***Neighbourhood chains***

SPSS macros by Kirill Orlov

kior@comtv.ru, ttnphns@gmail.com

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

All rights reserved

*Neighbourhood chains.* Out of data showing pairwise relationships within a set of objects there is extracted the information about which object is referred to “in the first place” or “most strongly” by each given object. This way, a trajectory of sequential references is being built. It is shown in form of a table (adjacency list) and a dendrogram.

# MACRO !KO\_NEITAB: TABLE OF NEIGHBOURHOODS

Version 1, June 2010. Tested on SPSS 13, 15, 17.

!KO\_neitab cols= *v1 to v10* /\*Matrix columns, all or needed, name-by-name or via “to”

/prox= DISSIM /\*Treat values as dissimilarities (DISSIM) or similarities (SIMIL)

/symm= /\*If asymmetric matrix needs be processed as symmetric:

/\*take lower riangle (LOWER), or upper triangle (UPPER),

/\*or average the two triangles (AVERAGE)

/diag= /\*Ignore diagonal elements (NOUSE, default) or allow them (USE)

/neighb= /\*What to choose (refer to) – nearest (NEAREST, default) or median (MEDIAN) neighbour

/tied= RANDOM /\*If rivals for “neighbour” role – choose the first one (FIRST default),

/\*choose randomly (RANDOM), choose one maximizing num. of neighbourhoods in the end (DILATAT),

/\*choose one minimizing num. of neighbourhoods in the end (CONTRACT)

/tiecrit= *.5* /\*Optionally: number >=0 for rivalry; if values differ

/\*no more than by this value they can become rivals for “neighbour” role

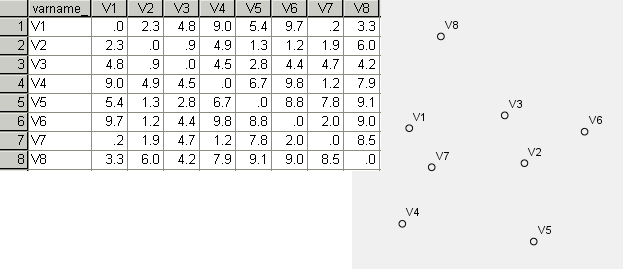
/source= NUM /\*If several matrices: numeric (NUM) or string (STR) is

/\*the identificatory variable *SOURCE*.

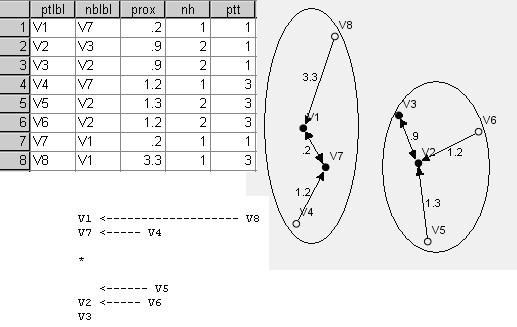
Minimal specification COLS, PROX.

The macro takes a square matrix of numbers which magnitudes show degree of some pairwise “affinity” within a set of objects, and modifies it into a “neighbourhood table” where, for each object, its neighbour is indicated (just one; it’s the nearest neighbour by default). A closed on itself chain of neighbours forms what is called here “heighbourhood”. In a neighbourhood, one can discern core objects, periferal and intermediate objects. In graph terminology, one could say, the macro takes a complete graph (a one all vertices of which are interconnected) in a form of adjacency matrix and reveals there one or several directed (i.e. with arrows) subgraphs where from each vertex just one arrow starts; those subgraphs are formatted by the macro as adjacency list.

Например, на следующем рисунке предствлена матрица сродства между объектами, или «точками», V1–V8. В данном примере это симметрическая матрица различий, но вообще матрица может быть и асимметрической и содержать любые данные (близости, оценки, ранжировки, вероятности). График справа, полученный многомерным шкалированием, приблизительно передает эти расстояния визуально на плоскости.



На рисунке ниже показано, что сделал макрос с этими данными. Получилась таблица из следующих колонок: точка (объект) по порядку; ее ближайший сосед; расстояние между ними; номер соседства; тип точки в соседстве (1=ядерная, 2=промежуточная, 3=периферийная). График справа (тот же, что был выше, но с нарисованными вручную стрелками) передает смысл этой таблицы визуально. Мы видим два *соседства* (понятие, имеющее смысл не столько кластера, сколько траектории), точки внутри которых цепочечно связаны одна с другой – своим ближайшим соседом. Пара точек в каждом соседстве, которые указывают *друг на друга* как на своего ближайшего соседа, составляют *ядро* соседства и называются ядерными. Точки, на которые никто не указывает, называются периферийными. Всякая точка, указывающая на иную точку чем указывающая на нее, именуется промежуточной. Основной принцип и задача, которую исполняет макрос: каждый объект должен отослать к одному и только одному соседу, в то время как к объекту может быть и несколько отсылок (более одной стрелки могут указывать на точку). Макрос не рисует графиков, а только **выдает таблицу соседств** в качестве нового массива данных. Однако другой макрос, [!KO\_NEIPLOT](#_МАКРОС_!NEIPLOT:_ГРАФ_СОСЕДСТВ), из такой таблицы рисует в текстовом режиме схему наподобие дендрограммы, показанную на рисунке внизу. Видно, что она схематически передает собой примерно ту же информацию, что и график с пририсованными стрелками.



!KO\_NEITAB имеет ряд опций. Вы можете ввести единственную матрицу (она же «источник») или множество источников сразу – для каждого будет выдана своя таблица соседств. Источник может быть матрицей симметрической или асимметрической и содержать сходства или различия. В качестве понятия «сосед» можно выбрать ближайшего соседа, дальнего соседа или срединного соседа. Вы можете учитывать либо игнорировать диагональные элементы, т.е. сродство объекта к самому себе. В случае, если на роль соседа претендует более одного объекта, можно предпочесть ту или иную стратегию выбора из них.

ПРИМЕР 1. Кто пойдет к кому за солью (мысленный эксперимент). Имеются данные об удаленности всех домов в деревне друг от друга. Если у хозяйки дома кончилась соль во время варки супа, то в какой дом она вероятно побежит занять щепотку? Требуется выяснить это для каждого из домохозяйств.

!KO\_neitab cols= a to n /prox= DISSIM.

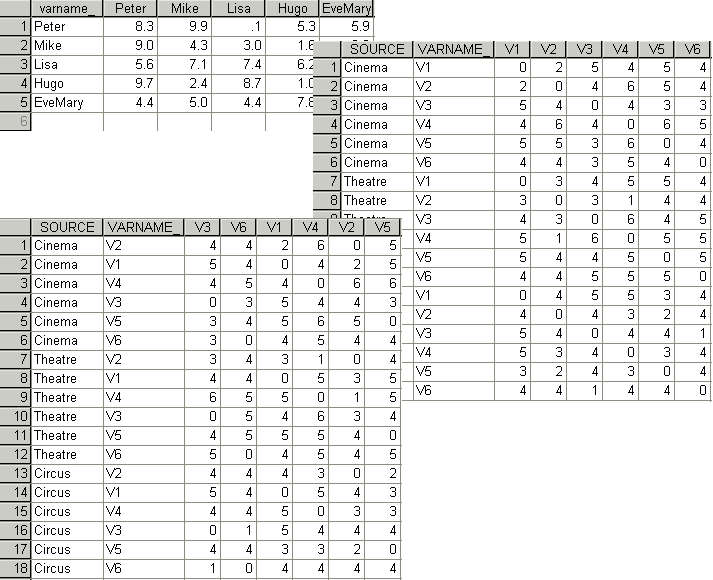
* Квадратная симметрическая матрица расстояний между домами находится в переменых с A до N. Макросу указано, что значения в матрице надо понимать как различия (а не сходства). Макрос выдаст «таблицу соседств», в которой для каждого дома указан ближайший к нему дом - именно в него побежит хозяйка за срочной щепоткой соли. Может оказаться, что данные распадутся на ряд «соседств». Соседство это совокупность домов, связанных между собой тем, что кто-то у кого-то попросит соль. Между соседствами такой связи нет, из одного соседства не обратится за солью в другое соседство.

***Строение входящих данных***

Данные представляют собой одну или более квадратных матриц размером не менее 3x3, содержащих валидные числовые значения (пропущенных данных должно не быть). Имена переменных, являющих собой столбцы матрицы, должны быть не длиннее *8 символов*. Должна присутствовать текстовая переменная VARNAME\_, именующая ряды матрицы теми же именами, что у переменных-столбцов, причем написаны они должны быть *в том же регистре* букв, что и имена переменных-столбцов[[1]](#footnote-1). Текстовая переменная ROWTYPE\_, обычная для оформления матричных данных в SPSS, необязательна.

Если матриц несколько, все они должны быть одного размера и располагаться одна под другой; и должна присутствовать переменная SOURCE, числовая или текстовая (значения - до *8 символов*), служащая идентификатором матриц, или источников.

Хотя квадратные матрицы, столбцы и ряды которых есть один и тот же набор объектов, обыкновенно делают диагонализованными – т.е. со столбцами и рядами, идущими в одном итом же порядке, - данный макрос этого не требует и допускает, чтобы имена столбцов слева направо шли не так, как имена рядов сверху вниз. Однако макрос требует, когда матриц несколько, чтобы порядок рядов у них всех был *один и тот же*[[2]](#footnote-2). Вот примеры входящих:

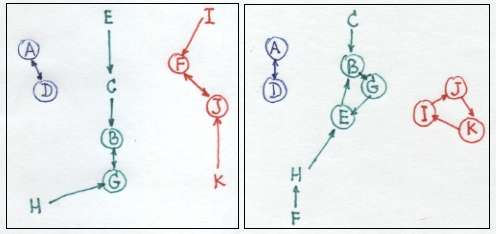


Сверху вниз:

* Единственный источник
* Несколько источников; они должны располагаться один под другим
* То же, но с неодинаковым порядком в рядах и столбцах. Допустимо, лишь бы порядок рядов у всех источников был одинаков

***Особенности результатов при разных входящих***

* При симметрической матрице и NEIGHB=NEAREST стрелки в соседстве укорачиваются с приближением к ядру, а ядро состоит всегда из двух точек (если DIAG=USE, то может состоять и из одной точки).
* Во всех остальных случаях стрелки в соседстве могут укорачиваться или удлиняться с приближением к ядру, а ядро может состоять и более чем из двух точек, они образуют тогда кольцо (если DIAG=USE, то ядро может состоять и из одной точки).



На рисунке – характерные графы соседств (ядерные точки соседства обведены): слева для симметрического источника и NEIGHB=NEAREST, справа для асимметрического источника либо NEIGHB=MEDIAN.

***Подкоманды***

**COLS**

Укажите поименно и/или ч-з *to* все или только нужные вам столбцы источника(ов). Вы можете написать имена в любом порядке, в этом порядке точкам (объектам) и будут присвоены порядковые номера. Порядок может иногда влиять на результаты (см. п/к TIED).

**SOURCE**

Эта подкоманда задается, если источников (матриц) несколько. Укажите SOURCE=NUM, если переменная-идентификатор источников *SOURCE* числовая. Укажите SOURCE=STR, если она текстовая (значения – не длиннее 8 символов). Если источников несколько, а вы умолчите или не зададите подкоманду, то макрос возьмет в обработку только одну, первую матрицу.

**PROX**

В этой обязательной подкоманде вы должны указать, как макросу рассматривать входящие данные – как «сходства» или как «различия». Если большее число отвечает большему сродству или большей важности или высокой оценке (и следовательно более короткой стрелке на графе), выберите «сходства»: PROX=SIMIL. Если большее число отвечает меньшему сродству или меньшей важности или низкой оценке (и следовательно более длинной стрелке на графе), то выберите «различия»: PROX=DISSIM. Значения в матрице могут быть любыми числами, в том числе отрицательными.

**SYMM**

По умолчанию макрос принимает и обрабатывает матрицу как она есть, симметрическая или асимметрическая. При этом за точку, от которой *исходит* стрелка (отсылка), он берет *ряд* матрицы, а за точку, на которую *укажет* стрелка, он берет *столбец* матрицы. Другими словами, для каждого объекта-ряда объект-сосед ищется среди столбцов. Это обстоятельство пользователю надо учитывать, готовя асимметрические данные: «субъекты» (например, те кто оценивает), должны быть рядами матрицы, а «мишени» (например, те кого оценивают) должны быть столбцами матрицы. Для симметрических матриц это обстоятельство не играет роли, т.к. в таких матрицах ряды и столбцы эквивалентны.

Опциональная подкоманда SYMM позволяет затребовать обработку матрицы как симметрической, даже если она асимметрическая. Макрос внутренне превратит асимметрическую матрицу в симметрическую, прежде чем анализировать ее. Есть 3 способа этого превращения на выбор. При SYMM=LOWER макрос использует нижний (поддиагональный) треугольник матрицы. При SYMM=UPPER он использует верхний (наддиагональный) треугольник. При SYMM=AVERAGE макрос усредняет нижний треугольник с верхним.

**DIAG**

По умолчанию/незаданию и при DIAG=NOUSE диагональные элементы матрицы, т.е. отношения объектов с самими собой, игнорируются макросом. Но в некоторых исследованиях бывает нужно учитывать и отношение объекта к самому себе тоже. Например, если членов коллектива просили оценить, в какой степени каждый из коллектива влияет на принятие общих решений, то очевидно следует регистрировать и оценку респондентом своего собственного влияния, наряду с оценками влияния каждого из остальных членов. DIAG=USE заставляет макрос анализировать диагональные элементы матрицы вместе с остальными элементами.

**NEIGHB**

Укажите, что понимать под тем единственным «соседом», кого каждая точка должна отыскать себе и к кому отослать.

NEAREST -(тж по умолчанию/незаданию) Ближайший сосед. Каждая точка (ряд источника) отыскивает максимальное (при PROX=SIMIL) или минимальное (при PROX=DISSIM) значение в своем ряду и связывается с точкой-столбцом, в котором лежит это значение.

MEDIAN -Срединный сосед. Каждая точка (ряд источника) отыскивает медианное значение в своем ряду и связывается с точкой-столбцом, в котором лежит это значение. Другими словами, точка выбирает себе в соседи точку, которая в наибольшей степени «и не близка, и не далека» от нее. Если число столбцов (при DIAG=NOUSE число столбцов-1) матрицы четное, то на роль медианного значения, естественно, претендуют два значения: макрос берет из них в таком случае меньшее.

Для того чтобы «соседом» был дальний сосед, укажите NEIGHB=NEAREST и задайте п/к PROX противоположным образом. Для того чтобы срединный сосед отвечал большему, а не меньшему из двух медианных значений, когда число столбцов таково, что медианных значений два, укажите NEIGHB=MEDIAN и задайте п/к PROX противоположным образом.

**TIED**

Если в ряду источника существуют одинаковые значения (или если вы принимаете близкие, хотя неодинаковые, значения за одинаковые с помощью подкоманды TIECRIT), то может возникнуть ситуация, когда на звание соседа по отношению к данной точке претендует более одной точки. Тогда поднимается вопрос, какую из нескольких претендующих точек выбрать в качестве соседа. П/к TIED предлагает стратегию выбора.

FIRST -(тж по умолчанию/незаданию) Соседом выбирается первая встреченная из претендующих точек.

RANDOM - Соседом выбирается точка из претендующих точек случайным образом.

DILATAT - Макрос выбирает соседа в попытках максимизировать в итоге число соседств и сделать ветви в соседствах более короткими. Это "расширяющий пространство" метод.

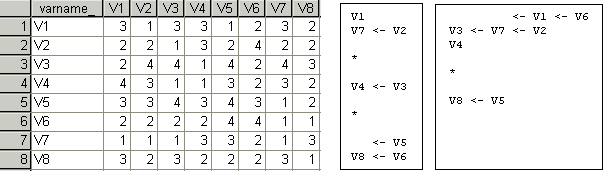
CONTRACT - Макрос выбирает соседа в попытках минимизировать в итоге число соседств и сделать ветви в соседствах более длинными. Это "сужающий пространство" метод.

Вариант FIRST испытывает сильное влияние на результат того, в каком порядке вы указали столбцы в п/к COLS. Это очевидно: если в столбцах V1 и V2 встретятся равные значения, оба претендующие быть выбранными в соседи, выбран будет столбец, который вы укажете в п/к COLS раньше. Вариант RANDOM зависит от датчика случайных чисел и будет давать разный результат от пуска к пуску макроса, если только вы не запретите зерну случайных чисел меняться; результаты при этом варианте также испытывают влияние порядка, в каком столбцы написаны в COLS. Варианты DILATAT и CONTRACT менее чувствительны к вышеупомянутому порядку в COLS, и тем не менее не совсем независимы от него, в общем случае. DILATAT и CONTRACT могут давать сильно различающиеся результаты (если, напомним, в рядах источника имеются ties, т.е. одинаковые значения), т.к. эти методы преследуют противоположные цели. Заметим, однако, что ни тот ни другой метод не гарантируют достижения самого «расширенного» и самого «суженного» из всех возможных пространств, - эти два алгоритма лишь пытаются максимизировать указанные свойства у графов.

ПРИМЕР 2. «Расширенный» и «суженный» версии графов.

!KO\_neitab cols= v1 to v8 /prox= DISSIM /symm= LOWER /DIAG= USE /tied= DILATAT.

!KO\_neitab cols= v1 to v8 /prox= DISSIM /symm= LOWER /DIAG= USE /tied= CONTRACT.



* Матрица асимметрическая, однако исследователь решил обработать ее как симметрическую на основе ее нижнего треугольника (SYMM=LOWER); он также решил не игнорировать диагональные значения (DIAG=USE). Данные в матрице дискретны, в рядах имеются повторяющиеся значения, поэтому вполне вероятно, что за роли соседей возникнет конкуренция; в этих условиях на результатах сказывается п/к TIED. В первом пуске исследователь заказал «расширение пространства», а во втором пуске – «сужение пространства».
* Результаты показаны справа – это дендрограммы, построенные макросом !KO\_NEIPLOT из полученных таблиц соседства. В одном случае получилось 3 соседства с короткими ветвями. В другом случае получилось только 2 соседства, в одном из которых к ядру (V3-V4) ведет длинная ветвь: в ней целых 2 промежуточных точки (V1 и V7). Обратите внимание, что некоторые ядра (V4 и V8 слева; V8 справа) состоят из единственной точки: такое возможно лишь при DIAG=USE.

**TIECRIT**

Эта подкоманда позволяет относиться к близким, но неодинаковым значениям во входящих данных как к одинаковым. По умолчанию/незаданию TIECRIT=0, что значит, что любые сколько-нибудь неодинаковые значения в ряду источника принимаются как есть за неодинаковые (а одинаковые, естественно, принимаются за одинаковые). Если вы зададите число больше нуля (задать можно любое число не меньше нуля), то значения, различающиеся *не больше чем на это число*, макрос сочтет одинаковыми (хотя они всего лишь близки), и следовательно, потенциально конкурирующими на звание «соседа».

ПРИМЕР 3.

Группу из 8 респондентов просили оценить свою личную привязанность друг к другу по 10-балльной шкале. Собранные данные – квадратная асимметрическая матрица, где каждый респондент есть ряд, а его оценки, данные остальным респондентам, располагаются в этом ряду.

!KO\_neitab cols= resp1 to resp8 /prox= SIMIL /tiecrit= 1 /tied= RANDOM.

* По умолчанию соседом респондента считается его ближайший сосед, - т.е. тот в данном случае, кого он оценил *выше* остальных, поскольку PROX=SIMIL. Диагональный элемент (оценка респондентом самого себя) по умолчанию игнорируется, поэтому неважно, какие числа набил исследователь в такие ячейки.
* TIECRIT=1, и значит, если максимальная оценка, выставленная неким респондентом, равна, допустим, 9 баллам, то тот, кому он выставил 8 баллов (т.е. на 1 меньше), если такой есть, оспорит право на соседство у того, кому выставлено 9 баллов, и следовательно между ними придется выбирать. Согласно TIED=RANDOM, выбран из них один будет случайным образом.

### Особые режимы

Макрос игнорирует взвешивание файла и не рассчитан на расщепленность файла.

# MACRO !KO\_NEIPLOT: GRAPH OF NEIGHBOURHOODS

Version 1, June 2010. Tested on SPSS 13, 15, 17.

!KO\_neiplot outfile= *'d:\exercise\nei.sps'* /\*Text file with the graphs (dendrograms) to save

/print= LABELS /\*Map points as numbers (NUMBERS, default) or labels (LABELS)

/width= *3* /\*Maximal width of number/label of a point: from 1 to 8 (default=8)

/field= *255* /\*Field width in the text file: from 40 to 255 (default=255)

/prox= /\*Optionally for scaling branches: values of variable *PROX* are

/\*dissimilarities (DISSIM) or similarities (SIMIL)

/source= STR /\*If several input tables: numeric (NUM) or string (STR) is

/\*the identificatory variable *SOURCE*

/cos= LINK /\*Optionally, if SOURCE=STR: produce matrix of cosines between tables,

/\*cosines being based on similarity of chain’s links (LINK) or of whole chain (CHAIN).

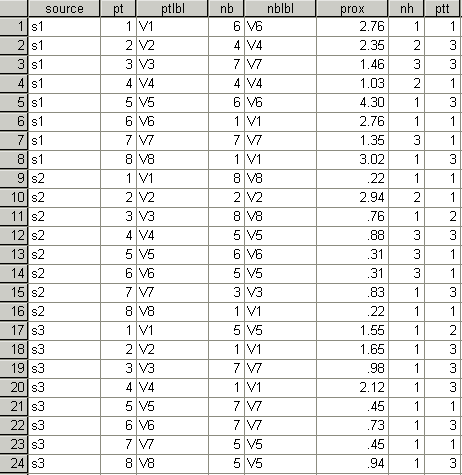
Minimal specification OUTFILE.

The macro draws (in a text format) a tree-like directed graph (points and arrows between points) thus visualizing information of one or more “neighbourhood tables” output by [!KO\_NEITAB](#_МАКРОС_!NEITAB:_ТАБЛИЦА_СОСЕДСТВ) macro. !KO\_NEIPLOT is made separate macro so that user be able to input neighbourhood tables created by hand as well. The dendrograms are sent not to Output window but to an ASCII file to hard disc. A dendrogram does not show relations between points that form core of a neighbourhood; only trajectories leading towards the core – the branches – are tiled with arrows.

Aside from saving dendrograms in an outer file the macro can, if there are more than one source (i.e. neighbourhood tables) compute and save as a new working dataset the square symmetric matrix of similarities (cosine) between the sources. This is a matrix of likeness between the sources regarding the trajectories, or graph structure, and it could be used for segmentation of assemblage of sources by means of cluster analysis or multidimensional scaling[[3]](#footnote-3).

***Строение входящих данных***

Данные должны представлять собой одну или несколько подшитых друг под другом «таблиц соседств». Таблица соседств показывает траекторию отсылок в наборе точек (неких объектов) – какая точка к какой отсылает (= имеет своим соседом). Точка может отсылать только к одной точке. Макрос !KO\_NEITAB выдает такие таблицы готовыми к введению в !KO\_NEIPLOT. Такие данные состоят из следующих переменных:

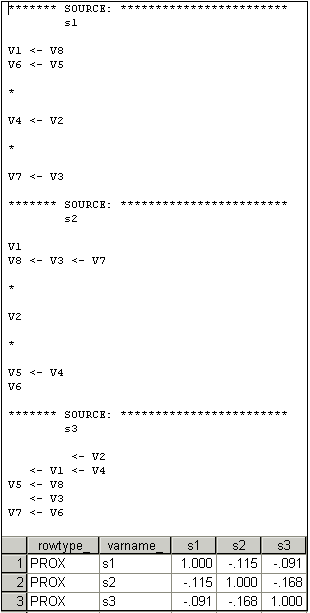


* *SOURCE* – числовая или текстовая переменная, обязательная в случае, если таблиц соседств, т.е. источников, несколько. Ширина числового или текстового значения – до *8 символов*. Эта переменная – идентификатор источников.
* *PT* – номер точки в источнике. Эта числовая переменная обязательна. Значения должны быть номерами 1, 2, …, число-точек-в-таблице и *идти сверху вниз по возрастанию*.
* *NB* – номер точки-соседа. Эта числовая переменная обязательна. Значения должны быть из набора номеров 1, 2, …, число-точек-в-таблице.
* *PTLBL* – ярлык точки в источнике. Эта текстовая переменная нужна, если вы задаете п/к PRINT=LABELS. *Во всех источниках ярлыки должны быть одни и те же*[[4]](#footnote-4). Значения - до *8 символов*.
* *NBLBL* – ярлык точки-соседа.
* *PROX* – величина сродства между точкой и ее соседом. Эта числовая переменная нужна, если вы задаете п/к PROX.
* *NH* – номер соседства в источнике. (Соседство это совокупность, цепь точек, связанных вместе отношениями точка–сосед, точка–сосед…) Эта числовая переменная обязательна.
* *PTT* – тип точки в соседстве: 1=ядерная, 2=промежуточная, 3=периферийная (ядерные точки имеют соседями друг друга, непосредственно или посредством кольца; периферийная точка не является соседом никому; промежуточная точка это неядерная точка, имеющая соседом другую точку чем та, которая имеет соседом ее). Эта числовая переменная обязательна.

Пропуски в данных недопустимы.

ПРИМЕР 1. Дендрограммы, соответствующие показанным на рисунке выше данным.

!KO\_neiplot outfile= 'd:\exercise\dendro.sps' /source= STR /print= LABELS /width= 2 /cos= CHAIN.



* Дендрограммы сохраняются как файл 'd:\exercise\dendro.sps'.
* Источников несколько и идентифицирующая их переменная SOURCE текстовая.
* Кроме создания дендрограмм, заказано подсчитать матрицу косинусов между источниками.

***Строение дендрограмм***

* Каждая дендрограмма отвечает одному соседству. Дендрограммы, или соседства, если их более одного в источнике, разделены звездочкой (\*).
* Дендрограмма состоит из ядра и одной или нескольких ведущих к ядру ветвей. Иногда бывает только ядро. Ядро находится на дендрограмме слева. Стрелки показываются только между точками внутри ветви и между точкой ветви и точкой ядра.
* Ядро это ядерные точки соседства. Когда их несколько, они размещаются одна под другой. Макрос не показывает стрелок между ядерными точками, однако порядок этих точек сверху вниз отвечает порядку отсылок между ними.
* Ветви состоят из промежуточных и периферийных точек. Последние находятся на правом фланге дендрограммы: от периферийной точки идет стрелка, но к ней не идет стрелка. К промежуточной точке идет стрелка и от нее идет стрелка; промежуточные точки могут и отсутствовать в ветви.
* Стрелка, якобы ни на что не указывающая, указывает на ту же точку, на какую указывает стрелка, находящаяся непосредственно под данной стрелкой.
* Если задать п/к PROX, стрелки будут разной длины – длина стрелки тогда отражает степень близости между точками.

***Подкоманды***

**OUTFILE**

Укажите, в кавычках или апострофах, путь/имя текстового файла для сохранения в нем рисованных графов-дендрограмм. Можете использовать любое подходящее для простого текстового файла расширение (TXT, DAT, SPS и проч.).

**SOURCE**

Эта подкоманда задается, если источников (таблиц соседств) несколько. Укажите SOURCE=NUM, если переменная-идентификатор источников *SOURCE* числовая. Укажите SOURCE=STR, если она текстовая (значения – не длиннее 8 символов). Если источников несколько, а вы умолчите или не зададите подкоманду, то макрос возьмет в обработку только одну, первую таблицу.

**PRINT**

По умолчанию/незаданию и при PRINT=NUMBERS точки на дендрограммах нумеруются (номера берутся из переменных *PT* и *NB*). Если у вас есть переменные *PTLBL* и *NBLBL* с ярлыками точек, вы можете указать PRINT=LABELS, чтобы точки обозначались этими ярлыками.

**WIDTH**

По умолчанию/незаданию этой подкоманды точке на дендрограмме дается место в ширину на 8 символов. Если реально номера или ярлыки точек, содержащиеся в ваших переменных *PT*/*NB* или *PTLBL*/*NBLBL* короче этого предела, вам имеет смысл указать реальную максимальную ширину здесь, чтобы дендрограммы не получились зрительно разреженными. Например, WIDTH=3, если номер или ярлык точки не превышает 3 символов.

**FIELD**

По умолчанию/незаданию этой подкоманды в сохраняемом текстовом файле используется поле максимальной ширины в 255 столбцов. Как правило это больше чем занимают дендрограммы и ведет к неоправданно большому размеру файла на жестком диске. Вы можете задать в FIELD любую ширину поля от 40 до 255. Например, вы можете прежде убедиться, что дендрограммы уложились в первые столько-то столбцов, а затем перезапустить макрос, установив FIELD= это число столбцов.

**PROX**

По умолчанию/незаданию этой подкоманды между соседними точками ветви дендрограммы укладывается стрелка единичной длины (🡨), т.е. показывается только, какая точка отсылает к какой. PROX позволяет варьировать длину стрелки в зависимости от количественной близости между соседними точками (п/к PROX требует наличия переменной *PROX* во входящих данных). Укажите PROX=DISSIM, если значения в *PROX* являются различиями; большее значение будет передано более длинной стрелкой. Укажите PROX=SIMIL, если значения в *PROX* являются сходствами; большее значение будет передано более короткой стрелкой. Макрос внутренне линейно перешкалирует все значения в переменной *PROX*, не относящиеся к ядерным точкам, в диапазон 1–25 и показывает на дендрограммах, соответственно, стрелками длиной от 1 до 25 черточек[[5]](#footnote-5).

**COS**

Эта подкоманда действует, если SOURCE=STR. Во избежание ошибки, значения в переменной *SOURCE* должны начинаться с буквы, чтобы они могли стать именами переменных. Опциональная п/к COS вычисляет и выводит в новое рабочее окно данных кв. симм. матрицу косинусов между источниками. Косинус – мера, подобная коэффициенту корреляции и колеблется между -1 и +1. Положительные значения говорят о преобладании сходства траекторий в отношении того, какова последовательность точек в них. Близкие к 0 значения говорят об отсутствии какого-л. сходства, а отрицательные значения говорят о преобладании противоположности, противонаправленности сравниваемых последовательностей. Косинусы учитывают только порядок точек в дендрограммах – они не учитывают длину стрелок между точками, т.е. значения переменной *PROX*. Выберите, на основе чего вычислять косинус:

LINK -Единицей сравнения является звено цепи, т.е. пара точка – ее сосед.

CHAIN -Единицей сравнения является всякая пара точек в цепи – как соседние точки, так и отдаленные точки одной цепи.

Увидим различия этих двух подходов на примере. Пусть надо сравнить граф I: A 🡪 B 🡪 C 🡪 **D** (точка D выделена жирным, чтобы показать, что она отсылает далее к самой себе, образуя одноточечное ядро соседства) с каждым из следующих:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | *Косинус* | |
| COS=LINK | COS=CHAIN |
| II | B 🡪 A 🡪 C 🡪 **D** | .143 | .692 |
| III | A 🡪 C 🡪 B 🡪 **D** | -.143 | .692 |
| IV | D 🡪 B 🡪 C 🡪 **A** | .286 | -.615 |
| V | D 🡪 C 🡪 B 🡪 **A** | -.857 | -.923 |

При COS=LINK все внимание поглощено порядком в соседних точках. В цепи II – по сравнению с цепью I – имеется одна инверсия (B🡪A), полтора совпадения (C🡪D, **D**) и одно отличие (A🡪C). А в цепи III имеется инверсия (C🡪B), два отличия (A🡪C, B🡪D), ½ совпадения (**D**). Очевидно, что III дальше от I чем II от I, в том смысле что в III уже заметна противонаправленность по отношению к I (косинус получился отрицательным). По сравнению с этим, IV имеет с I одно совпадение (B🡪C) и 2 ½ отличия, инверсий нет; в итоге IV сильнее похоже на I, чем II. Таким образом, при LINK косинус сильнее всего реагирует на перестановки между соседями. Используйте этот метод, если для вас X🡪Y vs Y🡪X есть более значимая или резкая неодинаковость, чем X🡪…🡪Y vs Y🡪…🡪X.

При COS=CHAIN учитывается порядок всех пар точек в цепи, как бы они далеко ни отстояли друг от друга. По сравнению с I в II, например, имеем в паре A-B инверсию, в A-C совпадение, в A-D совпадение, в B-C совпадение, в B-D совпадение, в C-D совпадение, кроме того, совпадает и ядерная точка (**D**), что идет за ½ совпадения. Итого 5 ½ совпадений и 1 инверсия. Косинус = 0.692. В отношениях с I такая же «статистика» у III, поэтому тот же косинус. Зато у IV, которое отличается от I перестановкой позиций далекоотстоящих друг от друга точек A и D, эта перестановка привела к значительному изменению порядка в цепи: A-B – инверсия, A-C – инверсия, A-D – инверсия, B-C – совпадение, B-D – инверсия, C-D – инверсия; плюс изменилась ядерная точка, что идет за ½ отличия. Итого 1 совпадение, 5 инверсий и ½ отличия. Косинус = -0.615. Таким образом, при CHAIN косинус сильнее всего реагирует на перестановки между отдаленными точками одной цепи. Используйте этот метод, если для вас X🡪…🡪Y vs Y🡪…🡪X есть более значимая или резкая неодинаковость, чем X🡪Y vs Y🡪X.

LINK и CHAIN различаются также отношением к ядерным точкам соседства, когда ядерных точек больше 2-х, т.е. когда ядро кольцевое. LINK учитывает последовательность точек в ядре, точно так же как везде: для LINK ядро X🡪Y🡪Z🡪X и ядро X🡪Z🡪Y🡪X это неодинаковые ядра. CHAIN не учитывает последовательности точек внутри ядра: для CHAIN оба только что приведенных ядра одинаковы.

Если вам трудно предпочесть один или второй метод получения косинуса, можете воспользоваться обоими и полученные две матрицы усреднить в одну.

### Особые режимы

Макрос игнорирует взвешивание файла и не рассчитан на расщепленность файла.

1. В SPSS ранее 12-й версии имена в VARNAME\_ пишите только заглавными буквами. [↑](#footnote-ref-1)
2. В действительности макрос смотрит на имена в VARNAME\_ только у первой матрицы и делает допущение, что их порядок у остальных матриц такой же как у первой, не проверяя этого. Фактически, вы можете оставить ячейки VARNAME\_ напротив всех матриц кроме 1-й пустыми, и это не скажется на результатах. [↑](#footnote-ref-2)
3. Для пересчета в расстояние угловых мер сходства (типа r Пирсона или косинуса, варьирующих между -1 и 1) обычно пользуются формулой sqrt[2\*(1-сходство)]. [↑](#footnote-ref-3)
4. Макрос не проверяет это, а просто использует во всех графах ярлыки, взятые из первого источника. [↑](#footnote-ref-4)
5. Диапазон 1–25 относится ко всему массиву, а не к каждому отдельному источнику. Вы можете вместо числа 25 задать свое число в качестве длины наидлиннейшей стрелки. Для этого в первой команде макроса, DEFINE, в SCALE=… !DEFAULT(25) замените 25 на другое число. [↑](#footnote-ref-5)