

Multidimensional scaling (MDS)

MDS: a medical example and (fabricated) data

There are fourteen people with multiple sclerosis (MS) who regularly attend an occupational therapy group. We know that there are numerous possible combinations of MS symptoms and some of the patients have been complaining that there is a tendency for occupational therapy staff to treat some of them inappropriately as rather distinct groups with similar needs on the basis of a few obvious symptoms. A pilot study is carried out to explore the extent to which this complaint appears justified, and to see how far staff perceptions of similarities are modified when they are given an opportunity to study the responses of the MS patients to a checklist of MS symptoms. Prior to the checklist intervention, mean staff ratings of similarities between MS pairs are obtained on a scale from 1 to 10. We start with a multi-dimensional scaling (MDS) analysis seeking two dimensions on a distance measure derived from these initial staff ratings of similarities between patient pairs.

The initial similarity ratings are shown in Table 11.4(a). Shortly after producing these, the staff are given time to study the symptom checklists completed by the patients. They are then asked once again to rate the similarity between each pair of patients. These similarities are shown in Table 11.4(b).

Table 11.4

Staff ratings of similarities between pairs of patients

(a) Before studying the symptom checklist (mds.similaritya.sav)

	Adam	Ana	Chila	Colin	Dee	Dom	Ewan	Holly	Jim	Krish	Laura	Mira	Pete	Sanjiv
Adam														
Ana	5													
Chila	4	10												
Colin	7	7	7											

Dee	4	7	8	6										
Dom	6	7	6	6	4									
Ewan	8	7	6	9	5	8								
Holly	3	7	8	6	9	4	5							
Jim	9	5	4	8	4	5	7	4						
Krish	7	6	5	6	3	8	8	3	6					
Laura	3	7	8	6	9	4	5	9	4	3				
Mira	3	8	9	5	7	5	5	7	3	4	7			
Pete	8	6	6	9	5	7	8	4	8	7	5	5		
Sanjiv	6	6	5	6	4	10	8	3	5	9	3	5	7	

(b) After studying the symptom checklist (mds.similarityb.sav)

	Adam	Ana	Chila	Colin	Dee	Dom	Ewan	Holly	Jim	Krish	Laura	Mira	Pete	Sanjiv
Adam														
Ana	7													
Chila	6	9												
Colin	9	7	6											
Dee	5	8	9	5										
Dom	9	8	7	9	6									
Ewan	7	9	8	7	7	8								
Holly	4	7	8	4	8	5	7							
Jim	4	6	8	4	9	5	6	8						
Krish	5	8	8	5	8	6	8	9	7					
Laura	8	9	8	8	7	9	8	6	6	7				
Mira	7	9	9	6	8	7	9	8	7	8	8			
Pete	6	9	9	7	9	7	8	7	7	7	8	8		
Sanjiv	9	8	7	9	6	10	8	6	5	6	9	7	7	

Obtaining a distance measure

We can use these similarity measures directly or we can convert them into distances.

We prefer to use the counts to derive a distance measure, since Multidimensional Scaling is based on the idea of producing a visual display of the distances between pairs of items. However, SPSS does enable you to use similarities (it calls them proximities) directly as we see in the next section.

The highest similarity score between any pair is 10, so if you subtract all the similarities in Table 11.4 from some number greater than 10, for instance from 12, then those with the lowest similarities will have the highest scores and we shall have a measure of distance. The distance of each person from him/herself must be taken as zero, by analogy with distances in space. Table 11.5 shows the result for the initial set

of similarity ratings. (For the later set the results can be found as mds.distanceb.sav on the book website.)

Table 11.5
'Distances' between pairs of patients derived from initial similarity ratings
(mds.distancea.sav)

	Adam	Ana	Chila	Colin	Dee	Dom	Ewan	Holly	Jim	Krish	Laura	Mira	Pete	Sanjiv
Adam	0													
Ana	7	0												
Chila	8	2	0											
Colin	5	5	5	0										
Dee	8	5	4	6	0									
Dom	6	5	6	6	8	0								
Ewan	4	5	6	3	7	4	0							
Holly	9	5	4	6	3	8	7	0						
Jim	3	7	8	4	8	7	5	8	0					
Krish	5	6	7	6	9	4	4	9	6	0				
Laura	9	5	4	6	3	8	7	3	8	9	0			
Mira	9	4	3	7	5	7	7	5	9	8	5	0		
Pete	4	6	6	3	7	5	4	8	4	5	7	7	0	
Sanjiv	6	6	7	6	8	2	4	9	7	3	9	7	5	0

We can use either of Tables 11.4(a) or 11.5 as the basis of a multidimensional scaling and the results will be equivalent.

Entering the data in SPSS

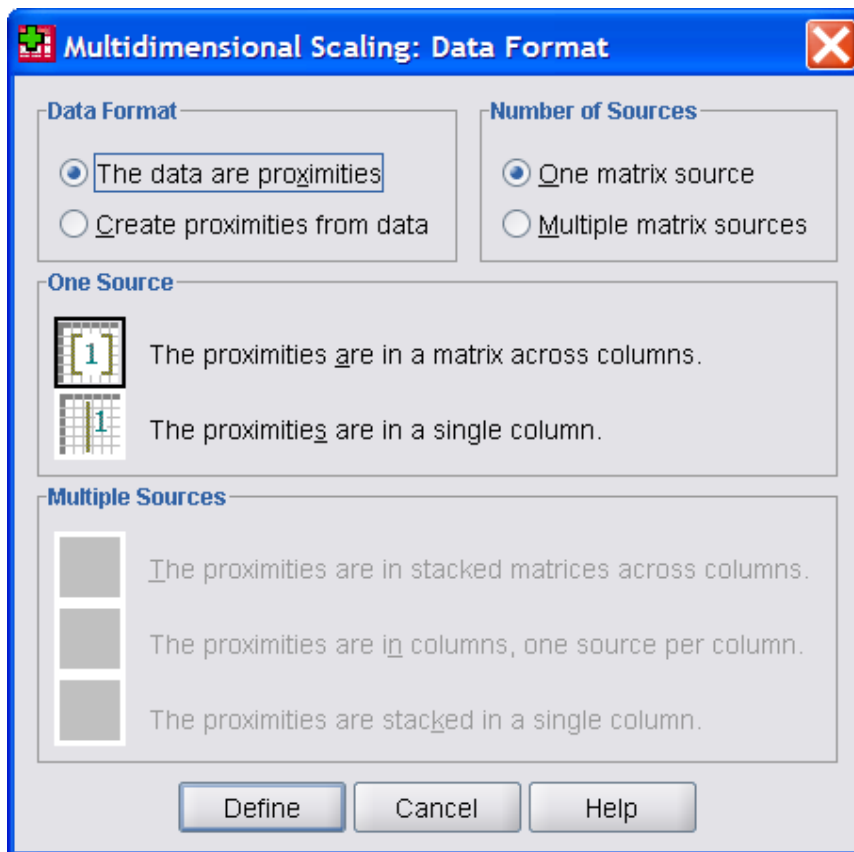
As in cluster analysis, we can get SPSS to calculate a distance matrix from a data matrix of cases and variables, and the same considerations apply to the choice of distance measure. Alternatively we can enter a distance matrix as a lower or upper triangle like the one in Table 11.1. We note, in passing, that SPSS offers MDS of non-metric (ordinal) data as well as scaling of metric data (distances or proximities on an interval scale), though we will not be looking at any ordinal examples.

First enter the data in Table 11.5 into the SPSS datasheet, with the patients' names as the variable names, and the upper triangle missing (note that the names do not appear as the first column as they do in Table 11.5). We can apply scaling to variables or

cases, but the default is variables and if we apply it to cases, then they appear as case 1, case 2 etc on the map, whereas variables are named. As this is the only difference in the results, we will let our patients be variables.

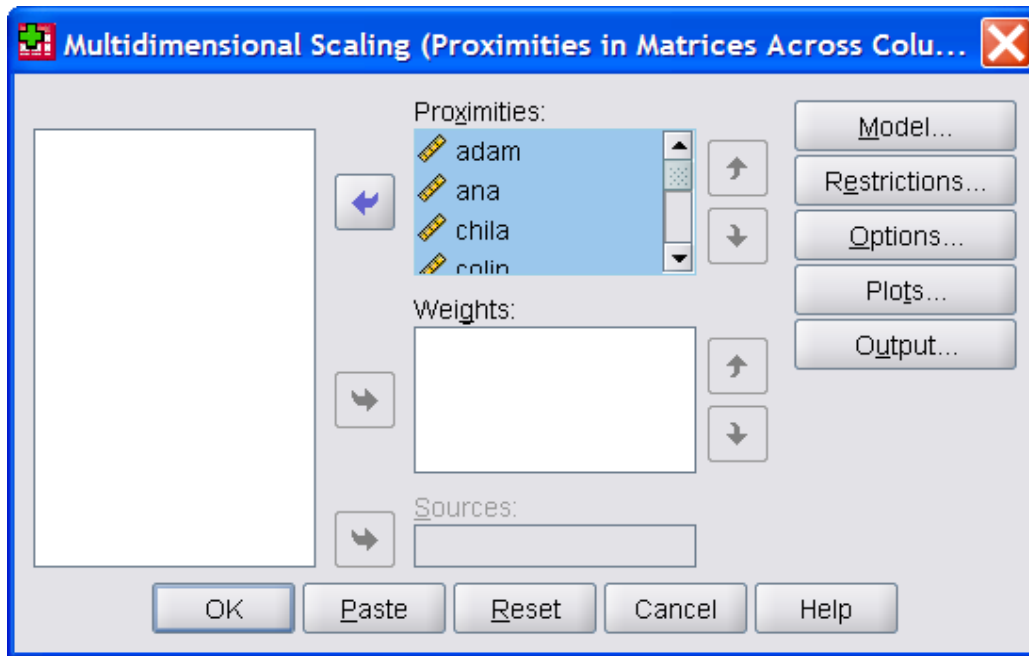
Requesting the analysis in SPSS

Once the data are entered, choose **Analyze**, then **Scale**, then **Multidimensional Scaling (PROXSCAL)**, to get SPSS Dialog Box 11.1.



SPSS Dialog Box 11.1. Defining the data layout

Our data are distances (SPSS includes distances under the term 'proximities'), we have one matrix, and our distances are entered as a lower triangle, across the columns, so accept the defaults and click **Define**. We get SPSS Dialog Box 11.2, and enter all the patients' names in the **Proximities** box using the arrow.



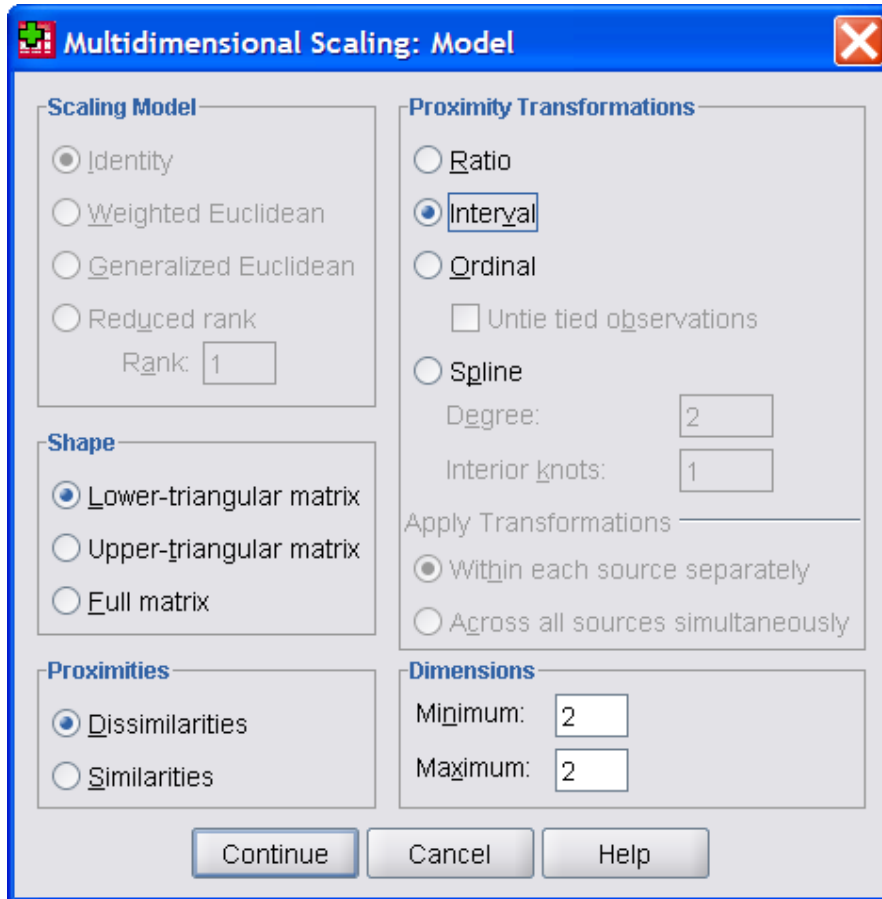
SPSS Dialog Box 11.2. Starting a Multidimensional Scaling

First click the Model button to get SPSS Dialog Box 11.3. We can accept most of the defaults here. Ours is a **Lower triangular matrix**, and our 'proximities' are distances or **Dissimilarities**. Usually a distance matrix will be square symmetric, because the distance from A to B is the same as that from B to A. However, if instead of obtaining similarity measures for pairs of our patients we had asked each patient to rate each of the others on a scale from 1 (our needs are very similar) to 5 (our needs are not at all similar), then we would have two measures of distance for each pair, one contributed by each member of the pair, and these two measures need not be the same. In a case like this it is necessary to enter the complete matrix, not just the upper or lower triangle, and to select Full matrix for the Shape.

For the moment, we are only interested in two dimensions, so leave both **Minimum** and **Maximum** at **2**. However, it would be more realistic perhaps to describe our data as **Interval** rather than ratio, since we subtracted numbers of interactions from an

arbitrary number greater than the maximum number. Even if we had used the similarity ratings directly as a similarity measure, however, there would still have been the same arbitrariness in deciding a number for each patient with themselves.

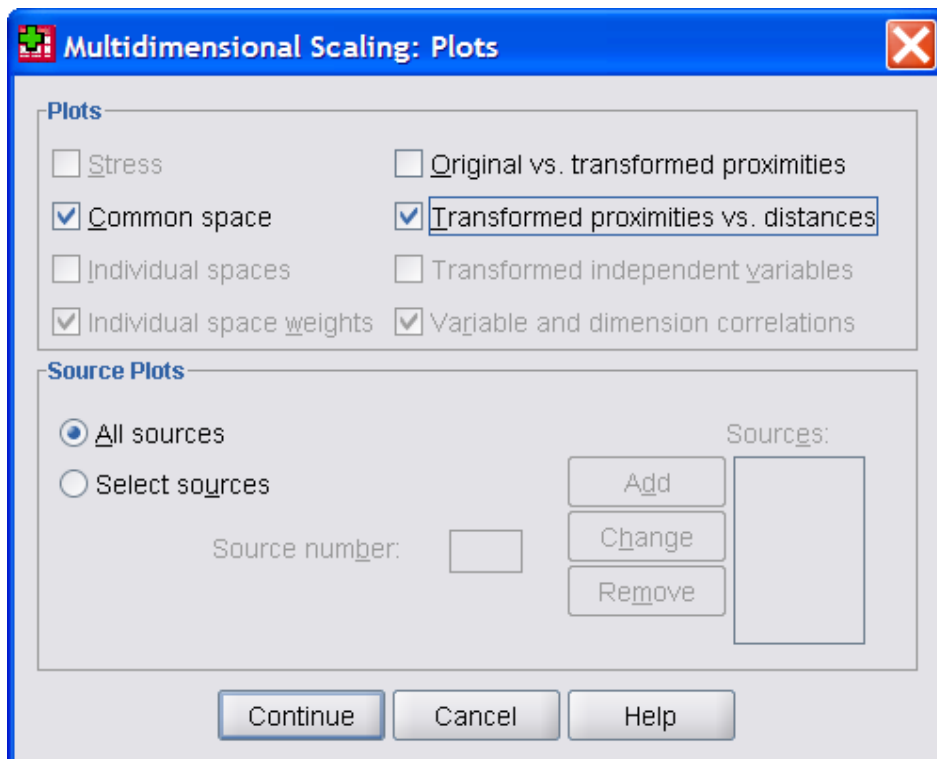
Click **Continue** to return to the main dialog box.



SPSS Dialog Box 11.3. Defining the model

We need not consider the **Restrictions** button, since we only consider problems with no restrictions, the default. (It is possible to impose values for some coordinates in the final map but we do not consider these problems). We can also ignore the **Options** button, which allows us to choose different ways of starting the process of trying to form the two-dimensional map from the distances, and also to choose different criteria to end the process. The defaults will be fine for us, and for most purposes.

Clicking the Plots button gives us SPSS Dialog Box 11.4, and if we opt for **Transformed proximities vs distances** as well as **Common space**, we shall get an additional idea of the goodness of fit. It's the Common space plot that is the map constructed from the matrix of distances. Note that the first plot (greyed out here) is called **Stress**. This becomes available if the Maximum is greater than the Minimum dimensions in SPSS Dialog Box 11.3, and we shall need this later. Click **Continue** to return to SPSS Dialog Box 11.2.

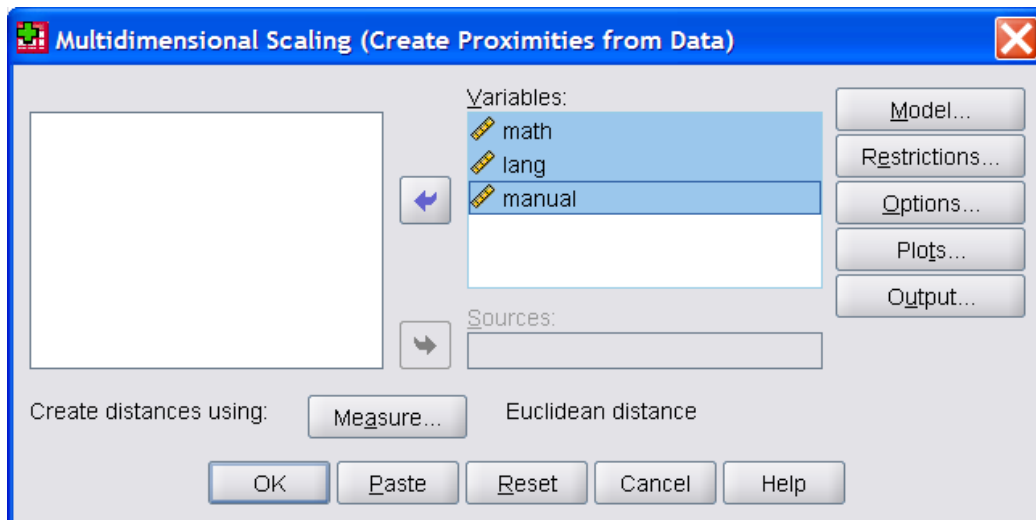


SPSS Dialog Box 11.4. Selecting plots

The last button, **Output**, allows us to save the coordinates of the points on the map to a file (these are the **Common space coordinates**). Further details of the fitting process can also be obtained but the defaults are enough for most problems. If we provided data on a set of variables and asked for SPSS to create the distance matrix

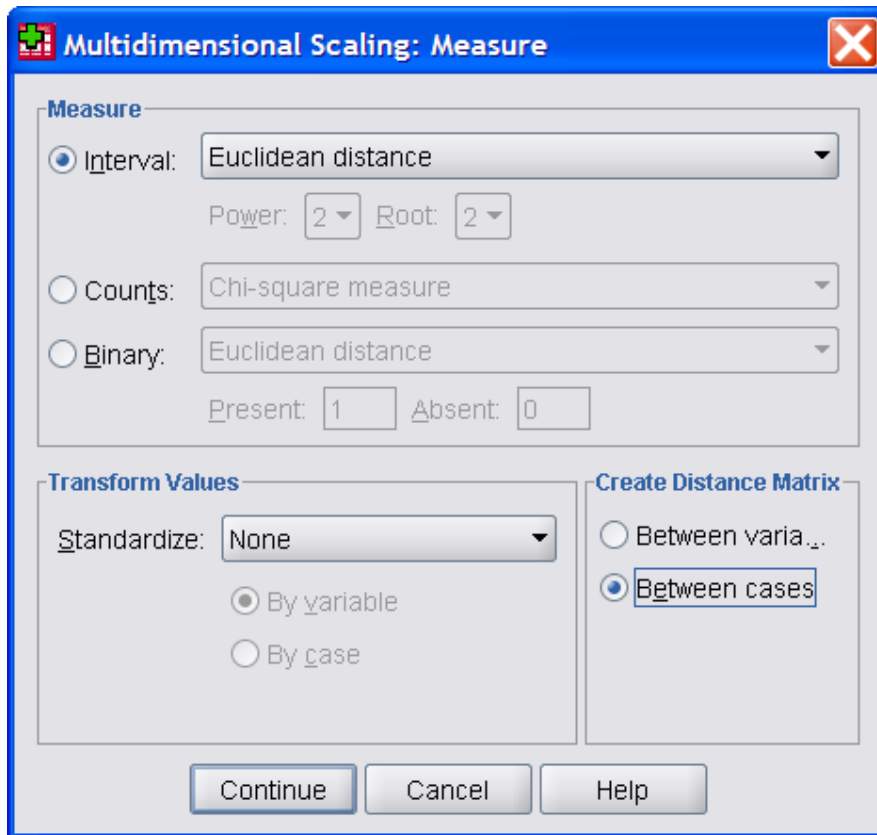
then it may be useful to click **Distances** in the **Display** group, but we do not need this as we submitted a matrix of distances. Click **OK** to get the results, which we consider in the next section. First we briefly describe how to proceed if you start with variables recorded for each case instead of with a distance or similarity matrix.

If instead of a distance or similarity matrix we have a set of variables recorded on each case, we can ask SPSS to create a distance matrix by selecting **Create proximities from data** in SPSS Dialog Box 11.1, then clicking **Define**. SPSS Dialog Box 11.5 appears (here we have used some non-proximity data from one of our Cluster Analyses as our example).



SPSS Dialog Box 11.5. Creating a distance matrix from data on variables.

Enter the variables into the **Variables** box using the arrow, and click the **Measure** button. SPSS Dialog Box 11.6, which then appears, is similar to SPSS Dialog Box 10.3, which we considered in the chapter on Cluster Analysis, and we must make the same kind of choice as we did there.



SPSS Dialog Box 11.6. Choosing a distance measure and deciding how the distance matrix should be created.

We must choose a distance measure appropriate for our data and decide whether standardisation is necessary. Standardisation should be used unless all variables are on the same scale. In this example, all three variables are on similar scales so we need not standardise. We also have to choose between applying multidimensional scaling to variables (the default) or cases. In our example here we have 52 cases (rows of the data matrix) and three variables (columns), so we need to calculate the distance matrix **Between cases**. Remember that, if you apply multidimensional scaling to variables, they are labelled on the map, whereas cases are only numbered, so we may often prefer to call the items of interest (whatever they are) variables, and arrange them as columns in the datasheet. But with 52 cases, as we have here, numbering them on the

map is more realistic than trying to use names which will unduly clutter the map.

Click **Continue** and proceed as we did when we already had a distance matrix.

Understanding the output tables

First in the output is a summary of the data (not shown here), which tells us how many objects were mapped (14 patients in our example) and how many distances were submitted ($14 \times 13 / 2 = 91$ in our example).

A table showing how well the data are fitted by a two-dimensional map follows, the first table in SPSS Output 11.1. For a good fit, we want to see low measures of stress (less than 0.15) and values close to 1 for Dispersion Accounted For (DAF) and Tucker's Coefficient of Congruence. The coordinates of the patients in the two-dimensional map follow, see the second table in SPSS Output 11.1.

Stress and Fit Measures

Normalized Raw Stress	.00491
Stress-I	.07006 ^a
Stress-II	.16614 ^a
S-Stress	.00881 ^b
Dispersion Accounted For (D.A.F.)	.99509
Tucker's Coefficient of Congruence	.99754

PROXSCAL minimizes Normalized Raw Stress.

a. Optimal scaling factor = 1.005.

b. Optimal scaling factor = .998.

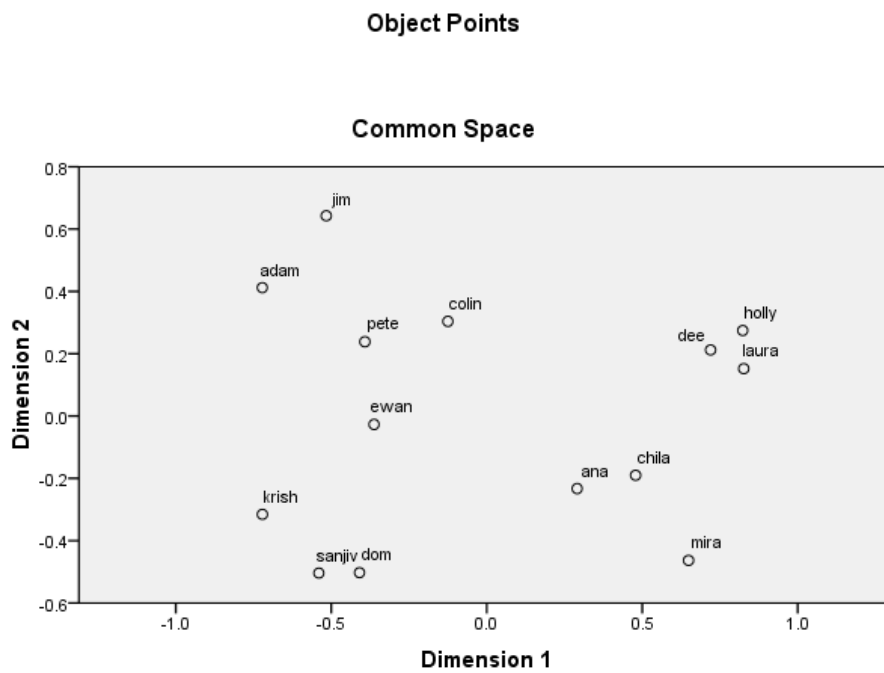
Final Coordinates

	Dimension	
	1	2
adam	-.722	.412
ana	.291	-.233
chila	.479	-.190
colin	-.125	.304
dee	.720	.212
dom	-.409	-.503
ewan	-.362	-.027
holly	.823	.275
jim	-.516	.643
krish	-.721	-.316
laura	.826	.152
mira	.649	-.463
pete	-.392	.238
sanjiv	-.540	-.504

SPSS Output 11.1. Measures of goodness of fit and coordinates in two dimensions

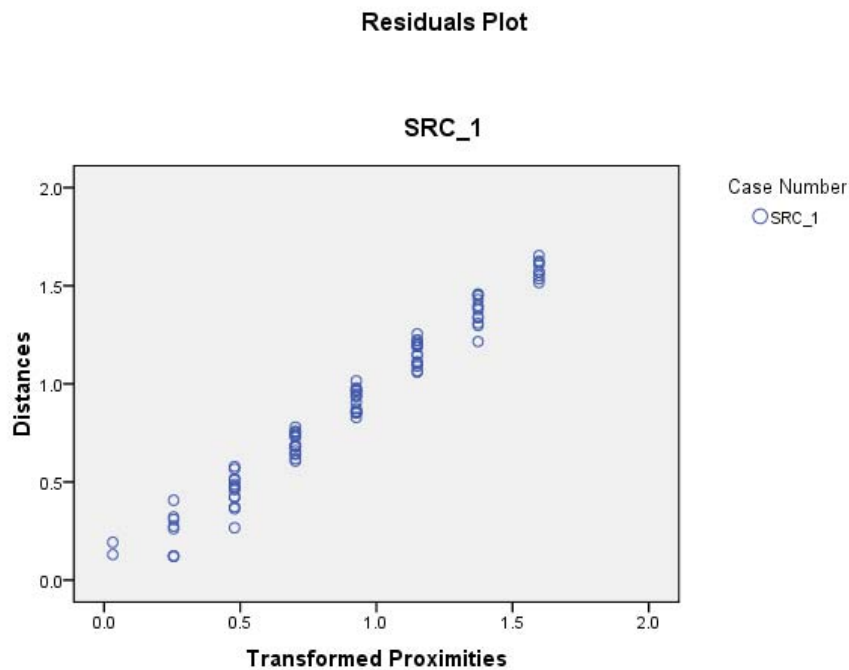
Understanding the graphical output

The coordinates may be useful but by far the most important part of the output is the plot, shown in SPSS Output 11.2. It is this visual display that we hope will give us some insight into our data. Dimension 1 may be interpreted as the presence of obvious physical symptoms, such as poor coordination and balance, involuntary movements, visual disturbance and numbness, and Dimension 2 as severity of symptoms. The tight group at the upper right of the map comprises people with obvious and severe physical symptoms. Toward the lower left is a group with few physical symptoms and whose symptoms are less severe. People toward the upper left have severe symptoms but not generally of an obvious physical kind (e.g., loss of sensation, facial pain, constipation, memory problems). Finally, at the lower right there is a group with obvious physical symptoms that are not yet severe.



SPSS Output 11.2. Two-dimensional map of the patients from distance matrix in Table 11.5

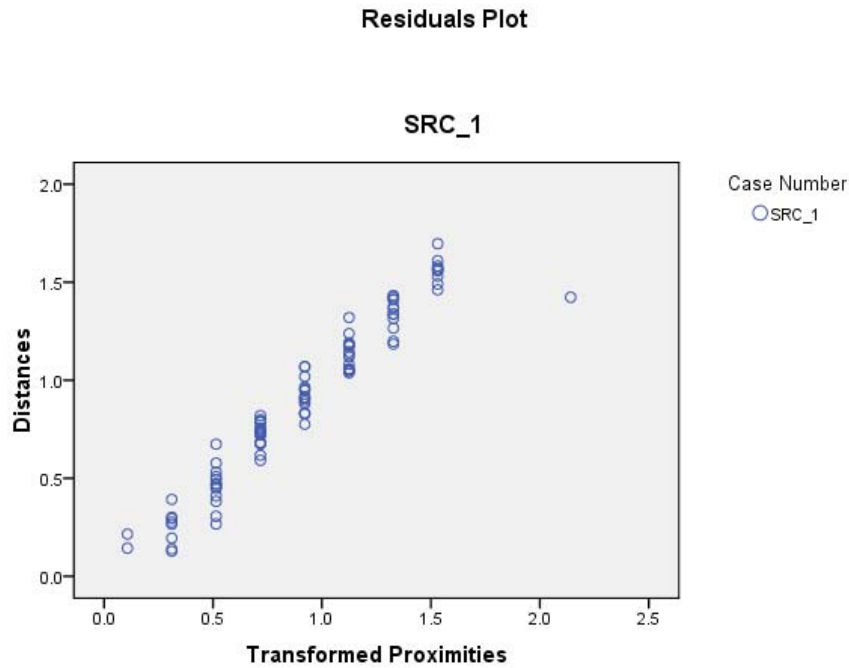
The final piece of output, shown in SPSS Output 11.3, is a plot of the distances from the matrix against the distances from the map. This would be a straight line if the fit were perfect. In the road map example, this would correspond to all distances between towns being the shortest straight line distances rather than actual road distances. Even if the fit is good, it can be useful to look at this plot in case one or two points are somewhat further than the rest from the line, which would indicate that one of the items mapped was fitted less well than the rest.



SPSS Output 11.3. Plot of actual and fitted distances

To illustrate the value of looking at the plot, we show in SPSS Output 11.4 a plot we obtained after we altered one of the distances in Table 11.5. We changed the distance from Adam to Ana from 7 to 12. The DAF value was still high at almost 0.99, but the

stress values are all higher and the point at the top right of the plot suggests that the lack of fit is confined to one of the distances.



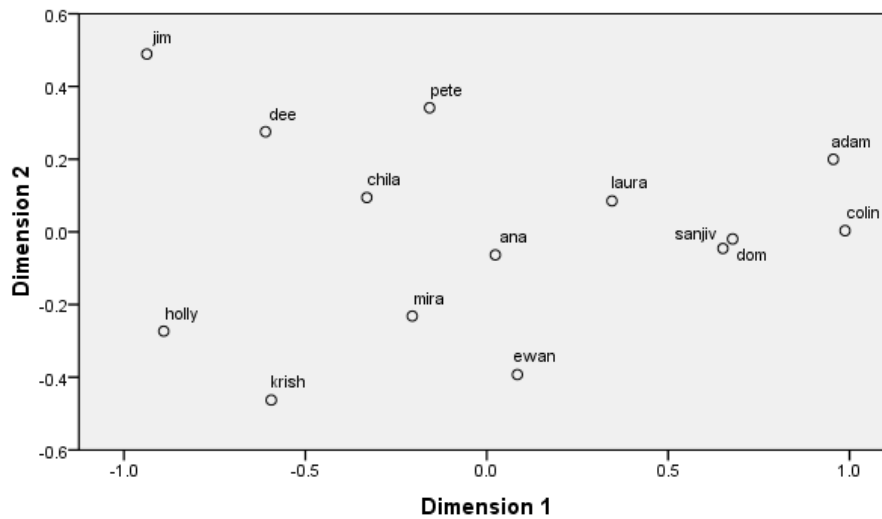
SPSS Output 11.4. Plot of actual and fitted distances when one distance in the matrix is altered

Understanding the post-intervention output

If we convert the similarities in Table 11.4(b) to distances, just as we obtained Table 11.5 from Table 11.4(a), we can produce a new map of the patients after the symptom checklist intervention. This is shown in SPSS Output 11.5. Again the stress and DAF values suggest a good fit.

Object Points

Common Space



SPSS Output 11.5. Two-dimensional map of the patients after the symptom checklist intervention

On the new map, people with different degrees of obvious physical symptoms and different degrees of severity of symptoms are now spread throughout. It appears that staff now use a wider range of symptoms when judging similarities and, presumably, needs in common among the MS patients.

As we said in the Introduction, interpreting the map dimensions is not a priority, but nevertheless if it can be done it may be useful. If we list the patients in their order along Dimension 1 we get Jim, Holly, Dee, Krish, Chila, Mira, Pete, Ana, Ewan, Laura, Sanjiv, Dom, Adam, Colin. At this point, we will imagine that we asked the patients whether this meant anything to them, and we will invent some responses that they might have given, in order to illustrate the kind of interpretation that might arise. When asked about the ordering along Dimension 1, the MS patients suggest that those

toward the left tend to be more vociferous than their peers, and, for Dimension 2, that those at the bottom tend to be among the most enthusiastic participators in therapeutic activities.

The scree diagram: considering more than two dimensions

If we have a sufficiently large set of items it is possible to make a check on whether our two-dimensional map is appropriate by graphing the stress against the number of dimensions for a range of dimensions, say from 1 to 4. This graph is called a scree diagram, and we hope to observe a sharp change in gradient at the most appropriate number of dimensions for our data. For a map derived from the road distances between cities we would expect that two dimensions would give the best representation. However, if we had the distances between some places in the Himalaya, where enormous changes in height occur over fairly short distances, we might expect that three dimensions would be needed to give an adequate representation.

The scree diagram: number of cases needed

When we collect data of the sort described in the introduction and for our experiment, there is no guarantee that a visual representation in two dimensions will display the inter-item distances without distortion. But with sufficient items in our distance matrix we can do a check with a scree diagram. The number of coordinates required to map (for instance) 14 patients in two dimensions is $2 \times 14 = 28$. In three dimensions we need $3 \times 14 = 42$, and for four we need $4 \times 14 = 56$. Our distance matrix for 14 patients gives us $14 \times 13/2 = 91$ distances, not a lot to use in estimating 56 coordinates. If we only had six patients ($6 \times 5/2 = 15$ distances), we would be unable to estimate $4 \times 6 = 24$ coordinates for a four-dimensional representation, or, indeed, $3 \times 6 = 18$

coordinates for a three-dimensional representation. This is why you can only check the scree diagram if you have a large enough set of items; at least nine ($9 \times 8 / 2 = 36$ distances), preferably more, for four dimensions (with $4 \times 9 = 36$ coordinates).

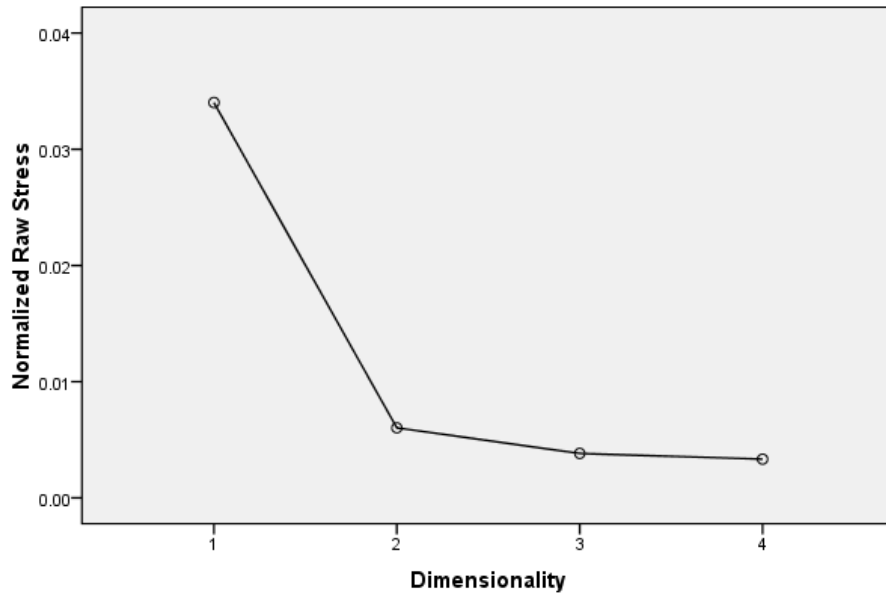
We obtained a scree plot for the patients' second distance matrix. In SPSS Dialog Box 11.3, replace the values for **Minimum** with **1** and for **Maximum** with **4**. In SPSS Dialog Box 11.4 we can now click the **Stress** plot.

The scree diagram: understanding the output

Map plots are produced for representations in 4, 3 and 1 dimension as well as in 2 dimensions. The plots in 4 and 3 dimensions are hard to read: probably this visual approach will only be useful if 2 dimensions (as in SPSS Output 11.5) suffice. The one-dimensional plot arranges the items in a line, also unlikely to be revealing. We therefore do not reproduce the map plots for 1, 3 and 4 dimensions.

SPSS Output 11.6 was produced as a result of clicking **Stress** in SPSS Dialog Box 11.4. We see from the graph, known as a scree diagram, that stress increases as the number of dimensions is reduced, showing that the fit becomes progressively less good. The biggest increase is from 2 to 1 dimensions. The sharp change in the slope at dimension 2 shows that a visual representation in two dimensions is the best for this distance matrix.

Scree Plot



SPSS Output 11.6. Stress for solutions in one to four dimensions