Rothamsted Research

Patron: Her Majesty The Queen

Harpenden, Hertfordshire AL5 2JQ
Telephone: ***(01582) 763133***
Fax: ***(01582) 760981***
Director of Research: ***Professor I Crute***

Plant & Invertebrate Ecology Division
Acting Head: ***Ian Denholm*** *PhD*

11 August 2002

Dear Colleague,

Thanks for your interest in the SADIE system of spatial analysis. Free software now exists for the SADIE methods to measure and detect clustering, applied to data in the form of counts at specified spatial locations; this documentation covers the software available for 32-bit operating systems such as Windows95 or WindowsNT. The copyright in this software is vested in the employer, Rothamsted Experimental Station, Harpenden, Herts. AL5 2JQ UK, of its author Joe Perry. The software was developed using Microsoft FORTRAN Powerstation; it is supplied as a .EXE file and as source code, although you will not need to use the latter to run the program. Further details are given in the papers referred to in the documentation. The software is supplied free on the condition that you accept the conditions of use outlined in the source code 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 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.

Please note that the SADIE website has recently been expanded and is at: http://www.iacr.bbsrc.ac.uk/pie/sadie. There you can find introductory material, reprints, downloads, a dynamic tutorial, etc.

Best Wishes,

Yours Sincerely,

Joe N. Perry

****

Professor J.N. Perry DSc FIBiol
Plant & Invertebrate Ecology Division
Rothamsted Experimental Station
Harpenden
Herts. AL5 2JQ
UK

email: joe.perry@bbsrc.ac.uk
Fax: +44 1582 760981
Phone: +44 1582 763133 (extn. 2375)

DOCUMENTATION FOR RBRELV13.EXE

The program **rbrelv13np.exe** analyzes the spatial pattern of data that are in the form of spatially-referenced counts. These are counts taken at specified spatial locations, for example numbers of moths in light-traps, numbers of plants in selected quadrats, where the two-dimensional location of the traps and the quadrats are known. It measures and detects the degree of clustering in the data, in the form of patches and gaps. The term cluster means a region of either relatively large counts close to one another in two-dimensional space (i.e. a patch), or of relatively small counts (i.e. a gap). The software uses new methods, termed red-blue techniques, as described in the paper: Perry, J.N., Winder, L., Holland, J.M. & Alston, R.D., (1999), Red-blue plots for detecting clusters in count data, *Ecology Letters*, **2**, 106-113. This software produces output that may be input into various other graphics packages, such as Surfer or Genstat, to produce coloured graphical displays and maps of the clustering in the data. In order to understand the output from the program it is essential that this paper be read. In addition, indices and randomization tests based on previous work are included, specifically those based on the distance to regularity and the distance to crowding, as described in the paper: Perry, J.N., (1998), Measures of spatial pattern for counts, *Ecology*, **79**, 1008-1017.

This is the non-parametric version, in which the counts of the observed data are transformed to twice the value of their ranks, prior to being read into SADIE. For example, the set of counts {0,0,1,2,4,9,16,63,904}, with a mean of 111, would then be transformed to {3,3,6,8,10,12,14,16,18}. This latter set is the non-parametric equivalent of the original. The mean is now 10, and there are four units with count smaller than this and four units with count larger. This new set of counts may be used as input to SADIE, exactly as the original counts would have been. With the new set, there are thus four potential patch units and four potential gap units and the ability of the new data to discriminate spatial pattern is therefore enhanced. Note that the arrangement of the original set, defined by the coordinates of its sample units is retained; it is just the counts that are different.

Hence, if the original data had coordinates:

x	*y*	*count*
1.0	1.0	0
2.0	2.0	0
1.0	2.0	1
2.0	1.0	2
2.0	2.0	4
.	.	.
.	.	.
5.0	6.0	904

the new, transformed, equivalent non-parametric data would have the same coordinates:

x	*y*	*count*
1.0	1.0	3
2.0	2.0	3
1.0	2.0	6
2.0	1.0	8
2.0	2.0	10
.	.	.
.	.	.
5.0	6.0	18

The idea behind the non-parametric approach is that it addresses the problem that for very skew data, there may be relatively few counts greater than the mean, and therefore an inherent difficulty for the method to detect clustering in the form of patchiness. It does this by ‘centering’ the data

about the median. The median of the (old) parametric data becomes the mean of the (new) non-prarametric equivalent data, so, by definition, there are as many values greater than the new mean as there are less than it. However, crucially, in transforming the data it retains the concept that there is information in the arrangement of the counts relative to one another.

Overview of the program structure

The program uses two files for input, reading from channels 5 and 8. The input file for channel 5 must have the name MS-DOS name: rbni5.dat; this contains the raw counts, together with their spatial co-ordinates. Similarly, the file for channel 8 must be called rbni8.dat; this contains two parameters that control program execution. Examples of files rbni5.dat and rbni8.dat accompany the software.

The program produces five files of output, writing to channels 6, 7, 9, 10 & 11. The file rbno6.dat from channel 6 contains a copy of the raw data, the transformed non-parametric set of counts that are to be used as input, the parameter values selected, plus some basic summary statistics of the data. The file rbno7.dat from channel 7 contains the minimal output required for an analysis. The file rbno9.dat from channel 9 contains output for further graphical analyses that must then be cut and pasted as input to some other package. The file rbno10.dat from channel 10 contains the briefest summary of the most important of those indices and probabilities output on channel 7. The file cluster.dat, from channel 11, contains ordered cluster indices in column 3 and corresponding x and y values in columns 1 and 2; you can ignore column 4; this file can be read straight into the SURFER mapping program for mapping and interpolation of the clustering indices. The core of the program is the transportation algorithm for determining the moves to regularity (this was adapted from code kindly supplied by Dr Les Proll of the University of Leeds); the output from this is unnecessary for reporting analyses and its action is made transparent to the user.

Please note: (i) both files rbni5.dat and rbni8.dat must be present in order to run the program, AND (ii) unless you are running the program under SADIEShell, none of the files rbno6.dat, rbno7.dat, rbno9.dat, rbno10.dat or cluster.dat must be present when the program is run. (IN THAT CASE YOU **MUST RENAME OR DELETE ANY EXISTING VERSIONS OF RBNO6.DAT, RBNO7.DAT RBNO9.DAT, RBNO10.DAT AND CLUSTER.DAT BEFORE RUNNING THE PROGRAM, OTHERWISE IT WILL FAIL IMMEDIATELY, WITHOUT GIVING ANY OBVIOUS ERROR MESSAGE.) If you are running the program under SADIEShell then the existing versions of rbno6.dat, rbno7.dat, rbno9.dat, rbno10.dat and cluster.dat will be overwritten, after a warning.**

How to use the program - input

With the above provisos, running the program is easy! This is what you must do.
First, put the *n* records in your data into file rbni5.dat, in the following form:

x co-ordinate 1	*y co-ordinate 1*	*count 1*
x co-ordinate 2	*y co-ordinate 2*	*count 2*
x co-ordinate 3	*y co-ordinate 3*	*count 3*
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
x co-ordinate n	*y co-ordinate n*	*count n*

where the *x* & *y* co-ordinates for each pair, (x_k, y_k), k=1,…,n, should be read in as real numbers and the count, c_k, k=1,…,n, should be read in, on the same line, as an integer, with no decimal point. No more than 2000 records can be analyzed in the version supplied. Note that with the non-parametric option there is no need for you to do anything to your data prior to input; the program generates the transformation to the non-parametric version of your data automatically.

Secondly, specify two parameters in the file rbni8.dat as follows. On line one, specify an integer seed, *iseed*, between 1 and 30,000, for the random number generator. Specifying the same seed in successive runs of the program will generate identical randomizations; specifying a different value will result in different randomizations. On line two, specify an integer value, *k5psim*, between 1 and 153 that will determine the number of randomizations done. The value of *k5psim* relates to how many blocks of 39 randomizations are performed. Within the program, the value of *k5psim* is multiplied by 39 to give the total number of randomizations performed; the result is denoted *nsims* in the output. If you can afford the time required to perform the randomizations, then I recommend that you use the largest value of *k5psim* possible, 153. Hence you should put the two values required into file rbni8.dat in the following form:

iseed
k5psim

How to use the program - output

The file rbno6.dat, created by the program, is a small file. It first tells you the number of records, *n*, there are in your data and then outputs them, giving each a unique reference number. It then gives, in a section marked 'New data', the transformed non-parametric set of counts that are to be used as input. Then the values of the parameters you have given in file rbni8.dat are output, together with *nsims*, the number of randomizations done. Next, some basic summary spatial statistics of the data are printed. First, the *x* and *y* co-ordinates of the centroid of the sample units (the 'middle' of the sample, defined as location *P*, with co-ordinates (x_p , y_p), where $x_p = \Sigma_k x_k / n$, and $y_p = \Sigma_k y_k / n$). Second, the *x* and *y* co-ordinates of the centroid of the counts (the spatial equivalent of the arithmetic mean, defined as location *C*, with co-ordinates (x_c , y_c), where $x_c = \Sigma_k c_k x_k / \Sigma_k c_k$, and $y_c = \Sigma_k c_k y_k / \Sigma_k c_k$). Third, the distance, referred to as δ, between these two centroids, i.e. the distance between *P* and *C*. (See also Perry *et al.* (1996) *Aspects of Applied Biology*, **46**, 95-102, or Perry (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 further discussion of the importance of the distance δ. Fourth, for comparison, the maximum distance between any two sample units is given. Next, some basic numerical summary statistics of the data are given: the sample mean, sample variance, the index of dispersion $(n\text{-}1)\, s^2 / m$, and the total number of individuals in the entire sample.

The file rbno7.dat, created by the program, is a medium-sized file that contains all the important results from the analysis of clustering and spatial pattern.

It first gives the sample mean of the *n* counts.

Next come some results from the analysis based on the distance to regularity. Firstly, *D*, the value for the observed data. Secondly, the value of P_a, the probability that the observed counts are arranged randomly among the given sample units. Thirdly, the mean distance to regularity over the randomizations, i.e. the quantity denoted as E_a in Perry (1998). Fourthly, the index, I_a, computed from $I_a = D / E_a$.

Next comes a longer section, devoted to the measurement and detection of clustering. First, several values are given for each unit, with notation that follows that in Perry *et al.* (1999). There is a row for each unit, ordered by the observed average flow distance, with inflow units, Υ_j, given above and outflow units, Υ_i, given below. The most important value is the standardized clustering index, ν_i or ν_j, given in column five. (An ordered copy of this information is given in the output file cluster.dat. Other columns are annotated and should be self-explanatory.) Second, the mean of these clustering indices over inflows and over outflows, $\bar{\nu}_i$ and $\bar{\nu}_j$, respectively, mentioned in the discussion of Perry *et al*. (1999) are given, together with the equivalent value for all flows. Thirdly, the results of the formal randomization tests of these mean clustering indices are given, also mentioned in the discussion of Perry *et al*., and again done separately for inflows, outflows and all flows. Fourthly, results are given relating to the distribution of clustering indices under the null hypothesis of a random distribution of the observed counts amongst the sample units, i.e. the clustering indices produced by the randomizations. These are as follows. Percentiles are given of the entire set of (NSIMS x number of counts in observed data) randomized clustering indices, so that the observed indices may be assessed against objective criteria. Then, percentiles are given for the distribution of the NSIMS values of the maximum clustering index (both for inflows and outflows), where the distribution is formed from the single value found for each randomization.

Finally, following a caveat that should be taken seriously, some results are given from the analysis based on the distance to crowding, if they are required, in similar format to that for regularity, described above.

The file rbno9.dat, created by the program, is a large file that contains various statistics and results to enable further graphical output. The file begins with the data required to graph the so-called 'initial-and-final' plot (see both papers referenced above). First, all the flows are given in a single

block, in arbitrary order, with the total number of flows at the end; then the flows are given again, but now on a unit by unit basis. Note that the flows referred to in these two sections of output have been almost always been scaled to achieve integer values - to get back to the actual values on the original scale of the counts just divide the scaled flows by *n*, the number of sample units.

Next, information is given, unit by unit, to allow the drawing of the so-called 'vector flow' plot in Fig. 6 of Perry *et al.* (1999). Again the flows are scaled as above. The important information is given in the final two columns, which give the *x*- and *y*- components of the vector for each unit, respectively.

Next, information is given concerning the observed distance to regularity, *D*, and the corresponding value for each of the *nsims* randomizations, as they were generated.

The next block of values in three columns facilitates the drawing of the most important contour map of clustering, such as that in Fig. 3 of Perry *et al.* (1999). The information is essentially the same as that given in file rbno7.dat, i.e. the standardized clustering indices, ν_i or ν_j, given as column three, together with the corresponding *x* and *y* co-ordinates of each unit in the first two columns of the block.

The next block of values gives information to draw the E.D.F. plots given as Figs. 4 & 5 of Perry *et al.* (1999). These plots are, however, probably less useful than those formal probability tests given in file rbno7.dat, that used the means of the clustering indices.

The next block repeats those same observed mean clustering indices, given in file rbno7.dat, and, immediately underneath, gives the equivalent values from each of the *nsims* randomizations. These are the raw values used in the comparison between observed and randomized values that was summarized in the formal randomization tests of clustering given in file rbno7.dat.

Finally, information is given concerning the observed distance to crowding, and the corresponding value for each of the *nsims* randomizations, as they were generated.

The file rbno10.dat, created by the program, is a very small file that contains the three most important indices and their probabilities under the null hypothesis of a random distribution of the observed counts amongst the sample units.

The file cluster.dat contains approximately ordered cluster indices in column 3 and corresponding x and y values in columns 1 and 2; you can ignore column 4. This file can be read straight into the SURFER mapping program for mapping and interpolation of the clustering indices.

Don't forget: when running the program, make sure that you do not already have files with the names rbno6.dat, rbno7.dat rbno9.dat, rbno10.dat or cluster.dat. If these already exist from previous runs, and you are not running under SADIEShell you should rename or delete them before each new run; if you are running under SADIEShell they will be overwritten after a warning.

That is all that is required. Good luck with the program!