***Clustering***

SPSS macros by Kirill Orlov

kior@akado.ru, ttnphns@gmail.com

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

All rights reserved

*Кластеризация*. Макросы для иерархического кластерного анализа (с опциями принуждения к предсуществующей структуре, преждевременной остановки, и другими), для вычисления расстояний между уже имеющимися группами/кластерами и для приписания новых объектов к ним. Макрос для инициирования центров кластеров в методе K-средних.

*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.

* [!KO\_HIECLU](#_МАКРОС_!HDIMPUT:_КОЛОДНАЯ_(HOT-DECK) выполняет агломеративный иерархический кластерный анализ (11 методов связывания на выбор). В качестве входящих требуется матрица близостей (расстояний или сходств) между объектами или группами их. Процедура предлагает ряд опций: есть возможность заставить алгоритм сначала собрать заданные пользователем кластеры, прежде чем сливать уже их, а также возможность прервать агломерацию на любом шаге и сохранить остаточную матрицу близостей.

* [!KO\_HIECLUEX](#_МАКРОС_!HIECLUEX:_ИЕРАРХИЧЕСКИЙ_1) выполняет иерархический кластерный анализ с тремя экзотическими методами связи: расстоянием Хаусдорфа, модифицированным расстоянием Хаусдорфа и точечно-центроидным кросс-расстоянием.
* [!KO\_ASSCLU](#_МАКРОС_!ASSCLU:_СБОРКА) собирает заданные пользователем кластеры и вычисляет расстояния между ними, а также приписывает новые объекты к этим существующим кластерам. В качестве входящих требуется матрица расстояний или сходств между объектами, существующая кластерная структура (группирование) и новые объекты, если есть.
* [!KO\_POINTCLUD](#_МАКРОС_!KO_POINTCLUD:_РАССТОЯНИЯ) вычисляет расстояния между объектами и кластерами (группами). В качестве входящих требуется матрица расстояний между объектами и существующая кластерная структура (группирование).
* [!KO\_DENDRO](#_МАКРОС_!DENDRO:_ДЕНДРОГРАММА_1) строит дендрограмму по агломеративной истории кластеризации.
* [!KO\_KMINI](#_МАКРОС_!KMINI:_НАЧАЛЬНЫЕ) создает или подбирает инициальные центры для кластерного анализа k-средних; есть несколько методов на выбор.

# МАКРОС !KO\_HIECLU: ИЕРАРХИЧЕСКИЙ КЛАСТЕРНЫЙ АНАЛИЗ

Version 3, Sep 2019 (Version 1, Feb 2014). Tested on SPSS Statistics 20, 22, 25.

!KO\_hieclu matrix= *var1 to var150* /\*Столбцы, образующие тело матрицы; можно через to

/seq= /\*Последовательность рядов/столбцов будет как у взятых столбцов (COL, тж п/у) или

/\*как у взятых рядов (ROW)

/id= *id* /\*Опционально: числовая переменная-идентификатор рядов/столбцов

/n= /\*Если кластеризуемые предметы это уже кластеры: переменная с внутрикластерными частотами

/square= /\*Для матрицы расстояний: на входе возвести в квадрат (YES) или не надо (NO, тж п/у)

/method= BAVERAGE /\*Метод связи: SINGLE, COMPLETE, BAVERAGE, EQBAVER , WAVERAGE, CENTROID,

/\*MEDIAN, WARD, MNSSQ, MIVAR, MNVAR

/preclu= *clu8* /\*Опционально: переменная кластерного членства, задающая предсуществующую структуру

/stop= /\*Опционально: остановить агломерацию раньше времени: STEP число, NCLU число, PROP доля

/\*IPROP доля, COEF значение, MAXN число, BOTHN число, ASS1, ASS2, ASS3 (см)

/save= *2 20* /\*Опционально: сохранить в новый массив переменные кластерного членства:

/\*одно кластерное решение (число) или диапазон решений (два числа) или все (ALL),

/\*или на момент остановки (ATSTOP)

/msave= /\*Опционально, если задано STOP: сохранить оставшуюся матрицу (имя файла)

/sched= *'d:\exercise\history.sav'* /\*Опционально: сохранить агломеративную историю

/\*(имя файла или заявленного массива)

/root= /\*Для некоторых методов (см.): извлечь корень из агломеративной истории (YES)

/\*или нет (NO, тж п/у)

/cumul= /\*В агломеративной истории кумулировать значения (YES) или нет (NO, тж п/у)

/print= /\*Распечатка: полная (FULL, тж п/у), без справки об ID рядов/столбцов в матрице (NOID),

/\*без агломеративной истории (NOSCHED), только сводка (SUMM).

Минимум надо задать MATRIX, METHOD.

Макрос делает иерархическую кластеризацию на основе матрицы попарных близостей (различий/расстояний или сходств). Предлагается на выбор 11 методов агломерации (связывания), включая «гибкие» методы средней связи. Учтите, что методы – центроидный, медианный, Уорда, суммы квадратов, прироста дисперсии, дисперсионный – требуют, ради геометрической правильности, матрицы квадратных евклидовых расстояний, в худшем случае – иных метрических расстояний (тоже оквадраченных), но уж никак не сходств.

Макрос несколько медленнее, чем команда CLUSTER в SPSS, но значительно превосходит ее в богатстве возможностей, встроенных в алгоритм. Именно, помимо обычной кластеризации объектов макрос позволяет:

* кластеризовать не только одиночные объекты, но и их группы (в ряде случаев это эквивалентно частотному взвешиванию) – см. подкоманду N;
* задать предсуществующую произвольную кластерную структуру, которую алгоритм обязан будет пошагово собрать, прежде чем продолжить агломерацию дальше – см. подкоманду PRECLU;
* остановить агломерацию на любом шаге, произвольном или по критерию, и сохранить остаточную матрицу близостей – см. подкоманду STOP;
* можно приписать новые объекты к существующим кластерам, а агломерацию между последними не делать;
* в макросе больше методов связи, чем в CLUSTER.

Макрос не строит дендрограмму или иных графиков. Для построения дендрограммы используйте отдельный макрос !KO\_DENDRO, работающий с агломеративной историей, которую можно сохранить. Также можно сохранить кластерные решения – переменные кластерного членства. В окно результатов макрос выдает агломеративную историю и сводку об анализе.

ПРИМЕР 1.

proximities v1 to v10 /view= case /measure= seuclid /matrix= out(\*) /print= none.

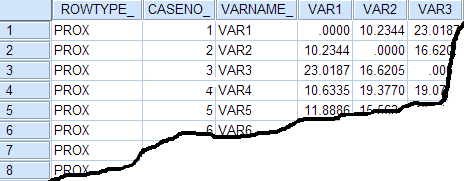
!KO\_hieclu matrix= var1 to var80 /method= COMPLETE /sched= 'd:\exercise\schedule.sav'.

!KO\_dendro sched= 'd:\exercise\schedule.sav'.

* Кластеризуются 80 наблюдений (по 10 переменным). Матрица квадратных евклидовых расстояний между наблюдениями вычислена командой PROXIMITIES и выведена как новый рабочий массив.
* Макрос !KO\_HIECLU делает кластеризацию методом полной связи и сохраняет агломеративную историю как внешний файл.
* Макрос !KO\_DENDRO берет этот файл и строит дендрограмму.

***Строение матрицы***

Массив данных должен быть матрицей попарных расстояний или сходств. Имена переменных – столбцов матрицы – до 8 байтов. Обязательно присутствие переменной VARNAME\_, именующей ряды в соответствие столбцам. Имена, являющиеся значениями этой переменной, должны быть написаны в том же регистре, как тождественные им имена среди имен переменных. Макрос не требует, чтобы ряды и столбцы шли в одинаковом порядке или чтобы их число и состав были полностью одинаковыми: макрос сам выберет из входящей матрицы одинаковые своими именами ряды и столбцы и соупорядочит их, чтобы таким образом составленная для анализа матрица имела квадратное диагонализованное строение. Переменная ROWTYPE\_ и прочие вспомогательные – не обязательны во входящей матрице.



***Подкоманды***

**MATRIX**

Укажите переменные рабочего массива, являющиеся столбцами собственно матрицы расстояний или сходств. Вы можете указать все или только нужные столбцы и в произвольном порядке. Можно использовать “to” для задания диапазоном.

*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?*.

Если ваши данные – **расстояния** (различия), то «диагональные» значения - т.е. данные в ячейках на пересечении одноименных столбцов и рядов – должны быть нулями, а прочие («внедиагональные») значения должны быть неотрицательны; большее число отвечает большему различию. Когда же ваши данные – **сходства**, то «диагональные» значения должны быть положительны, «внедиагональные» значения могут иметь любой знак; большее число (учитывая знак) отвечает большей схожести. Макрос не использует диагональные значения в кластеризации, но узнаёт данные как расстояния или сходства, глядя на них.

**SEQ**

Взяв входящую матрицу согласно MATRIX, макрос составляет из всех наличных в ней одноименных рядов и столбцов квадратную симметричную матрицу для анализа. Подкоманда SEQ диктует, в какой последовательности в ней будут идти ряды/столбцы:

COL - (тж. по умолчанию/незаданию) ряды и столбцы будут идти в порядке списка имен столбцов в подкоманде MATRIX.

ROW - ряды и столбцы будут идти в порядке расположения рядов (значений VARNAME\_) во входящей матрице.

Если матрица, введенная подкомандой MATRIX, квадратна и с соупорядоченными рядами и столбцами, подкоманда SEQ не играет роли.

ПРИМЕР 2. Изменение порядка объектов в матрице на иной, случайный.

compute sortvar= uniform(1).

sort cases by sortvar.

!KO\_hieclu matrix= var1 to var80 /seq= ROW /id= objid /method= BAVERAGE.

* Строки входящей матрицы сортируются по значениям случайной переменной SORTVAR.
* SEQ=ROW макроса составляет матрицу для анализа, в которой порядок рядов и столбцов такой, как у рядов входящей матрицы.
* Переменная-идентификатор объектов OBJID (приложенная к матрице) удерживает прежнюю идентификацию рядов/столбцов.
* Изменение порядка в матрице обычно не влияет на результаты; оно может повлиять, если в матрице имеются одинаковые (tied) значения близостей.

**ID**

Вы можете указать имя (до 8 байтов) числовой переменной-идентификатора рядов/столбцов. Ее значения будут использованы в агломеративной истории – в ее столбцах последней *Cluster1* и *Cluster2*, для обозначения сливающихся кластеров. В норме в этой переменной должно не быть дублирующихся значений и пропусков; их наличие не скажется на кластеризации, однако агломеративная история станет непригодной для последующего корректного построения дендрограммы. В переменной не разрешено значение: *-999*. По умолчанию/незаданию этой подкоманды идентификатором рядов/столбцов (и следовательно кластеров) будут их порядковые номера в составленной для анализа матрице.

*Обозначение кластеров в агломеративной истории*. Когда сливаются два кластера, кластер с меньшим кодом показывается в столбце Cluster1, а кластер с большим кодом – в столбце Cluster2; при этом новообразованный кластер получает меньший из двух кодов, т.е. код Cluster1, и дальше он фигурирует в агломеративной истории под этим кодом, пока не сольется с кластером, имеющим еще меньший код.

В агломеративной истории есть еще столбец *CluCode*, показывающий код новообразованного кластера: это его код в переменной кластерного членства (см. п/к SAVE), а не идентификаторный код, о котором шла речь выше и который фигурирует в столбцах Cluster1 и Cluster2.

**N**

Если рядами/столбцами входящей матрицы близостей выступают не единичные объекты, а целые кластеры или группы объектов, то необходимо проинформировать макрос о частотах в них. Укажите приложенную к матрице переменную (имя – до 8 байтов), показывающую внутрикластерное число объектов (значения – целые положительные числа). Если все значения =1, то это эквивалентно кластеризации единичных объектов, т.е. незаданию данной подкоманды.

Предполагается, что близости между кластерами (группами), содержащиеся в матрице, должны отвечать по природе тому методу кластеризации, какой вы собираетесь использовать (см. п/к METHOD, а также п/к MSAVE). Например, если собираетесь применить центроидный метод, входящее расстояние между двумя группами объектов в матрице ожидается быть квадратным расстоянием между центроидами этих групп. А если, более сложный пример, вы собираетесь применить метод Уорда, то входящее расстояние между двумя группами объектов в матрице ожидается быть удвоенным приростом суммы квадратов отклонений, происходящим от слияния этих двух групп.

Данная подкоманда несовместима с методами WAVERAGE, MNSSQ и MNVAR. Кластеризация объектов методами SINGLE, COMPLETE, EQBAVER, MEDIAN нечувствительна к этой подкоманде.

Используема ли п/к N с целью частотного *взвешивания* объектов – см. в параграфе «Некоторые вопросы» ниже.

**SQUARE**

Эта подкоманда только для расстояний (различий). Некоторые методы в макросе (CENTROID, MEDIAN, WARD, MNSSQ, MIVAR, MNVAR) требуют, чтобы входящие расстояния были квадратными. Если ваша матрица содержит неквадратные расстояния, вы можете затребовать их оквадратить на входе: SQUARE=YES.

**METHOD**

Укажите метод связывания в агломерации. Методы различаются тем, как определяется близость между любыми двумя кластерами на всяком шаге. Коэффициент слияния (показываемый в агломеративной истории) – это близость между двумя кластерами, слитыми на данном шаге.

SINGLE - метод **единичной связи,** или **ближнего соседа**. Близость между двумя кластерами это близость между их двумя ближайшими объектами. Это число – одно из значений входящей матрицы.

COMPLETE - метод **полной связи,** или **дальнего соседа**. Близость между двумя кластерами это близость между их двумя отдаленнейшими объектами. Это число – одно из значений входящей матрицы.

BAVERAGE [*число*] - метод **межгрупповой средней связи** (UPGMA). Близость между двумя кластерами это средняя арифметическая всех близостей между объектами того, с одной стороны, и объектами другого, с другой стороны. Об опциональном параметре *число* после ключевого слова – см. ниже.

EQBAVER [*число*] - метод **простой усредненный**, или **сбалансированной** **межгрупповой средней связи** (WPGMA). Близость между двумя кластерами это средняя арифметическая всех близостей между объектами того, с одной стороны, и объектами другого, с другой стороны; причем подкластеры, из которых недавно был слит каждый этот кластер, имеют уравненное влияние на ту близость, даже если подкластеры различались числом объектов. Об опциональном параметре *число* после ключевого слова – см. ниже.

WAVERAGE - метод **внутригрупповой средней связи** (MNDIS). Близость между двумя кластерами это средняя арифметическая всех близостей в их объединенном кластере.

CENTROID - метод **центроидный** (UPGMC). Близость между двумя кластерами это близость между их геометрическими центроидами: квадратное евклидово расстояние между ними.

MEDIAN - метод **медианный**, или **сбалансированный центроидный** (WPGMC). Близость между двумя кластерами это близость между их геометрическими центроидами (квадратное евклидово расстояние между ними); причем центроиды определены так, что подкластеры, из которых недавно был слит каждый этот кластер, имеют уравненное влияние на его центроид, даже если подкластеры различались числом объектов.

WARD - метод **Уорда**, или наименьшего **прироста суммы квадратов** (MISSQ). Близость между двумя кластерами это величина, на которую суммарный квадрат в их объединенном кластере станет больше, чем сложенный суммарный квадрат в двух этих кластерах: SS12-(SS1+SS2). (Между двумя одиночными объектами эта величина = квадратное евклидово расстояние / 2.)

MNSSQ - метод наименьшей **суммы квадратов**. Близость между двумя кластерами это суммарный квадрат в их объединенном кластере: SS12. (Между двумя одиночными объектами эта величина = квадратное евклидово расстояние / 2.)

MIVAR - метод наименьшего **прироста дисперсии**. Близость между двумя кластерами это величина, на которую средний квадрат в их объединенном кластере станет больше, чем взвешенно (числом объектов) усредненный средний квадрат в двух этих кластерах: MS12-(n1MS1+n2MS2)/(n1+n2) = [SS12-(SS1+SS2)]/(n1+n2). (Между двумя одиночными объектами эта величина = квадратное евклидово расстояние / 4.)

MNVAR - метод наименьшей **дисперсии**. Близость между двумя кластерами это средний квадрат в их объединенном кластере: MS12 = SS12/(n1+n2). (Между двумя одиночными объектами эта величина = квадратное евклидово расстояние / 4.).

Методы EQBAVER , MNSSQ, MIVAR, MNVAR отсутствуют в SPSS; см. о них напр. Podany J. New combinatorial clustering methods // Vegetatio, 1989, 81: 61-77.

Первые 5 методов допускают любые близости (любые сходства, расстояния). Последние 6 методов требуют расстояний; причем полностью корректно будет использовать только квадратные евклидовы расстояния, т.к. эти методы вычисляют центроиды в евклидовом пространстве. Вы можете оквадратить расстояния п/к SQUARE.

Методы центроидный, медианный, наименьшего прироста дисперсии – могут давать иногда так называемые *реверсии*: явление, когда сливаемые на некотором шаге два кластера оказываются ближе друг к другу, чем пары кластеров, слившихся ранее. Это потому, что эти методы не относятся к т.н. ультраметрическим.

Методы единичной связи и центроидный относятся к так называемым *сужающим пространство*, или «цепочечным». Это значит, что они склонны подсоединять к кластерам объекты поодиночке, поэтому имеют сравнительно плавный рост кривой «% закластеризованных объектов». Напротив, методы полной связи, Уорда, суммы квадратов, прироста дисперсии и дисперсии обычно уже на ранних шагах закластеризовывают значительную часть объектов в разные кластеры, и потом сливают уже эти кластеры – поэтому их кривая «% закластеризованных объектов» крута с первых шагов. Эти методы называются *расширяющими пространство*. Прочие же методы находятся посередине.

**Гибкие методы средней связи UPGMA и WPGMA** (Belbin, L. et al. A Comparison of Two Approaches to Beta-Flexible Clustering // Multivariate Behavioral Research, 1992, 27, 417–433.) задаются добавлением параметра в пределах -1≤ *число*<1 после слова BAVERAGE (гибкий UPGMA) и EQBAVER (гибкий WPGMA). Например: METHOD=BAVERAGE -0.1. Параметр 0 эквивалентен стандартным UPGMA и WPGMA. Параметр *вносит поправку* в вычисляемую межкластерную близость, зависящую от величины (разуплотненности) кластеров. Смысл параметра состоит в том, что он делает метод агломерации тем более расширяющим пространство, чем он более отрицателен, и делает метод агломерации тем более сужающим пространство, чем он более положителен. Следует использовать значения, близкие к 0, иначе качество снижается. На основании изучения способности методов UPGMA и WPGMA восстанавливать симулированные кластеры в евклидовом пространстве Belbin и коллеги заключили, что параметр немного ниже 0 (примерно -0.1 для UPGMA и -0.2 для WPGMA) в большинстве случаев оптимален, причем при наличии шумов/выбросов его полезно слегка сместить еще ниже.

**PRECLU**

Этой подкомандой можно задать предсуществующую кластерную структуру. Укажите числовую переменную кластерного членства (имя до 8 байтов), приложенную к массиву-матрице. Значения этой переменной есть коды предсуществующих кластеров или групп. Тогда макрос сначала проделает агломерацию раздельно для этих кластеров (словно они разные массивы), чтобы собрать каждый, а далее продолжит агломерацию этих кластеров уже друг с другом. Таким образом, макрос в любом случае начинает кластеризацию с отдельных объектов (или групп, при заданной п/к N), но в данном случае он это делает в режиме ограничения на слитие объектов.

Переменная предсуществующей кластерной структуры может иметь *пропуски*. Пропуски трактуются макросом как такие объекты, которые запрещено объединять *друг с другом* до того, как они соединятся с какими-л. кластерами или объектами-непропусками. Переменная должна не иметь значения: *-999*.

ПРИМЕР 3.

!KO\_hieclu matrix= var1 to var80 /method= BAVERAGE /preclu= agegroup /save= 2 10.

dataset name clusol.

* Кластеризуются 80 наблюдений, которые уже сгруппированы («предкластеризованы») согласно переменной AGEGROUP. Макрос будет кластеризовать с начала с целью достичь классификации AGEGROUP; достигнув ее, он продолжит агломерацию, объединяя уже эти группы.
* Заказано сохранить в новом массиве данных кластерные решения от 2 до 10 кластеров. Этот массив поименован CLUSOL.

**STOP**

Необязательная подкоманда, которой можно остановить (выйти из) агломерацию до ее окончания. Укажите критерий остановки:

STEP *целое число* - остановиться после этого по номеру шага агломерации.

NCLU *целое число* - остановиться по достижении этого числа кластеров (одиночный объект тоже считается кластером).

PROP *доля* - остановиться по достижении определенной доли закластеризованности объектов, т.е. доли объектов, переставших быть одиночными и нашедших себе по меньшей мере пару: укажите десятичную долю. Эта опция чувствительна к заданию п/к N, поскольку то есть частоты объектов, считающихся уже закластеризованными вместе.

IPROP *доля* - остановиться по достижении определенной доли закластеризованности рядов/столбцов матрицы: укажите десятичную долю. Эта опция позволена, только когда задана п/к N.

COEF *значение* - остановиться по достижении данного значения коэффициента слияния. Эта опция чувствительна к тому, как задана подкоманда CUMUL, а также ROOT.

MAXN *целое число* - остановиться как только слитие образовало кластер с этим (не менее) числом объектов.

BOTHN *целое число* - остановиться как только слились два кластера, каждый из которых имел не менее чем это число объектов.

Следующие три опции остановки требуют заданной подкоманды PRECLU и означают сделать только *сборку* (assemble) предсуществующей кластерной структуры и *приписание* (assign) новых объектов в эти существующие кластеры. «Новыми объектами» эти опции считают пропуски в PRECLU-переменной. Сделав сборку/приписание, процесс остановится: агломерации между кластерами не будет.

ASS1 - сделать только сборку предсуществующих кластеров; не делать приписания новых объектов к ним (последние останутся отдельными объектами).

ASS2 - сделать сборку предсуществующих кластеров и затем независимое приписание новых объектов к ним; в этом варианте новый объект (пропуск), будучи зачислен в ближайший к нему кластер, не влияет на приписания других новых объектов, т.к. характеристики кластеров до конца приписания удерживаются.

ASS3 - сделать сборку предсуществующих кластеров и затем прогрессивное приписание новых объектов к ним; в этом варианте новый объект (пропуск), будучи зачислен в ближайший к нему кластер, может влиять на дальнейшие приписания других новых объектов, т.к. приписания текуче изменяют характеристики кластеров.

Если пропусков в PRECLU-переменной не было, ASS2 и ASS3 эквивалентны ASS1: только сборка. Опции ASS1 и ASS2 возможно исполнить также отдельным макросом [!KO\_ASSCLU](#_МАКРОС_!ASSCLU:_СБОРКА).

ПРИМЕР 4. Приписать новые объекты к старым кластерам.

!KO\_hieclu matrix= var1 to var320 /method= CENTROID /preclu= clu /stop= ASS3 /id= clu /save= ATSTOP.

* Переменная CLU это старые кластеры – полученные в прошлой волне сегментации респондентов, плюс – в виде пропусков – новые респонденты. Матрица расстояний между всеми – старыми плюс новыми – респондентами вводится в анализ.
* STOP=ASS3 соберет сначала старые кластеры (это предсуществующая структура), после чего зачислит (прогрессивным вариантом) новых респондентов в ближайшие им кластеры (на основании центроидного подхода). Дальнейшей агломерации, т.е. между кластерами, не будет.
* Исследователь указал в качестве идентификатора кластеров саму переменную CLU – чтобы убедиться воочию, глядя на агломеративную историю, что сначала будут собраны старые кластеры, а потом к ним приписаны новые объекты.
* SAVE=ATSTOP сохранит кластерное решение на момент остановки процесса, т.е. после сборки и приписания: выданная переменная кластерного членства *CLUSTOP* в данном примере будет эквивалентна переменной ID, сохраненной вместе с ней, хотя в ней будут другие коды.

**SAVE**

Эта подкоманда для сохранения кластерных решений – переменных кластерного членства в новый безымянный рабочий массив. Укажите одно число (будет сохранено кластерное решение с этим числом кластеров) либо два числа, первое меньше второго (будет сохранен диапазон кластерных решений от … до … кластеров), либо ALL (тогда будут сохранены все кластерные решения). Коды кластеров в этих переменных – всегда последовательные числа1, 2, 3,…, а не коды переменной-идентификатора (см. п/к ID), которая тоже сохраняется заодно (если ID не было задано, сохраняются порядковые номера рядов/столбцов, *SEQNO\_*). Сохраняется и переменная *VARNAME\_* из входящей матрицы.

При заданной п/к STOP вы также можете указать SAVE=ATSTOP. В этом случае макрос сохранит две переменные кластерного членства: кластеры, получившиеся на шаге остановки кластеризации (*CLUSTOP*), и кластеры на предыдущем шаге (*CLUSTOP#*).

**MSAVE**

Эта подкоманда разрешена, если задана п/к STOP. Сохраняет матрицу, какая она на момент прерывания агломерации. Укажите внешний SAV-файл для сохранения, в кавычках или апострофах (указать пустой заявленный массив – нельзя). Вы можете в дальнейшем употребить эту матрицу снова, как входящую в !KO\_HIECLU, что будет являться продолжением прерванной кластеризации. Вы можете также использовать матрицу в других видах анализа, например многомерном шкалировании.

Значения в сохраняемой матрице ­– это близости между кластерами, как они были посчитаны примененным вами методом агломерации к моменту выхода из нее. Определение, что такое близость между кластерами при данном методе агломерации – см. в описании п/к METHOD. Заметьте, что при методах WARD и MNSSQ их выходящие близости умножены на 2, а при методах MIVAR и MNVAR умножены на 4: так требуется для сопоставимости с входящими близостями (в норме квадратными евклидовыми расстояниями при этих методах).

При матрице будет столбец *METHOD\_*, информирующий о методе, как памятка (его можно удалить). Еще при матрице будет столбец *N\_* с внутрикластерными частотами объектов. При продолжении кластеризации макросом !KO\_HIECLU надо будет указать имя этого столбца в п/к N. При матрице будет и столбец-идентификатор рядов/столбцов – это коды, обозначавшие кластеры в агломеративной истории: это либо коды из заданной вами переменной (п/к ID), либо порядковые номера в анализируемой матрице (столбец тогда будет называться *SEQNO\_*). Столбец *VARNAME\_* в сохраненной матрице – это *не* одноименный столбец из входящей матрицы, а созданный заново столбец из имен *var1 var2 var3*, etc.

Внимание. Чтобы корректно использовать сохраненную матрицу в родных процедурах SPSS Statistics, таких как команда CLUSTER, вам надо будет удалить столбцы *METHOD\_* и *N\_*, а также прописать ярлык значения ‘PROX’ в столбце *ROWTYPE\_*. Ярлык должен начинаться со слова DISSIMILARITY, если близости в матрице – различия, и со слова SIMILARITY, если близости в матрице – сходства.

ПРИМЕР 5. Продолжение прерванной кластеризации.

!KO\_hieclu matrix= var1 to var80 /method= CENTROID /stop= NCLU 20 /msave= 'd:\exercise\mat.sav'.

get file 'd:\exercise\mat.sav'.

!KO\_hieclu matrix= var1 to var20 /n= n\_ /method= CENTROID.

* Первая команда кластеризует 80 объектов и останавливается, когда осталось 20 кластеров. Остаточная матрица сохраняется.
* Матрица открывается, и вторая команда кластеризует до конца. Заметьте, что используется подкоманда /N. Результаты будут те же, как если бы все кластеризовать до конца за один раз.
* В норме следует использовать во второй команде прежний метод агломерации (в данном случае CENTROID), поскольку при каждом методе близости вычисляются по-своему. Отступления от этого правила возможны, но последствия должны быть тщательно обдуманы. Несовпадение методов сделает кластерный анализ не прерванным-продолженным, а эвристическим двустадийным.

**SCHED**

Подкоманда, сохраняющая агломеративную историю (по ней потом можно построить дендрограмму). Укажите внешний SAV-файл для сохранения, в кавычках или апострофах, либо имя объявленного ранее, но еще не существующего массива. Или можно указать SCHED=\* для выведения в новый безымянный массив, но тогда не пользуйтесь подкомандой SAVE.

**ROOT**

Эта подкоманда действует только для методов, связью в которых выступает квадратное расстояние (CENTROID, MEDIAN) или сумма квадратов отклонений (WARD, MNSSQ, MIVAR, MNVAR). Если ROOT=YES, макрос выдает агломеративную историю с извлеченным корнем из коэффициентов слияния. По умолчанию ROOT=NO, что отвечает реализации в SPSS команде CLUSTER. Разница лишь в показываемых и сохраняемых коэффициентах слияния в агломеративной истории; результаты кластеризации одни и те же. Помните, что для указанных 6 методов макрос ожидает, что входящие расстояния – квадратны.

**CUMUL**

Эта подкоманда регулирует способ отображения коэффициентов слияния в агломеративной истории. Она не влияет на результаты кластеризации, но важна для построения дендрограммы. Также, CUMUL=YES несовместимо с анализом сходств. По умолчанию/незаданию и при CUMUL=NO в качестве коэффициента слияния используются близость между двумя сливаемыми кластерами. При CUMUL=YES используются *накопленная* близость: коэффициент слияния на каждом шаге есть коэффициент слияния на предыдущем шаге плюс близость на данном шаге.

По традиции в методах, основанных на *приросте* неплотности, таких как метод Уорда или MIVAR, обычно показывают (в том числе на дендрограмме) кумулятивный коэффициент (CUMUL=YES), это скорее по причинам удобства, чем теоретическим. Так, при CUMUL=YES выдаваемый коэффициент слияния в методе Уорда есть общая внутрикластерная сумма квадратов, наблюдаемая на момент данного шага. В большинстве других методов (не Уорда и MIVAR) обычно показывают близость в сливаемой сейчас паре (CUMUL=NO). Если ROOT=YES, корень берется прежде кумулирования.

ПРИМЕР 6. Разные версии выдачи при методе Уорда.

!KO\_hieclu matrix= var1 to var80 /method= WARD /sched= 'd:\exercise\schedule.sav' /cumul= YES.

!KO\_hieclu matrix= var1 to var80 /method= WARD /sched= 'd:\exercise\schedule.sav' /root= YES.

* Первая команда дает агломеративную историю (и следовательно дендрограмму), как в SPSS процедуре Cluster.
* Вторая команда дает агломеративную историю (и следовательно дендрограмму), похожую на R команду hclust, в которой реализована так наз. “Ward-2” версия алгоритма (см. Murtagh F., Legendre P. Ward's hierarchical clustering method // Journal of Classification, 31, 2014).
* Результаты кластеризации одни и те же, но общий вид дендрограммы может несколько различаться.

**PRINT**

Выдача в окно результатов:

FULL - (тж. по умолчанию/незаданию) подробный отчет: сводка, информация об идентификации/порядке рядов/столбцов в анализированной матрице, агломеративная история.

NOID - не показывать информацию об идентификации/порядке рядов/столбцов в анализированной матрице.

NOSCHED - не показывать агломеративную историю.

SUMM - показать только сводку.

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами). Он не рассчитан на расщепленное состояние массива данных (SPLIT FILE). Макрос слушается команд, выбирающих наблюдения (SELECT IF, FILTER, USE), в том числе стоящих под командой TEMPORARY.

***Некоторые вопросы***

*Чем отличается подкоманда N от подкоманды PRECLU?* Если каждый ряд/столбец входящей матрицы это уже сам по себе кластер/группа, то надо указать число объектов в нем; это и делается с помощью переменной, указываемой в п/к N. Если же требуется заставить агломерацию собрать несколько рядов/столбцов в кластер в первую очередь, то надо объединить их кодом такого кластера в переменной, указываемой в п/к PRECLU. Обе подкоманды могут существовать вместе.

*Я хочу частотно взвесить объекты: чтобы каждая строка матрицы близостей означала несколько одинаковых объектов. Могу я для этого использовать подкоманду N?* Да, если применяете методы SINGLE, COMPLETE, BAVERAGE, EQBAVER, CENTROID, MEDIAN. Когда эти методы сливают два кластера, им безразлична плотность внутри этих кластеров. Поэтому п/к N будет эквивалентна явному размножению строк/столбцов матрицы, т.е. частотному взвешиванию объектов. Вообще, среди методов, для которым макрос разрешает использовать п/к N, существуют следующие группы методов:

* SINGLE, COMPLETE, EQBAVER, MEDIAN. Эти методы не обращают внимание на внутрикластерные частоты, когда сливают кластеры: они нечувствительны к частотному взвешиванию объектов. Поэтому использование или неиспользование п/к N не отражается на кластеризации объектов в кластеры.
* BAVERAGE, CENTROID. Эти методы чувствительны к внутрикластерным частотам, но нечувствительны к внутрикластерным плотностям сливаемых кластеров. Вы можете использовать п/к N для частотного взвешивания, ее использование сказывается на результатах кластеризации; результаты те же, как при явном размножении строк/столбцов.
* WARD, MIVAR. Эти методы чувствительны к внутрикластерным плотностям сливаемых кластеров. Задание п/к N с частотами не эквивалентно по результату кластеризации явному размножению строк/столбцов матрицы. Поэтому, если вам нужно частотное взвешивание объектов, делайте размножение строк/столбцов матрицы.

# МАКРОС !KO\_HIECLUEX: ИЕРАРХИЧЕСКИЙ КЛАСТЕРНЫЙ АНАЛИЗ С ЭКЗОТИЧЕСКИМИ МЕТОДАМИ СВЯЗЫВАНИЯ

Version 1, Sep 2019. Tested on SPSS Statistics 20, 22, 25.

!KO\_hiecluex matrix= *var1 to var150* /\*Столбцы, образующие тело матрицы; можно через to

/seq= /\*Последовательность рядов/столбцов будет как у взятых столбцов (COL, тж п/у) или

/\*как у взятых рядов (ROW)

/id= *id* /\*Опционально: числовая переменная-идентификатор рядов/столбцов

/fw= /\*Опционально: переменная с частотными весами

/method= CROSSPC /\*Метод связи: HAUSDORFF, MHAUSDORFF, CROSSPC

/stop= /\*Опционально: остановить агломерацию раньше времени: STEP число, NCLU число,

/\*PROP доля, IPROP доля, COEF значение, MAXN число, BOTHN число

/save= *2 20* /\*Опционально: сохранить в новый массив переменные кластерного членства:

/\*одно кластерное решение (число) или диапазон решений (два числа) или все (ALL),

/\*или на момент остановки (ATSTOP)

/msave= /\*Опционально, если задано STOP: сохранить оставшуюся матрицу (имя файла)

/sched= *'d:\exercise\history.sav'* /\*Опционально: сохранить агломеративную историю

/\*(имя файла или заявленного массива)

/print= /\*Распечатка: полная (FULL, тж п/у), без справки об ID рядов/столбцов в матрице (NOID),

/\*без агломеративной истории (NOSCHED), только сводка (SUMM).

Минимум надо задать MATRIX, METHOD.

Макрос делает иерархическую кластеризацию на основе матрицы попарных различий (расстояний). Предлагается на выбор 3 метода агломерации (связывания): расстояние Хаусдорфа, модифицированное расстояние Хаусдорфа, точечно-центроидное кроссрасстояние. Эти расстояния из разряда «расстояния между наборами» (set distances). Иерархические методы связи на базе этих расстояний мы условно назвали экзотическими, потому что стандартная имплементация на базе рекуррентной формулы Ланса–Уильямса (лежащей в основе макроса !KO\_HIECLU), не годится для них. !KO\_HIECLUEX спрограммирован поэтому как другой макрос. В остальном функциональность и опции !KO\_HIECLUEX похожи на таковые [!KO\_HIECLU](#_МАКРОС_!HDIMPUT:_КОЛОДНАЯ_(HOT-DECK).

Учтите, что метод точечно-центроидного кроссрасстояния требуют, ради геометрической правильности, матрицы евклидовых расстояний (неквадратных), иные расстояния/различия подойдут лишь в той мере, насколько они близки к евклидовым расстояниям в смысле способности образовывать евклидово пространство.

ПРИМЕР 1.

proximities v1 to v10 /view= case /measure= jaccard /matrix= out(\*) /print= none.

do repeat var= var1 to var80.

compute var= 1-var.

end repeat.

!KO\_hiecluex matrix= var1 to var80 /method= HAUSDORFF /sched= 'd:\exercise\schedule.sav'.

!KO\_dendro sched= 'd:\exercise\schedule.sav'.

* Кластеризуются 80 наблюдений (по 10 двоичным переменным). Матрица сходств Жаккара между наблюдениями вычислена командой PROXIMITIES и выведена как новый рабочий массив.
* Сходства Жаккара переведены в различия вычитанием из 1.
* Макрос !KO\_HIECLUEX делает кластеризацию методом Хаусдорфа и сохраняет агломеративную историю как внешний файл.
* Макрос !KO\_DENDRO берет этот файл и строит дендрограмму.

***Строение матрицы***

См. описание в макросе [!KO\_HIECLU](#_МАКРОС_!HDIMPUT:_КОЛОДНАЯ_(HOT-DECK). В !KO\_HIECLUEX требования к матрице те же, но матрица может быть только расстояниями (различиями): !KO\_HIECLUEX не рассчитан на сходства.

***Подкоманды***

**MATRIX**

Укажите переменные рабочего массива, являющиеся столбцами собственно матрицы расстояний. Вы можете указать все или только нужные столбцы и в произвольном порядке. Можно использовать “to” для задания диапазоном.

*Задание открытым диапазоном с помощью “?”*. То же, что в !KO\_HIECLU.

Матрица должна быть матрицей расстояний (различия). Ее «диагональные» значения - т.е. данные в ячейках на пересечении одноименных столбцов и рядов – должны быть нулями, а прочие («внедиагональные») значения должны быть неотрицательны; большее число отвечает большему различию.

**SEQ**

Эта подкоманда такая же, как в макросе [!KO\_HIECLU](#_МАКРОС_!HDIMPUT:_КОЛОДНАЯ_(HOT-DECK).

**ID**

Вы можете указать имя (до 8 байтов) числовой переменной-идентификатора рядов/столбцов. Ее значения будут использованы в агломеративной истории – в ее столбцах последней *Cluster1* и *Cluster2*, для обозначения сливающихся кластеров. Эта подкоманда такая же, как в макросе !KO\_HIECLU (читай там), за тем исключением, что не разрешает пропуски в переменной-идентификаторе.

*Обозначение кластеров в агломеративной истории*. Такое же, как в !KO\_HIECLU (читай там).

**FW**

Опциональная подкоманда для частотного взвешивания. Вы можете запросить макрос считать при кластеризации тот или иной ряд/столбец матрицы не одним объектом, а несколькими объектами – группой *одинаковых* объектов. Для этого укажите приложенную к матрице переменную (имя – до 8 байтов), показывающую число копий объектов (значения – целые положительные числа). Если все значения =1, то это эквивалентно кластеризации единичных объектов, т.е. незаданию данной подкоманды.

Коэффициент слияния при METHOD=HAUSDORFF нечувствителен к этой подкоманде.

Чем отличается данная подкоманда от похожей подкоманды N макроса !KO\_HIECLU – см. в параграфе «Некоторые вопросы» ниже.

**METHOD**

Укажите метод связывания в агломерации. Методы различаются тем, как определяется близость между любыми двумя кластерами на всяком шаге. Коэффициент слияния (показываемый в агломеративной истории) – это близость между двумя кластерами, слитыми на данном шаге.

HAUSDORFF - близость между двумя кластерами это **расстояние** **Хаусдорфа** между ними. Оно определяется так. Каждый объект первого кластера находит себе ближнего соседа во втором кластере, и из этих (минимальных) расстояний берется наибольшее. И обратно: каждый объект второго кластера находит себе ближнего соседа в первом кластере, и из этих (минимальных) расстояний берется наибольшее. Из двух взятых таким образом расстояний берется большее.

MHAUSDORFF - близость между двумя кластерами это **модифицированное расстояние Хаусдорфа** между ними. Оно определяется так. Каждый объект первого кластера находит себе ближнего соседа во втором кластере, и от этих (минимальных) расстояний берется среднее арифметическое. И обратно: каждый объект второго кластера находит себе ближнего соседа в первом кластере, и от этих (минимальных) расстояний берется среднее арифметическое. Из двух полученных таким образом значений берется большее.

CROSSPC - близость между двумя кластерами это **точечно-центроидное кроссрасстояние** между ними. Это сумма расстояний от объектов одного кластера до центроида второго, плюс сумма расстояний от объектов второго кластера до центроида первого, и разделить на совокупное число объектов в двух кластерах.

CROSSPC ожидает, что входящие расстояния – евклидовы (неквадратные). Если ваши расстояния неевклидовы, посчитанный результат может быть содержательно невалиден; а может случиться и ошибка взятия корня из отрицательного числа. HAUSDORFF/ MHAUSDORFF совместимы с любыми различиями.

Все три метода могут давать иногда так называемые *реверсии*: явление, когда сливаемые на некотором шаге два кластера оказываются ближе друг к другу, чем пары кластеров, слившихся ранее. Это потому, что эти методы не относятся к т.н. ультраметрическим.

**STOP**

Необязательная подкоманда, которой можно остановить (выйти из) агломерацию до ее окончания. Укажите критерий остановки:

STEP *целое число* - остановиться после этого по номеру шага агломерации.

NCLU *целое число* - остановиться по достижении этого числа кластеров (одиночный объект тоже считается кластером).

PROP *доля* - остановиться по достижении определенной доли закластеризованности объектов, т.е. доли объектов, переставших быть одиночными и нашедших себе по меньшей мере пару: укажите десятичную долю. Эта опция чувствительна к заданию п/к FW, поскольку «дубликаты» объекта считаются объектами, уже закластеризованными вместе.

IPROP *доля* - остановиться по достижении определенной доли закластеризованности рядов/столбцов матрицы: укажите десятичную долю. Эта опция позволена, только когда задана п/к FW.

COEF *значение* - остановиться по достижении данного значения коэффициента слияния.

MAXN *целое число* - остановиться как только слитие образовало кластер с этим (не менее) числом объектов.

BOTHN *целое число* - остановиться как только слились два кластера, каждый из которых имел не менее чем это число объектов.

**SAVE**

Эта подкоманда для сохранения кластерных решений – переменных кластерного членства в новый безымянный рабочий массив. Она точно такая, как в [!KO\_HIECLU](#_МАКРОС_!HDIMPUT:_КОЛОДНАЯ_(HOT-DECK).

**MSAVE**

Эта подкоманда разрешена, если задана п/к STOP. Сохраняет матрицу расстояний, какая она на момент прерывания агломерации. Укажите внешний SAV-файл для сохранения, в кавычках или апострофах. Эта подкоманда полностью аналогична таковой в [!KO\_HIECLU](#_МАКРОС_!HDIMPUT:_КОЛОДНАЯ_(HOT-DECK).

**SCHED**

Подкоманда, сохраняющая агломеративную историю (по ней потом можно построить дендрограмму). Эта подкоманда тождественна таковой в [!KO\_HIECLU](#_МАКРОС_!HDIMPUT:_КОЛОДНАЯ_(HOT-DECK).

**PRINT**

Выдача в окно результатов. Эта подкоманда тождественна таковой в [!KO\_HIECLU](#_МАКРОС_!HDIMPUT:_КОЛОДНАЯ_(HOT-DECK).

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами). Он не рассчитан на расщепленное состояние массива данных (SPLIT FILE). Макрос слушается команд, выбирающих наблюдения (SELECT IF, FILTER, USE), в том числе стоящих под командой TEMPORARY.

***Некоторые вопросы***

*Чем отличается подкоманда FW от подкоманды N макроса !KO\_HIECLU?* Обе подкоманды похожи тем, что задают переменную с частотами *n* объектов в группах (каждый ряд/столбец матрицы рассматривается тогда как группа объектов). П/к N не требует, чтобы *n* объектов были идентичны друг другу. П/к FW считает *n* объектов идентичными, копиями друг друга. Поэтому FW это подкоманда для частотного взвешивания; N функционально шире, но может быть использована в том числе для частотного взвешивания (см. подробнее в !KO\_HIECLU).

*Могу я проделать в !KO\_HIECLUEX прерванную-продолженную агломерацию, как в ПРИМЕРЕ 5 макроса !KO\_HIECLU?* Не совсем. Это из-за различия подкоманд FW и N. В !KO\_HIECLUEX вы тоже можете прервать агломерацию и сохранить матрицу, а затем агломерировать эту матрицу далее, указав /FW=*N\_*. Однако FW будет считать все *n* объектов (*n* – значение переменной *N\_*) *одинаковыми* объектами, тогда как на самом деле это просто объекты, собранные в кластер и не обязанные быть в общем случае одинаковыми.

# МАКРОС !KO\_ASSCLU: СБОРКА КЛАСТЕРОВ / ПРИПИСАНИЕ ОБЪЕКТОВ К КЛАСТЕРАМ

Version 2, Jan 2020 (Version 1, June 2015). Tested on SPSS Statistics 20, 22, 25.

!KO\_assclu matrix= *var1 to var150* /\*Столбцы, образующие тело матрицы; можно через to

/seq= /\*Последовательность рядов/столбцов будет как у взятых столбцов (COL, тж п/у) или

/\*как у взятых рядов (ROW)

/id= *id* /\*Числовая переменная-идентификатор объектов

/cluvar= *clu5* /\*Существующая переменная кластерного членства

/square= /\*Для матрицы расстояний: на входе возвести в квадрат (YES) или не надо (NO, тж п/у)

/method= BAVERAGE /\*Метод: SINGLE, COMPLETE, BAVERAGE, WAVERAGE, CENTROID,

/\*WARD, MNSSQ, MIVAR, MNVAR, HAUSDORFF, MHAUSDORFF, CROSSPC

/save= YES /\*Опционально: сохранить в новый массив обновленную переменную

/\*кластерного членства - YES или NO (тж п/у)

/msave1= *'d:\exercise\mx1.sav'* /\*Опционально: сохранить промежуточную собранную матрицу

/\*(имя файла)

/msave2= *'d:\exercise\mx2.sav'* /\*Опционально: сохранить окончательную собранную матрицу

/\*(имя файла)

/print= /\*Распечатка: полная (FULL, тж п/у), без справки об ID рядов/столбцов в матрице (NOID),

/\*только сводка (SUMM).

Минимум надо задать MATRIX, ID, CLUVAR, METHOD.

Этот макрос предназначен для установления близостей между известными кластерами (вообще, между любыми группами объектов, задаваемыми номинальной переменной), и для зачисления объектов, принадлежность которых к кластерам/группам неизвестна («новые объекты») в ближайшие им кластеры/группы.

Макрос берет матрицу близостей между объектами и переменную предсуществующей кластерной (групповой) структуры. Не делая фактически иерархического кластерного анализа (пошаговой агломерации), он **собирает** эти предсуществующие кластеры тем или иным методом связывания – результатом чего является матрица близостей между кластерами. Далее, если имеются «новые» объекты, которые требуется **приписать** (assign) к тем предсуществующим кластерам (каждый объект – к ближайшему для него кластеру), макрос делает это.

Всё, что делает данный макрос, может сделать и [!KO\_HIECLU](#_МАКРОС_!HDIMPUT:_КОЛОДНАЯ_(HOT-DECK) (см. опции ASS1 – сборка, и ASS2 – сборка/приписание там в подкоманде STOP). Но !KO\_ASSCLU быстрее, чем !KO\_HIECLU, и в этом его польза. !KO\_ASSCLU имеет несколько иную и более удобную форму выдачи результата, чем эквивалентный по результату пуск !KO\_HIECLU.

Приписание, выполняемое макросом !KO\_ASSCLU, эквивалентно по результату приписанию STOP=ASS2 макроса !KO\_HIECLU, т.е. речь идет о независимом приписании. Новые объекты !KO\_ASSCLU приписывает к кластерам одновременно и независимо друг от друга; на приписание каждого нового объекта к тому или иному предсуществующему кластеру оказывает эффект исключительно близость между одним этим объектом и предсуществующими кластерами в том их виде, как они вошли в анализ.

В отличие от макроса !KO\_HIECLU, данный макрос принимает матрицу близостей только между индивидуальными объектами, но не группами объектов (т.е. ряд/столбец матрицы не может представлять несколько объектов, пусть даже одинаковых).

ПРИМЕР 1. Зачислить новые объекты в старые кластеры или группы.

!KO\_assclu matrix= var1 to var630 /cluvar= oldclus /id= objid /method= SINGLE.

* VAR1 до VAR630 это матрица расстояний между объектами, среди которых есть старые (уже были разбиты на группы или кластеры) и новые (их требуется распределить по тем группам/кластерам – приписать к ближайшему). Переменная OLDCLUS задает эту кластерную структуру; в ней – коды кластеров, а новые объекты не имеют пока кодов, они пропуски.
* Приписание будет по методу ближайшего соседа. Макрос отчитается, к какому кластеру окажется приписан каждый новый объект. Переменная-идентификатор объектов – обязательна.

***Строение матрицы***

Такое же, как в макросе [!KO\_HIECLU](#_МАКРОС_!HDIMPUT:_КОЛОДНАЯ_(HOT-DECK). Близости могут быть, как и там, расстояниями ли сходствами. Это должны быть близости между индивидуальными объектами.

***Подкоманды***

**MATRIX, SEQ**

Эти подкоманды такие же, как в макросе [!KO\_HIECLU](#_МАКРОС_!HDIMPUT:_КОЛОДНАЯ_(HOT-DECK).

**ID**

Обязательная подкоманда, указывающая имя (до 8 байтов) числовой переменной-идентификатора объектов, т.е. рядов/столбцов входящей матрицы. Все значения переменной должны быть валидными (ср., в макросе !KO\_HIECLU это не обязательно). Переменная должна не иметь значения: *-999*. Повторяющиеся значения допустимы (но вряд ли могут понадобиться).

**CLUVAR**

Обязательная подкоманда для задания предсуществующей кластерной структуры объектов – числовая неконстантная переменная (имя до 8 байтов). Значения ее это коды кластеров/групп. Кластер может состоять и из одного объекта. Разрешены *пропуски* – они понимаются как новые объекты, которые необходимо будет приписать к предсуществующим кластерам/группам. Переменная должна не иметь значения: *-999*.

**SQUARE**

Эта подкоманда только для расстояний (различий). Некоторые методы в макросе (CENTROID, WARD, MNSSQ, MIVAR, MNVAR) требуют, чтобы входящие расстояния были квадратными. Если ваша матрица содержит неквадратные расстояния, вы можете затребовать их оквадратить на входе: SQUARE=YES. Эта опция игнорируется при METHOD=CROSSPC.

**METHOD**

Укажите метод определения расстояния между кластерами (при сборке), равно как между кластерами и новыми объектами (при приписании). Это те же методы, что существуют в макросах !KO\_HIECLU и !KO\_HIECLUEX, только тут их применение не пошаговое, а единовременное, но их смысл от этого не меняется.

SINGLE - метод **единичной связи,** или **ближнего соседа**. Близость между двумя кластерами это близость между их двумя ближайшими объектами. Это число – одно из значений входящей матрицы.

COMPLETE - метод **полной связи,** или **дальнего соседа**. Близость между двумя кластерами это близость между их двумя отдаленнейшими объектами. Это число – одно из значений входящей матрицы.

BAVERAGE - метод **межгрупповой средней связи** (UPGMA). Близость между двумя кластерами это средняя арифметическая всех близостей между объектами того, с одной стороны, и объектами другого, с другой стороны.

WAVERAGE - метод **внутригрупповой средней связи** (MNDIS). Близость между двумя кластерами это средняя арифметическая всех близостей в их объединенном кластере.

CENTROID - метод **центроидный** (UPGMC). Близость между двумя кластерами это близость между их геометрическими центроидами: квадратное евклидово расстояние между ними.

WARD - метод **Уорда**, или наименьшего **прироста суммы квадратов** (MISSQ). Близость между двумя кластерами это величина, на которую суммарный квадрат в их объединенном кластере станет больше, чем сложенный суммарный квадрат в двух этих кластерах: SS12-(SS1+SS2). (Между двумя одиночными объектами эта величина = квадратное евклидово расстояние / 2.)

MNSSQ - метод наименьшей **суммы квадратов**. Близость между двумя кластерами это суммарный квадрат в их объединенном кластере: SS12. (Между двумя одиночными объектами эта величина = квадратное евклидово расстояние / 2.)

MIVAR - метод наименьшего **прироста дисперсии**. Близость между двумя кластерами это величина, на которую средний квадрат в их объединенном кластере станет больше, чем взвешенно (числом объектов) усредненный средний квадрат в двух этих кластерах: MS12-(n1MS1+n2MS2)/(n1+n2) = [SS12-(SS1+SS2)]/(n1+n2). (Между двумя одиночными объектами эта величина = квадратное евклидово расстояние / 4.)

MNVAR - метод наименьшей **дисперсии**. Близость между двумя кластерами это средний квадрат в их объединенном кластере: MS12 = SS12/(n1+n2). (Между двумя одиночными объектами эта величина = квадратное евклидово расстояние / 4.).

HAUSDORFF - близость между двумя кластерами это **расстояние** **Хаусдорфа** между ними. Оно определяется так. Каждый объект первого кластера находит себе ближнего соседа во втором кластере, и из этих (минимальных) расстояний берется наибольшее. И обратно: каждый объект второго кластера находит себе ближнего соседа в первом кластере, и из этих (минимальных) расстояний берется наибольшее. Из двух взятых таким образом расстояний берется большее.

MHAUSDORFF - близость между двумя кластерами это **модифицированное расстояние Хаусдорфа** между ними. Оно определяется так. Каждый объект первого кластера находит себе ближнего соседа во втором кластере, и от этих (минимальных) расстояний берется среднее арифметическое. И обратно: каждый объект второго кластера находит себе ближнего соседа в первом кластере, и от этих (минимальных) расстояний берется среднее арифметическое. Из двух полученных таким образом значений берется большее.

CROSSPC - близость между двумя кластерами это **точечно-центроидное кроссрасстояние** между ними. Это сумма расстояний от объектов одного кластера до центроида второго, плюс сумма расстояний от объектов второго кластера до центроида первого, и разделить на совокупное число объектов в двух кластерах.

Методы SINGLE, COMPLETE, BAVERAGE, WAVERAGE, HAUSDORFF, MHAUSDORFF допускают любые близости (любые сходства, расстояния). Методы CENTROID, WARD, MNSSQ, MIVAR, MNVAR, CROSSPC требуют расстояний; причем полностью корректно будет использовать только евклидовы расстояния, т.к. эти методы вычисляют центроиды в евклидовом пространстве. Из этих 6-ти методов метод CROSSPC ожидает неквадратных евклидовых расстояний, а остальные 5 – квадратных евклидовых расстояний (оквадратить расстояния можно п/к SQUARE).

Заметьте, что методы EQBAVER и MEDIAN, присутствующие в !KO\_HIECLU, тут отсутствуют. Потому, что они возможны только как пошаговые. Вы можете собрать кластеры этими методами макросом !KO\_HIECLU.

**SAVE**

Подкоманда, выводящая в новый безымянный массив обновленную переменную кластерного членства CLUVAR, в которой «новые объекты» (бывшие пропуски, если были) приписаны к кластерам. Укажите YES (вывести) или NO (не выводить, тж. по умолчанию). Приписанные объекты будут находиться в хвосте переменной. Для приписанных объектов будут добавлены столбцы *PROX* (близость к кластеру приписания), *OBOB* (близость к ближайшему другому новому объекту), *CLUCLU* (близость кластера приписания к ближайшему другому кластеру – это близость, бывшая *до* актов приписания). Сравнивая эти три значения, вы можете решить, что лучше не зачислять новый объект в старый кластер, а создать новый кластер из двух новых объектов, а может быть, слить два старых кластера.

В массив также заодно сохраняется переменная-идентификатор (см. п/к ID) и переменная *VARNAME\_* из входящей матрицы.

ПРИМЕР 2.

!KO\_assclu matrix= var1 to var630 /cluvar= oldclus /id= objid /method= SINGLE /save= YES.

!KO\_hieclu matrix= var1 to var630 /preclu= oldclus /id= objid /method= SINGLE /stop= ASS2 /save= ATSTOP.

* Эти две команды эквивалентны. Они различаются только форматом сохраняемых в новый массив данных. !KO\_ASSCLU сохранил переменную кластерного членства *OLDCLUS*, в которой новые объекты приписаны к старым кластерам, причем новые объекты сдвинуты в хвост переменной. !KO\_HIECLU сохранил переменные кластерного членства *CLUSTOP#* и *CLUSTOP*. *CLUSTOP* эквивалентна *OLDCLUS*, только использует другие коды (последовательные натуральные числа) и новые объекты не сдвинуты в хвост.

**MSAVE1, MSAVE2**

Подкоманды, которыми можно, если это нужно, сохранить во внешний файл (укажите для каждой подкоманды свой путь/имя в кавычках или апострофах) собранную матрицу близостей между кластерами (группами). MSAVE1 сохраняет матрицу после сборки предсуществующих кластеров до приписания к ним новых объектов (*промежуточная* матрица). Новые объекты в этой матрице представлены своими рядами/столбцами, и они идут последними. MSAVE2 сохраняет матрицу после того, как и новые объекты будут включены в кластеры (*окончательная* матрица). Поэтому в ней нет рядов/столбцов, соответствующим новым объектам. Вы можете задать обе подкоманды, одну любую или ни одной. Если новых объектов (пропусков в переменной, указанной в CLUVAR) не было, промежуточная матрица будет тождественна окончательной.

MSAVE1 соответствует /STOP=ASS1 /MSAVE= … макроса !KO\_HIECLU, с той лишь разницей, что новые объекты идут в матрице последними. MSAVE2 эквивалентно /STOP=ASS2 /MSAVE=… макроса !KO\_HIECLU.

Вы можете в дальнейшем употребить собранные матрицы в других видах анализа, например многомерном шкалировании. А также ввести в макрос !KO\_HIECLU для дальнейшего кластеризации уже собранных кластеров. Вы не можете, однако, использовать собранную матрицу в !KO\_ASSCLU как входящую, потому что !KO\_ASSCLU принимает матрицу близостей между *объектами*, а не группами объектов.

Значения в сохраняемой матрице ­– это близости между кластерами, как они были посчитаны примененным вами методом (п/к METHOD). При методах WARD и MNSSQ выходящие близости умножены на 2, а при методах MIVAR и MNVAR умножены на 4: так требуется для сопоставимости с входящими близостями (квадратными евклидовыми расстояниями).

При матрице будет столбец *METHOD\_*, информирующий о методе, как памятка (его можно удалить). Еще при матрице будет столбец *N\_* с внутрикластерными частотами объектов. Столбец-идентификатор рядов/столбцов – это кластерная переменная, указанная в п/к CLUVAR. Если во входящей CLUVAR были пропуски («новые объекты»), то в промежуточной матрице (MSAVE1) будет также переменная, указанная в п/к ID и идентифицирующая новые объекты. Столбец *VARNAME\_* в сохраненной матрице – это *не* одноименный столбец из входящей матрицы, а созданный заново столбец из имен *var1 var2 var3*, etc.

Внимание. Чтобы корректно использовать сохраненную матрицу в родных процедурах SPSS Statistics, таких как команда CLUSTER, вам надо будет удалить столбцы *METHOD\_* и *N\_*, а также прописать ярлык значения ‘PROX’ в столбце *ROWTYPE\_*. Ярлык должен начинаться со слова DISSIMILARITY, если близости в матрице – различия, и со слова SIMILARITY, если близости в матрице – сходства.

ПРИМЕР 3. Сборка матрицы с приписанием новых объектов.

!KO\_assclu matrix= var1 to var150 /cluvar= pre /id= objid /method= BAVERAGE

/msave1= 'd:\exercise\ass\_mx1.sav' /msave2= 'd:\exercise\ass\_mx2.sav'.

* Макрос собирает предсуществующие кластеры (заданные переменной PRE) и, если есть новые объекты (пропуски в PRE) – присоединяет их к кластерам (к ближайшему кластеру). Метод – средней связи. Матрица близостей после сборки, но до этого присоединения, сохранена п/к MSAVE1, а окончательная матрица – п/к MSAVE2.

!KO\_hieclu matrix= var1 to var150 /preclu= pre /id= objid /method= BAVERAGE /stop= ASS2

/msave= 'd:\exercise\hie\_mx.sav'.

* Показано, как то же сделано макросом !KO\_HIECLU. STOP=ASS2 делает сборку кластеров и затем такое же приписание, какое делает макрос !KO\_ASSCLU. Сохраненная матрица тождественна той, что была сохранена подкомандой MSAVE2. Отличие касается только столбца, приложенного к матрице в качестве идентификатора. Там это был PRE, здесь это OBJID. Однако вы вправе указать ID=PRE в макросе !KO\_HIECLU.

!KO\_hieclu matrix= var1 to var150 /preclu= pre /id= objid /method= BAVERAGE /stop= ASS1

/msave= 'd:\exercise\hie\_mx.sav'.

* Этот пуск !KO\_HIECLU собирает матрицу без приписания к ней новых объектов. Сохраненная матрица тождественна той, была что сохранена подкомандой MSAVE1. Отличие касается столбца, приложенного к матрице в качестве идентификатора. Там это был PRE, здесь это OBJID. Однако вы вправе указать ID=PRE в макросе !KO\_HIECLU. Отличается также порядок рядов/столбцов: MSAVE1 макроса !KO\_ASSCLU переместила новые объекты в хвост.

**PRINT**

Выдача в окно результатов:

FULL - (тж. по умолчанию/незаданию) подробный отчет: сводка, информация об идентификации/порядке рядов/столбцов во входящей матрице, отчет о приписании новых объектов.

NOID - не показывать информацию об идентификации/порядке рядов/столбцов во входящей матрице.

SUMM - показать только сводку.

# 

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами). Он не рассчитан на расщепленное состояние массива данных (SPLIT FILE). Макрос слушается команд, выбирающих наблюдения (SELECT IF, FILTER, USE), в том числе стоящих под командой TEMPORARY.

# МАКРОС !KO\_POINTCLUD: РАССТОЯНИЯ МЕЖДУ ОБЪЕКТАМИ И КЛАСТЕРАМИ

Version 2, Jan 2022 (Version 1, May 2016). Tested on SPSS Statistics 20, 22, 26.

!KO\_pointclud matrix= var1 to var150 /\*Столбцы, образующие тело матрицы; можно через to

/seq= /\*Последовательность рядов/столбцов будет как у взятых столбцов (COL, тж п/у) или

/\*как у взятых рядов (ROW)

/id= id /\*Числовая переменная-идентификатор объектов

/groups= clu5 /\*Группы: либо одна категориальная переменная, либо набор двоичных переменных;

/\*можно через to

/square= /\*На входе возвести в расстояния в квадрат (YES) или не надо (NO, тж п/у)

/method= AVER /\*Метод (расстояние от объекта до чего): NEAR, FARTH, AVER, CENTROID,

/\*MEDOID

/self= /\*Для AVER, CENTROID: учитывать объект в своей группе: IN (тж п/у), или нет (OUT)

/saveid= /\*Для NEAR, FARTH, MEDOID: выдать id этих терминальных объектов:

/\*'путь/файл' или объявленный 'массив'

/print= /\*Распечатка: полная (FULL, тж п/у) или без справки об ID рядов/столбцов

/\*в матрице (NOID).

Минимум надо задать MATRIX, ID, GROUPS, METHOD.

Вычисляет расстояния между объектами и существующими (заданными пользователем) кластерами (или группами) объектов. На выходе – новый безымянный рабочий массив, показывающий расстояние от каждого объекта до каждого кластера/группы объектов.

Макрос берет матрицу расстояний (различий) между индивидуальными объектами и переменные, задающие предсуществующую кластерную (групповую) структуру среди них. Группы могут пересекаться или не пересекаться составом объектов.

ПРИМЕР 1.

!KO\_pointclud matrix= var1 to var630 /groups= oldclus /id= objid /method= NEAR.

* VAR1 до VAR630 это матрица расстояний между объектами. Переменная OLDCLUS задает некоторое группирование среди объектов, например, это разбиение объектов на кластеры. OBJID это переменная-идентификатор объектов.
* Макрос выдаст расстояния между объектами входящего массива и кластерами. Вычисляемое расстояние - до ближнего соседа (METHOD=NEAR). Это расстояние между объектом и ближайшим к нему объектом рассматриваемого кластера.

filter by selvar.

!KO\_pointclud matrix= var1 to var630 /groups= oldclus /id= objid /method= NEAR.

* В данном случае делается то же, что выше, но анализ составят лишь фильтрованные наблюдения массива.

***Строение матрицы***

Такое же, как в макросе [!KO\_HIECLU](#_МАКРОС_!HDIMPUT:_КОЛОДНАЯ_(HOT-DECK). Но это могут быть только расстояния/различия, не сходства. И это должны быть расстояния между индивидуальными объектами.

***Подкоманды***

**MATRIX, SEQ**

Эти подкоманды такие же, как в макросе [!KO\_HIECLU](#_МАКРОС_!HDIMPUT:_КОЛОДНАЯ_(HOT-DECK).

**ID**

Обязательная подкоманда, указывающая имя (до 8 байтов) числовой переменной-идентификатора объектов, т.е. рядов/столбцов входящей матрицы. Все значения переменной должны быть валидными. Повторяющиеся значения допустимы (но вряд ли могут понадобиться).

**GROUPS**

Подкоманда для задания предсуществующей кластерной/групповой структуры объектов. Вы можете задать ее одним из двух способов:

* Одна числовая категориальная переменная. Значения ее это коды групп. Сколько значений (их должно быть минимум два) – столько и групп. Группа может состоять и из одного объекта.
* Две или более числовых двоичных (с кодами 1 и 0) переменных, поименно и/или ч-з “to”. Каждая такая переменная соответствует группе, и код 1 означает принадлежность к ней. Объект может принадлежать одной группе, нескольким группам, или ни одной из групп. В каждой переменной должны быть единицы (т.е. группа не может быть пуста). Как частный случай, в переменной может не быть нулей (тогда эта группа представляет собой все объекты вместе взятые).

Имена переменных, задаваемых в GROUPS – до 8 байтов. Пропуски запрещены (по крайней мере в наблюдениях массива, взятых в анализ).

ПРИМЕР 2.

!KO\_pointclud matrix= var1 to var120 var161 /groups= total gr1 to gr4 /id= objid /method= AVER

/self= OUT.

* VAR1 до VAR120 и var161 есть выбранные пользователем столбцы матрицы расстояний между объектами. В анализ войдут только эти выбранные объекты. Двоичные (1 vs 0) переменные TOTAL и GR1 до GR4 задают группирования. Переменная TOTAL не содержит нулей, поэтому соответствует всей выборке наблюдений (в данном случае все отобранные в анализ объекты как одна группа).
* METHOD=AVER требует вычислить усредненное расстояние от каждого объекта до объектов, принадлежащих группам. Если сам объект принадлежит группе, с какой он сопоставляется, то он не участвует в усреднении (его нулевое расстояние до себя изымается из усреднения), поскольку SELF=OUT.

**SQUARE**

Эта подкоманда для того, чтобы оквадратить расстояния на входе, если это вам нужно: SQUARE=YES. Вычисление расстояния до центроида (METHOD=CENTROID), например, требует ввести квадратные расстояния.

**METHOD**

Укажите, какой вид расстояния между объектом и группой вы хотите. Такое расстояние будет получено между каждым индивидуальным объектом *i* (взятым в анализ наблюдением массива) и каждой группой объектов (тоже из взятых в анализ).

NEAR - расстояние до **ближнего соседа**. Это расстояние от объекта *i* до ближайшего к нему объекта, принадлежащего группе. Это число – одно из значений входящей матрицы.

FARTH - расстояние до **дальнего соседа**. Это расстояние от объекта *i* до отдаленнейшего от него объекта, принадлежащего группе. Это число – одно из значений входящей матрицы.

AVER - вычислить **усредненное** (среднее) расстояние от объекта *i* до объектов группы.

CENTROID - вычислить расстояние от объекта *i* до **центроида** (локуса средней арифметической) группы.

MEDOID - расстояние от объекта *i* до **медоида** группы. Это число – одно из значений входящей матрицы. Медоид – это такой объект в группе, суммарное расстояние от которого до всех объектов группы минимально. (Если в группе оказалось более одного медоида, макрос выберет из них объект с бо́льшим порядковым номером.)

Все методы, кроме CENTROID, допускают на входе всякие различия/расстояния. При методе CENTROID корректно будет использовать только евклидовы расстояния; другие метрические расстояния можно ввести разве что с эвристической целью. По технической причине метод CENTROID нуждается, чтобы входящие расстояния были квадратные. Вы можете возвести расстояния в квадрат на входе подкомандой SQUARE.

*Самоучастие*. Когда объект *i* принадлежит группе, расстояние его до которой должно быть выяснено, его участие в выяснении этого расстояния такое. При методах NEAR и FARTH объект *i* не участвует, т.е. он не может быть своим собственным соседом, кроме случая, когда он – единственный член группы (и тогда расстояние будет 0). При методе MEDOID объект *i* участвует в определении медоида и сам может оказаться медоидом. При методах AVER и CENTROID вы можете выбрать участие или неучастие – см. п/к SELF.

**SELF**

Эта подкоманда действует только при METHOD= AVER или CENTROID. Когда объект, расстояние которого до группы должно быть вычислено, сам принадлежит этой группе, то вы можете выбрать, участвовать ему (SELF=IN, тж. по умолчанию/незаданию) или не участвовать (SELF=OUT) в вычислении этого расстояния. Второй вариант означает временное изымание объекта из своей группы (leave-one-out) при вычислении расстояния его до нее. К примеру, если METHOD=AVER и SELF=OUT, то нулевое расстояние объекта до себя самого не входит в усреднение расстояний, а при SELF=IN – входит.

Если объект – единственный член группы, он не изымается, расстояние его до этой группы = 0.

**SAVEID**

Эта подкоманда игнорируется при METHOD= AVER или CENTROID. При METHOD= NEAR, FARTH или MEDOID полученные расстояния – это расстояния до определенных, реальных объектов внутри групп. Если вам надо идентифицировать эти объекты – знать их коды по переменной ID, воспользуйтесь SAVEID для получения массива, структурно похожего на основной выходящий массив и содержащий идентификаторные коды объектов на месте соответствующих расстояний. Например, при методе NEAR вы получите массив, показывающий, какие именно объекты оказались этими ближними соседями, расстояния до которых выяснил макрос.

При методах NEAR и FARTH, если более одного объекта группы являются ближним (или, соответственно, дальним) соседом по отношению к объекту *i*, то подкоманда выберет из них для показа объект с бо́льшим порядковым номером.

В SAVEID укажите путь/имя .SAV-файла в кавычках или апострофах – для сохранения файла на диск, или укажите имя объявленного вами ранее массива (имя опционально можно взять в кавычки/апострофы) – для вывода в новый массив.

ПРИМЕР 3.

dataset declare medoids.

!KO\_pointclud matrix= var1 to var630 /groups= oldclus /id= objid /method= MEDOID /saveid= medoids.

* Расстояния до медоидов групп будут в новом рабочем безымянном массиве, а сами медоиды (их идентифицация переменной OBJID) появятся в массиве MEDOIDS, объявленном заранее. Поскольку медоид в группе – всегда постоянный объект, то переменные в массиве MEDOIDS, показывающие медоиды, будут константы.

**PRINT**

Выдача в окно результатов:

FULL - (тж. по умолчанию/незаданию) стандартный отчет.

NOID - не показывать информацию об идентификации/порядке рядов/столбцов во входящей матрице.

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами). Он не рассчитан на расщепленное состояние массива данных (SPLIT FILE). Макрос слушается команд, выбирающих наблюдения (SELECT IF, FILTER, USE), в том числе стоящих под командой TEMPORARY.

# МАКРОС !KO\_DENDRO: ДЕНДРОГРАММА

Version 3, Sep 2019 (Version 1, Mar 2014). Tested on SPSS Statistics 20, 22, 25.

!KO\_dendro sched= \* /\*История агломерации: внешний sav файл (в кавычках или апострофах) или звездочка

/\*(если этот файл открыт как рабочий массив)

/revers= /\*Позволять реверсии (YES, тж п/у) или нет (NO)

/label= 'd:\exercise\distmx.sav' strid /\*Опционально: использовать эти ярлыки для объектов

/\*(имя файла/массива и имя задающей ярлыки переменной в нем)

/bycolor= 'd:\exercise\objects.sav' group /\*Опционально: придать цвета объектам

/\*(имя файла/массива и имя задающей цвета переменной в нем)

/aspect= /\*Внешний вид дерева: FENCE (тж п/у) или GRAPE

/size= /\*Опционально, размер графика: два числа в пикселях, ширина и высота

/format= /\*Формат ярлыков объектов.

Минимум надо задать SCHED.

Макрос строит дендрограмму по агломеративной истории в окне результатов. Агломеративная история выдается макросом !KO\_HIECLU или !KO\_HIECLUEX и она такая же, какую выдает SPSS команда CLUSTER. !KO\_DENDRO сделан отдельным макросом потому, что пользователь, возможно, захочет составить свою агломеративную историю.

Если подкоманды LABEL и BYCOLOR не заданы, макрос также выводит в новый безымянный массив две переменные: код объекта (*ID*) и уровень его слития на дендрограмме (*MERGELEV*). Объекты идут сверху вниз так, как на дендрограмме. Этот массив потенциально ценен для пользователя. Если LABEL или BYCOLOR заданы, выдается пустой массив.

ПРИМЕР 1.

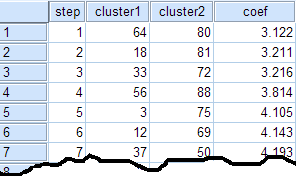
!KO\_hieclu matrix= var1 to var80 /method= COMPLETE /sched= \*.

!KO\_dendro sched= \*.

* Иерархическая кластеризация сохраняет агломеративную историю как безымянный массив.
* !KO\_DENDRO строит дендрограмму.

***Строение агломеративной истории***

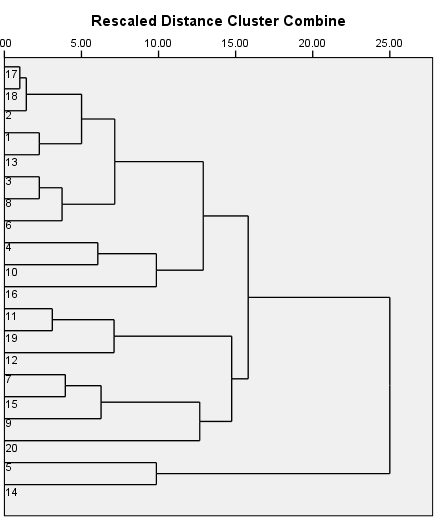
Этот SAV-файл должен иметь числовые переменные *CLUSTER1, CLUSTER2, COEF*. CLUSTER1 и CLUSTER2 содержат идентификационные коды кластеров; они отобразятся на дендрограмме как коды (ярлыки) объектов. COEF это коэффициент слияния – сходство или расстояние между сливающимися двумя кластерами cluster1 и cluster2 или соответствующий кумулятивный коэффициент (если коэффициент скумулирован, это должны быть расстояния, не сходства; см. п/к CUMUL макроса !KO\_HIECLU для большей информации).



Ряды истории соответствуют шагам агломерации. Имеется *правило*: когда объединяются два кластера cluster1=X и cluster2=Y, то объединенный кластер получает код X, т.е. код cluster1, под которым он фигурирует в дальнейшем. (Это означает, следовательно, что в переменной *CLUSTER2* коды не могут повторяться, а в *CLUSTER1* могут.)

***Вид дендрограммы***

На дендрограмме коэффициенты слияния ­– расстояния – перешкалированы в диапазон 1-25, как это делается в SPSS команде CLUSTER; сходства до этого обращаются в расстояния простым изменением знака. Вид дендрограммы:



Ярлыки объектов на дендрограмме это коды кластеров – значения столбцов *CLUSTER1* и *CLUSTER2* агломеративной истории. Вы можете заказать для них другие ярлыки, в том числе текстовые, подкомандой LABEL. Учтите, что зрительно хорошее размещение ярлыков может потребовать ручного редактирования построенного графика, для чего его надо открыть для редактирования, дважды щелкнув по нему.

***Подкоманды***

**SCHED**

Укажите путь к агломеративной истории. Если она рабочий массив, укажите звездочку. Если она внешний SAV-файл, укажите его путь/имя в кавычках или апострофах. Не указывайте имя массива в SPSS Statistics верс. ранее 26-й.

**REVERS**

Реверсией называется явление, когда кластер объединяется с другим на более низком уровне (величине расстояния), чем уровень, на котором он образовался. Некоторые методы иерархической кластеризации (центроидный, медианный, наименьшего прироста дисперсии) иногда могут давать реверсии на дендрограммах. Если вы хотите подавить отображение реверсий на дендрограмме, укажите REVERS=NO. Тогда реверсий не будет, но дендрограмма в местах реверсий не будет точно воспроизводить агломеративную историю.

**LABEL**

Опциональная подкоманда, которой можно подписать объекты другими ярлыками, нежели их дефолтные числовые коды, заимствованные из агломеративной истории. Укажите внешний SAV-файл (в кавычках/апострофах) или имя открытого массива, и после него – имя переменной в нем, содержащей ярлыки для объектов. Переменная может быть числовой или текстовой. Ярлыки произвольны, в том числе, разные объекты можно пометить одинаковыми ярлыками.

В упомянутом файле/массиве *должна быть* числовая переменная *ID*, значения которой – *неповторяющиеся*, *сортированные по возрастающей* коды объектов, какие использованы в переменных *CLUSTER1* и *CLUSTER2* агломеративной истории. (Полное совпадение не обязательно: например, в переменной *ID* может быть больше разных значений, чем присутствовало в агломеративной истории.)

**BYCOLOR**

Опциональная подкоманда, которой можно показать объекты разным цветом – например, по-разному закрасить объекты разных предзаданных групп. Укажите внешний SAV-файл (в кавычках/апострофах) или имя открытого массива, и после него – имя группирующей переменной в нем. Объекты, принадлежащие одной группе, получат на дендрограмме одинаковый цвет.

В упомянутом файле/массиве *должна быть* числовая переменная *ID*, значения которой – *неповторяющиеся*, *сортированные по возрастающей* коды объектов, какие использованы в переменных *CLUSTER1* и *CLUSTER2* агломеративной истории. (Полное совпадение не обязательно: например, в переменной *ID* может быть больше разных значений, чем присутствовало в агломеративной истории.)

ПРИМЕР 2.

!KO\_dendro sched= \* /label= 'd:\exercise\matrix.sav' caselab

/bycolor= 'd:\exercise\respondents.sav' agegroup.

* Строится дендрограмма по агломеративной истории, открытой как рабочий массив. Заказано показать респондентов, принадлежащих разным возрастным группам, разным цветом: использован файл, в котором имеется сортированный список респондентов *ID* (те же коды, что в агломеративной истории) и переменная *AGEGROUP*. Для подписей респондентов ярлыками переменной *CASELAB* использован в данном случае другой файл, но в нем также присутствует сортированная переменная *ID*.

**ASPECT**

Облик дендрограммы. По умолчанию/незаданию и ASPECT=FENCE дерево имеет вид забора, объекты «заземлены» на одной линии. При ASPECT=GRAPE объекты «подвешены» и дерево похоже на гроздь.

**SIZE**

Вы можете изменить дефолтный размер графика на какой хотите. Укажите два числа – ширина и высота, в пикселях.

**FORMAT**

Если ярлыки для подписи объектов являются числовыми кодами, как правило они целочисленные, и в этом случае подкоманда не требуется. Если ваши числовые коды не целочисленны, укажите формат с нужным количеством десятичных знаков, чтобы отобразить их все на дендрограмме. Например: FORMAT = f8.3.

***Особые режимы***

Макрос не рассчитан на особые режимы (взвешивание, расщепление и т.д.).

# МАКРОС !KO\_KMINI: НАЧАЛЬНЫЕ ЦЕНТРЫ ДЛЯ КЛАСТЕРИЗАЦИИ K-СРЕДНИХ

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

!KO\_kmini vars= v1 to v6 /\*Переменные (поименно и/или ч-з to)

/id= /\*Опционально: переменная-идентификатор наблюдений

/k= 5 /\*Создать это число кластерных центров

/method= GREP /\*Метод: RGC, RP, RUNFP, SIMFP, KMPP, GREP, WARD (см.)

/outfile= \* /\*Путь/имя файла для сохраниения центров, или звездочка

/print= YES /\*Распечать центры: YES (тж п/у) или NO.

Минимум надо задать VARS, K, METHOD.

Макрос создает разными методами инициальные центры для последующего кластерного анализа методом k-средних.

ПРИМЕР 1.

!KO\_kmini vars= v1 to v6 /k= 5 /method= RUNFP /outfile= 'd:\exercise\centers.sav'.

quick cluster v1 to v6 /criteria= cluster(5) mxiter(10) converge(0) /method= kmeans(noupdate)

/file= 'd:\exercise\centers.sav'.

* !KO\_KMINI набирает 5 наблюдений методом RUNFP на роль начальных кластерных центров и сохраняет их как файл.
* QUICK CLUSTER берет этот файл и делает кластеризацию k-средних (5 кластеров).

***Подкоманды***

**VARS**

Данные, наблюдения которых вы станете кластеризовать методом k-средних. Укажите поименно и/или ч-з to мерные переменные (имена до 8 байтов), составляющие эти данные. Макрос исключает пропуски списочно: если хотя бы в одной переменной VARS или в ID (если задано) наблюдение есть пропуск, оно не берется в анализ.

**ID**

Опционально укажите одну переменную-идентификатор наблюдений, если вы хотите, чтобы п/к PRINT показала, какие наблюдения были выбраны в качестве центров (в тех методах, которые делают центрами реальные наблюдения массива).

**K**

Сколько нужно центров (т.е. сколько кластеров вы станете выделять в кластерном анализе). Укажите целое положительное число больше 1 и меньше числа валидных наблюдений в данных VARS.

**METHOD**

Метод выбора или создания инициальных кластерных центров. Выберите:

RGC - *центроиды случайных подвыборок*. Данные разбиваются случайно на *k* неперекрывающихся составом наблюдений группы, и центроиды этих групп назначаются быть инициальными центрами. Т.о., центры вычисляются, а не выбираются из существующих наблюдений массива. Этот метод дает центры, близко расположенные друг к другу и к общему центроиду данных.

RP - *случайно выбранные точки*. *k* разных наблюдений данных случайно выбираются быть инициальными центрами.

RUNFP - *дальние точки (бегущий выбор)*. Первые *k* наблюдений принимаются за центры и затем в ходе пробежки остальных наблюдений массива прогрессивно делаются замены среди центров; цель замен – получить в итоге *k* наиболее отдаленных друг от друга точек в пространстве переменных. Эти точки (наблюдения), занимающие периферийные позиции в облаке данных, и есть инициальные центры. (Этот метод используется как дефолтный SPSS-командой QUICK CLUSTER. См. детали в SPSS Algorithms.)

SIMFP - *дальние точки (простой выбор)*. Первый центр выбирается как случайное наблюдение из массива. 2-й центр выбирается как наблюдение, максимально отдаленное от того центра. 3-й центр выбирается как наблюдение, максимально отдаленное от тех двух (от ближайшего из тех двух), - и так далее.

KMPP - *случайные дальние точки, или k-means++*. Первый центр выбирается как случайное наблюдение из массива. 2-й центр выбирается также случайно, но вероятность выбора наблюдения пропорциональна расстоянию (квадратному евклидовому) его до того центра. 3-й центр выбирается также случайно с вероятностью выбора, пропорциональной расстоянию до ближайшего из тех двух центров, - и так далее. (Arthur, D., Vassilvitskii, S.. K-means++: the advantages of careful seeding. // Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms. 2007., 1027–1035.)

GREP - *точки-представители групп*. Идея метода – набрать в качестве центров *k* наиболее представительных, «депутатских» наблюдений. Первым центром берется наблюдение, ближайшее к общему центроиду данных. Далее остальные центры набираются из точек данных так, что рассматривается каждая точка на предмет того, находится ли она ближе (и насколько, в терминах квадратного евклидового расстояния) к множеству точек, чем каждая из последних к какому-л. из существующих уже центров. Т.е. точка экзаменуется как кандидатура представлять некую группу точек, еще не достаточного хорошо представленную набранными центрами. Точка, наиболее представительная в этом плане, выбирается в качестве очередного центра. (Kaufman, L. Rousseeuw, P.J. Finding groups in data: an introduction to cluster analysis., 1990. См. также: Pena, J.M. et al. An empirical comparison of four initialization methods for the K-means algorithm // Pattern Recognition Lett. 20 (10), 1999, 1027-1040.)

WARD - *иерархическая кластеризация методом Уорда*. Делается кластерный анализ этим методом с выделением *k* кластеров; их центроиды и считаются *k* инициальными центрами. Т.о., центры вычисляются, а не выбираются из существующих наблюдений массива.

Методы RGC, RP, SIMFP, KMPP зависят от случайных чисел и от пуска к пуску могут менять свой результат. Используйте меню Transform – Random Number Generators или соответствующий синтаксис для регулирования зерна случайных чисел.

Метод RUNFP может быть чувствителен к порядку наблюдений в массиве; но метод GREP – не чувствителен (кроме случаев, когда в данных много одинаковых наблюдений, ties). Метод GREP может не набрать все k центров, если k велико относительно числа наблюдений в данных (n), особенно когда k>n/2. Макрос сообщит, если данные не позволяют этому методу набрать k центров. Метод GREP наиболее медленный метод, он вычисляет матрицу расстояний между всеми наблюдениями, поэтому не годится, если наблюдений многие десятки тысяч или миллионы.

Метод WARD может быть чувствителен к порядку наблюдений в массиве. Метод не годится, если наблюдений многие десятки тысяч или миллионы, т.к. он вычисляет матрицу расстояний между наблюдениями данных.

Если наблюдений слишком много, но вы желаете употребить методы такие как WARD или GREP, просто используйте случайную подвыборку наблюдений (командами SAMPLE или USE, например).

**OUTFILE**

Укажите в кавычках/апострофах путь/имя SAV-файла или имя заранее объявленного массива, для сохранения инициальных центров. Вы также можете указать звездочку – тогда центры будут выведены в новый безымянный массив данных. Не указывайте имя объявленного массива при METHOD=WARD.

Если вы не зададите подкоманду, центры не будут сохранены, но могут быть показаны (п/к PRINT).

**PRINT**

Распечатать в окно результатов инициальные центры: PRINT=YES (тж. по умолчанию). При PRINT=NO распечатки нет.

***Особые режимы***

Макрос не слушается взвешивания (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами). Он не рассчитан на расщепленное состояние массива данных (SPLIT FILE). Макрос слушается команд, выбирающих наблюдения (SELECT IF, FILTER, USE), в том числе стоящих под командой TEMPORARY.