# **Percolation Analysis Programs – Workflow and Descriptions**

The key percolation cluster programs are described below. The objective was to create a general tool to be readily useable for other databases, with minimal recoding. To this end key process parameters, and the identity of code files to be used are defined within two text files read in at initialisation.

The cluster generation and analysis programs were created as a series of individual programs that each carried out particular processes. This enabled rapid code development for each element of the overall process, made the processing time of some of the routines manageable, but importantly meant that reprocessing of particular data sets for particular purposes was possible without having to run the whole suite from the beginning.

The original suite of programs was provided by Elsa Arcaute, written in ‘R’ for the analysis of Domesday sites, in collaboration with data provided by Stuart Brookes. This suite had a similar structure to that below, but has been very extensively modified to meet the objectives of this study, and to provide a generic, parameter driven tool.

## **Recent** **Changes**

Changes have been made to the program suite for the following key changes:

1. Fractional values of the radius step value can now be accommodated, e.g. 0.1km
2. A ‘radius\_unit’ parameter has been added to the radius\_values.txt file. The value given will alter the unit value of the radius. A value of 1 means the radius is incremented in metre units, and a value of 1000 in km units. Other values are accommodated (note the assumption is that grid values are provided to the metre).

## **Directory Structure**

The programs and data sit in the following directory structure. This is relative to the working directory. The sub-directories need to be manually set up.

**IT IS NECESSARY TO MODIFY THE ROOT DIRECTORY WITHIN THE PROGRAMS**, currently this is:

setwd("D:/Iron\_Age\_Hillforts/Percolation")

***The ‘R’ program code sits in this root directory.***

/analysis\_results – output plots from the analysis, now principally the cluster size vs radius plots (earlier versions produced other related plots)

/maps – outputs from the mapping, showing clusters and nodes overlaid on given outline shape files, as well as output shape files with the point data

/shape\_files – given shape and related files for mapping plots

/source\_data – given source data, in .csv format (see below)

/working\_data – working data generated by the programs and used to pass intermediate data between them. This also contains the two configuration files: source\_file.txt and radius\_values.txt; see below

## **Source data**

This is a minimal csv file, with the format below:

PlcIndex,Easting,Northing

1,350350,233050

2,354700,260200

3,358700,238900

***PlcIndex*** is an assigned index number to distinguish the points. ***Easting*** and ***Northing*** are self-explanatory, and are to the metre. (Note that additional columns, if any, will be ignored).

## **Configuration files**

Two configuration files are needed. These are located in the /working\_data sub-directory and are in csv format.

***source\_file.txt***

source\_file,map\_name,shape\_file\_name,DTM\_file\_name

GB\_IoM\_coords.v1.csv,Britain,greatbritain.shp,UK\_hw.asc

***source\_file*** is the name of the csv source file

***map\_name*** is the name to be printed on the map plots

***shape\_file\_name*** is the name of the shape file to be used for the map plots

*DTM\_file\_name* is not currently used and the value can be left blank

***radius\_values.txt***

upper\_radius,lower\_radius,step\_value,limit,radius\_unit

40,2,1,50,1000

This file specifies the various parameters for running the percolation programs.

***upper\_radius*** is the upper value of the radius range to be used, in km - Integer

***lower\_radius*** is the lower value of the radius range to be used, in km - Integer

***step\_value*** is the step value to be used between these two values, in km (NOTE: this value can be decimal, e.g. 0.2)

***limit*** is the value above which distances will not be calculated between sites, in km (this is typically above the value at which all nodes are within a single cluster) – see later – Integer

***radius\_unit*** is the scale for computing the radius values. The assumption is that coordinates are in metres. A value of 1000 computes the radius in km, a value of 1 in metres etc.

Note that for large numbers of points, and large number of selected radius values, the computation time can be quite long. I recommend experimenting initially with only a small number of steps, e.g. with a small radius range, or with a large step value, in order to see which range of values is the most interesting. For Hillforts in Britain, I ran with 1 km steps between 2 and 40km, but also ran an analysis for 3-6km with 0.1km steps to investigate clusters in high density areas of Scotland and Wales. To run with 0.1km steps over the full range would not only take excessive time but also produce large numbers of very similar plots.

The *upper and lower thresholds* are values previously used in the analysis program for a computation described an earlier version of the work by Elsa Arcaute, are no longer used.

In the example above, percolation analysis will run for radius values between 40 and 2 km, with a step value of 1km (determined by the radius\_unit parameter, set to 1000). When computing the inter-site distances, values greater than 50km will not be stored. This restricts the intermediate file size and improves performance. (There is no point in computing the distance between sites in Scotland and Cornwall for example, when all sites are within a single cluster above a radius of say 40km).

## **Programs and running order**

Libraries required for each program are loaded at the start of each program. These need to be preloaded into the host system environment.

The programs for the percolation cluster generation and analysis suite are as follows:

***create\_nodes\_list\_d.v4xxx***

Creates the Inter-site distance matrix – given the dataset of sites and XY coordinates, this computes a partial matrix of inter-site distances. The limit is set by the ***limit*** parameter in the configuration file. The inter-site distance is computed using Pythagoras’ theorem.

The output file is located in ***/working\_data*** and is called: *nodes\_list\_d.txt*

It also creates a file listing the nodes that have been used in creating this list, which omits duplicates and null values etc. This is called *PlcIndex.csv*.

Duplicate and null entries are also listed in files with these names, but should have no values if the data is clean. (Historically there was a lot of cleaning to do, which is why this runs as a separate program).

***clustering\_script.v3xxx***

Cluster extraction – once the inter-site distance matrix has been created above, this program identifies all clusters for a specific radius. It steps through a range of radii, the range being set in the configuration file. The data is output as a text csv file called: *member\_cluster\_by\_radius.csv*

The output file is located in ***/working\_data***

This program is the heart of the analysis; the method of creating a graph where nodes are connected when the distance between them is the given radius or less, then extracting each cluster, remains as written by Elsa Arcaute.

***mapping\_clusters.v4xxx***

Mapping clusters – this program maps clusters for each radius on an outline map overlay. The same data can for example be mapped on a coastal outline, an outline with modern counties, or an outline with historic counties. The outline is defined by an ESRI shape file, and is defined in the configuration file by a text string. The outputs generated were originally a multi-page pdf file, but this proved unwieldy, and it was changed to generate png files, which are much easier to import into Microsoft Word and other applications, as well as to display. Point data is also exported as an ESRI shape file for import directly into ArcGIS etc. Outputs are placed in the ***/maps*** sub-directory.

Clusters are ranked in order of size for each different threshold radius. The top 15 clusters are assigned a colour code with Red the largest, Blue the next and so on. A legend is plotted with the assigned cluster number against the colour. The plots also include the map name and the name of the source file used to generate the data, as well as the radius value.

NOTE: the map projection is extracted from the given shape file and used for configuring the point plots and output shape files. This has not been extensively tested for different projection/ coordinate systems, and may need further development/modification for different input shape files. If necessary experiment by omitting the shape file plot by commenting out, and simply plot the points, in the first instance.

***cluster\_frequency\_script.v2xxx***

Percolation cluster size and rank analyses – this program processes the cluster data generated by the cluster extraction program to give various different frequency and ranking plots. This program evolved significantly as various different statistical analyses were experimented with, under the guidance of Mark Lake. The final version reflects the latest work by Elsa Arcaute. The outputs are placed in the /analysis\_results sub-directory.

The most useful plot is the cluster transition plot, showing the maximum cluster size (normalized) vs. radius, which shows the transitions.

# CONTACT DETAILS:

Simon Maddison

[simon.maddison@btinternet.com](mailto:smaddison@iee.org)

+44 797 089 2171