

Control Chart

A computer program
by Marti J. Anderson



Department of Statistics
University of Auckland
(2008)

DISCLAIMER

This FORTRAN program is provided without any explicit or implicit warranty of correct functioning. It has been developed as part of a university-based research program. If, however, you should encounter problems with this program, the author will be happy to help solve them. Researchers can use this program for scientific and research purposes, but intellectual property and copyright for the source code and program remains the property of Marti Jane Anderson.

Users of the program may refer to the present user's guide as follows:

Anderson, M.J. (2008). ControlChart: a FORTRAN computer program for calculating control charts for multivariate response data through time, based on a chosen resemblance measure. Department of Statistics, University of Auckland, New Zealand.

Author's contact details:

Dr Marti J. Anderson
Department of Statistics
University of Auckland
Private Bag 92019
Auckland, New Zealand
Tel: 64-9-373-7599 ext 85052
Fax: 64-9-373-7018
Email: mja@stat.auckland.ac.nz
Website: <http://www.stat.auckland.ac.nz/~mja>

I. Description

ControlChart is a computer program for calculating multivariate control charts as described by Anderson & Thompson (2004). It is designed specifically to analyse data consisting of multivariate samples (say, the abundances of multiple species in a community) taken from each of a number of sites of similar habitat type (such as multiple tide-pools, reefs, sites, plots, etc.) and for each of a series of time-points or visits (e.g., time 1, 2, 3, etc.), such as might be done, for example, as part of an ecological or environmental monitoring programme.

The essential idea is to identify sites that might, at a given point in time, deviate from what would be expected, given the temporal variability observed across all of the sites being monitored. For this, two possible measures may be used:

- (i) the deviation of a site's observed community at time t from the centroid obtained using a *baseline* set of observations from that site (e.g., the baseline might consist simply of the first 2 or 3 observations, for example);
- (ii) the deviation of a site's observed community at time t from the centroid obtained using all times up to and including time $(t - 1)$ from that site.

Whereas choosing (i) above focuses the analysis on detecting changes of either a sudden "pulse" or a more gradual "press" nature over longer periods of time, the choice of (ii) above allows the baseline to move with (or be updated by) more recent observations, so will be focused more on detecting "pulse" types of changes.

Next, to identify whether a given deviation is "outside the bounds" of what one might expect, given the temporal variability measured across all of the sites being monitored, a bootstrap algorithm is implemented. More particularly, the values obtained through time are sampled with replacement within each site, and the deviations re-calculated for the bootstrap sample. The 95th (or other) percentile of the distribution of deviations across all sites is obtained for each bootstrap sample. In other words, the bootstrapping is done through time within each site, but the 95th (or other) percentile is then calculated spatially, from the population of all sites. The mean of the bootstrap distribution of percentiles is then given in the output. One may then decide the appropriate upper bound to use for management decisions. For example, one might deem a given site to be "out of control" (i.e., to be deviating substantially enough from what is expected so as to be worthy of more detailed investigation) if its deviation exceeds the 95th percentile. A more conservative criterion would be to use the 90th percentile, which is also given in the output, along with the 75th and 50th percentiles of the bootstrap distribution, for reference.

The above bootstrapping approach is done under the null hypothesis of exchangeability of the multivariate observations through time at a given site, provided that site is indeed in a situation of being "in control" (by which we mean that the site is "bouncing around in multivariate space" in accordance with natural expected levels of temporal variation at the scale of measurement). Whereas the bootstrap distribution for a given percentile is not different for different time-points if one has chosen to use a baseline set of observations as a reference point (i.e., if one has chosen to use (i) above), the percentiles will be time-point specific if one has chosen to use (ii) above.

Note also that the deviation of an observation from its (time-based) centroid is calculated *in the space of the chosen resemblance measure*. Such deviations are *not* the same as distances from arithmetic averages of raw data *unless* there has been no transformation and Euclidean distances are being used as the basis of the analysis. See Anderson & Thompson (2004) for more details regarding the calculations underlying the method and also for further clarification regarding approaches (i) and (ii) above.

II. Input file

Suppose the data have the following characteristics:

- (i) there are a total of s sites (or reefs or tide-pools or plots, or some other spatial unit) that have been monitored through time;
- (ii) there are up to t time points (or visits) when each site has been sampled;

- (iii) the total number of site \times time sample units in the data set is N . It is possible that not all time points were sampled at every site. This is fine and causes no problems for the program.
- (iv) there are p variables (usually species) sampled for each of the N sample units.

The structure of the required input file has the following features:

- (i) The first column should contain integers that specify the individual sites. These need to be integers, not names. The integers need not be strictly sequential, but they should identify each site uniquely and be in increasing order.
- (ii) The second column should contain time points from 1 up to t , where t is the maximum number of visits. Some sites might not have been sampled at all times, which is fine. However, the integers used here should match up with one another for a given time point across all sites that were sampled at that time. For example, if a site was sampled at times 1, 3 and 4, then the integers 1, 3 and 4 should be used to identify these time points, to align its values through time with samples taken from other sites at those particular time points.
- (iii) The response data (of p variables) are provided in the form of a raw multivariate data matrix, with sample units as rows (there will be N of these in total) and variables (e.g., species) as columns. This matrix, having p columns, needs to occupy columns 3 to $(3 + p)$ in the input file.
- (iv) the file should be saved as ASCII *.txt, with no other column or row headings.

An example data file with $p = 4$ species variables might look like the following:

1	1	0	3	15	22
1	2	0	0	1	4
1	3	1	5	0	53
1	5	0	7	6	16
1	6	0	2	10	12
1	7	0	0	0	25
1	8	0	3	0	8
1	10	1	0	3	37
2	3	0	21	1	7
2	4	1	17	5	0
2	5	0	19	2	1
2	6	0	15	1	2
2	7	1	11	0	41
2	8	0	0	0	1
2	9	1	0	0	2
2	10	0	0	0	1

.

.

.

and so on...

Note how site 1 (identified by the integer “1” in column one) is missing an observation for time 4. This is fine. It simply means that when the program does the calculations of percentiles across all sites for time 4, this site will not be included. Next, notice that observations from site 2 did not actually start until time 3. This is also fine. Suppose, however, that the user were to ask the program to calculate deviations of observations from centroids that have been calculated using a baseline of 2 observations. In the case of site 2, the baseline will be calculated from time points 3 and 4 (the first two observations for that site) and deviations will be calculated from each subsequent time point to this baseline set.

To avoid dealing with long file names and paths to locate files, place the input file in the same location on your computer (i.e. the same directory) as the “ControlChart.exe” file, for use with the program. Double-click on the “ControlChart.exe” file to run the program. The user then has several options for transformation, standardization and choice of distance/dissimilarity measure.

ControlChart is written using dynamic allocation of memory and thus (theoretically) there are no limits to the size of the matrix (rows or columns) that is entered for analysis. Please contact the author if problems (other than the time required) are encountered in the analysis of large matrices.

The program outputs the control-chart bounds (percentiles) calculated from the bootstrap distributions, and also the deviations of individual sites from centroids (either a baseline or $(t - 1)$ centroid) through time, which can be imported into any graphics package (or Excel) for plotting (e.g., as shown, for example, in Fig. 5 of Anderson & Thompson 2004).

III. Questions Asked by the Program

The questions asked by the program are best demonstrated by an example. The data analysed here are from a monitoring programme of fish assemblages of the Great Barrier Reef, Australia in the form of annual samples of multiple reefs across the entire region, performed by the Australian Institute of Marine Science. Previous analyses have shown that inner, mid- and outer reefs have strongly differing communities, so should be analysed separately in control charts (e.g., Anderson & Thompson 2004). Here, we shall focus on the fish assemblages from the mid-shelf.

The data file has a total of $N = 168$ sample units (observations) for $p = 113$ fish species. There are $s = 18$ reefs and these were sampled up to $t = 10$ times (years), from 1992 to 2002. These data are contained in an ASCII text file called "reefmid.txt".

The questions asked by the program are given in Courier font, while responses given for this data set are in *Times bold italics*. Note that you can stop the program at any time by typing Ctrl-C (as for any DOS program).

```

CONTROL CHART
-----
A program for examining
deviations of multivariate observations
from centres calculated from previous observations

by M.J. Anderson, University of Auckland (2008)

Do you wish to perform bootstrapping?
1) No
2) Yes - bootstrap time within sites
2

Type the name of the input file containing your data
(Note: the first 2 columns will contain vectors with integers to
identify the following: Site, Visit)
reefmid.txt

Type a name for the output file of results (*.txt)
reefmid_output.txt

Do you wish to output the bootstrap distribution?
1) No
2) Yes
1

(Answer yes here only if you wish to obtain values for the distribution of percentiles under bootstrapping,
as were used, for example, to draw Fig. 4 of Anderson & Thompson 2004.)

Choice of transformation:
1) none
2) square-root
3) fourth-root

```

- 4) $\ln(x)$
- 5) $\ln(x+1)$
- 6) $\log_{10}(x)$
- 7) $\log_{10}(x+1)$
- 8) presence/absence

I

Choice of standardization:

- 1) none
- 2) standardise by row (sample) sums
- 3) standardise by column (variable) sums
- 4) double standardise by row and column sums
- 5) standardise each variable to z-scores (normalize)
- 6) standardise each variable by dividing by its s.d.
- 7) standardise each variable by dividing by its range

I

Choice of distance measure:

- 1) Bray-Curtis dissimilarity
- 2) square root of Bray-Curtis
- 3) Euclidean distance
- 4) Orloci's Chord distance
- 5) Chi-square metric
- 6) Chi-square distance
- 7) Hellinger distance
- 8) Gower dissimilarity
- 9) Gower, excluding double zeros
- 10) Canberra distance
- 11) square root of Canberra distance
- 12) CY dissimilarity
- 13) Deviance based on the binomial
- 14) Deviance per observation (scale invariant)
- 15) Kulczynski dissimilarity
- 16) McArdle-Gower dissimilarity
- 17) Manhattan (city block) distance
- 18) Jaccard dissimilarity

I2

Note: Although the Bray-Curtis distance measure is used by many ecologists for species abundance data, the Chi-squared distance measure, Orloci's chord distance, or other measure may also be appropriate in different situations. For example, the chi-square distance measure has been used extensively and is the basis of correspondence analysis. It will tend to emphasize compositional changes rather more than changes in abundance, compared to the Bray-Curtis measure. The CY distance measure is also useful; it is somewhat akin to using the Bray-Curtis measure on $\log(x+1)$ transformed data, but no log transformation is necessary. The Gower distance is somewhat akin to using the Canberra measure, but where each variable is standardized by its range. See Legendre and Legendre (1998) for an extensive review of various distance measures. Recall also that if you require a different measure, you may input a distance matrix directly. Note that in this case it needs to be a full ($N \times N$) matrix.

You have chosen the CY dissimilarity measure.

This measure modifies zero values by adding a constant before taking logs.

Type the value you want for this constant (e.g. 0.1, 0.9, etc.).

Note: The published value for this is 0.1.

0.9

As described in Anderson & Thompson (2004), it is expected that 0-1 differences should be expected by chance quite frequently in temporal data from the same site. A choice of 0.9 here downplays the importance of presence/absence compared to changes in relative abundance.

How many species variables (columns) are there?

(Note: NOT including the first 2 columns)

113

How many total observations (rows) are there?

168

How many sites are there (total)?

18

What is the maximum number of observation times (visits)?

10

Would you like centroids to be calculated from

1) all times up to time t-1

2) a single specified time

2

Up to (and including) what time would you like centroids to be calculated from?

2

Here, we have chosen to obtain deviations from centroids based on the first two time points of sampling at a given site.

How many random bootstrap samples do you wish to have?

10000

It pays to make this a big number!

Type an integer for the seed of the random number generator

5

Choose any integer you wish. Note that you can obtain identical bootstrap outputs in multiple runs of the program if you choose the same integer here.

```
Currently working on bootstrap sample      1
Currently working on bootstrap sample      2
Currently working on bootstrap sample      3
Currently working on bootstrap sample      4
.
.
.
Currently working on bootstrap sample      10000
```

These results have been sent to the output file.

End of the program.

Press q to quit.

q

The program has completed the analysis and sent results to the named output file. In this case, the output file looks like this:

```
-----
Results CONTROL CHART
-----

Input file name: reefmid.txt
No transformation
No standardization
Analysis based on CY dissimilarities
Value for zero replacement was 0.900
No. of species variables = 113
Total no. of observations = 168
Total no. of sites = 18
Max. no. of observation times (visits) = 10
No. of bootstrap samples = 10000
Integer used for the random seed = 5
```

 Results of the bootstrap

Percentile	Value from bootstrap sample of deviations
95	0.16218571
90	0.14770885
75	0.12627114
50	0.10527276

 No. of observations from which
 percentiles were calculated = 132

 Deviations of observations at time t from centroids
 Centroids were calculated for all times
 up to and including time 2

Site	Time	Obs no.	Deviation
4	4	3	0.06726548
4	6	4	0.10337195
4	7	5	0.09800895
4	8	6	0.08378520
4	9	7	0.10118570
4	10	8	0.08233357
5	3	11	0.07149408
5	4	12	0.10180876
5	5	13	0.11410653
5	6	14	0.10218293
5	7	15	0.15154988
5	8	16	0.11630598
5	9	17	0.14891548
5	10	18	0.09315203
6	3	21	0.09310003
6	4	22	0.13950167
6	5	23	0.10850035
6	6	24	0.11825887
6	7	25	0.19337059
6	8	26	0.17090723
6	9	27	0.19760917
6	10	28	0.14944520
7	4	31	0.09662873
7	5	32	0.11703239
7	6	33	0.11659862
7	7	34	0.18380744
7	8	35	0.16280748
7	10	36	0.10045354
18	3	39	0.09467641
18	4	40	0.08917604
18	5	41	0.11968060
18	6	42	0.12016174
18	7	43	0.15121179
18	8	44	0.11460294
18	9	45	0.11011522
18	10	46	0.10177913
19	4	49	0.13134553
19	5	50	0.14461365
19	6	51	0.09810552
19	7	52	0.12258216
19	8	53	0.08150804
19	9	54	0.13436732
19	10	55	0.12619708
20	4	58	0.08211859

20	5	59	0.09363814
20	6	60	0.08417814
20	7	61	0.14367580
20	8	62	0.11134214
20	9	63	0.10938898
20	10	64	0.12338818
24	4	67	0.13728833
24	5	68	0.08586960
24	6	69	0.10361088
24	7	70	0.11227389
24	8	71	0.14153848
24	9	72	0.11140704
24	10	73	0.16259101
25	3	76	0.10694117
25	4	77	0.10286705
25	5	78	0.08531659
25	6	79	0.10916936
25	7	80	0.14377057
25	8	81	0.13366025
25	9	82	0.13842624
25	10	83	0.11707174
26	3	86	0.12375220
26	4	87	0.12909157
26	5	88	0.11187260
26	6	89	0.12421987
26	7	90	0.14493405
26	8	91	0.10925723
26	9	92	0.10212369
26	10	93	0.14003780
27	3	96	0.11751239
27	4	97	0.12012013
27	5	98	0.11976533
27	6	99	0.13695727
27	7	100	0.12598386
27	8	101	0.12593194
27	9	102	0.10440283
27	10	103	0.15820132
28	4	106	0.10560530
28	5	107	0.10790151
28	6	108	0.13022638
28	7	109	0.10469326
28	8	110	0.15122647
28	9	111	0.13858097
28	10	112	0.15785123
33	4	115	0.10525336
33	5	116	0.11055358
33	6	117	0.11620237
33	7	118	0.05426410
33	8	119	0.08914463
33	9	120	0.11678442
33	10	121	0.11266152
34	4	124	0.10357569
34	5	125	0.13685649
34	6	126	0.09310614
34	7	127	0.10324995
34	8	128	0.12097363
34	9	129	0.08738304
34	10	130	0.08094021
35	3	133	0.09871826
35	4	134	0.09679506
35	5	135	0.11030515
35	6	136	0.10369482
35	7	137	0.14737467
35	8	138	0.10771301
35	9	139	0.10363974
35	10	140	0.09842656

42	4	143	0.13509222
42	5	144	0.11268537
42	6	145	0.16509949
42	7	146	0.19926005
42	8	147	0.17316933
42	9	148	0.15461207
42	10	149	0.15512769
43	3	152	0.09438087
43	4	153	0.09277409
43	5	154	0.16369656
43	6	155	0.12331654
43	7	156	0.15906115
43	8	157	0.13610611
43	9	158	0.14982890
43	10	159	0.14072075
44	3	162	0.10719500
44	4	163	0.12418876
44	6	164	0.15001060
44	7	165	0.15718292
44	8	166	0.14395441
44	9	167	0.13918759
44	10	168	0.15770292

From the above results, it is straightforward to plot the deviations from the baseline centroid through time for each reef. We might use, for example, the 95th percentile from the bootstrap distribution as an upper bound on the deviation values, which has a value of 0.162, based on the CY dissimilarity measure. Thus, any deviations exceeding this value are outside the control-chart bounds and would warrant further investigation to discover potential factors that might be responsible for the change at those time points. For example, the reefs numbered 6 and 7 both exceeded this upper bound at time 7, and remained outside the bounds of what had been previously seen for those sites until time 10.

If a $(t - 1)$ centroid is used instead of a baseline centroid, then the control-chart bounds are specific to individual time-points (they usually start quite high, then settle down to something reasonably consistent), but all other aspects of the interpretation of output from the program are as described above.

IV. References

Anderson, M.J. and Thompson, A.A. (2004). Multivariate control charts for ecological and environmental monitoring. *Ecological Applications* 14: 1921-1935.