг от друга кластеры. [!KO\_DAVBOULV](#_МАКРОС_!KO_DAVBOULV:_КРИТЕРИИ) берет к качестве входящих переменные, а [!KO\_DAVBOULM](#_МАКРОС_!KO_DAVBOULM:_КРИТЕРИИ) – матрицу расстояний.
    - **Кубический кластерный критерий** похож на Calinski–Harabasz. Он (под вопросом) стандартизован для сравнения результатов, получаемых на разных данных. Предпочитает сферические кластеры. [!KO\_CCCRITV](#_МАКРОС_!KO_CCCRITV:_КУБИЧЕСКИЙ) берет к качестве входящих переменные, а [!KO\_CCCRITM](#_МАКРОС_!KO_CCCRITM:_КУБИЧЕСКИЙ) – матрицу расстояний.
    - **Log SS Ratio** родствен Calinski–Harabasz, только вместо нормирования *B/W* использует логарифм. Этот критерий выдается макросами [!KO\_CALHARV](#_МАКРОС_!CALHARV:_КРИТЕРИИ) и [!KO\_CALHARM](#_МАКРОС_!CALHARM:_КРИТЕРИЙ).
    - **Log Det Ratio** – логарифм обращенной лямбды Уилкса; это MANOVA-критерий, принимающий во внимание объемное свойство облака данных. Вычисляется [!KO\_CCCRITV](#_МАКРОС_!CCCRITV:_КУБИЧЕСКИЙ).
  + Кластерные критерии, исповедующие одномерный подход: анализ идет по каждой переменной. Это закрепленные признаки: данные не рассматриваются как лежащие в пространстве, где их можно бы произвольно повернуть.
    - **Ratkowsky–Lance** рассчитан как на мерные признаки (где базируется на идее дисперсионного анализа), так и категориальные признаки (базируется на идее хи-квадрат-критерия). [!KO\_RATLAN](#_МАКРОС_!KO_RATLAN:_КРИТЕРИЙ) берет в качестве входящих переменные.
    - **AIC** и **BIC** кластерные критерии тоже допускают мерные и категориальные признаки. Эти индексы связаны с идеей вариационной энтропии. Они налагают штраф на избыток кластеров и потому позволяют предпочесть обоснованно экономное (немного кластеров) решение. [!KO\_AICBIC](#_МАКРОС_!KO_AICBIC:_ИНФОРМАЦИОННЫЕ) берет в качестве входящих переменные.
  + Кластерные критерии на основе идеологии «кофенетической» корреляции (корреляции между подобием объектов и их попаданием в один и тот же кластер).
* **Точечно-бисериальная корреляция** это обычный линейный r Пирсона. [!KO\_RPBCLU](#_МАКРОС_!RPBCLU:_ТОЧЕЧНО-БИСЕРИАЛЬНА) берет в качестве входящих матрицу близостей.
* **Гамма** Гудмана-Краскела это непараметрическая, монотонная корреляция. [!KO\_GAMMACLU](#_МАКРОС_!SILHOU:_СИЛУЭТ-СТАТИСТИКА) берет в качестве входящих матрицу близостей.
* **C-Index** определяет, насколько кластерное разбиение близко к (недостижимому) идеальному в данных условиях. Этот критерий эквивалентен перешкалированному r Пирсона. [!KO\_CINDEX](#_МАКРОС_!KO_CINDEX:_C-INDEX) берет в качестве входящих матрицу близостей.
* Другие критерии:
* **Dunn** ищет кластерное решение с максимально разграниченными, разведенными кластерами, по возможности примерно одинаковой величины (диаметра). Макрос вычисляет разные версии этого критерия. [!KO\_DUNN](#_МАКРОС_!KO_DUNN:_КРИТЕРИЙ) берет в качестве входящих матрицу близостей.
* **McClain–Rao** это отношение среднего однокластерного расстояния к среднему межкластерному расстоянию среди объектов. [!KO\_RPBCLU](#_МАКРОС_!RPBCLU:_ТОЧЕЧНО-БИСЕРИАЛЬНА) берет в качестве входящих матрицу близостей.
* **PBM** это эклектичный критерий, учитывающий суммы отклонений (не квадратов их) от центроидов и раздвинутость центроидов. [!KO\_DAVBOULV](#_МАКРОС_!DAVBOULV:_КРИТЕРИЙ) берет к качестве входящих переменные, а [!KO\_DAVBOULM](#_МАКРОС_!KO_DAVBOULM:_КРИТЕРИИ) – матрицу расстояний.
* **Силуэт-статистика** (макрос вычисляет несколько ее версий) умеет оценивать качество кластеризации каждого отдельного объекта, а не только целого кластерного решения. Критерий замеряет оправданность приписания объектов к своим кластерам. [!KO\_SILHOU](#_МАКРОС_!SILHOU:_СИЛУЭТ-СТАТИСТИКА_2) берет в качестве входящих матрицу близостей. [!KO\_SILDEV](#_МАКРОС_!SILDEV:_СИЛУЭТ-СТАТИСТИКА) берет в качестве входящих переменные и вычисляет одну из версий критерия.

***Внутренние кластерные критериии***

**Таблица 1**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| критерий | базовая идея | тип переменных  (если входящие –переменные) | тип близостей (если входящие –близости) | к изменению числа объектов N | к одинаковому линейному преобразованию переменных (мерных) | к неодинаковому линейному преобразованию переменных (мерных) | к умножению близостей на число | к прибавлению к близостям числа | пространственные предпочтения (в условиях евклидовых расстояний или мерных переменных) | | | | решение лучше, когда значение критерия | примечания |
| к форме распределения в (округлых) кластерах: предпочитает колоколообразное или плоское\*\* | к форме кластеров: предпочитает сферические или эллипсоидные\*\*\* | к повороту всего облака относительно центра данных: предпочитает… | к поворотам отдельных эллипсоидных кластеров относительно своих центров: предпочитает конфигурацию…\*\*\*\* |
| SSw | ANOVA в пространстве\* | мерные | евкл расст | чувствит | только умножен на констатну | чувствит | только умножен на констатну | чувствит | нечувст | нечувст | нечувст | нечувст | нижний локоть |  |
| Calinski–Harabasz | ANOVA в пространстве | мерные | евкл расст | чувствит | нечувст | чувствит | нечувст | чувствит | нечувст | нечувст | нечувст | нечувст | выше |  |
| Davies–Bouldin | ANOVA в пространстве | мерные | евкл расст | нечувст | нечувст | чувствит | нечувст | чувствит | нечувст | нечувст | нечувст | нечувст | ниже | !KO\_DAVBOULV имеет опцию L1-отклонений |
| PBM | Эклектична: сумма отклонений, межцентроидные расстояния | мерные | евкл расст | едва чувствит | только умножен на констатну | чувствит | чувствит | чувствит | колокол | эллипс | нечувст | стопка, кольцо | выше | штраф «избытку» кластеров  !KO\_DAVBOULV имеет опцию L1-отклонений |
| CCC | ANOVA в пространстве | мерные | евкл расст | чувствит | нечувст | чувствит | нечувст | чувствит | нечувст | сферич | нечувст | стопка | выше |  |
| Log SS Ratio | ANOVA в пространстве | мерные | евкл расст | едва чувствит | нечувст | чувствит | нечувст | чувствит | нечувст | нечувст | нечувст | нечувст | верхний локоть |  |
| Log Det Ratio | MANOVA | мерные | - | едва чувствит | нечувст | нечувст | - | - | нечувст | эллипс | нечувст | стопка | верхний локоть |  |
| Ratkowsky–Lance | одномерный\* ANOVA; одномерный хи-квадрат | мерные; номинальные | - | нечувст | нечувст | нечувст | - | - | нечувст | эллипс | равные дисперсии вдоль осей | цепь | выше |  |
| AIC, BIC | Информационная энтропия (одномерная) | мерные; номинальные | - | чувствит | только смещен на константу | только смещен на константу | - | - | нечувст | эллипс | равные дисперсии вдоль осей | цепь | ниже | штраф «избытку» кластеров |
| Dunn | близость кластеров сравнительно с их «диаметрами» | - | любые | (ориг. версия)  чувствит | - | - | нечувст | чувствит | (ориг. версия) плоск | (ориг. версия) нечувст | (ориг. версия) нечувст | (ориг. версия) стопка, кольцо | выше | !KO\_DUNN вычисляет несколько версий критерия |
| Point-biserial r | кофенетическая корреляция | - | любые | нечувст | - | - | нечувст | нечувст | плоск | сферич | нечувст | стопка, кольцо | выше |  |
| Gamma | кофенетическая корреляция | - | любые | нечувст | - | - | нечувст | нечувст | не проверено | не проверено | не проверено | не проверено | выше |  |
| C-index | внутренняя плотность сравнительно с идеальной | - | любые | нечувст | - | - | нечувст | нечувст | плоск | сферич | нечувст | стопка, кольцо | ниже |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| McClain–Rao | внутренняя плотность относительно межкластерной | - | любые | нечувст | - | - | нечувст | чувствит | колокол | эллипс | нечувст | стопка, кольцо  (едва) | ниже, нижний локоть |  |
| Silhouette | оправданность приписания к своему кластеру, а не соседнему | - | любые | нечувст (ориг. версия; DEVIAT);  чувствит (прочие версии) | нечувст (DEVIAT) | чувствит (DEVIAT) | нечувст | чувствит | колокол (ориг. версия; DEVIAT);  плоск (NEAR; FARTH) | эллипс (ориг. версия; DEVIAT; NEAR); сферич (FARTH) | нечувст | стопка (ориг. версия; DEVIAT);  стопка, кольцо (NEAR); цепь, звезда (FARTH) | выше | вычисляется и для каждого объекта  !KO\_SILHOU вычисляет несколько версий критерия |

\*«в пространстве» означает учет многомерной природы данных, но игнорирование ковариационных связей. «одномерный» означает игнорирование многомерной/пространственной природы данных: индекс складывается или усредняется из слагаемых, вычисленных для отдельных переменных.

\*\*при тех же внутрикластерных дисперсиях и неизменных межцентроидных расстояниях.

\*\*\*при тех же внутрикластерных дисперсиях; эллипсоидные кластеры все параллельны друг другу.

\*\*\*\*близко расположенные кластеры: стопка vs цепь; кольцо vs звезда



**Внутренние кластерные критерии** или индексы (internal clustering criteria) существуют для оценки внутренней валидности разбиения объектов на группы (кластеры или иные классы).

Внутренняя валидность: общая идея

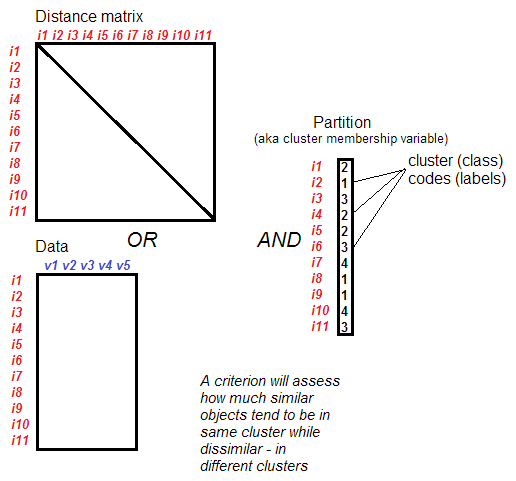
Внутренняя валидность разбиения набора объектов это его оправданность с позиции той информации о наборе объектов, которая использовалась процедурой, делавшей разбиение. Следовательно, внутренняя валидность отвечает на вопрос, «успешно» или «вполне» ли была учтена в деле разбиения та характеризующая объекты информация. (Напротив, внешняя валидность разбиения это то, как хорошо отвечает разбиение той информации о наборе объектов, которая не использовалась в деле разбиения.)

Внутренняя валидность: операционально

Внутренняя валидность у группирования тем выше, чем в большей степени похожие объекты попали в одну группу, а непохожие – в разные группы. Другими словами, точки одногруппные должны быть, в своей массе, больше похожи друг на друга, чем точки разногруппные. Или сформулировать в терминах плотности: чем более плотны группы внутри и чем меньшая плотность вне их (либо чем дальше группы раздвинуты), тем выше внутренняя валидность. Разные кластерные критерии, в зависимости от их формулы, *по-разному реализуют и акцентируют* этот интуитивный принцип, проверяя внутреннюю валидность.

Что входит

Разбиение (группирование) объектов и массив – данные (наблюдения X переменные) либо матрица близостей между объектами. Массив поставляет информацию о подобии между объектами.



Разбиение/группирование: какое

Внутренние кластерные критерии приложимы отнюдь не только к результатам кластеризации. Любое разбиение на классы *любого* происхождения (кластерный анализ, классификация машинная, ручная), если эти группы не пересекаются *составом* элементов (пространственно же классы могут и пересекаться), может быть проверено на внутреннюю валидность этими индексами. Критерии, представленные в данном документе, рассчитаны на неиерархические классификацию, т.е. группы не делятся на подгруппы в оцениваемом разбиении.

Применение: сравнение разных k

Чаще всего внутренние кластерные критерии применяются для сравнения кластерных разбиений с *разным числом кластеров k*, полученных одним и тем же методом кластеризации (или иного метода группирования) на основе одного и того же массива входящих (одной и той же матрицы близости или одних и тех же данных). Цель такого сравнения – выбрать лучшее k, т.е. разбиение с наиболее валидным числом кластеров. В таком контексте внутренние кластерные критерии иногда еще называют правилами остановки (stopping rules) кластеризации. Подробнее см. дальше.

Применение: сравнение разных методов

Вы можете сравнивать разбиения (с одним или разным числом k кластеров/классов), полученные и *разными способами* (например, разными методами кластерного анализа) на основе одного и того же массива входящих. Вообще, критерию безразлично, каким – одинаковым или нет – методом получены сравниваемые группирования, вы даже можете сами не знать, каким. Если вы сравниваете разные методы под одним и тем же значением k, вы этим выбираете «лучший» метод (при данном k).

Применение: сравнение не одинаковых наборов объектов

Это возможно. Надо понимать, что для кластерного критерия объекты “i” в массиве – просто безымянные строки. Поэтому будет корректно сравнить величиной критерия разбиения P1 и P2, частично или полностью состоящие из разных объектов. При этом k в разбиениях может быть одно или разное. Однако если P1 и P2 состоят из неодинакового числа объектов, можно пользоваться критерием, если он нечувствителен к числу объектов N.

Применение: с разными вариантами входящих (не одинаковыми признаками или не одинаковыми матрицами близостей)

Это возможно, но это тонкий и проблемный момент (потому что начинает выбиваться из идеи внутренней валидности). Речь идет о прямом сопоставлении значениев критерия val1 и val2, где val1 получено от входящих данных (переменных или близостей) X1 и разбиения P1, а val2 получено от входящих данных X2 и разбиения P2. Именно:

(1) Можно сравнить разбиения с одним и тем же k и полученные одним и тем же методом, но различающиеся употребленной *мерой близости* между объектами. Например, одно разбиение может быть результатом кластеризации матрицы евклидовых расстояний (норма L2), другое – матрицы манхэттенских расстояний (норма L1), третье – матрицы расстояний Минковского с нормой L3. Нет ничего формально противозаконного в таком сравнении, если вы готовы допустить, что разные типы расстояний, вычисленных на одних и тех же данных, *непосредственно сравнимы* в вашем случае. Если же они, эти меры, для вас имеют систематическое различие, которое хочется нивелировать (к примеру, разный уровень или разный разброс в величинах) – сделайте соответствующую «стандартизацию» матриц до вычисления кластерного критерия. Рассматривая вопрос о преобразовании матрицы близости, полезно также осведомиться о том, как тот или иной кластерный критерий реагирует на преобразования элементов матрицы. «Универсальные» критерии вроде точечно-бисериальной корреляции или C-индекса не реагируют на прибавление константы к близостям, поэтому общий уровень величины дистанций в матрице им не важен.

(2) Также можно в принципе сравнивать разбиения с одним и тем же k и полученные одним и тем же методом, но различающиеся набором признаков, *переменными в данных*. Здесь следует повторить все те же предостережения насчет *сравнимости* для вас значений этих разных наборов переменных: если они несравнимы (например, уровнем или разбросом) – позаботьтесь привести их к сравнимости. Также, кластерным критериям, как правило, небезразлично число переменных: будет неправильно в общем случае прямо сравнивать значение критерия, полученного на данных с 2-мя переменными, со значением, полученным на данных с 5-ю переменными.

(3) Отдельно скажем о линейном преобразовании переменных, таком как z-стандартизация. Можно ли сравнивать кластерным критерием разбиения (с одним и тем же k), одно из которых получено на сырых данных, а второе на этих же переменных стандартизованных? Ответ на вопрос зависит от конкретного критерия. Если критерий нечувствителен к *неодинаковому* линейному преобразованию переменных, тогда – можно.

Сравнение разных k: два типа критериев

Чаще всего внутренние кластерные критерии используются для выбора оптимального числа кластеров *k*. (Все эти кластерные разбиения с разным k должны были быть вами получены и присутствовать в массиве данных как переменные кластерного членства; т.е. критерий оценивает уже существующие, готовые разбиения.) Смотрят на график, где по оси X идут решения с разным числом кластеров в возрастающим или убывающим порядке, – например, k от 2 до 20, а по оси Y отложена величина критерия.

Есть критерии-*экстремумы* (extremum criteria) и критерии-*сгибами* (elbow criteria). У критериев-экстремумов чем значение выше (или наоборот, ниже – в зависимости от конкретного критерия), тем лучше разбиение; следовательно, абсолютно лучшее k соответствует максимальному (или минимальному) значению, когда k пробегает последовательные значения. У критериев-сгибов их значение с увеличением k монотонно растет (или наоборот, падает – в зависимости от конкретного критерия), и абсолютно лучшее k соответствует месту перепада этой тенденции, где дальнейшее увеличение k уже мало сопровождается ростом (падением) критерия. Преимущество критериев-экстремумов перед критериями-сгибами состоит в том, что по любым двум k можно судить, какое из них лучшее k; таким образом, критерии-экстремумы применимы для сравнения не только серии *последовательных* значений k. Критерии-сгибы не позволяют сравнивать несмежные k и вообще пары k, потому что неясно, по какую «сторону» от этих двух k или же между ними находится сгиб (elbow). Это существенный недостаток критериев-сгибов.

Сравнение разных k: приоритет остроты перед экстремумом

Надо сказать, что на практике острота перегиба – пика или сгиба – имеет большое значение и для критериев типа экстремумов. На графике профиля значения такого критерия под разными последовательными k следует обращать внимание не только на *max* (или *min*, в зависимости от конкретного критерия) значение в профиле, но на резкие перепады тенденции, не обязательно совпадающие с max. Если разбиение с данным k сильно лучше, чем разбиение с k-1 и с k+1, т.е. наличествует пик (peak), то это сильный довод в пользу этого k, даже если существуют на графике области k, где в общем критерий «лучше». Даже односторонний перепад (сгиб, elbow) может оказаться предпочтительнее абсолютному max для критерия-экстремума. Основание для этих советов следующее.

Дело в том, что разные кластерные критерии имеют свои *небольшие и носящие фоновый, присущий* характер склонности (bias) в отношении числа кластеров: некоторые «предпочитают» много кластеров, другие – мало кластеров. Причем проявление этих тенденций зависит и от особенностей данных: почти невозможно придумать массивы данных с разным числом кластеров k, которые были бы равновалидны другу другу одновременно для всех возможных критериев[[1]](#footnote-1). Симуляционные эксперименты, генерирующие определенные k кластеров, показывают, что все критерии-экстремумы временами «ошибаются», когда кластеры достаточно близки: ошибаются в том смысле, что общее max значение не совпадает с заявленным человеком числом сгенерированных кластеров. Если же обращать внимание на пики и сгибы, а не на max, то критерии «ошибаются» в таких экспериментах реже. (Надо, впрочем, сознавать ограниченность таких симуляционных экспериментов в оценке тенденциозностей кластерных критериев: ведь кластерный критерий не призван обнаружить задуманную при порождении кластерную структуру, он просто оценивает четкость структуры, как она вышла, а выйти при случайном порождении она могла не совсем такой, какой задумана.) По идее кластерные критерии, помогающие выбрать лучшее k, должны иметь нулевой базовый фаворитизм в отношении k. К сожалению, это едва ли достижимый идеал.

Некоторые критерии (например, BIC или PBM) сознательно предпочитают решения с небольшим числом кластеров, тогда говорят, что они «штрафуют избыток кластеров». C-Index, напротив, явно склонен поощрять решения с большим числом кластеров.

Критерий vs глаз

Если данные интервальные, кластеры нередко определимы зрительно на диаграммах рассеяния в пространстве переменных или же их главных компонент. Однако глаз имеет свои собственные предвзятости (апофения) и является одним из, и не лучшим, кластерным критерием. Нередко тот или иной кластерный критерий, основанный на статистической формуле, «вскроет» кластеры, не заметные глазу, интерпретация которых впоследствии подтвердит их валидность содержательно.

Выбор критерия: природа данных

Некоторые критерии (1) требуют в качестве входящих *массив данных* (наблюдения x переменные), и на кластеры/классы должны быть разбиты наблюдения. Одни из таких критериев требуют мерных, количественных переменных; другие – категориальных переменных или смесь мерных и категориальных. Некоторые критерии могут быть оптимальны для двоичных переменных. Другого типа критерии (2) основаны на анализе *матрицы близостей* между объектами. Зачастую таким критериям неважно, что выступает предметами, разбитыми на кластеры – наблюдения/респонденты или переменные/признаки, поскольку матрица близостей может существовать между предметами любой природы. Некоторые из критериев типа (2) требуют близостей определенных, например, евклидовых расстояний. Другим – безразличен тип близостей; эти последние называются универсальными критериями. (Впрочем, вопрос об «универсальности» тоньше, чем кажется, ведь эти критерии делают, к примеру, суммирование близостей, и возникает теоретический вопрос, всякие ли близости можно суммировать.) Некоторые критерии (3) могут быть вычислены эквивалентно как из данных (мерных), так и из матрицы (евклидовых расстояний).

Примечание: если вы кластеризовали наблюдения массива данных методом K-средних, а оцениваете качество решения критерием, требующий матрицы близостей (а не массива данных), то вычислите матрицу (квадратных) евклидовых расстояний (матрицы разных близостей можно в SPSS получить и вывести командами PROXIMITIES или CLUSTER) и вводите ее.

Число объектов, или холмистость

Есть критерии, реагирующие на (пропорционально одинаковое) увеличение или уменьшение численности в кластерах. Это кажется естественным, поскольку добавление объектов в кластеры усиливает рельефность формы распределения данных, когда кластеры не совпадают, и значение критерия ожидаемо улучшится. Но есть и критерии, которые не реагируют на такое изменение N: хотя этим критериям важно, чтобы плотность в кластерах была выше плотности вне кластеров, они не поощряют усиление плотности за счет увеличения в кластерах числа объектов.

Пространственная форма

Если критерий требует мерных данных или евклидовых расстояний, то кластеры могут иметь тут или иную конфигурацию в пространстве. Здесь разные кластерные критерии обладают своими предпочтениями, т.е. могут умеренно поощрять кластеры, имеющие в кластерном решении определенную пространственную форму или взаиморасположение. Этот достаточно сложный вопрос можно разбить на три подвопроса: чувствителен ли критерий, и как, (1) к форме очертаний кластера (круглые или вытянутые или изогнутые); (2) к повороту продолговатых кластеров относительно друг друга, т.е. вокруг своих центроидов; (3) к повороту всего облака данных вокруг общего центра (в пространстве мерных переменных).

Замечание к (1): бывает впечатление *ложного* предпочтения круглых кластеров. Ни один кластерный критерий не требует, чтобы кластеры не перекрывались своими краями в пространстве, однако большинство методов кластерного анализа дают кластеры как раз пространственно непересекающиеся. В этих условиях (кластерам не позволено физически взаимонакладываться) круглые кластеры могут притиреться в пространстве ближе друг к другу, чем продолговатые кластеры с неконтролируемой повернутостью, отчего последние попросту имеют меньше шансов встретиться или сформироваться кластеризацией в реальных исследуемых данных – где, как известно, кластеры обычно расположены друг к другу впритык. Из-за этого явления критерии, *не* чувствительные к очертанию кластера, такие как Calinski-Harabasz, чаще *наталкиваются* на «хорошие» решения с круглыми, нежели с продолговатыми кластерами. Это не значит, что эти критерии сами по себе предпочитает круглые кластеры.

Форма распределения в кластерах

Есть критерии, отдающие предпочтение кластерам с равномерным, плоским распределением внутри (например, гипершар), и критерии, отдающие предпочтение кластерам с колоколообразным распределением внутри (как нормальное распределение); другим же критериям форма распределения плотности в кластере не важна.

Размерность пространства

Еще один непростой вопрос – реакция разных кластерных критериев на увеличение размерности пространства, которое «расправляют» собой данные, разбитые на кластеры. Этот вопрос связан, в том числе и с проклятием размерности (curse of dimensionality), «висящим» над евклидовыми расстояниями, на которых основано много кластерных критериев.

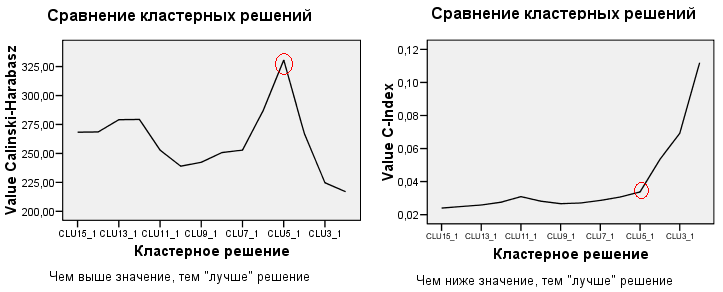
Статистическая значимость

Внутренние кластерные критерии не сопровождаются вероятностным p-значением, т.к. не делают заключений о популяции, а занимаются только наличным массивом. Разумеется, хорошее кластерное решение в виде высокого значения критерия может быть следствием случайных особенностей конкретной выборки (overfitting). Кросс-валидация эквивалентными массивами (в виде проверок стабильности или проверок генерализабельности) будет всегда в помощь.

**ПРИМЕР**. Применение двух кластерных критериев для решения о числе кластеров в кластерном анализе. Вот пять изрядно контактирующих кластеров; глаз не сразу распознаёт их.



С этим облаком точек проделали иерархический кластерный анализ методом средней связи на базе евклидовых расстояний, и сохранили все разбиения от 15-кластерного до 2-кластерного. Затем применены 2 кластерных критерия, Calinski–Harabasz и C-Index, в попытке выбрать, какое же решение лучшее.



Как видно на графике слева, Calinski–Harabasz довольно легко (в данном примере) справился с задачей, указав на 5-кластерное решение как на абсолютно лучшее. C-Index, однако, рекомендует 15-ти или 9-кластерное решения (C-Index тем «лучше», чем ниже). Тем не менее следует это игнорировать, а обратить внимание на излом, который дает C-Index на уровне 5 кластеров: 5-кластерное решение еще хорошее, но 4-кластерное уже гораздо хуже. Поэтому лучшее решение – 5-кластерное даже на правом графике.

Разумеется, надо понимать, что если в ваших данных кластерная структура почти совсем отсутствует, то никакие критерии не помогут выбрать «правильное» кластерное решение, ибо его нет. В таких случаях на графиках не будет изломов, а будут сравнительно гладкие линии, возрастающие, снижающиеся или горизонтальные – в зависимости от данных и от критерия.

# МАКРОС !KO\_CALHARV: КРИТЕРИИ CALINSKI–HARABASZ, SSW и LOG SS RATIO (ВХОДЯЩИЕ – ПЕРЕМЕННЫЕ)

Version 3, Jul 2018 (Version 1, May 2001). Tested on SPSS Statistics 17, 20, 22.

!KO\_calharv vars= *v1 v3 to v10* /\*Количественные переменные, по которым сравнить кластерные решения

/\*(поименно и/или ч-з to)

/missing= VARIABLE /\*Исключать наблюдения с пропущенными данными в VARS списочно

/\*(LISTWISE, тж п/у) или бережливо (VARIABLE)

/clusol= *clu10\_1 clu9\_1 clu8\_1 clu7\_1 clu6\_1 clu5\_1 clu4\_1 clu3\_1 clu2\_1*

/\*Переменные, являющие собой кластерные решения (поименный список).

Минимум надо задать VARS, CLUSOL.

**Критерий Calinski–Harabasz**

Этот индекс есть F Фишера, примененный к многомерным данным (поскольку он не учитывает ковариационные отношения, его называют многомерным псевдо-F Фишера). Он есть отношение междукластерной неплотности, понимаемой как дисперсия центров кластеров вокруг центра всей конфигурации точек, к внутрикластерной неплотности, понимаемой как средневзвешенная дисперсия членов кластеров вокруг своих центров. Чем выше величина критерия, тем лучше кластерное разбиение. Формула (Calinski, R.B., Harabasz, J. A dendrite method for cluster analysis // Communications in Statistics, 1974, 3, 1–27.):

где *W* это сумма квадратов отклонений внутри кластеров в данном *k*-кластерном решении (эта сумма равна следу совокупной внутрикластерной матрицы рассеяния **W**); *B* это сумма квадратов отклонений кластерных центроидов от общей средней (эта сумма равна следу межкластерной матрицы рассеяния **B**); *n* – число объектов в данных.

Будучи основан на идее дисперсионного анализа, критерий рассчитан прежде всего на выпуклые кластеры, но ему все равно ­– круглые они или продолговатые (эллипсоидные). Он также нечувствителен к повороту данных в пространстве: как повороту всего облака, так и повороту отдельных кластеров относительно своих центров. Он нечувствителен к форме распределения в кластерах (колокол vs равномерное): хотя критерий и воплощает идею ANOVA, он не делает теста значимости и потому не настаивает на нормальном распределении.

Критерий чувствителен к числу объектов в данных, поэтому сравниваемые кластерные разбиения должны состоять из равного общего числа объектов *n*. Хотя зависимость от *n* линейна у данного критерия, она зависит от *k*, поэтому ее нелегко устранить.

При прочих равных условиях критерий склонен отдавать предпочтение кластерным разбиениям с приблизительно равновеликими по количеству объектов кластерами (поскольку на фоне многообъектных кластеров малообъектные кластеры мало что вносят в *B*, однако они входят в число *k*, и это уменьшает числитель формулы, дисперсию центроидов).

**Критерий SSW**

Это просто совокупная внутрикластерная сумма квадратов отклонений, *W*. Критерий годится только для сравнения разбиений одних и тех же данных с разным числом кластеров *k*. На графике, где по оси абсцисс последовательные k отложены по возрастанию или убыванию, «лучшее» разбиение соответствует сгибу («локтю»), после которого кривая для бо́льших k уже сравнительно полога. Это популярный, но примитивный показатель.

**Критерий Log SS Ratio**

логарифм отношения между меж- и внутрикластерными суммами квадратов отклонений (Hartigan, J.A. Clustering algorithms. New York: Wiley, 1975).

Этот критерий обычно похож графиком на SSW, только перевернут. SSW помечает лучшее разбиение «нижним локтем», а Log SS Ratio – «верхним локтем». В остальном критерий похож на Calinski–Harabasz (только это критерий-сгиб, не критерий-экстремум). Его преимущество перед Calinski–Harabasz в том, что он почти не зависит от общего числа объектов *n*.

***Подкоманды***

**VARS**

Переменные (количественные), по которым требуется оценивать кластерные разбиения наблюдений. В них должно не быть значения: *-999.*

**MISSING**

По умолчанию и при MISSING=LISTWISE наблюдения, имеющие пропуск хотя бы в одной переменной VARS, исключаются из анализа. MISSING=VARIABLE исключает пропуски из каждой переменной независимо и использует в калькуляциях все валидные значения переменных. Индекс не вычисляем (выйдет в результаты как sysmis), если по исключении пропусков окажется, что хотя бы по одной из VARS кластеров меньше 2-х.

**CLUSOL**

Поименный список группирующих наблюдения переменных (минимум одна) - кластерные разбиения, или «переменные кластерного членства» (cluster membership variables); имена переменных должны быть *не длиннее 8 байтов*. Переменные рекомендуется писать в порядке увеличения или уменьшения числа кластеров в них. Минимальное число кластеров в переменной – 2 (иначе результат выйдет как system missing).

Значения группирующих переменных (коды кластеров) могут быть любыми числами: каждое уникальное значение обозначает один кластер. Макрос исключает из анализа, отдельно из каждой группирующей переменной, наблюдения, имеющие пропуски или значение 0. Исследователь может этим пользоваться, если хочет проигнорировать в анализе какие-л. целые кластеры или отдельные объекты: надо лишь сделать коды этих кластеров/объектов в данной группирующей переменной user- или system-missing или заменить на 0; число оставляемых для анализа кластеров должно быть не меньше 2-х. Таким образом, сравниваемые разбиения могут, в принципе, состоять из разного числа объектов (насколько допустимо такое сравнение, зависит от критерия).

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива (SPLIT FILE). Он не слушается временных операций (стоящих под TEMPORARY). Для отбора наблюдений в анализ вы можете использовать фильтрацию (FILTER или USE) и SELECT IF. Кроме того, вы можете исключать из анализа разные наблюдения в разных кластерных решениях, приписав им missing или нулевые коды (см. об этом в п/к CLUSOL).

# МАКРОС !KO\_CALHARM: КРИТЕРИИ CALINSKI–HARABASZ, SSW и LOG SS RATIO (ВХОДЯЩИЕ – МАТРИЦА)

Version 3, Jul 2018 (Version 1, Jul 2000). Tested on SPSS Statistics 17, 20, 22.

!KO\_calharm matrix= *var1 to var100* /\*Переменные, составляющие матрицу расстояний (поименно и/или ч-з to)

/square= NO /\*Возводить ли расстояния в квадрат (YES) или не возводить (NO, тж п/у),

/\*т к они уже квадратные

/clusol= *clu10\_1 clu9\_1 clu8\_1 clu7\_1 clu6\_1 clu5\_1 clu4\_1 clu3\_1 clu2\_1*

/\*Переменные, являющие собой кластерные решения (поименный список).

Минимум надо задать MATRIX, CLUSOL.

Этот макрос вычисляет критерии Calinski-Harabasz, SSW, Log SS Ratio, как макрос !KO\_CALHARV, но принимает в качестве входящих не переменные-признаки, а квадратную симметрическую матрицу расстояний между объектами. Ради геометрической правильности требуется, чтобы это были квадратные евклидовы расстояния или евклидовы расстояния (макрос может оквадратить их сам). Если это другие расстояния/различия чем евклидовы, данные критерии корректны ровно настолько, насколько расстояния близки к евклидовым. Во всяком случае, им следует быть метрическими. В любом случае матрица должна состоять из расстояний, не сходств.

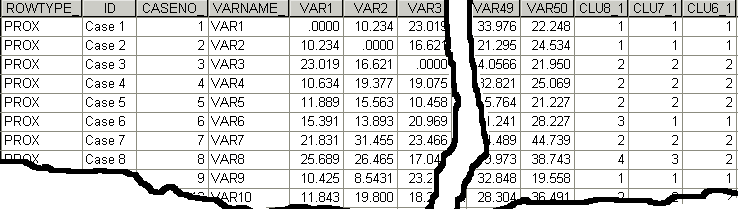
Calinski-Harabasz и Log SS Ratio нечувствителены к пропорциональному изменению близостей (квадратных или неквадратных входящих расстояний).

***Строение матрицы***

Матрица близостей должна быть квадратной симметричной. Макрос не требует, чтобы матрица содержала столбцы ROWTYPE\_ и VARNAME\_ и прочие вспомогательные. К матрице в тот же массив должны быть подшиты переменные кластерного членства. В массиве должно не быть переменной с именем *CASENUM#*.

ПРИМЕР 1.

!KO\_calharm matrix= var1 to var50 /clusol= clu8\_1 clu7\_1 clu6\_1.



* Тело квадратной симметрической матрицы расстояний между 50-ю объектами задано столбцами VAR1 до VAR50 (и соответственно таким же числом рядов). Переменные кластерных решений – CLU8\_1 CLU7\_1 CLU6\_1, значения в них – коды кластеров.
* Макрос вычислит 3 значения индекса – один для каждого кластерного решения, выведет их в новый файл и построит график.

***Подкоманды***

**MATRIX**

Укажите переменные, являющиеся столбцами собственно матрицы расстояний. Матрица должна быть квадратной симметрической и не может содержать пропуски. Вы можете указать список поименно и/или диапазоном через “to”. Тело матрицы, находящееся в рабочем массиве, должно быть квадратным и симметричным, и вы должны ввести подкомандой MATRIX все столбцы тела.

*Задание открытым диапазоном с помощью “?”*. В некоторых случаях возникает нужда указать диапазон, заключенный между парой переменных, которые сами не входят в диапазон. Используйте для этого “?” по краям. Например, *?VARNAME\_ to ENDVAR?* означает все переменные, находящиеся в массиве между переменными *VARNAME\_* и *ENDVAR*, не включая их самих. Для обозначения диапазона, открытого с одного конца, используйте “?” только у этого конца. Например: *?VARNAME\_ to VAR100* или *VAR1 to ENDVAR?*.

**SQUARE**

Макрос относится к входящим расстояниям как к квадратным евклидовым. Если ваша матрица это неквадратные евклидовы расстояния (или меры, которые вы трактуете как неквадратные евклидовы расстояния), то вам следует указать макросу, чтобы он возвел их в квадрат: SQUARE=YES.

**CLUSOL**

Поименный список группирующих наблюдения переменных – см. описание подкоманды выше, в макросе [!KO\_CALHARV](#_МАКРОС_!CALHARV:_КРИТЕРИИ).

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива (SPLIT FILE). Он не слушается временных операций (стоящих под TEMPORARY). Не применяйте перед макросом SELECT IF, т.к. порушите квадратность матрицы. Вы можете, однако, использовать для отбора наблюдений фильтрацию (FILTER или USE). Кроме того, вы можете исключать из анализа разные наблюдения в разных кластерных решениях, приписав им missing или нулевые коды (см. об этом в п/к CLUSOL).

# МАКРОС !KO\_DAVBOULV: КРИТЕРИИ DAVIES-BOULDIN и PBM (ВХОДЯЩИЕ – ПЕРЕМЕННЫЕ)

Version 3, Jun 2018 (Version 1, Sep 2012). Tested on SPSS Statistics 17, 20, 22.

!KO\_davboulv vars= *v1 v3 to v10* /\*Количественные переменные, по которым сравнить кластерные решения

/\*(поименно и/или ч-з to)

/clusol= *clu10\_1 clu9\_1 clu8\_1 clu7\_1 clu6\_1 clu5\_1 clu4\_1 clu3\_1 clu2\_1*

/\*Переменные, являющие собой кластерные решения (поименный список)

/norm= /\*Использовать евклидову (L2, тж п/у) или манхэттенскую (L1) норму.

Минимум надо задать VARS, CLUSOL.

**Критерий Davies–Bouldin**

Этот критерий (Davies, D.L., Bouldin, D.W. A cluster separation measure // IEEE Transactions on Pattern

Analysis and Machine Intelligence, 1979, 1(2), 224-227) есть отношение внутрикластерной неплотности к расстоянию между кластерными центрами. Формула:



где *s* это среднеквадратическое отклонение в кластере от его центроида, мера его распыленности; *d* это евклидово расстояние между центроидами кластеров *i* и *j*. Каждый кластер *i* (всего кластеров *k*) сопоставляется с каждым из остальных, и максимальное отношение суммы двух *s* к *d* записывается. Эта величина характеризует степень необособленности кластера *i*. Величина критерия, характеризующая кластерное разбиение в целом, есть средняя этих *k* значений. Чем ниже величина критерия, тем лучше кластерное разбиение.

Основанный, как и Calinski-Harabasz, на идее ANOVA в пространстве, данный критерий подобен ему большинством своих свойств в отношении к пространственной конфигурации и к внутрикластерному распределению. Критерий нечувствителен к числу объектов *n*, поэтому им можно сравнивать группирования с разным числом объектов.

В отличие от Calinski-Harabasz критерий Davies-Bouldin не склонен к равновеликим по численности точек кластерам; зато он имеет тенденцию отдавать предпочтение решения с равноудаленными друг от друга кластерами (что связано с тем, что разделенность кластеров меряется не от общего центра конфигурации, а от кластера-соседа).

**Критерий PBM**

Этот критерий, названный как акроним имен его авторов (Pakhira, M.K., Bandyopadhyay, S., Maulik, U. Validity index for crisp and fuzzy clusters // Pattern Recognition, 2004, 37, 487-501) есть произведение трех множителей:

где есть сумма отклонений (не сумма квадратов отклонений) объектов от общего центроида выборки (всего объектов *n*); есть сумма отклонений *n* объектов от центроидов кластеров – каждый объект от центроида своего кластера; это расстояние между парой кластерных центроидов, – таким образом, третий множитель это наибольшее расстояние между кластерными центроидами. Идея критерия следующая. Второй множитель выражает внутрикластерную компактность, а третий множитель – межкластерную разведенность; чем одновременно выше оба, тем лучше кластерное разбиение. Однако с ростом дробности (числа кластеров, *k*) имеется естественная тенденция к увеличению обоих множителей и, поскольку желательно ее снять, введен сдерживающий множитель 1/*k*. Таким образом критерий «штрафует» за избыток кластеров. Если , макрос не вычислит значение критерия. Чем выше значение критерия, тем лучше разбиение.

Критерий чувствителен к линейному преобразованию переменных; если оно одинаковое для всех переменных – то реагирует линейно. Имеется некоторое предпочтение к эллипсоидным кластерам перед круглыми. Критерий слабо чувствителен к числу объектов *n*.

***Подкоманды***

**VARS**

Переменные (количественные), по которым требуется оценивать кластерные разбиения наблюдений. Макрос исключает наблюдения, имеющие пропуски, списочно.

**CLUSOL**

Поименный список группирующих наблюдения переменных – см. описание подкоманды выше, в макросе [!KO\_CALHARV](#_МАКРОС_!CALHARV:_КРИТЕРИИ).

**NORM**

Эта подкоманда позволяет заказать критерии на базе манхэттенских, вместо евклидовых, расстояний. По умолчанию и при NORM=L2 (см. формулы критериев) *ew* и *et* это евклидовые отклонения объекта, *s* это стандартное, т.е. среднеквадратическое отклонение (посчитанное на “df=n”) в кластере от его центроида, а *d* это евклидовое расстояние между центроидами кластеров. При NORM=L1 *ew* и *et* это манхэттенские отклонения объекта, *s* это среднее абсолютное отклонение в кластере от его центроида, а *d* это манхэттенское расстояние между центроидами кластеров.

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива (SPLIT FILE). Он не слушается временных операций (стоящих под TEMPORARY). Для отбора наблюдений в анализ вы можете использовать фильтрацию (FILTER или USE) и SELECT IF. Кроме того, вы можете исключать из анализа разные наблюдения в разных кластерных решениях, приписав им missing или нулевые коды (см. об этом в п/к CLUSOL).

# МАКРОС !KO\_DAVBOULM: КРИТЕРИИ DAVIES-BOULDIN и PBM (ВХОДЯЩИЕ – МАТРИЦА)

Version 2, Jun 2018 (Version 1, Sep 2012). Tested on SPSS Statistics 17, 20, 22.

!KO\_davboulm matrix= *var1 to var100* /\*Переменные, составляющие матрицу расстояний (поименно и/или ч-з to)

/square= NO /\*Возводить ли расстояния в квадрат (YES) или не возводить (NO, тж п/у),

/\*т к они уже квадратные

/clusol= *clu10\_1 clu9\_1 clu8\_1 clu7\_1 clu6\_1 clu5\_1 clu4\_1 clu3\_1 clu2\_1*

/\*Переменные, являющие собой кластерные решения (поименный список).

Минимум надо задать MATRIX, CLUSOL.

Этот макрос вычисляет критерии Davies-Bouldin и PBM, как макрос !KO\_DAVBOULV (с NORM=L2), но принимает в качестве входящих не переменные-признаки, а квадратную симметрическую матрицу расстояний между объектами. Ради геометрической правильности требуется, чтобы это были квадратные евклидовы расстояния или евклидовы расстояния (макрос может оквадратить их сам). Если это другие расстояния/различия чем евклидовы, Davies-Bouldin и PBM корректны ровно настолько, насколько расстояния близки к евклидовым. Во всяком случае, им следует быть метрическими. Ситуация и ввод матрицы точно такие, как в макросе [!KO\_CALHARM](#_МАКРОС_!CALHARM:_КРИТЕРИЙ).

Davies-Bouldin нечувствителен, а PBM чувствителен к пропорциональному изменению близостей (квадратных или неквадратных входящих расстояний).

***Подкоманды***

**MATRIX**

Укажите переменные, являющиеся столбцами собственно матрицы расстояний. Матрица должна быть квадратной симметрической и не может содержать пропуски. Вы можете указать список поименно и/или диапазоном через “to”. Тело матрицы, находящееся в рабочем массиве, должно быть квадратным и симметричным, и вы должны ввести подкомандой MATRIX все столбцы тела.

*Задание открытым диапазоном с помощью “?”*. В некоторых случаях возникает нужда указать диапазон, заключенный между парой переменных, которые сами не входят в диапазон. Используйте для этого “?” по краям. Например, *?VARNAME\_ to ENDVAR?* означает все переменные, находящиеся в массиве между переменными *VARNAME\_* и *ENDVAR*, не включая их самих. Для обозначения диапазона, открытого с одного конца, используйте “?” только у этого конца. Например: *?VARNAME\_ to VAR100* или *VAR1 to ENDVAR?*.

**SQUARE**

Макрос относится к входящим расстояниям как к квадратным евклидовым. Если ваша матрица содержит неквадратные евклидовы расстояния (или меры, которые вы трактуете как неквадратные евклидовы расстояния), то вам следует указать макросу, чтобы он возвел их в квадрат: SQUARE=YES.

**CLUSOL**

Поименный список группирующих наблюдения переменных – см. описание подкоманды выше, в макросе [!KO\_CALHARV](#_МАКРОС_!CALHARV:_КРИТЕРИИ).

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива (SPLIT FILE). Он не слушается временных операций (стоящих под TEMPORARY). Не применяйте перед макросом SELECT IF, т.к. порушите квадратность матрицы. Вы можете, однако, использовать для отбора наблюдений фильтрацию (FILTER или USE). Кроме того, вы можете исключать из анализа разные наблюдения в разных кластерных решениях, приписав им missing или нулевые коды (см. об этом в п/к CLUSOL).

# МАКРОС !KO\_CCCRITV: КУБИЧЕСКИЙ КЛАСТЕРНЫЙ КРИТЕРИЙ, КРИТЕРИЙ LOG DET RATIO (ВХОДЯЩИЕ – ПЕРЕМЕННЫЕ)

Version 2, Jul 2018 (Version 1, Jul 2016). Tested on SPSS Statistics 17, 20, 22.

!KO\_cccritv vars= *v1 v3 to v10* /\*Количественные переменные, по которым сравнить кластерные решения

/\*(поименно и/или ч-з to)

/clusol= *clu10\_1 clu9\_1 clu8\_1 clu7\_1 clu6\_1 clu5\_1 clu4\_1 clu3\_1 clu2\_1*

/\*Переменные, являющие собой кластерные решения (поименный список)

/fast= YES /\*Быстрый режим: YES или NO (тж п/у).

Минимум надо задать VARS, CLUSOL.

**Кубический кластерный критерий**

Этот индекс основан на отношении внутрикластерной суммы квадратов отклонений к полной сумме квадратов отклонений, *W/T*, величине, известной как лямбда Уилкса. В многомерных данных лямбда учитывает и ковариационные отношения, но в случае кубического кластерного критерия (CCC) они игнорируются, поэтому можно говорить о многомерной псевдо-лямбде Уилкса (как в случае критерия Calinski-Harabasz мы говорим о псевдо-F Фишера). 1-W/T есть R-squared = Eta-squared. CCC имеет формулу (Sarle, W.S. Cubic clustering criterion // SAS Technical Report A-108, 1983, SAS Institute):

где

(**W** есть совокупная внутрикластерная матрица рассеяния; **T** есть матрица рассеяния всей выборки)

*RsqE* есть ожидаемое значение *Rsq* в условиях, когда наши данные были бы бескластерно-однородны и представляли собой прямоугольный брикет из соприкасающихся кубиков. Таким образом, CCC это логарифм отношения наблюдаемого *W/T* к ожидаемому в отсутствие кластеров *W/T*. *Stab* это эмпирически подобранный стабилизатор (см. ниже).

где *n* число объектов в данных; *k* это число кластеров; *p* это размерность данных, которая равна числу ненулевых сингулярных чисел ковариационной матрицы данных **T***/(n-1)*. Величины *u* есть эти сингулярные числа нормированные, т.е. деленные на константу *c*, которая определяется следующим образом:

где *P* есть произведение первых *p\** сингулярных чисел. Перебирая *p\*= p, p-1, p-2…,* вычисляют повторно *с* и нормируют им сингулярные числа; останавливают выбор на таком *p\** и следовательно таких *u* для формулы, чтобы *p\** было меньше *k*, а *up\** было не меньше 1.

есть поправка, позволяющая сравнивать значения CCC в данных с разной размерностью (числе переменных) и разным числом объектов. Таким образом, CCC – один из немногих кластерных индексов, который при разработке попытались (насколько успешно – вопрос) стандартизовать для разных массивов данных. Еще он отталкивается от конкретной «нулевой» гипотезы: безкластерные данные исходят из многомерного равномерного распределения («гиперкоробка»). Сложность формул обязана не базовой идее (*W/T*) критерия, а стандартизации его значения. CCC может быть положительным или отрицательным. Чем выше значение, тем лучше кластерное решение, т.е. тем выраженнее кластеры.

Замечу, что моя проверка данного критерия, выдаваемая макросом, заставляет усомниться в достаточной стандартизованности CCC. В частности, критерий оказался весьма чувствителен к числу объектов в данных, поэтому, по моему мнению, сравниваемые кластерные разбиения должны состоять из равного общего числа объектов *n*.

Как основанный на идее дисперсионного анализа CCC лучше всего подходит для обнаружения выпуклых кластеров. Он напоминает Calinski-Harabasz по ряду свойств, но, в отличие от него, предпочитает круглые кластеры продолговатым (при размерности пространства *p* выше 2 или 3), а из последних – параллельные в виде стопки. Почему CCC имеет такие предпочтения, становится понятно, если проанализировать формулу *RsqE*, как эта величина реагирует на пространственные мутации облака.

**Критерий Log Det Ratio**

где **T** и **W** есть, соответственно, полная матрица рассеяния и совокупная внутригрупповая матрица рассеяния. В оригинале (Scott, A.J., Symons, M.J. Clustering methods based on likelihood ratio criteria // Biometrics, 1971, 27, 387-397) критерий еще умножен на *n*, но это излишне. Если один или оба определителя нулевые, критерий не вычисляется. Данный показатель выражает идеологию MANOVA, т.е. откровенно чувствителен к форме облака данных и к форме совокупного (совмещенного центроидами) кластера. Это понятно, ведь определитель det связан с объемом многомерной фигуры. Значение критерия увеличено, если кластеры продолговаты и параллельны, но совокупно занимают большой объем (т.е. не «цепочка» кластеров). Впрочем, критерий относится к типу критериев-сгибов, а не критериев-экстремумов. Лучшее число кластеров – такое k, дальнейшее увеличение которого сопровождается лишь пологим ростом значения критерия.

***Подкоманды***

**VARS**

Переменные (количественные), по которым требуется оценивать кластерные разбиения наблюдений. Макрос исключает наблюдения, имеющие пропуски, списочно.

**CLUSOL**

Поименный список группирующих наблюдения переменных – см. описание подкоманды выше, в макросе [!KO\_CALHARV](#_МАКРОС_!CALHARV:_КРИТЕРИИ).

**FAST**

Эта подкоманда действует, только если в CLUSOL указано более одной переменной. FAST=YES ускоряет работу макроса, что может быть заметно в случае больших данных (много переменных VARS и много кластерных решений CLUSOL). Не применяйте FAST=YES, если в каких-л. переменных CLUSOL наличествуют пропущенные значения или нули. По умолчанию/незаданию FAST=NO.

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива (SPLIT FILE). Он не слушается временных операций (стоящих под TEMPORARY). Для отбора наблюдений в анализ вы можете использовать фильтрацию (FILTER или USE) и SELECT IF. Кроме того, вы можете исключать из анализа разные наблюдения в разных кластерных решениях, приписав им missing или нулевые коды (см. об этом в п/к CLUSOL).

# МАКРОС !KO\_CCCRITM: КУБИЧЕСКИЙ КЛАСТЕРНЫЙ КРИТЕРИЙ (ВХОДЯЩИЕ – МАТРИЦА)

Version 1, Jul 2016. Tested on SPSS Statistics 17, 20, 22.

!KO\_cccritm matrix= *var1 to var100* /\*Переменные, составляющие матрицу расстояний (поименно и/или ч-з to)

/square= NO /\*Возводить ли расстояния в квадрат (YES) или не возводить (NO, тж п/у),

/\*т к они уже квадратные

/clusol= *clu10\_1 clu9\_1 clu8\_1 clu7\_1 clu6\_1 clu5\_1 clu4\_1 clu3\_1 clu2\_1*

/\*Переменные, являющие собой кластерные решения (поименный список)

/fast= YES /\*Быстрый режим: YES или NO (тж п/у).

Минимум надо задать MATRIX, CLUSOL.

Этот макрос вычисляет кубический кластерный критерий (CCC), как макрос !KO\_CCCRITV, но принимает в качестве входящих не переменные-признаки, а квадратную симметрическую матрицу расстояний между объектами. Ради геометрической правильности требуется, чтобы это были квадратные евклидовы расстояния или евклидовы расстояния (макрос может оквадратить их сам). Если это другие расстояния чем евклидовы, критерий CCC корректен ровно настолько, насколько расстояния близки к евклидовым. Во всяком случае, им следует быть метрическими. Ситуация и ввод матрицы точно такие как в макросе [!KO\_CALHARM](#_МАКРОС_!CALHARM:_КРИТЕРИЙ).

Индекс нечувствителен к пропорциональному изменению близостей (квадратных или неквадратных входящих расстояний).

В отличие от !KO\_CCCRITV данный макрос не вычисляет критерий Log Det Ratio. Если нужен этот критерий, создайте из матрицы расстояний данные «объекты × переменные», т.е. координаты, и используйте !KO\_CCCRITV. Координаты создаются из матрицы расстояний метрическим многомерным шкалированием, например, методом Торгерсона, или PCoA; при этом надо затребовать создать переменных побольше, чтобы исчерпать всю или почти всю дисперсию. Простой макрос PCoA можно найти среди моих матричных функций “MATRIX – END MATRIX functions”.

***Подкоманды***

**MATRIX**

Укажите переменные, являющиеся столбцами собственно матрицы расстояний. Матрица должна быть квадратной симметрической и не может содержать пропуски. Вы можете указать список поименно и/или диапазоном через “to”. Тело матрицы, находящееся в рабочем массиве, должно быть квадратным и симметричным, и вы должны ввести подкомандой MATRIX все столбцы тела.

*Задание открытым диапазоном с помощью “?”*. В некоторых случаях возникает нужда указать диапазон, заключенный между парой переменных, которые сами не входят в диапазон. Используйте для этого “?” по краям. Например, *?VARNAME\_ to ENDVAR?* означает все переменные, находящиеся в массиве между переменными *VARNAME\_* и *ENDVAR*, не включая их самих. Для обозначения диапазона, открытого с одного конца, используйте “?” только у этого конца. Например: *?VARNAME\_ to VAR100* или *VAR1 to ENDVAR?*.

**SQUARE**

Макрос относится к входящим расстояниям как к квадратным евклидовым. Если ваша матрица содержит неквадратные евклидовы расстояния (или меры, которые вы трактуете как неквадратные евклидовы расстояния), то вам следует указать макросу, чтобы он возвел их в квадрат: SQUARE=YES.

**CLUSOL**

Поименный список группирующих наблюдения переменных – см. описание подкоманды выше, в макросе [!KO\_CALHARV](#_МАКРОС_!CALHARV:_КРИТЕРИИ).

**FAST**

Эта подкоманда действует, только если в CLUSOL указано более одной переменной. FAST=YES ускоряет работу макроса, что может быть заметно в случае больших данных (большая матрица MATRIX и много кластерных решений CLUSOL). Не применяйте FAST=YES, если в каких-л. переменных CLUSOL наличествуют пропущенные значения или нули. По умолчанию/незаданию FAST=NO.

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива (SPLIT FILE). Он не слушается временных операций (стоящих под TEMPORARY). Не применяйте перед макросом SELECT IF, т.к. порушите квадратность матрицы. Вы можете, однако, использовать для отбора наблюдений фильтрацию (FILTER или USE). Кроме того, вы можете исключать из анализа разные наблюдения в разных кластерных решениях, приписав им missing или нулевые коды (см. об этом в п/к CLUSOL).

# МАКРОС !KO\_RATLAN: КРИТЕРИЙ RATKOWSKY–LANCE

Version 2, Jul 2016 (Version 1, Jul 2012). Tested on SPSS Statistics 17, 20, 22.

!KO\_ratlan scavars= *V1 V3 to V10* /\*Количественные переменные, по которым сравнить кластерные решения (можно /\*ч-з to)

/catvars= /\*Категориальные переменные, по которым сравнить кластерные решения (можно ч-з to)

/clusol= *CLU10\_2 CLU9\_2 CLU8\_2 CLU7\_2 CLU6\_2 CLU5\_2 CLU4\_2 CLU3\_2 CLU2\_2*

/\*Переменные, являющие собой кластерные решения (поименный список)

/univar= NO /\*Сохранить и значение критерия для каждой переменной: YES или NO (тж п/у).

Минимум надо задать CLUSOL и хотя бы одно из SCAVARS, CATVARS.

Данный индекс (Ratkowsky, D.A., Lance, G.N. A criterion for determining the number of groups in a classification // Australian Computer J, 1978, 10, 115-117.) оценивает степень рассеяния кластеров отдельно по каждой переменной, и значения индекса от всех переменных усредняются в одно значение. Критерий подходит как для мерных, так и для номинальных переменных.

Для каждой мерной переменной *q* индекс есть (нормированное числом кластеров *k*) корреляционное отношение Эта:

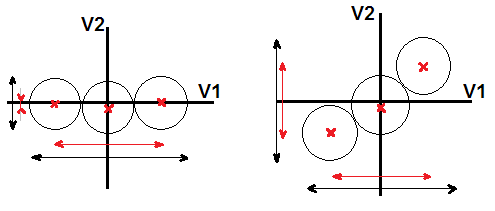
(*B* и *T* есть соответственно междукластерная и полная сумма квадратов отклонений)

Для каждой номинальной переменной *r* индекс есть (нормированный числом кластеров *k*) ассоциативный показатель V Крамера:

(*χ2* это мера связи хи-квадрат между этой переменной (с числом категорий *l*) и кластерной переменной; *n* есть размер выборки)

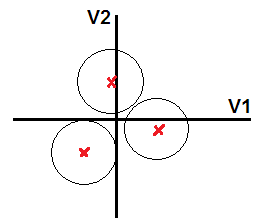
Совокупный индекс есть просто средняя арифметическая этих величин от всех переменных, *p1* мерных и *p2* номинальных; иначе говоря, Ratkowsky-Lance есть

Чем выше значение, тем лучше разбиение. Поскольку критерий является усреднительно-одномерным, он чувствителен к повороту облака данных в пространстве мерных переменных. На рисунке слева три кластера различаются центроидами вдоль V1 и почти совпадают центроидами вдоль V2. По переменной V2 разброс центроидов относительно изменчивости данных очень мал и критерий по этой переменой близок к 0. Т.к. совокупный Ratkowsky-Lance есть простая невзвешенная средняя между значением критерия по V1 и значением критерия по V2, то это совокупное значение для рисунка слева будет мало, - ниже, чем таковое для рисунка справа, где облако данных повернуто так, что центроиды получаются достаточно сильно разбросаны по обеим переменным. Ratkowsky-Lance поощряет данные, повернутые так, чтобы дисперсия мало различалась вдоль осей-переменных.

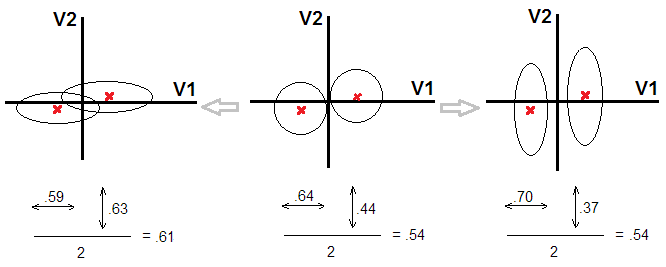


Таким образом, совокупный Ratkowsky-Lance не подойдет вам, если для вас повернутые в пространстве данные – это «те же» данные, что до поворота. Вы можете хотеть заменить мерные переменные их главными компонентами до применения критерия; поворот в главные компоненты переведет конфигурации типа правой на рисунке, где главные оси диагональны, в такую как слева.

Если кластеры разбросаны в пространстве мерных переменных равномерно, что обычно связано с некоррелированностью переменных, тогда совокупный Ratkowsky-Lance мало меняется при поворотах, как например при таких данных:



Критерий Ratkowsky-Lance подвержен также влиянию формы кластеров и их повернутости относительно своих центроидов. На рисунке ниже два круглых кластера в пространстве мерных переменных V1 и V2 расходятся сильнее по переменной V1 (критерий = .64), чем по V2 (критерий = .44); общий критерий = .54.



Слева эти кластеры вытянуты в эллипсы (с сохранением прежней двумерной внутрикластерной дисперсии) вдоль V1, представляя собою «цепь», а справа – вытянуты вдоль V2, представляя собой «стопку». Справа величина критерия не изменилась (.54), но слева – возросла до .61. Это несколько неожиданно для глаза, который счел бы самым лучшим кластерное решение справа, где кластеры совсем не соприкасаются. Но для критерия Ratkowsky-Lance такое в порядке вещей и имеет свой смысл. Ratkowsky-Lance акцентирует (поощряет) съеживание (уплотнение) кластера вдоль той переменной, по которой кластеры близки (своими центрами); если же кластеры далеки, их внутреннее съеживание вдоль переменной, их разделяющей, почти не поощряется (т.к. оно в определенном смысле уже «излишне»).

С увеличением размерности пространства критерий начинает регулярно предпочитать эллипсоидные кластеры сферическим. Он нечувствителен к форме распределения в кластерах (колокол vs равномерное). Критерий нечувствителен к числу объектов в данных, поэтому сравниваемые кластерные разбиения могут состоять из неравного общего числа объектов *n*.

Если переменная дихотомическая, то нет разницы, считать ее мерной или категориальной (поскольку в этом случае Eta и Cramer’s V одинаковы). Было показано, что Ratkowsky-Lance очень хорошо определяет число кластеров в дихотомических данных (Dimitriadou, E., Dolnicar, S., Weingassel, A. An examination of indexes for determining the number of clusters in binary data sets // Psychometrika, 2002, 67(1), 137-160.).

***Подкоманды***

**SCAVARS, CATVARS**

В SCAVARS укажите мерные переменные. В CATVARS укажите категориальные переменные. Оба списка можно писать поименно и/или ч-з to. Вы должны задать хотя бы один список из двух. Макрос исключает пропущенные значения «списочно»: если хотя бы в одной переменной наблюдение есть пропуск, оно исключается из калькуляций по всем переменным.

**CLUSOL**

Поименный список группирующих наблюдения переменных – см. описание подкоманды выше, в макросе [!KO\_CALHARV](#_МАКРОС_!CALHARV:_КРИТЕРИИ).

**UNIVAR**

Опциональная подкоманда, которой можно сохранить (UNIVAR=YES) в выходящий массив значения критерия для каждой переменной, а не только усредненное значение. Учтите, что критерий вычисляется, только если он вычислим для всех переменных, иначе выводится пропуск.

Благодаря этой опции вы можете видеть и сравнивать вклад различных переменных в расхождения между кластерами.

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива (SPLIT FILE). Он не слушается временных операций (стоящих под TEMPORARY). Для отбора наблюдений в анализ вы можете использовать фильтрацию (FILTER или USE) и SELECT IF. Кроме того, вы можете исключать из анализа разные наблюдения в разных кластерных решениях, приписав им missing или нулевые коды (см. об этом в п/к CLUSOL).

# МАКРОС !KO\_RPBCLU: ТОЧЕЧНО-БИСЕРИАЛЬНАЯ КОРРЕЛЯЦИЯ и McCLAIN-RAO

Version 2, Jul 2018 (Version 1, Sep 2001). Tested on SPSS 17, 20, 22.

!KO\_rpbclu matrix= *var1 to var200* /\*Переменные, составляющие матрицу сходств или расстояний

/\*(поименно и/или ч-з to)

/clusol= *clu6 clu5 clu4 clu3 clu2*

/\*Переменные, являющие собой кластерные решения (поименный список).

Минимум надо задать MATRIX, CLUSOL.

**Критерий точечно-бисериальная корреляция**

Идея коррелирования непохожести объектов с их классификацией очевидна и проста, и называется кофенетической корреляцией. Когда классификация это разбитие на кластеры, то речь идет о замере согласованности между непохожестью объектов друг на друга, с одной стороны, и фактом, принадлежат эти объекты разным кластерам или же одному и тому же кластеру, с другой стороны. Если более похожие объекты попадают в один кластер, а более непохожие попадают в разные кластеры, корреляция будет высока, значит и данное кластерное решение – хорошее. Привлекательность корреляционных индексов перед такими мерами как дисперсионный индекс Calinski–Harabasz заключается в их универсальности: они могут применяться как к расстояниям, так и к сходствам как мерам близости между объектами, – причем к любым мерам, независимо от того, согласуются ли последние с геометрией евклидового (и вообще метрического) пространства или нет.

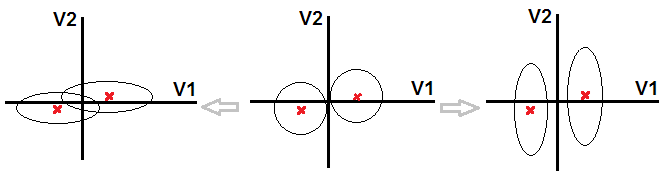
Точечно-бисериальный коэффициент *rpb* корреляции это обыкновенный *r* Пирсона между величиной близости (похожести) и фактом «в одном кластере / в разных кластерах». Он показывает относительную разницу между средней величиной внутрикластерных парных близостей и средней величиной межкластерных парных близостей:

где *Mb* это среднее расстояние между разнокластерными объектами (таких расстояний *fb*), а *Mw* это среднее расстояние между однокластерными объектами (таких расстояний *fw*); всего *f* расстояний. *s* это стандартное отклонение в распределении всех расстояний. Точечно-бисериальный r варьирует между -1 до 1. Чем выше значение, тем лучше кластерное разбиение. *Rpb* имеет умеренно выраженное фоновое предпочтение к среднему (в районе 4-10) числу кластеров.

Этот индекс нечувствителен к прибавлению к близостям константы (вообще, к линейному преобразованию близостей). Например, на следующем рисунке показаны два кластера, а правее – они же после прибавки 10 мм ко всем попарным расстояниям. Многие другие кластерные критерии сочтут, что кластерность справа менее убедительна, чем та, что слева, что согласуется и со зрительным впечатлением. Но точечно-бисериальная корреляция сочтет оба решения одинаково качественными. Подобная «геометрическая» контр-интуитивность критерия может иногда считаться его недостатком.



В евклидовом пространстве с использованием евклидовых расстояний значение *rpb* будет тем выше, чем правее конфигурация на следующем рисунке (расстояние между центроидами и внутрикластерные дисперсии тут удержаны постоянными).



Т.е. критерий поощряет эллипсоиды, если они образуют стопку, не цепь. Однако с увеличением размерности пространства кластерные разбиения, состоящие из эллипсоидных кластеров, начинают стабильно проигрывать разбиениям, состоящим из круглых кластеров. Таким образом, rpb в общем симпатизирует больше сферичным кластерам.

Критерий точечно-бисериальная корреляция годится для расстояний и для сходств (сходства макрос внутренне переводит линейно в расстояния).

Критерий нечувствителен к числу объектов в данных, поэтому сравниваемые кластерные разбиения могут состоять из неравного общего числа объектов *n*.

**Критерий McClain-Rao**

Определяется как отношение среднего расстояния между разнокластерными объектами к среднему расстоянию между однокластерными объектами (McClain, J.O., Rao,V.R. Clustisz: A program to test for the quality of clustering of a set of objects // Journal of Marketing Research, 1975, 12(4), 456-460):

где *Sb* и *Sw* это суммы расстояний между разнокластерными объектами и между однокластерными объектами, соответственно; остальные обозначения см. выше. Если *Sb*=0, индекс не вычисляется. Формально чем ниже значение, тем лучше кластерное разбиение, но критерий присуще сильно благоволит бо́льшему k, поэтому основной смысл есть смотреть на перегиб вверх (нижний «локоть»).

Критерий McClain–Rao годится для расстояний и для сходств (сходства макрос внутренне переводит линейно в расстояния).

Критерий имеет незничительную тенденцию поощрять эллипсоидные кластеры, расположенные стопкой или кольцом. Прибавление к близостям константы изменяет критерий линейно. Критерий нечувствителен к числу объектов в данных, поэтому сравниваемые кластерные разбиения могут состоять из неравного общего числа объектов *n*.

***Подкоманды***

**MATRIX**

Укажите переменные, являющиеся столбцами собственно матрицы расстояний (на диагонали должны быть нули, внедиагональные элементы должны быть неотрицательны, большее число отвечает большему различию) или матрицы сходств (на диагонали должны быть положительные числа, внедиагональные элементы могут иметь любой знак, большее число отвечает большей схожести). Матрица должна быть квадратной симметрической и не может содержать пропуски. Учтите, что макрос не проверяет корректность матрицы: позаботьтесь, чтобы она отвечала вышеописанным требованиям. Строение матрицы показано выше в макросе [!KO\_CALHARM](#_МАКРОС_!CALHARM:_КРИТЕРИЙ). Тело матрицы, находящееся в рабочем массиве, должно быть квадратным и симметричным, и вы должны ввести подкомандой MATRIX все столбцы тела.

*Задание открытым диапазоном с помощью “?”*. В некоторых случаях возникает нужда указать диапазон, заключенный между парой переменных, которые сами не входят в диапазон. Используйте для этого “?” по краям. Например, *?VARNAME\_ to ENDVAR?* означает все переменные, находящиеся в массиве между переменными *VARNAME\_* и *ENDVAR*, не включая их самих. Для обозначения диапазона, открытого с одного конца, используйте “?” только у этого конца. Например: *?VARNAME\_ to VAR100* или *VAR1 to ENDVAR?*.

**CLUSOL**

Поименный список группирующих наблюдения переменных – см. описание подкоманды выше, в макросе [!KO\_CALHARV](#_МАКРОС_!CALHARV:_КРИТЕРИИ).

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива (SPLIT FILE). Он не слушается временных операций (стоящих под TEMPORARY). Не применяйте перед макросом SELECT IF, т.к. порушите квадратность матрицы. Вы можете, однако, использовать для отбора наблюдений фильтрацию (FILTER или USE). Кроме того, вы можете исключать из анализа разные наблюдения в разных кластерных решениях, приписав им missing или нулевые коды (см. об этом в п/к CLUSOL).

# МАКРОС !KO\_GAMMACLU: ГАММА-СТАТИСТИКА

Version 2, Jan 2022 (Version 1, Jul 2012). Tested on SPSS Statistics 20, 22, 26.

!KO\_gammaclu matrix= *var1 to var200* /\*Переменные, составляющие матрицу сходств или расстояний

/\*(поименно и/или ч-з to)

/rescrnd= *3* /\*Опционально: перешкалировать/округлить континуальные близости, ради скорости

/clusol= *clu6 clu5 clu4 clu3 clu2*

/\*Переменные, являющие собой кластерные решения (поименный список).

Минимум надо задать MATRIX, CLUSOL.

Данный индекс есть гамма Гудмана–Краскела – статистика, выражающая собой непараметричекую ранговую корреляцию. В отличие от точечно-бисериальной корреляции гамма не оперирует самой величиной попарных близостей, а оценивает вероятность того, что два случайно взятых однокластерных объекта будут иметь большее сходство, чем два случайно взятых разнокластерных объекта:



где *w+* это число случаев (комбинаций) в данных, когда два однокластерных объекта ближе друг к другу, чем два разнокластерных объекта; *w-* это число случаев в данных, когда два однокластерных объекта дальше друг от друга, чем два разнокластерных объекта. Гамма варьирует между -1 до 1. Чем выше значение, тем лучше кластерное разбиение.

Поскольку гамма является мерой монотонной корреляции, она нечувствительна ко всякому монотонному изменению близостей. Поэтому, в частности, неважно, квадратные или неквадратные расстояния вы вводите. Подобная нечувствительность к величине близостей в определенном геометрическом смысле контринтуитивна и может иногда считаться недостатком.

В евклидовом пространстве с использованием евклидовых расстояний отношение гаммы к эллипсоидным кластерам по сравнению со сферичными сходно с таковым у точечно-бисериальной корреляции (см. описание !KO\_RPBCLU): с увеличением размерности пространства усиливается предпочтение к сферичным кластерам.

Критерий гамма годится для расстояний и для сходств.

Критерий нечувствителен к числу объектов в данных, поэтому сравниваемые кластерные разбиения могут состоять из неравного общего числа объектов *n*.

Макрос !KO\_GAMMACLU довольно медленный и поэтому не рекомендуем для данных с тысячами объектов. Но подкоманда RESCRND может ускорить его работу.

***Подкоманды***

**MATRIX**

Укажите переменные, являющиеся столбцами собственно матрицы расстояний (на диагонали должны быть нули, внедиагональные элементы должны быть неотрицательны, большее число отвечает большему различию) или матрицы сходств (на диагонали должны быть положительные числа, внедиагональные элементы могут иметь любой знак, большее число отвечает большей схожести). Матрица должна быть квадратной симметрической и не может содержать пропуски. Учтите, что макрос не проверяет корректность матрицы: позаботьтесь, чтобы она отвечала вышеописанным требованиям. Строение матрицы показано выше в макросе [!KO\_CALHARM](#_МАКРОС_!CALHARM:_КРИТЕРИЙ). Тело матрицы, находящееся в рабочем массиве, должно быть квадратным и симметричным, и вы должны ввести подкомандой MATRIX все столбцы тела.

*Задание открытым диапазоном с помощью “?”*. В некоторых случаях возникает нужда указать диапазон, заключенный между парой переменных, которые сами не входят в диапазон. Используйте для этого “?” по краям. Например, *?VARNAME\_ to ENDVAR?* означает все переменные, находящиеся в массиве между переменными *VARNAME\_* и *ENDVAR*, не включая их самих. Для обозначения диапазона, открытого с одного конца, используйте “?” только у этого конца. Например: *?VARNAME\_ to VAR100* или *VAR1 to ENDVAR?*.

**CLUSOL**

Поименный список группирующих наблюдения переменных – см. описание подкоманды выше, в макросе [!KO\_CALHARV](#_МАКРОС_!CALHARV:_КРИТЕРИИ).

**RESCRND**

Вычисление гаммы, если данные (в нашем случае это значения матрицы близостей) континуальны, может занять много времени даже при матрице умеренного размера. Уменьшение числа различных значений в матрице существенно ускорит процесс. П/к RESCRND уменьшает число различных значений во входящей матрице, дискретизируя их методом перешкалирования-округления. Укажите цифру: 1, 2, 3, 4 или 5. Это количество десятичных цифр, которое макрос оставит значениям в матрице после линейного перешкалирования их в диапазон 0–1. Самая быстрая работа макроса будет при RESCRND=1, поскольку различных значений при такой дискретизации останется не больше 11. При RESCRND=2 останется не больше 101 таковых. При RESCRND=5 число различных значений останется не больше 100001, так что время работы макроса может быть большим. Максимальное время работы будет (при континуальных входящих) в случае незадания/умолчания подкоманды, т.к. дискретизация тогда не производится вообще.

Между скоростью и разборчивостью существует очевидный компромисс (trade-off). Когда число различных значений среди близостей сделано слишком малым, гамма теряет чувствительность к тому, как маленькие изменения в расстояниях/сходствах между объектами коррелируют с кластеризованностью. Вообразите предельный случай, когда разных значений в матрице близостей – всего два. Тогда индекс будет выражать ассоциацию максимально «грубую»: между переменной «маленькое расстояние vs большое расстояние» и переменной «в одном кластере vs в разных кластерах». Но, с другой стороны, очень малые перепады расстояний/сходств не могут иметь никакого практического значения для кластеризации. Вообразите, что исходные расстояния в матрице имеют разброс между 0.2 и 9.7. На этом фоне пара расстояний, например, 3.56146 и 3.56148, из матрицы – это практически неразличимые расстояния; едва ли есть смысл полагать их разными, способными по-разному определять кластерную судьбу объектов. Следовательно, оправдана будет умеренная дискретизация: она, скорее всего, почти не повляет на величину гаммы, но заметно убыстрит работу. RESCRND=3 в большинстве случаев – приемлемый компромиссный вариант.

Макрос показывает число различных значений, оставшихся в матрице близостей после дискретизации, в переменной *N#$.*

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива (SPLIT FILE). Он не слушается временных операций (стоящих под TEMPORARY). Не применяйте перед макросом SELECT IF, т.к. порушите квадратность матрицы. Вы можете, однако, использовать для отбора наблюдений фильтрацию (FILTER или USE). Кроме того, вы можете исключать из анализа разные наблюдения в разных кластерных решениях, приписав им missing или нулевые коды (см. об этом в п/к CLUSOL).

# МАКРОС !KO\_CINDEX: C-INDEX

Version 1, Sep 2001. Tested on SPSS 11.5, 13, 15.

!KO\_cindex matrix= *var1 to var200* /\*Переменные, составляющие матрицу сходств или расстояний

/\*(поименно и/или ч-з to)

/clusol= *clu6 clu5 clu4 clu3 clu2*

/\*Переменные, являющие собой кластерные решения (поименный список)

/fast= YES /\*Быстрый режим: YES или NO (тж п/у).

Минимум надо задать MATRIX, CLUSOL.

C-Index (Hubert, L.J., Levin, J.R. A general statistical framework for assessning categorical clustering in free recall // Psychol. Bull., 1976, 83, 1072-1080) является универсальным критерием, вроде точечно-бисериальной корреляции. Он показывает, насколько наблюдаемая внутрикластерная плотность – сумма внутрикластерных парных близостей – относительно близка к максимальной предельной таковой (как правило фактически недостижимой) при данном их числе.

где *Sw* это суммарное расстояние между однокластерными объектами (таких расстояний *fw*); *Smin* это сумма *fw*наименьших расстояний между объектами (неважно, межкластерное расстояние или внутрикластерное), *Smax* это, аналогично, сумма *fw*наибольших расстояний между объектами. C-Index варьирует между 0 и 1. Чем ниже значение, тем лучше кластерное разбиение.

C-Index линейно тождествен (только с отрицательным коэффициентом наклона) точечно-бисериальной корреляции (см. !KO\_RPBCLU), перешкалированной относительно своего эмпирического максимума. Т.е., если *rpb*=.57, допустим, но в данных и при данном рассматриваемом разбиении *rpb* не может превысить, скажем, .95, тогда перешкалированный *rpb*=.57/.95. Критерий C-Index есть величина, эквивалентная этой; таким образом, перешкалированная корреляция и C-Index – содержательно, логически одно и то же. C-Index имеет тенденцию предпочитать больше кластеров, чем *rpb*. Ибо – в ходе кластеризации – пока кластеры маленькие (и значит, их еще много), они имеют больше шансов близко приближаться к тому «идеалу», который имеет в виду данный критерий.

Как и точечно-бисериальная корреляция, C-Index нечувствителен к линейному преобразованию близостей. Подобная нечувствительность к величине близостей в определенном геометрическом смысле контринтуитивна и может иногда считаться недостатком.

В евклидовом пространстве с использованием евклидовых расстояний отношение индекса к эллипсоидным кластерам по сравнению со сферичными сходно с таковым у точечно-бисериальной корреляции (см. описание !KO\_RPBCLU): с увеличением размерности пространства усиливается предпочтение к сферичным кластерам. По моим осторожным впечатлениям в симуляционных пробах, C-Index сильнее наказывает за цепочечность (см. рисунок в описании !KO\_RPBCLU), чем поощряет за стопочность, а точечно-бисериальный r, наоборот, сильнее поощряет второе, чем наказывает за первое. В отношении же круглых кластеров оба критерия часто дают очень похожие результаты.

Критерий C-index годится для расстояний и для сходств (сходства макрос внутренне переводит линейно в расстояния).

Критерий нечувствителен к числу объектов в данных, поэтому сравниваемые кластерные разбиения могут состоять из неравного общего числа объектов *n*.

Макрос !KO\_CINDEX сравнительно медленный и не рекомендуем для большого (тысячи) количества объектов. П/к FAST существенно ускоряет работу.

***Подкоманды***

**MATRIX**

Укажите переменные, являющиеся столбцами собственно матрицы расстояний (на диагонали должны быть нули, внедиагональные элементы должны быть неотрицательны, большее число отвечает большему различию) или матрицы сходств (на диагонали должны быть положительные числа, внедиагональные элементы могут иметь любой знак, большее число отвечает большей схожести). Матрица должна быть квадратной симметрической и не может содержать пропуски. Учтите, что макрос не проверяет корректность матрицы: позаботьтесь, чтобы она отвечала вышеописанным требованиям. Строение матрицы показано выше в макросе [!KO\_CALHARM](#_МАКРОС_!CALHARM:_КРИТЕРИЙ). Тело матрицы, находящееся в рабочем массиве, должно быть квадратным и симметричным, и вы должны ввести подкомандой MATRIX все столбцы тела.

*Задание открытым диапазоном с помощью “?”*. В некоторых случаях возникает нужда указать диапазон, заключенный между парой переменных, которые сами не входят в диапазон. Используйте для этого “?” по краям. Например, *?VARNAME\_ to ENDVAR?* означает все переменные, находящиеся в массиве между переменными *VARNAME\_* и *ENDVAR*, не включая их самих. Для обозначения диапазона, открытого с одного конца, используйте “?” только у этого конца. Например: *?VARNAME\_ to VAR100* или *VAR1 to ENDVAR?*.

**CLUSOL**

Поименный список группирующих наблюдения переменных – см. описание подкоманды выше, в макросе [!KO\_CALHARV](#_МАКРОС_!CALHARV:_КРИТЕРИИ).

**FAST**

Эта подкоманда действует, только если в CLUSOL указано более одной переменной. FAST=YES ускоряет работу макроса, что может быть заметно в случае больших данных (большая матрица MATRIX и много кластерных решений CLUSOL). Не применяйте FAST=YES, если в каких-л. переменных CLUSOL наличествуют пропущенные значения или нули. По умолчанию/незаданию FAST=NO.

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива (SPLIT FILE). Он не слушается временных операций (стоящих под TEMPORARY). Не применяйте перед макросом SELECT IF, т.к. порушите квадратность матрицы. Вы можете, однако, использовать для отбора наблюдений фильтрацию (FILTER или USE). Кроме того, вы можете исключать из анализа разные наблюдения в разных кластерных решениях, приписав им missing или нулевые коды (см. об этом в п/к CLUSOL).

# МАКРОС !KO\_DUNN: КРИТЕРИЙ DUNN (НЕСКОЛЬКО ТИПОВ)

Version 2, Jul 2016 (Version 1, Sep 2012). Tested on SPSS Statistics 17, 20, 22.

!KO\_dunn matrix= *VAR1 TO VAR200* /\*Переменные, составляющие матрицу сходств или расстояний, можно ч-з to

/clusol= *CLU5\_1 CLU4\_1 CLU3\_1 CLU2\_1*

/\*Переменные, являющие собой кластерные решения (поименный список)

/distb= /\*Формулировка межкластерного расстояния (см.): SINGLE (тж п/у), COMPLETE, BAVERAGE,

/\*CENTROID, CROSSPC, HAUSDORFF

/diamw= /\*Формулировка внутрикластерного "диаметра" (см.): MXDIST (тж п/у), AVDIST, AVDEV.

Минимум надо задать MATRIX, CLUSOL.

Этот критерий (Dunn, J.C. Well separated clusters and optimal fuzzy partitions // J Cybernetics, 1974, 95-104) является отношением, в данном кластерном решении из *k* кластеров, наименьшего расстояния между кластерами, рассмотренными попарно, к наибольшему внутрикластерному «диаметру» среди всех кластеров.

Чем больше эта пропорция (могущая быть от нуля и выше), тем из более разделенных, раздвинутых кластеров состоит кластерное разбиение (следовательно, тем оно лучше).

В оригинальной формулировке за межкластерное расстояние *dij* принято наименьшее расстояние между объектами кластера *i* и объектами кластера *j*, т.е. расстояние «ближайшего соседа» между *i* и *j*. За диаметр же кластера *diami* принято наибольшее расстояние внутри кластера *i*, т.е. расстояние между его крайними точками. Такой индекс предпочитает именно решения с пространственно более разделенными кластерами, в ущерб решениям с кластерами пусть и плотными внутри, но с межкластерным «шумом» (в частности, если кластеры соприкасаются или взаимонакладываются хотя бы одной точкой, это сразу портит кластерное решение).

Кроме оригинальной версии индекса были предложены позднее (Bezdek, C., Pal, N.R. Some new indexes of cluster validity // IEEE Trans. Systems, Man and Cybernetics − B, 28 (1998), 301–315. См. также: Vendramin, L., Campello, R.J.G.B., Hruschka, E.R. On the comparison of relative clustering validity criteria // Proceedings of the 2009 SIAM International Conference on Data Mining, 2009, 733-744.) и другие варианты, с иной дефиницией для *dij* и *diami*, но той же формулой. Макрос !KO\_DUNN реализует все такие варианты критерия, рассмотренные в Vendramin et al. – см. подкоманды DISTB и DIAMW. Эти альтернативные формулировки критерия (вместе называемые иногда *generalized Dunn index,* GDI) имеют свои склонности и терпимее относятся к межкластерному шуму. Важно заметить, что значения критерия Dunn разных типов нельзя сравнивать между собой. Сравнивать можно разные классификации внутри одного типа критерия.

Как бы ни формулировать *dij* и *diami*, базовая формула критерия, в которой из всех *dij* берется одно минимальное и из всех *diami* одно максимальное, логически предполагает влияние величины неодинаковости среди *dij* и среди *diami* на критерий. Присутствие в кластерном решении одного непропорционально большого по физическому размеру кластера либо тесная сдвинутость двух кластеров на фоне хорошо раздвинутых остальных – портит картину (понижает критерий). Итак, Dunn «предпочитает» разбиения с приблизительно равноразделенными приблизительно равновеликими кластерами.

Индекс нечувствителен к пропорциональному преобразованию близостей, но реагирует на прибавление константы.

Критерий Dunn годится для расстояний и для сходств (сходства макрос внутренне переводит линейно в расстояния).

Критерий (по крайней мере в оригинальной версии) чувствителен к числу объектов в данных, поэтому сравниваемые кластерные разбиения должны состоять из равного общего числа объектов *n*.

***Подкоманды***

**MATRIX**

Укажите переменные, являющиеся столбцами собственно матрицы расстояний (на диагонали должны быть нули, внедиагональные элементы должны быть неотрицательны, большее число отвечает большему различию) или матрицы сходств (на диагонали должны быть положительные числа, внедиагональные элементы могут иметь любой знак, большее число отвечает большей схожести). Матрица должна быть квадратной симметрической и не может содержать пропуски. Учтите, что макрос не проверяет корректность матрицы: позаботьтесь, чтобы она отвечала вышеописанным требованиям. Строение матрицы показано выше в макросе [!KO\_CALHARM](#_МАКРОС_!CALHARM:_КРИТЕРИЙ). Тело матрицы, находящееся в рабочем массиве, должно быть квадратным и симметричным, и вы должны ввести подкомандой MATRIX все столбцы тела.

*Задание открытым диапазоном с помощью “?”*. В некоторых случаях возникает нужда указать диапазон, заключенный между парой переменных, которые сами не входят в диапазон. Используйте для этого “?” по краям. Например, *?VARNAME\_ to ENDVAR?* означает все переменные, находящиеся в массиве между переменными *VARNAME\_* и *ENDVAR*, не включая их самих. Для обозначения диапазона, открытого с одного конца, используйте “?” только у этого конца. Например: *?VARNAME\_ to VAR100* или *VAR1 to ENDVAR?*.

**CLUSOL**

Поименный список группирующих наблюдения переменных – см. описание подкоманды выше, в макросе [!KO\_CALHARV](#_МАКРОС_!CALHARV:_КРИТЕРИИ).

**DISTB**

Выберите определение междукластерного расстояния *dij*:

SINGLE - (тж по умолчанию/незаданию) расстояние единичной связи, или ближнего соседа. Это наименьшее расстояние между точками (объектами) одного кластера и точками (объектами) второго кластера. Соответствует оригинальной версии критерия.

COMPLETE - расстояние полной связи, или дальнего соседа. Это наибольшее расстояние между точками одного кластера и точками второго кластера.

BAVERAGE - расстояние средней связи. Это усредненное расстояние между точками одного кластера, с одной стороны, и точками второго кластера, с другой стороны.

CENTROID - расстояние межцентроидное. Это прямое расстояние между центроидами двух кластеров в евклидовом пространстве.

CROSSPC - кроссрасстояние точечно-центроидное. Это сумма расстояний от точек одного кластера до центроида второго, плюс сумма расстояний от точек второго кластера до центроида первого, и разделить на общее число точек в двух кластерах.

HAUSDORFF - расстояние Хаусдорфа, которое определяется так. Каждая точка первого кластера находит себе ближнего соседа во втором кластере, и из этих (минимальных) расстояний берется наибольшее. Так же делается наоборот: каждая точка второго кластера находит себе ближнего соседа в первом кластере, и берется наибольшее из этих расстояний. Из двух взятых таким образом расстояний берется большее.

Варианты CENTROID и CROSSPC ожидают, что входящие близости – евклидовые расстояния (неквадратные). Если ваши расстояния неевклидовы или если близости есть сходства, посчитанный результат может быть содержательно невалиден; а может случиться и ошибка взятия корня из отрицательного числа.

**DIAMW**

Выберите определение внутрикластерного диаметра *diami*:

MXDIST - (тж по умолчанию/незаданию) наибольшее расстояние в кластере, т.е. дистанция между самыми разудаленными точками (объектами) в нем. Соответствует оригинальной версии критерия.

AVDIST - среднее расстояние в кластере, т.е. усредненное расстояние между всеми точками попарно в нем.

AVDEV - удвоенное среднее отклонение от центроида кластера, т.е. усредненное расстояние между точками кластера и его центроидом, умноженное на 2. Данная опция ожидает, что входящие близости – евклидовые расстояния (неквадратные). Если ваши расстояния неевклидовы или если близости есть сходства, посчитанный результат может быть содержательно невалиден; а может случиться и ошибка взятия корня из отрицательного числа.

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива (SPLIT FILE). Он не слушается временных операций (стоящих под TEMPORARY). Не применяйте перед макросом SELECT IF, т.к. порушите квадратность матрицы. Вы можете, однако, использовать для отбора наблюдений фильтрацию (FILTER или USE). Кроме того, вы можете исключать из анализа разные наблюдения в разных кластерных решениях, приписав им missing или нулевые коды (см. об этом в п/к CLUSOL).

# МАКРОС !KO\_SILHOU: СИЛУЭТ-СТАТИСТИКА (НЕСКОЛЬКО ТИПОВ)

Version 3, Jul 2017 (Version 1, Oct 2010). Tested on SPSS Statistics 17, 20, 22.

*Этот макрос требует SPSS Statistics 17 или выше*

!KO\_silhou matrix= *VAR1 to VAR250* /\*Переменные, составляющие матрицу сходств или расстояний, можно ч-з to

/clusol= *CLU6\_2 CLU5\_2 CLU4\_2 CLU3\_2 CLU2\_2*

/\*Переменные, являющие собой кластерные решения (поименный список)

/type= /\*Тип силуэт-статитстики: AVER (тж п/у), NEAR, FARTH, FARNEAR, DEVIAT (см.)

/caps= *'nei#' 'sil#'* /\*Две приставки в имена переменных: 1-я в переменую ближнего кластера, 2-я в силуэт-переменную

/single= *0* /\*Силуэт-значение для одинокого объекта: значение или SYSMIS (пропуск) или

/\*IFMERGE или ASSEEDS (см.)

/widorig= *-1* /\*Опционально для графика silhouette width: уровень отсчета (п/у 0)

/widcoll= /\*Опционально для графика silhouette width: объединять кластеры с процентом ниже (число, п/у 5).

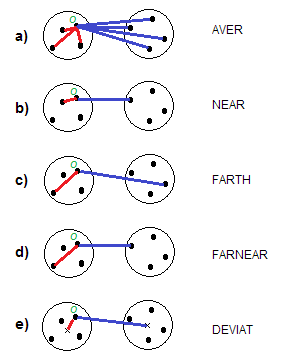
Минимум надо задать MATRIX, CLUSOL, CAPS, SINGLE.

Силуэт-статистика – универсальный кластерный критерий, измеряющий степень оправданности того, что объекты приписаны данным кластерам, а не ближайшим соседним кластерам. Особая ценность силуэт-статистики состоит в том, что она вычисляется уже на уровне каждого объекта и, следовательно, является показателем, позволяющим судить, насколько хорошо кластеризован данный конкретный объект. Что касается качества кластерного решения в целом, то обычно мерой этого является просто средняя арифметическая силуэт-значений всех объектов данного кластерного решения. Для объекта *o* силуэт-статистика есть (Kaufman, L., Rousseeuw, P. Finding groups in data: an introduction to cluster analysis. New York, 1990):

где *a(o)* это удаленность объекта *o* от прочих объектов того же кластера; *b(o)* это удаленность объекта *o* от объектов другого кластера, ближнего к объекту *o*. Имеется в виду, что удаленность *b(o)* вычисляется между нашим объектом и объектами каждого чужого кластера (всего кластеров *k*). Наименьшая из этих посчитанных *k-*1 удаленностей и будет *b(o)*, взятое в формулу. Силуэт-статистика может колебаться от -1 (теоретически наихудшая кластеризция объекта) до 1 (теоретически наилучшая кластеризация объекта) и ее смысл таков: в какой степени мы проигрываем (или выигрываем) в хорошести классификации объекта, перемести мы его из его настоящего кластера в другой, ближний к нему.

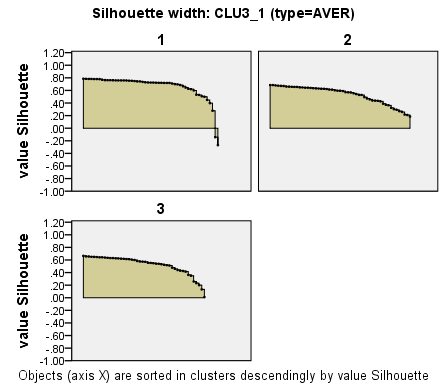
В зависимости от того, что понимать под удаленностью, можно получить несколько типов силуэт-статистики:

1. Удаленность как *среднее* расстояние: *a(o)* это среднее расстояние от объекта *o* до прочих объектов своего кластера; *b(o)* это среднее расстояние от объекта *o* до объектов ближнего к *o* другого кластера. Это силуэт-статистика средней связи; именно она была предложена Kaufman и Rousseeuw, и обычно ее имеют в виду, когда говорят о силуэт-статистике. Макрос вычисляет по умолчанию именно этот тип.
2. Удаленность как *наименьшее* расстояние: *a(o)* это расстояние от объекта *o* до ближнего объекта в своем кластере; *b(o)* это расстояние от объекта *o* до ближнего объекта в ближнем к *o* другом кластере. Это силуэт-статистика ближайшего соседа.
3. Удаленность как *наибольшее* расстояние: *a(o)* это расстояние от объекта *o* до дальнего объекта в своем кластере; *b(o)* это расстояние от объекта *o* до дальнего объекта в ближнем к *o* другом кластере. Это силуэт-статистика дальнего соседа.
4. Удаленность *a(o)* это расстояние от объекта *o* до дальнего объекта в своем кластере; *b(o)* это расстояние от объекта *o* до ближнего объекта в ближнем к *o* другом кластере. Это смешанный вариант.
5. Удаленность как расстояние до геометрического *центра*: *a(o)* это расстояние от объекта *o* до центроида своего кластера; *b(o)* это расстояние от объекта *o* до центроида ближнего к *o* другого кластера. Это силуэт-статистика отклонения.



Надо подчеркнуть, что значения силуэт-статистик *разных типов* *нельзя* сравнивать между собой: у них разный характерный уровень значений. Например, у типа ближнего соседа значения обычно выше, чем у других типов, а у смешанного типа обычно ниже всех, и часто вовсе отрицательны. Сравнивать значения можно у разных разбиений только внутри одного и того же типа силуэт-статистики. Внутри типа принцип сохраняется: чем выше значение, тем лучше кластеризованность.

Макрос вычисляет значения силуэт-статистики для объектов предложенных кластерных разбиений и выводит эти переменные в рабочий массив данных. Он также выводит переменные, показывающие код ближнего кластера. График усредненной по всему кластерному разбиению силуэт-статистики строится, если кластерных разбиений было предложено в CLUSOL более одного. Если кластерное разбиение в CLUSOL единственное, макрос построит для него *график ширины силуэта* (silhouette width plot): значения силуэт-статистики в каждом кластере отдельно выстроены по убывающей, образуя «силуэт» кластера. Чем выше и толще силуэт, тем лучше кластеризованы объекты, составляющие кластер. Вы можете, открыв график для редакции, добавить в него горизонтальную привязочную линию, показывающую среднее значение силуэт-статистики, характеризующее данное кластерное решение в целом (установите Reference line на Mean).



Индекс нечувствителен к пропорциональному преобразованию близостей, но реагирует на прибавление константы. Критерий (в оригинальной версии, TYPE=AVER) нечувствителен к числу объектов в данных, поэтому сравниваемые кластерные разбиения могут состоять из неравного общего числа объектов *n*. Предпочтения критерия к форме распределения и к пространственной конфигурации кластеров – см. **Таблицу 1**.

Критерий силуэт-статистика годится для расстояний и для сходств (сходства макрос внутренне переводит линейно в расстояния).

*Использование силуэт-значения для улучшения кластеризации*. То, что для каждого объекта известен ближний ему кластер, открывает возможность к улучшению кластеризации. Объект(ы), у которого силуэт-статистика оказалась слишком низка по сравнению с другими объектами его кластера, пользователь может перезаписать в ближайший этому объекту кластер (код его макрос выдает), затем перезапустить макрос и посмотреть, увеличилась ли от этого перемещения силуэт-статистика у этого объекта и средняя силуэт-статистика всего кластерного решения. Так можно проделывать несколько раз с одними или разными объектами (получается своего рода переместительный кластерный алгоритм, нацеленный улучшить классификацию без изменения числа кластеров). Средняя силуэт-статистика перестает скоро улучшаться, но некоторое время может еще уменьшаться ее дисперсия, т.е. увеличиваться толщина силуэтов некоторых кластеров.

Макрос создает временную переменную CASENUM# во входящем массиве, поэтому убедитесь, что такой переменной у вас нет, во избежание перезаписи.

***Подкоманды***

? \_

**MATRIX**

Укажите переменные, являющиеся столбцами собственно матрицы расстояний (на диагонали должны быть нули, внедиагональные элементы должны быть неотрицательны, большее число отвечает большему различию) или матрицы сходств (на диагонали должны быть положительные числа, внедиагональные элементы могут иметь любой знак, большее число отвечает большей схожести). Матрица должна быть квадратной симметрической и не может содержать пропуски. Учтите, что макрос не проверяет корректность матрицы: позаботьтесь, чтобы она отвечала вышеописанным требованиям. Строение матрицы показано выше в макросе [!KO\_CALHARM](#_МАКРОС_!CALHARM:_КРИТЕРИЙ). Тело матрицы, находящееся в рабочем массиве, должно быть квадратным и симметричным, и вы должны ввести подкомандой MATRIX все столбцы тела.

*Задание открытым диапазоном с помощью “?”*. В некоторых случаях возникает нужда указать диапазон, заключенный между парой переменных, которые сами не входят в диапазон. Используйте для этого “?” по краям. Например, *?VARNAME\_ to ENDVAR?* означает все переменные, находящиеся в массиве между переменными *VARNAME\_* и *ENDVAR*, не включая их самих. Для обозначения диапазона, открытого с одного конца, используйте “?” только у этого конца. Например: *?VARNAME\_ to VAR100* или *VAR1 to ENDVAR?*.

**CLUSOL**

Поименный список группирующих наблюдения переменных – см. описание подкоманды выше, в макросе [!KO\_CALHARV](#_МАКРОС_!CALHARV:_КРИТЕРИИ).

**TYPE**

Тип силуэт статистики:

AVER - (тж по умолчанию/незаданию) средней связи. Это классический вариант силуэт-статистики, предложенный ее изобретателями; его можно считать самым универсальным.

NEAR - ближнего соседа. Поскольку этот вариант позитивно реагирует на увеличение *n* в кластерах, он может поощрять решения с меньшим числом кластеров, приблизительно равнонаполненных объектами.

FARTH - дальнего соседа. Поскольку этот вариант негативно реагирует на увеличение *n* в кластерах, он может поощрять решения с бо́льшим числом кластеров.

FARNEAR - смешанного типа: дальний сосед в своем кластере, ближний – в соседнем кластере. Этот вариант может поощрять решения с бо́льшим числом кластеров.

DEVIAT - отклонения от центроида. Этот вариант (известный также как “Simplified Silhouette index”) ожидает, что входящие близости – евклидовые расстояния (неквадратные). Если ваши расстояния неевклидовы, посчитанный результат может быть содержательно невалиден; а может случиться и ошибка взятия корня из отрицательного числа, в последнем случае результаты не будут вычислены вовсе. Если близости – сходства, этот вариант неупотребим.

**CAPS**

Имена выводимых макросом в рабочий массив переменных составляются из имен списка CLUSOL и приставки. Укажите две приставки – первая для переменных, показывающих код ближнего кластера, а вторая для переменных со значением силуэт-статистики. Приставки можно взять в кавычки или апострофы (это желательно, если в конце приставки вы указываете разделитель-точку). Имейте в виду, что при повторном пуске макроса прежде выведенные в массив переменные с этими же именами заменятся, если вы не измените приставки на новые.

**SINGLE**

Это обязательная подкоманда, которая влияет, если в кластерном решении есть объекты-одиночки (т.е. кластеры, состоящие из одного объекта). Она задает, какое значение силуэт-статистики придать таким объектам. У них значение *a(o)* неопределено, т.к. других объектов в кластере нет. Возникает вопрос, какое значение статистики *sil* приписать одиночкам. Выберите:

*значение* - укажите произвольное значение *sil* между -1 и 1, оно будет придано всем одиночным объектам *s*. Авторы оригинального критерия рекомендовали значение 0, оно и используется во многих пакетах по умолчанию. Но так именно для оригинальной версии TYPE=AVER. Для других версий 0 не обязательно является оптимальным советом[[2]](#footnote-2).

SYSMIS - не назначать значения, значение *sil(s)* будет пропуском (system missing). Следовательно, объекты-одиночки не принимутся в расчет при вычислении среднего значения *sil* всего кластерного решения, которое будет в этом случае равно среднему *sil* только объектов-неодиночек. Поэтому используйте эту опцию, если вас не заботит величина доли объектов, объединенных в кластеры в кластерном решении.

IFMERGE - каждому одиночному объекту *s* вычисляется свое значение критерия по следующей логистической функции:

где *b(s)* это это *b* объекта *s* (см. формулу критерия); и это средняя и ст. отклонение значений *a* среди объектов-неодиночек, а это среднее силуэт-значение среди них же. Коэффициент *w* принят за (вы можете изменить *w* на иное значение в первой команде тела макроса DEFINE… /COEF… !DEFAULT). Смысл этой формулы следующий. Одиночный объект получит *sil*, близкое к -1 (наихудшее возможное значение), если он находится близко к каким-л. кластерам или одиночным объектам: *b(s)* мало; и он получит *sil*, близкое к усредненному *sil* неодиночных объектов (т.е. индифферентное значение), если он находится далеко от всех кластеров/объектов: *b(s)* велико. Т.е. одиночка наказывается низким *sil*, когда он мог бы быть слит с кем-нибудь (но он не слит). Если же он отщепенец (не мог бы быть слит ни с кем), то его можно игнорировать и считать, что он не портит кластерное решение, и поэтому ему назначится значение *sil*, близкое к среднему закластеризованных объектов. При *b(s)=* одиночный объект получит *sil* посредине между -1 и .

ASSEEDS - (эта опция только для TYPE=AVER или DEVIAT) всем одиночным объектам *s* назначается единое значение *a(s)*, после чего *sil(s)* объекта вычисляется по формуле критерия. Значение *a(s)* определяется так: в каждом кластере-неодиночке вычисляется среднее попарное расстояние между его объектами; средневзвешенная этих средних, где вес = 1/(*n* в кластере), есть скаляр *a(s)* (при TYPE=DEVIAT полученное *a(s)* делится на 2, поскольку в этом случае это расстояние до центроида). Смысл подхода следующий. Одиночный объект *s* рассматривается в качестве зачинателя нового, своего кластера, и поэтому пребывает как бы в ожидании объекта-пары для себя, чтобы этот новый кластер возник. Потребное *a(s)* это ожидаемое расстояние между *s* и парным ему пока несуществующим объектом, и оно принимается за среднее расстояние в уже существующих кластерах. Вес, обратный численности в кластере, означает, что в вычислении *a(s)* нам важнее «мнение» кластеров с малым числом объектов, потому что допустимо решить, что такие кластеры – более «молодые по времени образования» и, следовательно, еще не образованные кластеры, зачинателями которых выступают объекты-одиночки, больше похожи на них, нежели на «старые» (многообъектные) кластеры.

**WIDORIG**

Эта подкоманда действует, только когда в CLUSOL одна переменная. Подкоманда задает отсчетный уровень (origin) силуэт-значения на графике ширины силуэта. Укажите число от -1 до ниже чем 1. По умолчанию принято 0. Точки (объекты) со значением ниже отсчетного, если такие есть, будут находиться ниже горизонтальной линии, соответствующей отсчетному уровню. Данная подкоманда не влияет на результат, но лишь на внешность упомянутого графика. Вы всегда можете изменить отсчетный уровень и в уже готовом графике, открыв его для редактирования и задав origin оси Y.

**WIDCOLL**

Эта подкоманда действует, только когда в CLUSOL одна переменная. Подкоманда задает порог численности в кластерах, ниже которого силуэты этих малых кластеров на графике ширины силуэта будут сведены в одну категорию “Other (Silhouette mix)”. Укажите число от 0 до 100. Это число – процент численности в кластере от общей численности объектов. По умолчанию принято за 5. Т.е. кластеры с числом объектов ниже 5% от всей выборки каждый будут на графике смешаны в одно изображение “Other (Silhouette mix)”. Если хотите видеть раздельно силуэты всех кластеров, даже одиночных объектов, задайте WIDCOLL=0.

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива (SPLIT FILE). Макрос не слушается временных операций (стоящих под TEMPORARY). Не применяйте перед макросом SELECT IF, т.к. порушите квадратность матрицы. Вы можете, однако, использовать для отбора наблюдений фильтрацию (FILTER или USE). Кроме того, вы можете исключать из анализа разные наблюдения в разных кластерных решениях, приписав им missing или нулевые коды (см. об этом в п/к CLUSOL).

***Некоторые вопросы***

*График ширины силуэта появляется пустой*. Формат переменной кластерного членства, указанной в CLUSOL, не отвечает у вас числу десятичных цифр кодов кластеров. Обычно коды – целые числа, но если у вас числа дробные, то целочисленный формат не позволит графику отобразиться. Задайте переменной дробный формат, отображающий цифры после запятой.

# МАКРОС !KO\_SILDEV: СИЛУЭТ-СТАТИСТИКА (ТИП DEVIAT, ВХОДЯЩИЕ – ПЕРЕМЕННЫЕ)

Version 1, Jul 2017. Tested on SPSS Statistics 17, 20, 22.

*Этот макрос требует SPSS Statistics 17 или выше*

!KO\_sildev vars= *V1 to V10* /\*Количественные переменные, по которым сравнить кластерные решения (можно ч-з to)

/clusol= *CLU6\_2 CLU5\_2 CLU4\_2 CLU3\_2 CLU2\_2*

/\*Переменные, являющие собой кластерные решения (поименный список)

/caps= *'nei#' 'sil#'* /\*Две приставки в имена переменных: 1-я в переменую ближнего кластера, 2-я в силуэт-переменную

/single= SYSMIS /\*Силуэт-значение для одинокого объекта: значение или SYSMIS (пропуск) или IFMERGE (см.)

/widorig= /\*Опционально для графика silhouette width: уровень отсчета (п/у 0)

/widcoll= /\*Опционально для графика silhouette width: объединять кластеры с процентом ниже (число, п/у 5).

Минимум надо задать VARS, CLUSOL, CAPS, SINGLE.

Данный макрос вычисляет силуэт-статистику типа отклонения от центроида (TYPE=DEVIAT, см. макрос !KO\_SILHOU), известный также как “Simplified Silhouette index”. В качестве входящих он берет не матрицу расстояний, а переменные. Результат тот же, что выдает !KO\_SILHOU.

Макрос создает временную переменную CASENUM# во входящем массиве, поэтому убедитесь, что такой переменной у вас нет, во избежание перезаписи.

***Подкоманды***

**VARS**

Переменные (количественные), по которым требуется оценивать кластерные разбиения наблюдений. Макрос исключает наблюдения, имеющие пропуски, списочно.

**CLUSOL**

Поименный список группирующих наблюдения переменных – см. описание подкоманды выше, в макросе [!KO\_CALHARV](#_МАКРОС_!CALHARV:_КРИТЕРИИ).

**CAPS**

Эта подкоманда такая же, как в [!KO\_SILHOU](#_МАКРОС_!SILHOU:_СИЛУЭТ-СТАТИСТИКА_2).

**SINGLE**

Это обязательная подкоманда, которая влияет, если в кластерном решении есть объекты-одиночки (т.е. кластеры, состоящие из одного объекта). Она задает, какое значение силуэт-статистики придать таким объектам. У них значение *a(o)* неопределено, т.к. других объектов в кластере нет. Возникает вопрос, какое значение статистики *sil* приписать одиночкам. Выберите:

*значение* - укажите произвольное значение *sil* между -1 и 1, оно будет придано всем одиночным объектам *s*. Авторы оригинального критерия рекомендовали для него значение 0, но это не обязательно оптимальный совет для типа DEVIAT. (Подробнее – см. п/к SINGLE в макросе !KO\_SILHOU.)

SYSMIS - не назначать значения, значение *sil(s)* будет пропуском (system missing). Следовательно, объекты-одиночки не примутся в расчет при вычислении среднего значения *sil* всего кластерного решения, которое будет в этом случае равно среднему *sil* только объектов-неодиночек. Поэтому используйте эту опцию, если вас не заботит величина доли объектов, объединенных в кластеры в кластерном решении.

IFMERGE - каждому одиночному объекту *s* вычисляется свое значение критерия по следующей логистической функции:

где *b(s)* это это *b* объекта *s* (см. формулу критерия); и это средняя и ст. отклонение значений *a* среди объектов-неодиночек, а это среднее силуэт-значение среди них же. Коэффициент *w* принят за (вы можете изменить *w* на иное значение в первой команде тела макроса DEFINE… /COEF… !DEFAULT). Смысл этой формулы следующий. Одиночный объект получит *sil*, близкое к -1 (наихудшее возможное значение), если он находится близко к каким-л. кластерам или одиночным объектам: т.к. *b(s)* мало; и он получит *sil*, близкое к усредненному *sil* неодиночных объектов (т.е. индифферентное значение), если он находится далеко от всех кластеров/объектов: *b(s)* велико. Т.е. одиночка наказывается низким *sil*, когда он мог бы быть слит с кем-нибудь (но он не слит). Если же он отщепенец (не мог бы быть слит ни с кем), то его можно игнорировать и считать, что он не портит кластерное решение, и поэтому ему назначится значение *sil*, близкое к среднему закластеризованных объектов. При *b(s)=* одиночный объект получит *sil* посредине между -1 и .

Данный макрос не поддерживает опции SINGLE=ASSEEDS.

**WIDORIG**, **WIDCOLL**

Эти подкоманды такие же, как в [!KO\_SILHOU](#_МАКРОС_!SILHOU:_СИЛУЭТ-СТАТИСТИКА_2).

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива (SPLIT FILE). Макрос не слушается временных операций (стоящих под TEMPORARY). Для отбора наблюдений в анализ вы можете использовать фильтрацию (FILTER или USE) и SELECT IF. Кроме того, вы можете исключать из анализа разные наблюдения в разных кластерных решениях, приписав им missing или нулевые коды (см. об этом в п/к CLUSOL).

***Некоторые вопросы***

*График ширины силуэта появляется пустой*. См. тот же вопрос в [!KO\_SILHOU](#_МАКРОС_!SILHOU:_СИЛУЭТ-СТАТИСТИКА_2).

# МАКРОС !KO\_AICBIC: ИНФОРМАЦИОННЫЕ КРИТЕРИИ AIC и BIC

Version 1, March 2007. Tested on SPSS 11.5, 13, 15.

!KO\_aicbic scavars= *zq8\_1 to zq8\_8* /\*Мерные переменные (рекомендуется стандартизованные),

/\*поименно и/или ч-з to

/catvars= *q3 q12* /\*Категориальные переменные, поименно и/или ч-з to

/clusol= *clu6 clu5 clu4 clu3 clu2*

/\*Переменные, являющие собой кластерные решения (поименный список).

Минимум надо задать CLUSOL и хотя бы одно из SCAVARS, CATVARS.

Akaike Information Criterion (AIC) и Schwarz Bayesian Information Criterion (BIC) широко применяются для оценки информативности статистистических моделей. Информативность тем выше, чем меньше число параметров и чем точнее они предсказывают. SPSS процедура двухэтапного кластерного анализа (TwoStep Cluster Analysis) вычисляет синтетические – позволяющие учесть вместе непрерывные (мерные) и категориальные переменные – индексы AIC и BIC для определения исходного (ориентировочного для алгоритма) числа кластеров при кластеризации в режиме «авто». Макрос использует формулы AIC и BIC, позаимствованные из описания TwoStep кластерного анализа[[3]](#footnote-3) (IBM SPSS Statistics 20 Algorithms, IBM Corporation, 2011. См. тж.: Fraley, C., Raftery, A.E. How many clusters? Which clustering method? Answers via model-based cluster analysis // Computer Journal, 1998, 4, 578–588).

где *m* это число параметров в рассматриваемом *k*-кластерном решении и *Li* это лог-правдоподобие кластера *i* в нем.

где *ni* это численность в кластере *i*, *E1q* это энтропия кластера *i* по мерной переменной *q* (всего мерных переменных *p1*), *E2r* это энтропия кластера *i* по номинальной переменной *r* (всего номинальных переменных *p2*); *lr* это число категорий в переменной *r*. Энтропия кластера по переменной:

(*sq2* это посчитанная на “df=n” дисперсия в переменной *q* в кластере *i*; *c* же константа, добавленная во избежание логарифмирования нуля и принятая за дисперсию переменной *q* в целой выборке, так что если входящие переменные стандартизованы, то *c*=1)

(*ntr* это число объектов в категории *t* переменной *r*, в кластере *i*)

Как видно, функция правдоподобия складывается из изменчивости (энтропии) непрерывных переменных и равнораспределенности (энтропии) категориальных переменных; просуммированная от всех кластеров данного разбиения, она выражает собой внутрикластерную неплотность рассматриваемого кластерного решения.

AIC и BIC родственны друг другу и часто дают близкие результаты в сравнении качества разбиений. Оба критерия (т.к. штрафуют за «избыток» параметров) при прочих равных условиях отдают предпочтение малому числу кластеров (причем эта тенденция у BIC сильнее), поэтому рекомендуются в случаях, когда есть запрос «желательно поменьше кластеров». Чем ниже значения индекса, тем лучше кластерное решение.

Оба критерия нечувствительны к неодинаковости дисперсий у мерных переменных; от общего же уровня величины дисперсий мерных переменных они зависят уровнем (т.е. линейно с наклоном 1).

Как и критерий Ratkowsky-Lance, данные индексы можно назвать одномерно-обобщающими, а не многомерными, поскольку они вычисляют свою главную меру (энтропию в данном случае) по каждой переменной отдельно, прежде чем свести в одну величину. Отношением к пространственным свойствам кластеров и к форме распределения в кластерах эти критерии похожи на Ratkowsky-Lance. Следует, однако, учесть, что Ratkowsky-Lance чувствителен к неодинаковости дисперсий мерных переменных, поскольку обрабатывает их «как есть». AIC/BIC, как уже сказано, имплицитно уравнивают дисперсии этих переменных (сжимают/растягивают облако данных вдоль измерений-переменных до достижения равенства дисперсий). Тем не менее, если говорить о пространственных свойствах кластеров *входящих* данных, тенденции тут у AIC/BIC похожи на таковые Ratkowsky-Lance. AIC/BIC несколько устойчивее к повороту всего облака, чем Ratkowsky-Lance, но тоже предпочитают, чтобы дисперсия была одинакова вдоль осей-переменных.

Критерии чувствительны к числу объектов в данных, поэтому сравниваемые кластерные разбиения должны состоять из равного общего числа объектов *n*.

***Подкоманды***

**SCAVARS, CATVARS**

В SCAVARS укажите мерные переменные. В CATVARS укажите категориальные номинальные переменные (числовые). Оба списка можно писать поименно и/или ч-з to. Вы должны задать хотя бы один список из двух. Макрос исключает пропущенные значения «списочно»: если хотя бы в одной переменной наблюдение есть пропуск, оно исключается из калькуляций по всем переменным.

Не имеет значения, одинаковы ли дисперсии в мерных переменных, т.к. формула критерия уравнивает их дисперсии. Однако общий уровень дисперсии мерных переменных сказывается на значении критерия. Поэтому их рекомендуется вводить *стандартизованными* – все дисперсии приведены к 1 (стандартизовать переменные можно с помощью Descriptive Statistics – Descriptives). В этом случае вклад мерных и категориальных переменных можно будет считать в общем сбалансированным между двумя типами переменных.

**CLUSOL**

Поименный список группирующих наблюдения переменных – см. описание подкоманды выше, в макросе [!KO\_CALHARV](#_МАКРОС_!CALHARV:_КРИТЕРИИ).

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива (SPLIT FILE). Он не слушается временных операций (стоящих под TEMPORARY). Для отбора наблюдений в анализ вы можете использовать фильтрацию (FILTER или USE) и SELECT IF. Кроме того, вы можете исключать из анализа разные наблюдения в разных кластерных решениях, приписав им missing или нулевые коды (см. об этом в п/к CLUSOL).

1. Для примера: пусть есть в пространстве 2-х переменных 2 круглых кластера (расстояние между ними единица), 3 таких кластера (треугольник из них, расстояние между ними единица), 4 таких кластера (квадрат из них, расстояние между соседними единица). Особенности расположения кластеров в этих трех конфигурациях не одинаковые (в 2-кластерной облако данных вытянуто; в 4-кластерной существуют межкластерные расстояния больше единицы), что затрудняет считать три конфигурации одинаково внутренне валидными по некоей «универсальной» внутренней валидности. Как раз такой универсальной внутренней валидности и не существует. Одни кластерные критерии отреагируют на вышеозначенную неодинаковость в конфигурациях, отдав предпочтение этой или той (и это-то и будет входить в понятие тенденциозности, bias, критерия в отношении k), тогда как другие – нет. [↑](#footnote-ref-1)
2. Если имплицитный смысл задания *sil(s)=*0 для всякого одиночного объекта *s* состоит в том, чтобы определить *a(s)* и *b(s)* как удаленности *s* от *воображаемых* двух ближайших к нему кластеров *A* и *B*, которые равноудалены от него: *a(s)=b(s)*, и *потому-то* *s* остался одиночкой (ибо оправданность приписать его к *A* такая же точно, как оправданность приписать его к *B*), – тогда, по-видимому, значение *sil(s)=*0 логически корректно – как нейтральное, т.е. выражающее такой «нерешенный» статус *s* – для любого типа силуэт-статистики, в котором *a(.)* и *b(.)* это одноименные функции расстояния (т.е., напр., оба *a(.)* и *b(.)* это расстояния до ближнего соседа, или оба они расстояния до центроида, и т.п.). Среди пяти типов TYPE только TYPE=FARNEAR содержит *a(.)* и *b(.)*, которые есть разноименные функции. Таким образом, только для TYPE=FARNEAR *sil(s)=*0 явно некорректная установка. [↑](#footnote-ref-2)
3. Это не значит, что значения индексов, выдаваемые процедурой TwoStep cluster analysis, всегда совпадут с таковыми, вычисленными макросом. Дело в том, что SPSS алгоритм TwoStep Cluster вычисляет AIC или BIC по ходу кластеризации (это «прикидочные» значения), а не после кластеризации (т.е. из готового кластерного разбиения). Кроме того, AIC/BIC не единственный показатель, на чем основано авто-определение числа кластеров в TwoStep. [↑](#footnote-ref-3)