

SADIE software for measuring and testing for spatial association and dissociation

Dear Colleague,

This document is designed as an aid to the use of my Fortran program SADIEA.FOR to measure and test for spatial association and dissociation for data in the form of counts.

Note that this approach, while valid, is now out-of-date, and a better, new method is available, based on the correlation between cluster indices.

The program is still only a test version and may produce more output than you want. You will have to delete some output if this is the case. The major reference for the work is:

Perry, J.N. (1998). Measures of spatial pattern and spatial association for counts of insects. pp. 21-33 in: *Population and Community Ecology for Insect Management and Conservation*. (eds. J. Baumgartner, P. Brandmayr & B.F.J. Manly). Balkema, Rotterdam. *Proceedings of the Ecology and Population Dynamics Section of the 20th International Congress of Entomology, Florence, Italy, 25-31 August 1996*. ISBN 90 5410 930 0.

For an example of its use, see: Korie, S., Perry, J.N., Mugglestone, M.A., Clark, S.J., Thomas, C.F.G. & Mohamad Roff, M.N. (2000). Spatiotemporal Associations in Beetle and Virus Count Data. *Journal of Agricultural, Biological & Environmental Statistics*, **5**, 214-239.

Other discussion is in: Perry, J.N. (1997) Spatial association for counts of two species. *Acta Jutlandica*, **72**, 149-169.

The data for which it is designed comprise counts, for two populations, sampled simultaneously at identical locations.

The program was developed on a PC using Microsoft Fortran Powerstation. It runs under Windows '95 and 'NT

Documentation for SADIEA.FOR

The program **sadiea.for** analyzes the spatial association of data that are in the form of counts of two populations at specified spatial locations, for example numbers of moths of two species caught simultaneously in light-traps, or numbers of a host and its parasitoid caught in the same water traps, or numbers of diseased plants and some count of its pathogen made in the same sample unit. Alternatively it may be used for the association between numbers of the *same* species caught in the same traps on a pair of *successive* occasions.

The techniques and notation follow closely those outlined in the paper Perry, J.N. (1998). Measures of spatial pattern and spatial association for counts of insects. pp. 21-33 in: *Population and Community Ecology for Insect Management and Conservation*. (eds. J. Baumgartner, P. Brandmayr & B.F.J. Manly). Balkema, Rotterdam. *Proceedings of the Ecology and Population Dynamics Section of the 20th International Congress of Entomology, Florence, Italy, 25-31 August 1996*. ISBN 90 5410 930 0; you will find it difficult to understand this documentation without reference to that paper, so please ask me if you need a copy of this or the other papers to which it refers.

The paper describes a range of techniques that are still under development. Therefore, the program produces a number of output statistics, not all of which you may want to use. Briefly, the program (i) computes various spatial statistics for the observed data, separately for each of the sets; (ii) computes five statistics for the association between the two observed sets of data; (iii) uses (i) to find randomizations of set 1 with similar spatial patterns as those observed and recomputes the five association statistics between each of these randomized sets of data and the observed data of set 2, to form a randomization distribution; from this an index and test is produced for each of the five statistics; (iv) repeats (iii) but this time for randomizations of set 2 and associations with observed set 1; (v) combines the results found in (iii) and (iv) of the two randomizations (of set 1 to observed set 2 and of set 2 to observed set 1) to give a summary set of five indices and significance tests, which simplifies interpretation if appropriate.

The following describes the assignment of channels to input or output.

<i>Channel No.</i>	<i>Input/Output</i>	<i>Description</i>
5	Input	The raw data concerning the counts and their locations
6	Output	The briefest summary of the raw data for checking purposes
7	Output	Contains annotated full results and detailed calculations
8	Input	A few extra input parameters required to run the program, explained below
9	Output	The briefest annotated summary of results for the five statistics
10	Output	Further unannotated output, for use as input by graphics routines in other programs such as Genstat; these are still under development

How to use the program

I will send you by email examples of input files nsai5.dat (channel 5) and nsai8.dat (channel 8) together with their associated output files, nsao6.dat (channel 6), nsao7.dat (channel 7), nsao9.dat (channel 8), nsao10.dat (channel 10) and a copy of the program as a .EXE file: nsadiea.exe.

This is what you must do to run the program:

First, put the n records comprising your data into a file assigned to channel 5, in the following form:

<i>xcoord1</i>	<i>ycoord1</i>	<i>count1, species 1</i>	<i>count1, species 2</i>
<i>xcoord2</i>	<i>ycoord2</i>	<i>count2, species 1</i>	<i>count2, species 2</i>
<i>xcoord3</i>	<i>ycoord3</i>	<i>count3, species 1</i>	<i>count3, species 2</i>
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
<i>xcoordn</i>	<i>ycoordn</i>	<i>countn, species 1</i>	<i>countn, species 2</i>

where the x & y coordinates should be read in as real numbers and the count as an integer, with no decimal point. No more than 2000 records can be analyzed in the version supplied.

Next, specify various input parameters in an input file assigned to channel 8. These are:

- (i) *ISEED* (an integer) seed for the random number generator
- (ii) *NSIMS* (an integer) the number of simulations to be done; a value of at least 200 and preferably 500 is recommended, but no greater than 1000.
- (iii) *EPSILO* (a decimal number) a parameter controlling how closely you require the distance to regularity, D , of the randomized patterns to match that of the observed data. A reasonable value for this is probably 0.05; a very stringent value would be 0.01. The smaller the value the more accurately the program conditions on the observed patterns, but the longer it takes to run.
- (iv) *EPSIL2* (a decimal number) a parameter controlling how closely you require the value of δ , the distance from the centroid of the counts to the centroid of the sample units, for the randomized patterns to match that of the observed data. Usually, if δ is small, a value of *EPSIL2* as large as 1.0 will be sufficient, but usually a smaller value of *EPSIL2* is required. Unfortunately, it is difficult to give simple advice concerning precisely when each of these conditions holds. However, to aid a choice, the program outputs an approximate advised value on channel 6. It is suggested that if you want advice on the value *EPSIL2* you run the program as a test with *nsims*=1, which should be very quick, and inspect the output on channel 6, then amend the value of *NSIMS* and *EPSIL2* in this channel 8 file for the actual run.

Note that if the randomizations fail to yield values that match δ and/or D then information on the convergence of the iterative procedure will be output on channel 6 in the file nsao6.dat. A common fault occurs if the value of δ is so small that randomized arrangements cannot easily be found to match it; this can be a problem especially for sparse data with many zeroes. In that case it is recommended that the value of *EPSIL2* be inflated to unity or even larger.

(v) *ITOT1*

(vi) *ITOT2*

these last two parameters are two *small* counts that 'represent' the ratio of the mean count of species 1 to the mean count of species two. These values are given for approximate scaling. This is recommended instead of the full scaling done by default, to overcome problems of large integers that may cause integer programming problems in subroutine TRANSP, etc. The outcome is that values of set2 are scaled by *ITOT1*, and vice-versa. As an example, if the total for set1 was 1391 and for set2 was 1532, then quite sensible results may be obtained

with ITOT1=10 and ITOT2=11.

You do this by specifying these values on consecutive lines of the file, thus:

ISEED
NSIMS
EPSILO
EPSIL2
ITOT1
ITOT2

If you do not specify some of these parameters they are set to default values.

The output

The five statistics

The five statistics calculated are:

- (i) I_t , a statistic based upon the distance to regularity of the scaled totals for the two sets.
- (ii) f , a statistic based upon a summation of the strength of the flows at each sample unit.
- (iii) I_z , a statistic based upon a summation of the direction of the flows at each sample unit.
- (iv) I_m , a statistic that combines the information in Z and f by a suitable scaling.
- (v) C , a statistic that combines flow strength and direction at each sample unit, prior to summation. **IGNORE THIS AND DELETE REFERENCES TO C FROM YOUR OUTPUT - I DO NOT RECOMMEND THE USE OF C - I JUST HAVEN'T BOTHERED TO DELETE IT FROM THE OUTPUT YET.**

I_t is probably most useful if there is noticeable association or dissociation that is caused through the proximity of counts of different species to one another, whilst not necessarily being expressed in coincident units, as explained in the text on p.11 concerning Table 4 in the *Acta Jutlandica* draft. Values of this distance to regularity-based index less than unity indicate dissociation; values greater than unity indicate association.

f is a statistic most closely related to the correlation coefficient, that operates parallel lists of flows, themselves derived directly from the size of the original counts, and which takes no account of spatial information. f is calculated as the sum of contributions over each of the sample units. The value of f is scaled by normalization according to the mean and standard error of the randomized values to give a value that has a standard normal distribution, under the null hypothesis of random placement of one population with respect to the other. Values of the resulting *FLIOSC* index less than zero indicate dissociation; values greater than zero indicate association.

I_z is a statistic that is concerned totally with spatial information, that utilises the strength of the vector flow not the actual flow, and only utilises this as a weight to modulate the directional information. It shares some of the properties of I_t in that it is influenced by proximal as well as coincident units. I_z is not normally considered on its own. It also is calculated as the sum of contributions over each of the sample units, and is scaled similarly to f .

I_m is derived from a combination of the previous two indices, f and I_z . When these two values are scaled, as described below, the value of I_m is essentially the distance from the origin of the projection onto the line: scaled f = scaled I_z .

Currently, I recommend you use I_t and I_m . ***I would not use any of the output for index C, and indeed this will be deleted from later versions.***

Tests

The recommended randomization tests of the above null hypothesis are based on the above indices. They are one-tailed tests, against the alternatives either of association or dissociation, whichever is appropriate, given the value of the index. Full self-explanatory output concerning the randomization probabilities is given on channel 7. The probability printed out in the summary output on channel 9 concerns the proportion of the randomization distribution that had greater values than that observed, and therefore gives for each of the five indices a significance test for association. For a 5% test, if the probability quoted is < 0.05 you may reject in favour of the alternative that there is significant association, whereas if the probability quoted is > 0.95 you may reject in favour of the alternative that there is significant dissociation. For a different size of test adjust the critical probability accordingly, e.g. for a 1% test reject in favour of significant association if probability quoted is < 0.01 and in favour of dissociation if > 0.99 .

Summary output on channel 9

Please note that the values output are simple arithmetic averages of the values gained for each randomization (respectively, of set 1 against observed set 2, and of set 2 against observed of set 1). You should check to see that the results for these two sets of randomizations are sufficiently similar to allow this averaging to be valid. I believe that only in extreme cases would this not be the case, and so far I have encountered none such.

When running the program, make sure that you do not already have files with the names nsao6.dat, nsao7.dat, nsao9.dat or nsao10.dat. If these already exist from previous runs, you should rename or delete them before each new run.

Good luck with the program! I will try to answer any problems you may encounter. Please read the following conditions for use of this program carefully, and sign that you accept them.

Joe N. Perry
12 October 2001

Conditions of use

The term software means all or any part of the code supplied. The copyright in this software is vested in Rothamsted Experimental Station AL5 2JQ UK the employer of its author J.N.Perry. The software is distributed free of charge and supplied to recipients only on the following conditions, all of which must be agreed by the recipients before any use of the software is made.

The software is supplied free on the condition that you accept the conditions of use outlined under the terms of the GNU General Public License version 2 as published by the Free Software Foundation, Inc., 59 Temple Place – Suite 330, Boston, MA 02111-1307, USA. No warranty is given with this distribution and no support is offered for those using this software. While every effort has been made to ensure that this software is free of defects no guarantee can be given as to its accuracy and no liability is accepted by the author J.N. Perry or his employer The BBSRC or IACR Rothamsted or The Lawes Trust for any damage or loss of any form caused by its use.

I hope that the software will be used mainly for research purposes and that recipients will acknowledge its supply in any publication which arises from its use. I would be interested to receive a copy of any such publication.

In accepting free copies of the software **sadica.exe**, the recipient is deemed to accept the above conditions of use.