Percolation Analysis – Archaeological applications at widely different spatial scales

-Maddison, Simon -Schmidt, Sophie C.

19 Juli, 2019

This paper provides a summary of the Percolation Analysis method, and then provides two detailed case studies where the technique is applied at significantly different geographical scales. This will demonstrate not only the potential for the technique within archaeology, but also how it can be applied at different spatial scales with different objectives appropriate to the specific problem in question. The technique, originally developed in physics and more recently adopted in Geography, is a way of identifying groupings or clusters, purely based on spatial separation using Euclidian distance. In this paper, percolation Analysis is being applied for investigating the distribution of Hillforts in Britain and Ireland, and Domesday sites in England, identifying clusters and groupings at a national scale. The investigations for British Hillforts and Domesday sites have shown this to be a fruitful starting point for investigating possible socio-political entities, and patterns of deep history. The technique has now also been applied to features at a sub-regional level, in Saxony-Anhalt, middle Germany, with the different objective of identifying settlement sites along a 13km strip excavation. Aim here was to arrive at estimates of settlement sizes, which in turn can inform landscape archaeological surveys on the decision on attributing single finds to an already known site or registering a new site.

### Background

This proposal outline covers a paper for submission to the Journal of the CAA, and a presentation at the German Chapter of the CAA to be held in Wilhelmshaven in September 2019.

## Introduction

Percolation Analysis has been recently applied in Archaeology for identifying clusters and groupings at a national scale, for example investigating the distribution of Hillforts in Britain and Ireland, and Domesday sites in England (references to papers at CAA 2018/2019, book chapter, paper in Atlas Conference etc.). The technique, originally developed in physics and more recently adopted in Geography (references), is a way of identifying groupings or clusters, purely based on spatial separation using Euclidian distance. The above investigations for British Hillforts and Domesday sites have shown this to be a fruitful starting point for investigating possible socio-political entities, and patterns of deep history. The technique has now also been applied to features at a sub-regional level, in Saxony-Anhalt, middle Germany, with the different objective of identifying settlement sites along a 13km strip excavation. Aim here was to arrive at estimates of settlement sizes, which in turn can inform landscape archaeological surveys on the decision on attributing single finds to an already known site or registering a new site. This paper provides a summary of the Percolation Analysis method, and then provides two detailed case studies where the technique is applied at significantly different geographical scales. This will demonstrate not only the potential for the technique within archaeology, but also how it can be applied at different spatial scales with different objectives appropriate to the specific problem in question.

## Methodology

Percolation Analysis is a technique for identifying clusters within a set of spatially arranged points. The computation is based on the Euclidian distance between those points. Percolation analysis is as a way of mathematically describing clusters of spatially arranged points and analysing their related behaviour. A ready example is the propagation of fire through a forest, where fire will spread from tree to tree within a cluster, when the trees are close enough together, but at a critical density when all the trees form a single cluster it can spread to the forest limits (reference). A cluster is based on a defined distance threshold, so that for any given point all neighbouring points falling within this threshold are part of the cluster. The test is then re-applied for each of these neighbours in turn, and any further points meeting this criterion are also part of the cluster. This technique can be applied at any scale, from the molecular to the geographical and beyond. Originally developed for describing the physical properties of materials, Percolation Analysis has recently been used in geography to identify metropolitan areas, based on population density. The City Clustering Algorithm (CCA) has been developed out of percolation theory by Rozenfeld et al. (2008) based on distance within a cellular lattice and further developed by Rozenfeld, et al. (2011) to use the Euclidean distance between points. The technique is illustrated in Figure tba, where an arbitrary point is selected as a start and any point falling within a defined threshold distance ‘l’ becomes part of the cluster and the process is then re-applied to each of these points in turn, until the cluster can extend no further. This threshold distance is also known as the ‘percolation radius’ and this term is generally used below. Arcaute et al. (2016) have adopted this technique for defining urban areas, using the density of street interconnections rather than population. They have also developed analytical techniques for identifying transition points in cluster growth as the distance threshold is progressively increased, as discussed further below. Note that the method has evolved from statistical techniques in materials science, where it is applied to sets of spatially distributed points which are identical and where there is no interest in distinguishing between the points. However when used in archaeology there is potentially a very great interest in the individual points, reflecting as they do distinct archaeological entities, as will be shown later.

### Comparison to other clustering algorithms

A number of spatial clustering algorithms are known in spatial data mining (Neethu - Surendran 2013, Varghese et al 2014), used in archaeology has been mostly the k-means non-hierarchical clustering algorithm, developed by Kintight and Ammermann in 1982 (Baxter 1994 [2015a], 148). This algorithm has been applied widely in the last 40 years in archaeology (to do: many examples see Baxter 2015b), probably because of its ease in use and understanding: At the beginning the whole data set is considered as one cluster and the centre of this cluster is determined. The point furthest from this centre is then used as a centroid for a new cluster and attracts those points to its cluster that are nearer to it than to the original cluster (Baxter 1994, [2015a], 149). Step by step the algorithm determines new cluster centers and allocates each point into one of k clusters in such a way, that the sum-squared error (SSE) is minimized. The SSE is the sum of the squared distances from each point to the center of the cluster to which it is assigned (Kintight 1990, 185). By minimizing global sum-squared errors the algorithm tends to create circular clusters (Kintight 1990, 190). Another point of critique has been the need to define the number of clusters (k) that are being created beforehand. To navigate some problems of k-means and associated algorithms, which have been highlighted in the years from Kintight (1990) to Ducke (2015) and in parts mitigated by Baxter (2015b), Caspari and Jendryke developed Achsphere as an archaeological cluster algorithm (Caspari – Jendryke 2017). They integrate structural parameters, which were recorded together with the spatial information of the studied features. Because they analysed mostly round structures, the features were weighted using the diameter, though other parameters might be used. (Caspari – Jendryke 2017, 182). This weighing of the influence of the features after archaeological pre-consideration was implemented to mitigate the effect of different densities in the same data set (Caspari – Jendryke 2017, 182). Around each feature a sphere is constructed, which size is based on the weighing factor. Two features will belong to the same cluster, if their spheres touch or overlap. The clustering algorithm starts at an arbitrary node, checks all neighbours (touching spheres) and stops only, when no new neighbours are found. Inputs needed for the algorithm are a minimum number of points for a cluster and the the weight (Caspari – Jendryke 2017, 184). It is considered by the authors not to be a “generic cluster algorithm but a specific method to divide landscape archaeological datasets into sub-entities for further analysis assuming that spherical structuring dominates the data.” (Caspari – Jendryke 2017, 186).

Maybe some words on DBSCAN here? I need to read up on it still…

Other algorithms, such as Nearest Neighbour, Ripley’s K-function, the F- and G-functions are clustering detection algorithms, that tell about whether the point distribution is spatially clustered but don’t directly give an attribution of points to a certain cluster. We believe percolation analysis to be a worthy contribution to the toolkit of clustering algorithms for archaeologists, because it is easily understandable and applicable. Available in the free and open source scripting language R it is easily replicable for any researcher, who might want to implement it.

Ducke formulated the following parameters for an useful spatial cluster detector: Its mathematical core should be simple, so that archaeologists can understand why a certain input leads to a certain output, it should assume as little as possible about the clustering structure (i.e., the number, sizes and locations of clusters), it should be robust against noise, i.e., it should just ignore the odd scattering of sherds around the actual clusters, and it should not assume that every point is indeed part of a cluster, it should either assign each point to a cluster or classify it as noise, and finally, it should always produce the same result when given the same input data. (Ducke 2015, 357-358).

We think percolation analysis fits well these desiderata, as its mathematical core is simple; size, number and location of clusters is explored not predetermined, it will include outliers only at very late stages of the clustering process (which is easily detected and may be excluded from analysis results if so chosen) and… well… do we always get the very same results? Hm, I think as the first seed is always chosen randomly, we may not get the very same results every time. What are your experiences here?

## Case Studies

### Hillforts in Britain

Working with data from the ‘Atlas of Hillforts’ (reference) it was decided to use Percolation Analysis to establish any ‘natural’ or intrinsic groupings of hillforts that might exist in Britain and Ireland. Once identified, these groupings could then be compared with topography and geographical regions, as well as other historical data sets, and further analysed with data from the Atlas. The Atlas dataset comprises 3xxx confirmed sites in Britain and xxx in all of Ireland. Percolation Analysis was applied in radius increments of 1km. The cluster transitions for Britain are shown in Figure xxx, which plots the normalized maximum cluster size (with a normalized size range of between 0 and 1.0) against percolation radius. For large radius values all sites will form a single cluster while for progressively smaller radii, the clusters start to break up, and this transition diagram is an indicator of where potentially interesting clusters become visible. For small radius values no clusters form at all with the most notably obvious transition occurring at 34 km, as discussed below. Geographical plots of clusters are shown in Figures xxx. Clusters are ranked by size and allocated a colour, with red being the largest cluster and blue the next, below rank 15 all sites are plotted as grey. Above 35km all sites form a single large cluster, except for the Hebrides, Shetland Islands, Isle of Man and Scillies. Apart from south-east Scotland, the most interesting transitions occur in the 6-13km range. At 34km (Figure xxx) the predominantly Scottish cluster includes sites in England as far south as a line roughly between Morecambe and Flamborough Head. Within England as the radius values reduce, sites in the Pennines and the east progressively break out of the bigger cluster, and at 14km, the south-west peninsula forms its own cluster in Cornwall and part of Devon, and other clusters appear in the south-east. The plot for 12km (Figure xxx) shows for example Cornwall and Devon/part of Somerset as individual clusters, and a cluster along the Chilterns. The plot for 9km (Figure xxx) shows clusters in north-west Wales, the Clwydian Range, south-west Wales, the Gower, central Wales and the Marches and two clusters on the north-west and the south-east of the Severn Valley, the latter being the Cotswolds group. Some of these clusters have been the subject of more detailed analysis and three of them are discussed below. An important point is that there is nothing necessarily special about particular values of the percolation radius, so that transitions will not necessarily occur for the same values in different regions or countries. There are various possible explanations for this, such as the variability in regional and local topography. In areas of high density such as south-west Wales and south-east Scotland, then a finer grained analysis with radius increments of 0.1km reveals more localised cluster patterns in the range 3-7km radius values. Further study is required to establish what significance these might have. (add figure??)

### Strip excavation in Germany

The second case study is located in Saxony-Anhalt, middle Germany. During the early 2000s the street “Bundesstraße 6” was built near the town of Köthen. The area of the street, approach roads, exits and retention reservoirs were machine test trenched by the Heritage Management to find archaeological sites. Along the ca. 13 km long and mostly 40 m broad part of the street over 6000 features were uncovered (Fahr in prep.), but this paper will focus on features from the later Bronze Age (group “Saalemündungsgruppe”). For this periods the aim is to find the clusters which delineate a site with the help of percolation analysis. The radius will estimate the distance between features that belong together into one site.

There are 1006 settlement features cataloged as belonging to this period.

Because of the test trenching it is assumed, that all features and sites along the transect have been found, even though there are gaps between the actual excavation areas and many features remain undated. Machine trenching has been shown to be the most effective way to find sites (Hey 2006).

The term “site” has been at the center of a lively debate in survey archaeology over the last 40 years. Until the 70’s it has mostly been defined as a place (Dunnell 1992, 23–24) - a find spot, which is similar to German archaeology, where the term used, “Fundplatz”, means exactly that. In the German archaeological inventories, though, a site may contain several culturally distinct loci, “Fundstellen” (Dauber 1950, 96), which are usually cataloged as a settlement, a burial ground, a hoard or as a single find spot of a period (Eggert 2005, 56). Two questions arise, when deciding how to classify a site as a settlement: Qualitatively, what kind of finds or features define a settlement (Linke 1976, 8), and quantitatively, of how far apart two “find spots” are supposed to be. If two sherds of the same time period are found 50 m apart - are these two sites or belong they to one site? In every larger settlement or landscape archaeological study this topic is being discussed and in the end a decision felled, which the authors agree, is kind of arbitrary and should actually rely on empiric studies (Malmer 1962, 258; Mischka 2007, 49–50). Since the middle of the 1970s there has been the discussion of a non- or off-site approach to anglophone archaeology. The terms were introduced by Thomas (1975) and Foley (1981) and describe the analysis of finds “between sites”. This approach understands the archaeological record as a continual distribution in space, with spots of higher densities, the so-called “sites”, and areas of lower density (Wobst 1983, 39).

At this point the continuously excavated data set of the street and the clustering analysis percolation can be fruitfully connected: As percolation works with radii around points it may assist in quantitatively assessing which distances between features are common and what distance features don’t seem to belong to each other anymore. This question needs to be answered for each archaeological culture seperately as we cannot assume similar settlement distributions through time(Schirren 1997, 30).

Assuming a connection between underground features and surface finds enables us to transfer the distances of the features to those of finds and may inform landscape archaeological surveys on the decision on attributing single finds to an already known site or registering a new site. Of course, settlement features only comprise the “build space” of a settlement. Though features in-between sites may exist, like traps or temporary storage facilities, places like working areas or fields which may be detected by off-site surface find distributions, won’t be included.

### Percolation

As this analysis is concerned with a very fine level of detail, the starting point and the step size of the percolation algorithm was chosen to be 1m. The size of a prehistoric village in middle Germany can’t be assumed to be larger than a kilometer in diameter, therefore the largest radius used was 500m.

There are several levels of clustering. Features of the late Bronze Age create a steep curve in the beginning, which shows a close clustering at low percolation radii. Longer stretches of flattening might be interpreted as the algorithm running out of features to add to the clusters. We thereby find the edges of settlements and can define their size. The step-wise rises at higher percolation radii needs explanation as well. To provide these, the results will be compared to a large scale excavation of a late bronze age site in middle Germany.

I highlighted those points that mark the most distinctive steps in the distribution. We can see the radii after which the algorithm doesn’t change the cluster size for some time, showing that here stretches of space between features exist. The turning point of the mean cluster sizes is at the radius of 68m.

### Interpretation

The late bronze age features are mostly concentrated in the western half and on the eastern-most part of the transsect. As can be seen by the rapidly rising graph describing the mean cluster size in relation to the percolation radius, the features cluster closely to each other. First steps in the distribution can be seen at 46 m, 68 m and 88 m. These are the distances, at which the clusters “stabilise” and don’t grow with growing percolation radii. As mentioned above, the mean width of the excavated strip was 40m. Therefore it is assumed, that the first “stabilisation” is due to the lack of features beyond the edges of the excavation area.

The largest excavation site of the late bronze age in middle Germany is to be found in Zwenkau(-Eythra) in Saxony, which was excavated in advance of coal mining. At Zwenkau several single farmsteads delimited by houses and fences spread over the whole excavation area, which measures at its greatest dimensions 1100 x 825 m (Huth/Stäuble 1998, 194.213). The farmsteads in Zwenkau don’t overlap (Huth/Stäuble 1998, 214) and the distances to each other are quite diverse. Nonetheless a number of them seem to scatter at at intervalls of ca. 65-75 m. (cf. Huth/Stäuble 1998, 211, fig. 6). The farmsteads themselves are not all the same size, but the longest sides show to be about 50 - 85m long (cumulating to sizes of 2000 - 2500 m²) (Huth/Stäuble 1998, 213.). It is assumed not all of them existed at the same time, but that there were always several isolated farmsteads in a scattered distribution to each other (Huth/Stäuble 1998, 214).

At the analysed transect excavation we also deal with a depth in time, as there are overlapping features and several ditches, which most probably did not exist at the same time, but can be interpreted as evidence of small-scale settlement shifts. We can therefore assume that at least partly the high density of settlement features belonging to the late bronze age might be interpreted as evidence of several phases or local re-settlement. Nonetheless we can notice that the turning points of the radius vector by 68m falls well into the range of the farmstead sizes of Zwenkau-Eythra and might be a possible farmstead site delimiter by the percolation algorithm.

We note: As long as the next feature is 68m away, features seem to belong to the same cluster. We can therefore assume, late bronze age finds at a distance of 68m to each other to belong to the same site.

Mapping the clusters reveals a more complete picture: As the percolation algorithm moves from node to node, the cluster groups are much larger than the radii. Should “in-between” finds be missing, still two findspots belonging to the same site might not identified as to belong to the same site.

Nonetheless the question posed by Malmer and Mischka (Malmer 1962; Mischka 2007) on how to define the distance between to findspots before delimiting a new site has been answered in an empirical fashion as demanded.

# Conclusion

Percolation analysis offers…

* great at different scales
* different research questions might be answered:
* cluster sizes, density of points inside a cluster
* …

# Acknowledgements

##### pagebreak

# References

Dauber 1950: A. Dauber, Der Forschungsstand als innere Gültigkeitsgrenze der Fundkarte. in: H. Kirchner (Hrsg.) Ur-und Frühgeschichte als historische Wissenschaft: Festschrift zum 60. Geburtstag von Ernst Wahle (Heidelberg 1950) 94–111.

Dunnell 1992: R.C. Dunnell, The notion site. in: J. Rossignol/L. Wandsnider (Hrsg.) Space, time, and archaeological landscapes. Interdisciplinary contributions to archaeology (New York 1992) 21–41 <<https://link.springer.com/chapter/10.1007/978-1-4899-2450-6_2>>(26.07.2017).

Eggert 2005: M.K. Eggert, Prähistorische Archäologie. Konzepte und Methoden 2(Tübingen und Basel 2005).

Foley 1981: R. Foley, Off-site archaeology: an alternative approach for the short-sited. in: I. Hodder/G. Isaac/N. Hammond (Hrsg.) Pattern of the past. Studies in honour of David Clarke (Cambridge 1981) 157–184.

Hey 2006: G. Hey, Scale and Archaeological Evaluation: What are We Looking For? in: G. Lock/B.L. Molyneaux (Hrsg.) Confronting Scale in Archaeology. Issues of Theory and Practice (New York 2006) 114–127.

Huth/Stäuble 1998: C. Huth/H. Stäuble, Ländliche Siedlungen der Bronzezeit und der älteren Eisenzeit. Ein Zwischenbericht aus Zwenkau. in: H. Küster/A. Lang/A. Schauer (Hrsg.) Archäologische Forschungen in urgeschichtlichen Siedlungslandschaften. Festschr. G. Kossack. Regensburger Beitr. Prähist. Arch. (5) (Regensburg 1998) 185–230.

Linke 1976: W. Linke, Frühes Bauerntum und geographische Umwelt. Ein historisch-geographische Untersuchung westfälischer und nordhessischer Bördenlandschaften. Bochumer Geographische Arbeiten (278) (Paderborn 1976).

Malmer 1962: M.P. Malmer, Jungneolithische Studien. Acta Archaeologica Lundensia (8, 2) (Bonn 1962).

Mischka 2007: D. Mischka, Methodische Aspekte zur Rekonstruktion Prähistorischer Siedlungsmuster. Landschaftsgenese vom Ende des Neolithikums bis zur Eisenzeit im Gebiet des südlichen Oberrheins. Freiburger Archäologische Studien (5) (Rahden/Westf. 2007).

Schirren 1997: C.M. Schirren, Studien zur Trichterbecherkultur in Südostholstein. Universitätsforschungen zur Prähistorischen Archäologie (42) (Bonn 1997).

Thomas 1975: D.H. Thomas, Nonsite Sampling in Archaeology: Up the Creek Without a Site? in: J.W. Mueller (Hrsg.) Sampling in Archaeology (Tucson 1975) 61–81.

Wobst 1983: H.M. Wobst, We Can’t See the Forest for the Trees: Sampling and the Shapes of Archaeological Distributions. in: J.A. Moore/A.S. Keene (Hrsg.) Archaeological Hammers and Theories. Studies in Archaeology (New York 1983) 38–85 <<http://www.sciencedirect.com/science/book/9780125059800>>.

##### pagebreak

### Colophon

This report was generated on 2019-07-19 14:04:30 using the following computational environment and dependencies:

#> ─ Session info ──────────────────────────────────────────────────────────  
#> setting value   
#> version R version 3.4.4 (2018-03-15)  
#> os Ubuntu 16.04.6 LTS   
#> system x86\_64, linux-gnu   
#> ui X11   
#> language de\_DE   
#> collate de\_DE.UTF-8   
#> ctype de\_DE.UTF-8   
#> tz Europe/Berlin   
#> date 2019-07-19   
#>   
#> ─ Packages ──────────────────────────────────────────────────────────────  
#> package \* version date lib source   
#> assertthat 0.2.1 2019-03-21 [1] CRAN (R 3.4.4)  
#> backports 1.1.4 2019-04-10 [1] CRAN (R 3.4.4)  
#> bookdown 0.10 2019-05-10 [1] CRAN (R 3.4.4)  
#> callr 3.2.0 2019-03-15 [1] CRAN (R 3.4.4)  
#> cli 1.1.0 2019-03-19 [1] CRAN (R 3.4.4)  
#> crayon 1.3.4 2017-09-16 [1] CRAN (R 3.4.4)  
#> desc 1.2.0 2018-05-01 [1] CRAN (R 3.4.4)  
#> devtools 2.0.2 2019-04-08 [1] CRAN (R 3.4.4)  
#> digest 0.6.18 2018-10-10 [1] CRAN (R 3.4.4)  
#> evaluate 0.13 2019-02-12 [1] CRAN (R 3.4.4)  
#> fs 1.3.1 2019-05-06 [1] CRAN (R 3.4.4)  
#> glue 1.3.1 2019-03-12 [1] CRAN (R 3.4.4)  
#> htmltools 0.3.6 2017-04-28 [1] CRAN (R 3.4.4)  
#> knitr 1.22 2019-03-08 [1] CRAN (R 3.4.4)  
#> magrittr 1.5 2014-11-22 [1] CRAN (R 3.4.4)  
#> memoise 1.1.0 2017-04-21 [1] CRAN (R 3.4.4)  
#> percolation \* 0.0.0.9000 2019-07-17 [1] local   
#> pkgbuild 1.0.3 2019-03-20 [1] CRAN (R 3.4.4)  
#> pkgload 1.0.2 2018-10-29 [1] CRAN (R 3.4.4)  
#> prettyunits 1.0.2 2015-07-13 [1] CRAN (R 3.4.4)  
#> processx 3.3.1 2019-05-08 [1] CRAN (R 3.4.4)  
#> ps 1.3.0 2018-12-21 [1] CRAN (R 3.4.4)  
#> R6 2.4.0 2019-02-14 [1] CRAN (R 3.4.4)  
#> Rcpp 1.0.1 2019-03-17 [1] CRAN (R 3.4.4)  
#> remotes 2.0.4 2019-04-10 [1] CRAN (R 3.4.4)  
#> rlang 0.3.4 2019-04-07 [1] CRAN (R 3.4.4)  
#> rmarkdown 1.12 2019-03-14 [1] CRAN (R 3.4.4)  
#> rprojroot 1.3-2 2018-01-03 [1] CRAN (R 3.4.4)  
#> sessioninfo 1.1.1 2018-11-05 [1] CRAN (R 3.4.4)  
#> stringi 1.4.3 2019-03-12 [1] CRAN (R 3.4.4)  
#> stringr 1.4.0 2019-02-10 [1] CRAN (R 3.4.4)  
#> testthat 0.11.0 2015-10-14 [3] CRAN (R 3.2.2)  
#> usethis 1.5.0 2019-04-07 [1] CRAN (R 3.4.4)  
#> withr 2.1.2 2018-03-15 [1] CRAN (R 3.4.4)  
#> xfun 0.7 2019-05-14 [1] CRAN (R 3.4.4)  
#> yaml 2.2.0 2018-07-25 [1] CRAN (R 3.4.4)  
#>   
#> [1] /home/sophie/R/x86\_64-pc-linux-gnu-library/3.4  
#> [2] /usr/local/lib/R/site-library  
#> [3] /usr/lib/R/site-library  
#> [4] /usr/lib/R/library