

Documentation for

W O M B A T S

**in the MDS (X) series of programs.
a specially-written program for pre- and post
processing of raw data and conversion to
appropriate measures of dis/similarity¹**

1. OVERVIEW

Concisely: WOMBATS (Work Out Measures Before Atttempting To Scale), does just what its acronym says and computes from a rectangular data matrix a set of (dis)similarity measures suitable for input to other MDS(X) programs.

1.1 WOMBATS in brief

The WOMBATS program takes as input a rectangular matrix either of raw data, and computes a measure of (dis)similarity between each pair of variables in the matrix. These measures are output in a format suitable for input to other MDS(X) programs. This format is chosen by the user.

2. DESCRIPTION OF THE PROGRAM

The following section describes briefly those aspects of the program pertinent to its use. For a fuller discussion, see chapter 2 of 'The User's Guide' (Coxon 1982).

Section 2.1 describes the type of data suitable for input, and its presentation to the program and section 2.2 the range of measures available. Section 2.3 describes further options including those available for outputting the results.

¹ D:\SIGMA\FULLGP93\WOMBATS\WOMBATS.ANN

2.1 Data

The basic form of input data for the WOMBATS program is a rectangular matrix in which the rows represent cases (or subjects) and the columns, variables (or stimuli).

The number of rows in the matrix is specified by the user on the N OF CASES card or, (alternatively, on the N OF SUBJECTS card). The number of columns fields is given on either of the N OF VARIABLES or on the N OF STIMULI card. In these cards 'N' may of course be replaced by either 'NO' or '#'. The data are read by the program when it encounters a READ MATRIX card and the INPUT FORMAT card should describe one row of the data matrix.

If the data to be input are for some reason in a matrix where the rows represent variables and the columns cases, then the user should specify MATFORM(O) on the PARAMETERS card.

The chosen measures are calculated between the entities designated as variables. This will be the case whatever value is taken by the parameter MATFORM. If the user wishes measures to be calculated between cases rather than between variables, then see section 2.3.1 below.

N.B. The program expects data to be input as reals. The INPUT FORMAT statement must therefore be specified to read F - type numbers, even if the numbers do not contain a decimal point.

2.1.1 Levels of Measurement

The users must specify, for each of the variables in the analysis, the level of measurement at which it is assumed to be. Five levels are recognised by the program. The recognised levels are ratio, interval, ordinal, nominal and dichotomous. If a particular variable is not assigned to a particular level by the user, then the program assigns it by default to the ordinal level.

Each of the measures in the program assumes that the variables on which they are operating to have the properties of a particular level of measurement. If an attempt is made to compute a measure which assumes a level of measurement higher than that at which the variables have been declared to lie, then the program will fail. No restriction is placed, obviously, on the attempt to calculate measures which assume levels lower than those declared.

The user signals the measurement level of the variables to the program by means of the LEVELS card, peculiar to the WOMBATS program. This card contains the command LEVELS in columns 1 to 6 and in column 16 *et seqq.* one or more of the keywords RATIO, INTERVAL, NOMINAL, DICHOTOMOUS or ORDINAL.

(Obviously, since the program defaults to ordinal, there is no need actually to specify this last keyword). In parentheses following each specified keyword are listed the variables which are to be assumed to be at that level of measurement. In these parentheses ALL and TO are recognized. The following are valid examples of a LEVELS declaration.

col 1	col.16
LEVELS	INTERVAL (1, 2, 5, 7,), NOMINAL (3, 4, 6, 8)
LEVELS	RATIO (ALL)
LEVELS	NOMINAL (1 TO 4), INTERNAT (7 TO 11)

In the last example, the spelling mistakes are not important since only the first four letters of the keywords are significant. In the same example, variables 5 and 6 are presumed by default to be at the ordinal level.

2.1.2 Missing Data

Variables that include missing data are a problem. The user may specify, for each variable in which there are missing data, one code which the program will read as specifying a missing datum. Users will note however that an attempt to calculate certain measures between variables containing missing data will fail if missing data are present. The measures for which this is the case are indicated in the discussion of the available measures in section 2.2.1.

The user signals the occurrence of missing data by means of the MISSING card. This card has the command MISSING in columns 1 to 7 and in column 16 *et seqq.* the value(s) to be regarded as signifying missing data are listed. In parentheses following each missing data value is a list of the variables for which that value represents a missing datum. In these parentheses the forms ALL and TO are recognised. The following are valid examples of MISSING declarations.

col 1	col.16
MISSING	-9.(1, 2, 7, 9), 99.(3, 4, 6, 8)
MISSING	0. (ALL)
MISSING	.1(1 TO 7), -.1(8 TO 16)

2.2 ANALYSIS

The aim of the WOMBATS program is to calculate for each pair of variables in the analysis a measure of the (dis)similarity between them. Having described the data to the program, the user must then choose the measure to be calculated.

WOMBATS allows 26 measures.

The required measures are chosen by means of the MEASURE card. This card contains the command MEASURES in columns 1 to 8 and in column 16 *et seqq.* one of the keywords referring to the available measures described below. Only one measure is computed in each TASK of the run. If more than one measure is required on the same set of data, then a separate TASK NAME is necessary.

2.2.1 Available measures

It is convenient to consider the available measures in WOMBAT under their respective assumed levels of measurement.

2.2.1.1 Dichotomous measures

Sixteen measures of agreement between dichotomous variables are included in WOMBATS. These correspond to those described in 'The User's Guide'pp.24-27. Missing data are allowed in all these measures.

In this section, the following notation will be crucial. Consider two dichotomous variables which we will assume to measure whether the objects under consideration do or do not possess a particular attribute. The co-occurrence(or frequency) matrix of these two variables looks as follows.

		Variable 1	
		1	0
Variable 2	1	a	b
	0	c	d

The cell 'a' is the number of times that the attributes 1 and 2 co-occur, 'b', the number of times attribute 2 is present when attribute 1 is not, 'c' is the number of times attribute 1 is present and 2 is not and 'd' is the number of objects possessing neither attribute 1 nor attribute 2. All the measures of agreement to be considered in this section result from the combination of these quantities in some way.

The measures available for the comparison of dichotomous variables are denoted by the 'keywords' D1, D2, ..., D16 and it is these 'keywords' that appear in column 16 *et seqq.* of the MEASURES card. The numbers correspond to the indices named in Coxon (1982, Table 2.4)

For example, the card

Col.1	col.16
MEASURES	D15

will select Yule's Q as the measure to be calculated

<u>Command</u>	MEASURES D1
<u>Type</u>	Similarity measure
<u>Range</u>	low = 0, high = 1

<u>Name</u>	Jaccard's coefficient
<u>Formula</u>	$\frac{a}{(a+b+c)}$

<u>Description</u>	Excludes 'd'. Represents the probability of a pair of objects exhibiting both of a pair of attributes when only those objects exhibiting one or other are considered. It is possible that a division by zero may occur in the calculation of this measure.
--------------------	--

<u>Command</u>	MEASURES D2
<u>Type</u>	Similarity measure
<u>Range</u>	low = 0, high = 1

$$\frac{a}{(a+b+c+d)}$$

<u>Name</u>	Russell and Rao's measure
-------------	---------------------------

<u>Formula</u>	
----------------	--

<u>Description:</u>	Represents the probability of a pair of objects in a pre-selected set exhibiting both of a pair of attributes.
---------------------	--

* * *

<u>Command</u>	MEASURES D3
<u>Type</u>	Similarity measure
<u>Range</u>	low = 0, high = 1
<u>Name</u>	Sokal's measure

$$\frac{(a + d)}{(a + b + c + d)}$$

Formula

Description

Includes 'd' in numerator and denominator. Represents the probability of a matching of two attributes.

Command

MEASURES D4

Type

Similarity measure

Range

low = 0, high = 1

$$\frac{2a}{(2a + b + c)}$$

Name

Dice's measure

Formula

Description

Gives the positive matches 'a' twice as much importance as anything else. Excludes entirely the mismatches. It is thus possible that a division by zero may occur in the calculation of this measure.

Command

MEASURES D5

Type

Similarity measure

Range

low = 0, high = 1

$$\frac{2(a + d)}{(2(a + d) + b + c)}$$

Name

no name

Formula

Description

Includes 'd' in both numerator and denominator. The matches (a and d) are given twice as much weight as the mismatches.

* * *

<u>Command</u>	MEASURES	D6
<u>Type</u>	Similarity measure	
<u>Range</u>	low = 0, high = 1	

$$\frac{a}{(a + 2(b + c))}$$

<u>Name</u>	no name
<u>Formula</u>	

<u>Description</u>	Excludes `d' entirely. The matches (b and c) are accorded twice as much weight as the matches. It is possible that a division by zero may occur in the calculation of this measure.
--------------------	---

<u>Command</u>	MEASURES	D7
<u>Type</u>	Similarity measure	
<u>Range</u>	low = 0, high = 1	

$$\frac{(a + d)}{(a + d + 2(b + c))}$$

<u>Name</u>	Rogers and Tanimoto's measure
<u>Formula</u>	

<u>Description</u>	Includes `d' in numerator and denominator. The mismatches (b and c) are accorded twice as much weight as the matches.
--------------------	---

<u>Command</u>	MEASURES	D8
<u>Type</u>	Similarity measure	
<u>Range</u>	low = 0, high = a + b + c + c + d - 1	

Name

Kulczynski's measure

$$\frac{a}{b+c}$$

Formula

Description

Excludes 'd' entirely. This measure is the simple ratio of the positive matches (a) to the mismatches (cf. D9). it is possible that a division by zero could occur in the calculation of this measure and an undefined statistic occur. The maximum value otherwise is as stated.

Command

MEASURES D9

Type

Similarity measure (Sokal & Sneath)

Range

low = 0, high = a + b + c + d - 1

$$\frac{(a+d)}{(b+c)}$$

Name

no name

Formula

Description

This measure is the simple ratio of all matches (positive and negative) to the mismatches (cf D8). The statistic may be undefined, due to a zero divisor. The maximum finite value is as stated.

* * *

Command

MEASURES D10

Type

Similarity measure

Range

low = 0, high = 1

Name

Kulczynski's measure

$$\frac{1}{2} \left(\frac{a}{a+c} + \frac{a}{a+b} \right)$$

Formula

Description

Excludes `d' entirely. This measure is a weighted average of the matches to one or other of the mismatches. This statistic may be undefined.

Command

MEASURES D11

Type

Similarity measure

Range

low = 0, high = 1

$$\frac{1}{4} \left(\frac{a}{a+c} + \frac{a}{a+b} + \frac{d}{b+d} + \frac{d}{c+d} \right)$$

Name

no name

Formula

Description

Includes `d' in numerator and denominator. This is the analogue of D10 with mismatches included.

Command

MEASURES D12

Type

Similarity measure

Range

low = 0, high = 1

$$\frac{a}{\sqrt{(a+c)(a+b)}}$$

Name

Ochiai's measure

Formula

Description

Excludes `d' from numerator. It uses the geometric mean of the marginals as a denominator. This statistic may have a zero divisor.

<u>Command</u>	MEASURES	D13
<u>Type</u>	Similarity measure	
<u>Range</u>	low = 0, high = 1	

$$\frac{ad}{\sqrt{(a+c)(a+b)(b+d)(c+d)}}$$

<u>Name</u>	no name
<u>Formula</u>	
<u>Description</u>	Includes 'd' in numerator and denominator. It uses the geometric mean of the marginals as a denominator and will return a value of 0 iff either a or d is empty.

* * *

<u>Command</u>	MEASURES	D14
<u>Type</u>	Similarity measure	
<u>Range</u>	low = -1, high = +1	

$$\frac{(a+d)-(b+c)}{(a+b+c+d)}$$

<u>Name</u>	Hamann's coefficient
<u>Formula</u>	
<u>Description</u>	Simply the difference between the matches and the mismatches as a proportion of the total number of entries. A value of 0 indicates an equal number of matches to mismatches. Some thought should be given to the interpretation of any negative coefficients before scaling the results.

* * *

<u>Command</u>	MEASURES	D15
<u>Type</u>	Similarity measure	

Range low = -1, high = +1

$$\frac{(ad) - (bc)}{(ad + bc)}$$

Name Yule's Q

Formula

Description This is the original measure of dichotomous agreement, designed to be analogous to the product-moment correlation. A value of 0 indicates statistical independence. Some thought should be given to the interpretation of any negative coefficients before scaling the results. This statistic may be undefined.

* * *

Command MEASURES D16

Type Similarity measure

Range low = -1, high = +1

$$\frac{(ad - bc)}{\sqrt{(a + c)(a + b)(b + d)(c + d)}}$$

Name Pearson's Phi

Formula

Description A value of 0 indicates statistical independence. Some thought should be given to the interpretation of any negative coefficients before scaling the results. The statistic may be undefined if any one cell is empty.

* * *

2.2.1.2 Nominal measures

Five measures are available in WOMBATS for the measurement of nominal agreement between variables. Four of these are based on the familiar chi-square statistic. The other is the Index of Dissimilarity.

2.2.1.2.1 Chi-square based measures

The following procedure is used to evaluate the chi-square statistic that forms the basis of four of the available measures.

Consider two variables \underline{x} and \underline{y} . We form the table whose row elements are the values taken by (or the categories of) the variable \underline{x} and whose column elements are the values (categories) taken by variable \underline{y} . (Obviously, since this is a nominal measure, these values have no numerical significance). The entries of this table are the number of cases which take on particular combinations of values of \underline{x} and \underline{y} i.e. the number of cases that fall into the particular combinations of categories.

The value of the chi-square statistic is calculated by comparing the actual distribution of these values in the cells of the table to that distribution which would be expected by chance (statistical independence occurs when $p(i,j) = p(i) \times p(j)$). Thus, the higher the value of the statistic, the more the actual distribution diverges from the chance or expected one (0).

In the case of there being missing data in the original matrix, then the whole row or column corresponding to that value is deleted. Caution should be exercised if there are many missing data and particularly if these are unequally distributed around the variables since the value of the statistic is dependent on the number of values it considers and strictly speaking chi-square measures based on largely different numbers of cases are not comparable.

The other measures in this section seek to overcome the dependence of chi-square on the number of cases by norming it. The norming factor differs for each statistic. The following notation will be useful in the section on nominal measures.

- N will indicate the number of cases
- r will stand for the number of rows in the matrix i.e. the number of categories (values) taken by variable y and
- c will stand for the number of columns i.e. the number of categories in variable y .

<u>Name</u>	Chi - square
<u>Command</u>	MEASURES CHISQUARE
<u>Type</u>	Similarity measure
<u>Range</u>	low = 0, high = $N \times \min(r,c)$
<u>Comment</u>	A value of 0 indicates statistical independence. The maximum value is dependent on the value of N.

* * *

<u>Name</u>	Phi
<u>Command</u>	MEASURES PHI
<u>Type</u>	similarity measure
<u>Range</u>	low = 0, high = $\#(\min(r,c)-1)$
<u>Comment</u>	The phi coefficient is chi-square normed to be independent of N. Reaches a maximum for 2 x 2 tables in which case it reduces to the product-moment correlation. It may, however, exceed 1 when the minimum of r and c is greater than 2.

* * *

<u>Name</u>	Cramer's V
<u>Command</u>	MEASURES CRAMER
<u>Type</u>	similarity measure
<u>Range</u>	low = 0, high = 1
<u>Comment</u>	Cramer's coefficient is chi-square normed to be independent of N <u>and</u> of the number of r and c. Reaches a maximum for non-square tables.

* * *

<u>Name</u>	Pearson's Contingency coefficient C
<u>Command</u>	MEASURES PEARSON
<u>Type</u>	similarity measure
<u>Range</u>	low = 0, high = +1
<u>Comment</u>	Pearson's coefficient is chi-square normed to be independent of N, originally developed as a measure for contingency tables. Cannot reach its maximum of 1 for non-square tables.

* * *

2.2.1.2.2 The index of dissimilarity

The remaining statistic in this section is the index of dissimilarity. In the case of the chi-square measures, the implicit comparison is between the actual (bi-variate) distribution and the expected (chance) one. In the case of the index it is two (univariate) distributions that are compared.

Consider again the table that is formed by cross-tabulating the values of variable \underline{x} and those of variable \underline{y} . If the two variables had identical distributions then all the off-diagonal cells would be empty. The index of dissimilarity is simply the proportion of cases that appear in these off-diagonal cells and may be thought of as the proportion of changes needed to change the one distribution into the other. The index does not require equal numbers of values in the variables.

<u>Name</u>	Index of dissimilarity
<u>Command</u>	MEASURES ID
<u>Type</u>	dissimilarity
<u>Range</u>	low = 0, high = 100

2.2.1.2 Ordinal level measures

At present, there are three measures of ordinal agreement in WOMBATS, all related to the basic tau (τ) measure of Kendall (19..). τ_b , τ_c and Goodman and Kruskal's gamma (γ). There are two important distinctions in considering these measures. First, we need to know if they measure weak or strong monotonic agreement

between the variables and secondly how they treat tied values in them. This second distinction can be crucial since much ordinal level data, being composed of a relatively small number of categories, will contain a large proportion of tied data values.

Consider first just one variable, say the variable x scored in ordinal categories. Now consider each pair of cases i and j . For this pair of cases or individuals it can be the case that:

- a) the value of i is greater than ($>$) that of j or
- b) the value of i is less than ($<$) that of j or
- c) the values are equal.

Now consider the case of two variables. For each pair of individuals there are nine possible states of affairs:

- a) case i , case j on x and i , j on y
- b) π_i , π_j on x and $i+j$ on y
- c) π_i , π_j on x and $i=j$ on y

- d) case i + case j on x and i , j on y
- e) $\pi_i + \pi_j$ on x and $i+j$ on y
- f) $\pi_i + \pi_j$ on x and $i=j$ on y

- g) case $i =$ case j on x and i , j on y
- h) $\pi_i = \pi_j$ on x and $i+j$ on y
- i) $\pi_i = \pi_j$ on x and $i=j$ on y

It is possible to represent this in a table thus:

	,	+	=
,	a	b	c
+	d	e	f
=	g	h	i

But it should be remembered that the entries in this table refer to pairs of cases and not, as in previous sections, to single cases. Thus in this case, the sum $(a + b + c + d + e + f + g + h + i)$ is equal to $N(N - 1)/2$ and not to N as in the case of, say, chi-square.

The numerator of all the ordinal measures here considered is the sum of the concordant pairs that is those pairs of cases that are higher (lower) on both variables namely $(a$ and $e)$ minus the sum of the discordant pairs i.e. those which are higher on one variable and lower on the other (or *vice versa*) namely $(b$ and $D)$. The difference in the measures is in that denominators.

<u>Name</u>	Goodman and Kruskal's gamma (γ)
<u>Command</u>	MEASURES GAMMA
<u>Type</u>	similarity measure
<u>Range</u>	low = -1, high = +1

$$\frac{(a + e) - (b + d)}{a + b + d + e}$$

Formula

<u>Comment</u>	Measures the weak monotonic agreement between the variables, taking the ratio of the difference between concordant
----------------	--

and discordant pairs to their sum. It thus ignores the ties completely. For this reason it is possible that the value be undefined (i.e. there may be no cases). If there are no ties then the index reduces to Yule's Q (D15). Some thought should be given to the interpretation of the negative values before the results are scaled.

<u>Name</u>	Kendall's tau-b (τ_b)
<u>Command</u>	MEASURES TAUB
<u>Type</u>	similarity measure

$$\frac{(a+e)-(b+d)}{\sqrt{\langle (a+e)+(b+d)+(c+f+i) \rangle} \cdot \sqrt{\langle (a+e)+(b+d)+(g+h+i) \rangle}}$$

<u>Range</u>	low = -1, high = +1
--------------	---------------------

Formula

<u>Comment</u>	Measures strong monotonic agreement in the variables, relating the difference between concordant and discordant pairs of the geometric mean of the quantities arrived at by adding in the ties to the denominator. This should be used only for square tables.
----------------	--

* * *

<u>Name</u>	Kendall's tau-c (τ_c)
<u>Command</u>	MEASURES TAUC
<u>Type</u>	similarity measure

$$\frac{(a+e)-(b+d)}{(N_2((m-1))/m)/2}$$

<u>Range</u>	low = -1, high = +1
--------------	---------------------

Formula

<u>Comment</u>	In the formula, m stands for the lesser of the number of rows and columns in the original matrix. The statistic
----------------	---

may be used for non-square tables and reduces, in the case of square ones to tau-b.

* * *

2.2.1.4 Interval level measures

The interval level measures currently available in WOMBATS all belong to the product-moment family of measures. They are covariance, the product-moment correlation and Euclidean distance.

Consider the conventional scatter-plot of, a number of cases measured on two variables. These cases may be represented as points in a space, the two dimensions of which are the variables concerned. (The statement holds for more than two variables, of course.) The Euclidean distance between the cases is the straight line distance between the points which represent them. The correlation between each pair of points is simply the cosine of the angle between the two vectors drawn from the origin to the points concerned and the covariance is that same cosine multiplied by the length of the vectors.

<u>Command</u>	<u>MEASURES</u>	<u>DISTANCE</u>
<u>Type</u>	dissimilarity	
<u>Range</u>	low = 0, high = maximum variance in the variables	
<u>Comments</u>	If the ranges of the variables involved are markedly different, then some attempt at rescaling (i.e. normalisation) should be made so that differences in a highly valued variable do not swamp out differences in one of humbler dimensions.	

* * *

<u>Command</u>	<u>MEASURES</u>	<u>COVARIANCE</u>
<u>Type</u>	similarity	
<u>Range</u>	low = 0, high = highest variance	

<u>Comments</u>	The interpretation given to the negative values should be carefully thought out before scaling.	
<u>Command</u>	MEASURES	CORRELATION
<u>Type</u>	similarity	
<u>Range</u>	low = 1, high = 1	
<u>Comments</u>	The negative values may need to be given some thought before the results of this calculation are scaled.	

2.3 FURTHER OPTIONS

2.3.1 Measures between cases

It may be that the user wishes to have the measures calculated between the cases (subjects, individuals) in the analysis rather than the variables. This is accomplished simply by specifying on the PARAMETERS card, the keyword ANALYSIS, followed in brackets by the figure 1.

This command has the effect of calculating the measures between the entities designated as cases and is independent of the MATFORM parameter.

2.3.2 Multiple Analysis

Only one measure may be calculated at each TASK NAME. In order to calculate more than one measure on the same data at one time, more than one TASK NAME should be contained in one run. If this is done then care should be taken to REWIND the data. The TASK NAME card also resets PARAMETER values to their original (default) values and it is necessary to reset these on subsequent runs.

3. OUTPUT OPTIONS

The measures are output by default as a lower triangular matrix suitable for input to other programs in the MDS(X) library. There is no need to signal this output with a control card. Other options are available which match different conventions in other programs (see below) and in this case it is necessary to specify the output format for the measures.

3.1. Output Format

The format of the output is controlled by the control card OUTPUT FORMAT which has those words in columns 1-13 and in columns 16 *et seqq.*, a valid FORTRAN statement in brackets. In this it is obviously similar to the INPUT FORMAT card. The FORTRAN statement in brackets should read the longest row of the output matrix. There will be as many measures as there are variables and the statement should allow for any minus signs and for a sufficient number of decimal places to be printed out. For most purposes two are sufficient.

3.2 Alternative output forms

By request, the measures may be output as an upper triangular or as full (symmetric) matrix. This is accomplished by use of the keyword OUTPUT on the PARAMETERS card. The specification OUTPUT(1) gives an upper triangle and OUTPUT(2) a full matrix. This parameter does not affect the operation of the OUTPUT FORMAT command.