***Clustering tendency***

SPSS macros by Kirill Orlov

kior@akado.ru, ttnphns@gmail.com

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

All rights reserved

*Тенденция кластерности*. Блок-диагонализация матрицы расстояний между объектами позволяет предварительно судить до кластерного анализа, есть ли в данных кластеры и сколько их.

*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.

# МАКРОС !KO\_BLOCKDIAG: БЛОК-ДИАГОНАЛИЗАЦИЯ МАТРИЦЫ РАССТОЯНИЙ

Version 1, Mar 2023. Tested on SPSS Statistics 22, 27, 30.

!KO\_blockdiag matrix= *VAR1 to VAR80* /\*Столбцы, образующие тело матрицы расстояний (можно ч-з "to")

/id= /\*Опционально: числовая переменная-идентификатор наблюдений (рядов)

/method= VAT /\*Метод: VAT (тж п/у) или MDS

/poster= YES /\*"Постеризация" после перестановки: YES или NO (тж п/у)

/plot= /\*Тепловаяю карта: GREY (тж п/у) или RGREY;

/\*после можно добавить слово LABEL; либо NONE

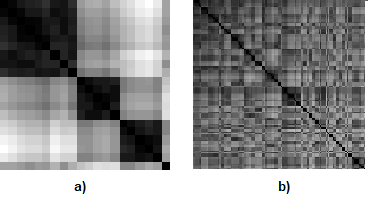
/bounds= /\*Пределы для тепловой шкалы на карте: AUTO (тж п/у),

/\*OBS или min max

/mds= /\*При METHOD=MDS: трансформация в MDS: SPLINE (тж п/у), ORDINAL, INTERVAL.

Минимум надо задать MATRIX.

Пусть есть матрица расстояний (различий) между объектами. Блок-диагонализация квадратной симметричной матрицы расстояний – это переупорядочивание ее рядов/столбцов – т.е. объектов – так, чтобы объекты с малыми расстояниями между ними разместились в матрице рядом друг с другом (т.е. их индексы стали близки). На тепловой карте такая переставленная матрица выглядит диагонально-блочной. Она тем более отчетливо, контрастно будет диагонально-блочна, чем сильнее присутствует тенденция к кластерности среди объектов, т.е. чем сильнее расстояния распадаются сами собой на «внутрикластерные» (малые) и «междукластерные» (большие). По тепловой карте, таким образом, можно приблизительно судить о наличии или отсутствии кластеров в данных, а также о числе кластеров, не делая самого кластерного анализа. Каждый кластер на тепловой карте после блок-диагонализации матрицы выглядит как блок на диагонали. Если кластеров нет, блок-диагонализация не покажет на тепловой карте ясных блоков.



**Рис. 1**. Тепловая карта блок-диагонализованной матрицы расстояний: a) с четкими кластерами в данных, b) без кластеров в данных.

Макрос делает блок-диагонализацию (переставляет ряды/столбцы) входящей матрицы расстояний, сохраняет полученную матрицу как новый безымянный массив данных и рисует тепловую карту.

**Алгоритм**

При METHOD=VAT макрос делает переупорядочивание рядов/столбцов матрицы расстояний алгоритмом VAT. VAT algorithm (“Visual Assessment of [Cluster] Tendency”) подробно описан в [1], также в [2,3]. Он тесно связан с алгоритмом построения минимального остовного дерева Прима взвешенного графа, и через это – имплицитно родствен иерархической кластеризации методом единичной связи, или ближайшего соседа [2]. Общей чертой этих алгоритмов является пошаговое наращивание остовного дерева/кластера/блока присоединением ближайших элементов.

Метод iVAT (“improved VAT”) [3] является надстройкой над VAT и по результату эквивалентен применению алгоритма Флойда–Уоршалла в варианте «определение легчайших проходов» к переупорядоченной матрице, выданной VAT. iVAT определенным образом заменяет некоторые расстояния в матрице другими ее расстояниями, уменьшая разнообразие расстояний в матрице. Идея iVAT проста: если две точки парно далеко друг от друга, но они опосредуются точечной цепью, все звенья (расстояния) в которой малы, то следует признать, что эти две точки «на самом деле» близки. Эффект iVAT-замены тот, что на тепловой карте (1) контраст между междукластерными и внутрикластерными расстояниями усилится, помогая визуально обнаружить кластеры-блоки; (2) повысится обнаруживаемость кластеров цепочечной структуры (включая сильно вытянутые, древовидные, кольцевидные). iVAT метод исполняется данным макросом, когда METHOD=VAT /POSTER=YES.

При METHOD=MDS макрос делает переупорядочивание рядов/столбцов матрицы расстояний, применяя многомерное шкалирование (SPSS-команда PROXSCAL) с взвешиванием расстояний. Это идея автора макроса (что не означает, что что она – новшество). Каждое расстояние *dij*, элемент матрицы, получает вес (важность) , где *Ri*– ранг значения *dij* в ряду *i*, *Rj* – ранг значения *dij* в ряду *j*, *R* – ранг значения *dij* в треугольнике матрицы. PROXSCAL исполняет картирование (ординацию) в пространстве размерности 1. Ряды/столбцы матрицы расстояний упорядочиваются вслед возрастанию полученных координат по этому измерению. Идея этого метода в том, чтобы принудить расстояния распределиться вдоль диагонали, отдавая в этом приоритет малым расстояниям. В методе VAT величина расстояния определяет очередность его подключения к остовному дереву. В методе MDS величина расстояния определяет его важность во влиянии на ординацию. В обоих методах конечным результатом становится то, что сгустки малых дистанций образуют блоки, нанизанные на диагональ матрицы.

METHOD=MDS /POSTER=YES применяет алгоритм Флойда–Уоршалла в варианте «определение легчайших проходов» к переупорядоченной матрице, выданной MDS-методом. Это производит такой же эффект, как iVAT после VAT.

Sources

1. Bezdek, J.C., Hathaway, R.J. VAT: a tool for visual assessment of (cluster) tendency // Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 – 2002 – Volume 3 – p. 2225–2230. [DOI:10.1109/IJCNN.2002.1007487]
2. Havens, T.C., Bezdek, J.C., Keller, J.M, Popescu, M., Huband, J.M. Is VAT really single linkage in disguise? // Annals of Mathematics and Artificial Intelligence. – 2009 – Vol. 55, article 237 [DOI 10.1007/s10472-009-9157-2]
3. Havens, T.C., Bezdek, J.C. An efficient formulation of the Improved Visual Assessment of Cluster Tendency (iVAT) Algorithm // IEEE Transactions on Knowledge and Data Engineering – 2012 – vol. 24, no. 5, p. 813-822 [DOI: 10.1109/TKDE.2011.33]

**Ограничения**

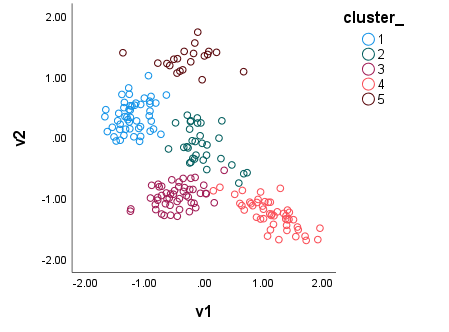
При METHOD=MDS составленная для анализа матрица – до 700 рядов/столбцов. При METHOD=VAT ограничений макрос не накладывает, но рекомендуемый размер матрицы – не более 1000, иначе тепловая карта будет рисоваться слишком долго. Вы всегда можете сделать анализ на случайной подвыборке наблюдений, если данные большие.

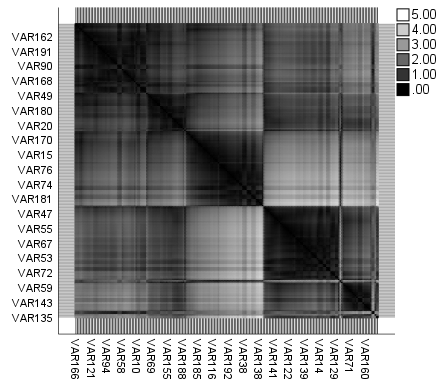
ПРИМЕР 1.

proximities v1 v2 /view= case /measure= seuclid /matrix= out(\*) /print= none.

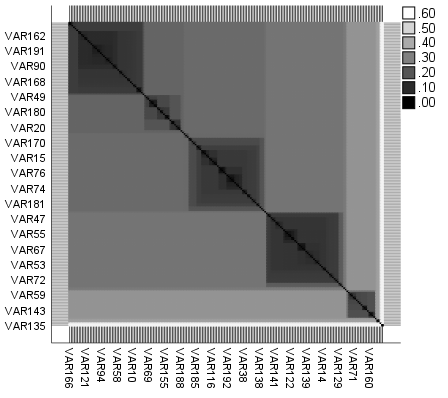
dataset name dist.

!KO\_blockdiag matrix= VAR1 to VAR199.





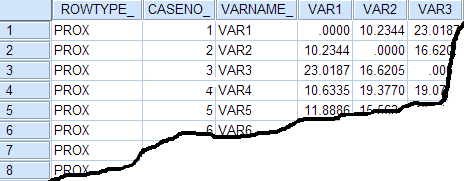
!KO\_blockdiag matrix= VAR1 to VAR199 /poster= YES.



* PROXIMITIES вычисляет из данных квадратные евклидовы расстояния между наблюдениями и сохраняет матрицу в новый массив, который называется *DIST*.
* Макрос исполняет блок-диагонализацию методом VAT и строит теплокарту. На ней можно распознать пять кластеров.
* Во втором пуске добавляется опция «постеризации». Пять кластеров проявились более отчетливо.

***Строение матрицы***

Массив данных должен быть матрицей попарных расстояний (различий, не 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 (различий). Вы можете указать все или только нужные столбцы и в произвольном порядке. Можно использовать “to” для задания диапазоном. Если у вас сходства, а не различия, то преобразуйте предварительно их в различия так, как сочтете нужным.

*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?*.

Т.к. ваши данные – 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.

ПРИМЕР 2.

temporary.

sample 0.2.

!KO\_blockdiag matrix= VAR1 to VAR100 /method= MDS /poster= YES .

* Команда SAMPLE временно (под TEMPORARY) отбирает случайно 20% рядов матрицы расстояний.
* Макрос берет столбцы *VAR1* до *VAR100*. Матрица, составленная макросом для анализа, состоит из рядов/столбцов, являющихся пересечением двух списков - отобранных рядов и отобранных столбцов.
* Макрос исполняет блок-диагонализацию методом MDS и «постеризацию» перед построением теплокарты.

**ID**

Опциональная числовая переменная-идентификатор наблюдений (объектов). Имя переменной до 8 байтов длиной. В переменной не может быть пропущенных значений.

**METHOD**

Укажите метод блок-диагонализации:

VAT - (тж. по умолчанию/незаданию) метод VAT.

MDS - многомерное шкалирование со взвешиванием расстояний исполняется SPSS-процедурой PROXSCAL. Она доступна в SPSS Statistics Professional Edition или в модуле Categories.

Оба метода часто дают очень похожие результаты.

**POSTER**

По умолчанию/незаданию и при POSTER=NO макрос делает только блок-диагонализацию, т.е. перестановку рядов/столбцов. При POSTER=YES он делает после этого «постеризацию». Тепловая карта при «постеризации» выглядит более контрастно, и обычно на ней легче увидеть кластеры (блоки). Кроме того, несколько повышается обнаруживаемость кластеров цепочечной структуры (включая сильно вытянутые, древовидные, кольцевидные). Иногда, впрочем, «постеризация» скрывает кластеры, которые близки друг к другу.

«Постеризация» при методе VAT известна как iVAT (“improved VAT”). «Постеризация» касается только теплокарты, она не влияет на сохраняемую матрицу. «Постеризация» не есть постеризация полученного изображения теплокарты, она есть контрастирование расстояний в матрице, по которой непосредственно строится теплокарта.

**PLOT**

Тепловая карта (heatmap) рисуется серо-полутоновой. При PLOT=GREY (тж. по умолчанию/незаданию) чем выше значение элемента, тем он ярче, а при PLOT=RGEY – наоборот, тем он темнее. После кл. слова вы можете добавить второе кл. слово LABEL, для оярлычения клеток значениями элементов. PLOT=NONE не рисует тепловую карту.

**BOUNDS**

Эта подкоманда не действует при PLOT=NONE. Задает границы для яркостной шкалы.

AUTO - (тж. по умолчанию/незаданию) позволить SPSS автоматически определить подходящие границы.

OBS - границы точно совпадают с наблюдаемыми минимальным и максимальным значениями в матрице.

*min* *max* - укажите границы вручную в виде двух чисел – минимума и максимума. Указывайте значения, в общем и целом сопоставимые со значениями матрицы.

Задание границ вручную означает, что яркостное отображение на рисунке вами фиксировано относительно величины содержащихся в матрице значений. Становится возможным сравнивать разные матрицы между собой по тону непосредственно.

**MDS**

Эта подкоманда действует при METHOD=MDS. Вы можете выбрать, метрическое или неметрическое многомерное шкалирование PROXSCAL использовать. Укажите INTERVAL (метрическое), ORDINAL (неметрическое) или SPLINE (промежуточное между метрическим и неметрическим). По умолчанию/незаданию, MDS=SPLINE.

***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.