CASE STUDY

Percolation Analysis – Archaeological Applications at Widely Different Spatial Scales

M. Simon Maddison[*] and Sophie C. Schmidt[†]

This paper provides a summary of the percolation method of cluster analysis, and then describes two detailed archaeological case studies where the technique is applied at significantly different geographical scales. This demonstrates 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. The paper includes a discussion comparing percolation analysis with other clustering techniques already established within archaeology. The first case study uses percolation analysis as an exploratory tool for investigating the distribution of hillforts in Britain, identifying clusters and groupings at a national scale. The purpose is to identify possible socio-political entities for further investigation. In the second case study, the technique is applied to excavated features at a sub-regional level, in Saxony-Anhalt, central Germany, with the different objective of identifying settlement sites along a 13 km strip excavation. The aim here was to arrive at estimates of settlement sizes, which in turn can inform landscape archaeological surveys for deciding on the attribution of single finds to an already known site or registering a new site. Percolation analysis is shown to be an effective tool for analysing archaeological data sets at widely different geographical scales and should become a standard part of the archaeologist's spatial analysis toolkit.

1. 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 (Maddison 2019) and Domesday sites in England in 1086 AD (Arcaute et al. forthcoming). The technique, originally developed in physics and more recently adopted in geography (Arcaute et al. 2016) is a way of identifying groupings or clusters, purely based on spatial separation using Euclidian distance. The above analyses for British Hillforts and Domesday sites have shown this to be a fruitful starting point for investigating possible past 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, central Germany, with the different objective of identifying settlement sites along a 13 km strip excavation. The aim here was to arrive at estimates of settlement sizes, in order to inform landscape archaeological surveys for deciding either on the attribution of single finds to an already

known site or registering a new site. This paper provides a summary of the percolation analysis method, compares it to other known cluster algorithms in archaeology 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, particularly as an exploratory tool, but also how it can be applied at different spatial scales with distinct objectives appropriate to the specific problem in question.

2. 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. It was first developed in physics in the 1940s to describe polymer formation processes and the percolation of a liquid through a solid body; it has since been used for a wide range of applications and analyses (e.g. Frisch and Hammersley 1963; Stauffer and Aharony 1991). A ready example is the propagation of blight through an orchard, where disease 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 orchard limits (Frisch and Hammersley 1963). A cluster is based on a defined

* Independent Researcher, GB

† Universität zu Köln, DE

Corresponding author: M. Simon Maddison
(simon.maddison@btinternet.com)

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.

Recently it has 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; it has been further developed by Rozenfeld et al. (2011) to use the Euclidean distance between points. The technique is illustrated in **Figure 1**, where an arbitrary point is selected as a start and any point falling within a defined threshold distance 'r' becomes part of the cluster; 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 them. 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.

The algorithm, which was developed by Elsa Arcaute in R, has been adapted for archaeology by Simon Maddison and is now available as an R package under DOI 10.17605/OSF.IO/7EXTC. A detailed description of the code package is included at the end of this paper.

2.1. Comparison to other Clustering Algorithms

A number of spatial clustering algorithms are known in spatial data mining (Neethu and Surendran 2012; Varghese, Unnikrishnan and Jacob 2014). The one mostly used in archaeology has been the k-means non-hierarchical clustering algorithm. It was developed in the 1950s and 60s and popularised in archaeology by Kintigh and Ammerman in 1982 (Baxter 2015b: 148; Kintigh and Ammerman 1982). This algorithm has been applied widely in archaeology during the last 40 years (Baxter 2015c; e.g. Enloe, David and Hare 1994; for overviews see: Koetje 1994; Ladefoged and Pearson 2000; Lemke 2013; Savage 1997), probably because of its ease of use and understanding (Ducke 2015: 360): initially the whole data set is considered as one cluster and the centre of this cluster is determined. The point farthest 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. Step by step the algorithm determines new cluster centres 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 centre of the cluster to which it is assigned (Kintigh 1990: 185). By minimizing global sum-squared errors the algorithm tends to create circular clusters (Kintigh 1990: 190). This problem has been described by Baxter as the "well-advertised 'problem' of k-means". He continues "it tends to produce circular clusters that don't necessarily
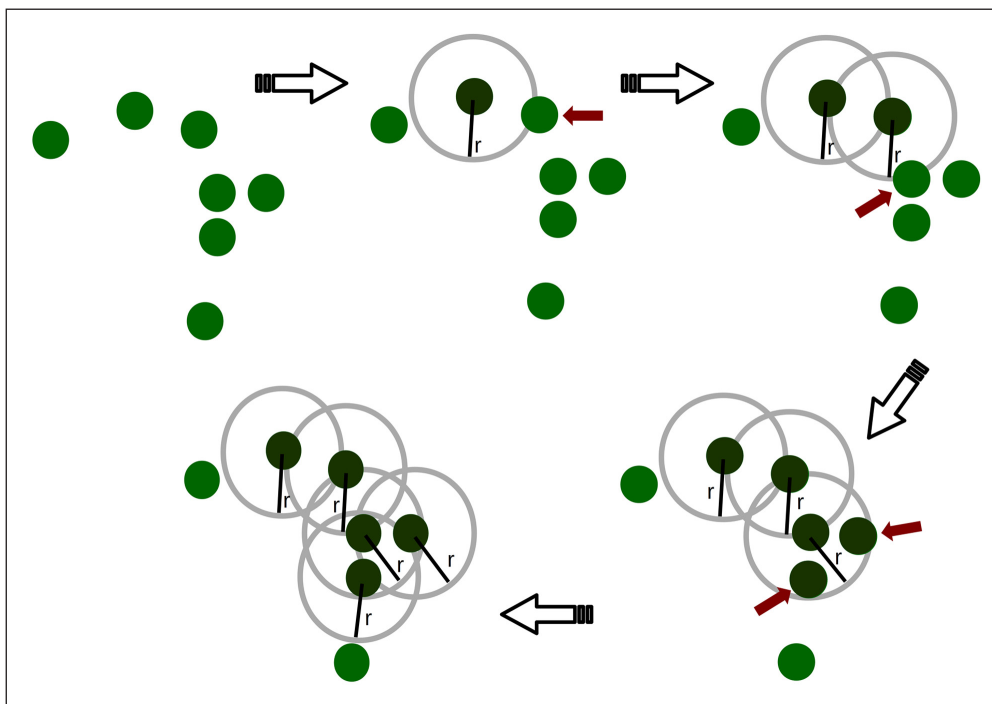
Figure 1: Percolation of a small cluster of points. A radius is drawn around a random point; if another point falls into this radius, it is now "infected" and in turn a radius is drawn around it. If no more points fall within that radius, the cluster is defined.

respect the observable 'geometry' of a point scatter, [this] can be understood as a consequence of its close relationship to a model-based methodology that is optimal if clusters are spherical and of equal size." (Baxter 2015c: 3). Ducke focuses on the centroid-centred approach of this model and sees the result of k-means as a complete and overlap-free partitioning of the space, which corresponds to Voronoi-diagrams. Voronoi-diagrams partition space according to the Euclidean distance and draw borders exactly in the middle between two centroids (Nakoinz and Knitter 2016: 98). According to Ducke, this is a "limitation" (Ducke 2015: 360–61). Baxter, on the other hand, concludes: "If clusters are tightly-defined and well-separated there is not a problem; a representation in terms of convex hulls rather than Voronoi diagrams will emphasise that there is plenty of empty unoccupied space" (Baxter 2015c: 18). Two other points of critique have been the need to define the number of clusters (k) that are being created beforehand and that noise is not excluded from cluster allocation, problems Ducke and Baxter agree on (Baxter 2015c: 18; Baxter 2015a: 161; Ducke 2015: 361).

To address some of the problems of k-means and associated algorithms, as highlighted from Kintigh (1990) to Ducke (2015) and in part mitigated by Baxter (2015c), Jendryke and Caspari (2017) developed Archsphere as an archaeological cluster algorithm. They integrate structural parameters, which were recorded along with the spatial information of the features under investigation. Because they mainly analysed round structures, the features were weighted using the diameter, though other parameters might be used (Jendryke and Caspari 2017: 182). This weighting of the influence of the features after archaeological pre-consideration was implemented to mitigate the effect of different densities in the same data set (Jendryke and Caspari 2017: 182). Around each feature a sphere is constructed, whose size is based on the weighting factor. Two features will be in 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 weight (Jendryke and Caspari 2017: 184). This is not considered by the authors 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" (Jendryke and Caspari 2017: 186).

Less well known and somewhat similar to percolation analysis, in using a radius as measurement for spatial distance, are the 'density-based spatial clustering of applications with noise', coined DBscan (Ester et al. 1996) and the local Ripley's K-function (Getis and Franklin 1987). DBscan, similarly to percolation, uses a radius, which is called the Eps (ε) -neighbourhood of a point, to determine whether other points fall into the same cluster (Ester et al. 1996: 227). Points that lie directly in the Eps-neighbourhood of the starting point are called 'density-reachable'. They will belong in the cluster and themselves become points from which to draw an Eps-neighbourhood. Other points reached by those, are called 'density-connected'. This differs from percolation in that core points and border points

are differentiated. Core points are defined as having more points in their Eps-neighbourhoods than border points (Ester et al. 1996: 228). The minimum number of neighbouring points needed to define a core point is one of the inputs needed for the DBscan algorithm. DBscan begins with an arbitrary point (seed), looks for density-reachable points and if the criterium of minimal number of points for the creation of a core point is met, connects all points that are density-reachable and then moves to the next point. Noise consists of points that cannot be connected to other points in this way (Ester et al. 1996: 229). If the minimal number of points for a core point is set to only two, the DBscan algorithm has empirically the same output as percolation analysis. In archaeology-related contexts DBscan has been used e.g. by Argote-Espino et al. (2012). The discussion of its usefulness varies: Whereas Ducke (2015: 363–364) favours it over k-means, Baxter (2015c: 16–19) is not convinced it offers better solutions.

Ripley's K function is best known as a multi-scalar cluster detection algorithm. The measure K describes the distribution of inter-point distances (Ripley 1976: 260). Its graph shows the cumulative average number of points lying in a certain radius (window) of a typical data point. This is compared with an expected point count under the assumption of a poisson process to determine at which radii the empirical distribution either clusters, is regularly spaced or random (Baddeley, Rubak and Turner 2015: 203–208; Ripley 1976). This **global** Ripley's K-function has been used widely in archaeology (e.g. Bevan and Conolly 2006; Palmisano 2013; Sayer and Wienhold 2013). Its variant, the **local** K-function analysis, focuses on where significant clustering occurs by rewriting the function for each point i (Getis and Franklin 1987). This was further modified by adding a Monte Carlo simulation, which may indicate significance of clustering at a point i at a certain scale (Smith 2016: I.4-22), using the hypothesis of complete spatial randomness (CSR). The output is a matrix of p-values at each point and each distance value, which may be mapped (Smith 2016: I.4-23–24). In distinction to percolation analysis, the density of points for a certain radius is calculated and a probability of belonging to a cluster given. It has not yet been used extensively in archaeology; examples are Crystal Safadi's M.Sc. thesis (Safadi 2013) and Pillot and Saligny (2013).

Other algorithms, such as nearest neighbour and the F- and G-functions, are clustering detection algorithms that indicate whether the point distribution is spatially clustered, but they do not directly give an attribution of points to a certain cluster.

Ducke formulated the following parameters for a 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, for example it should just ignore the odd scattering of sherds around the clusters; 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 satisfies these desiderata well, as its mathematical core is simple; the size, number and location of clusters is explored and 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 desired); and the results for a given dataset are always the same. It is an effective and practical spatial analysis tool for the archaeologist (Maddison 2020) and especially interesting for exploratory research.

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 readily accessible for any researcher who might wish to utilise it. A description and link to the code are provided at the end of this paper.

## 3. Case Studies

3.1. Hillforts in Britain

Working with data from the 'Atlas of Hillforts' (Lock and Ralston 2017), it was decided to use percolation analysis to explore and establish any 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 3007 confirmed sites in Britain and 347 in all of Ireland. Percolation analysis was applied in radius increments of 1km. The cluster transi-

tions for Britain are shown in **Figure 2**, which plots the normalized maximum cluster size (with a normalized size range of between 0 and 1.0) against percolation radius. Selected cluster plots are overlaid for illustration. 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, whilst the most notably obvious transition occurs at 34 km as most sites join a single very large cluster.

Geographical plots of clusters are shown in **Figures 3** to **5**. 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, and sites not in a cluster (i.e. noise) have a cross symbol. Above 35 km radius value all sites form a single large cluster, except for the Isle of Man and the island groups Hebrides, Shetlands, and Scillies. Apart from south-east Scotland, the most interesting transitions occur in the 6–14 km range. Within England, as the radius values reduce, sites in the Pennines and the east progressively break out of the bigger cluster, and at 14 km (**Figure 3**), the southwest peninsula forms its own cluster in Cornwall and part of Devon, and other clusters appear in the southeast. The plot for 12 km (**Figure 4**) shows for example Cornwall and Devon/part of Somerset as individual clusters, and a cluster along the Chilterns. The plot for 9 km (**Figure 5**) shows clusters in north-west Wales, the Clwydian Range, south-west Wales, the Gower, central
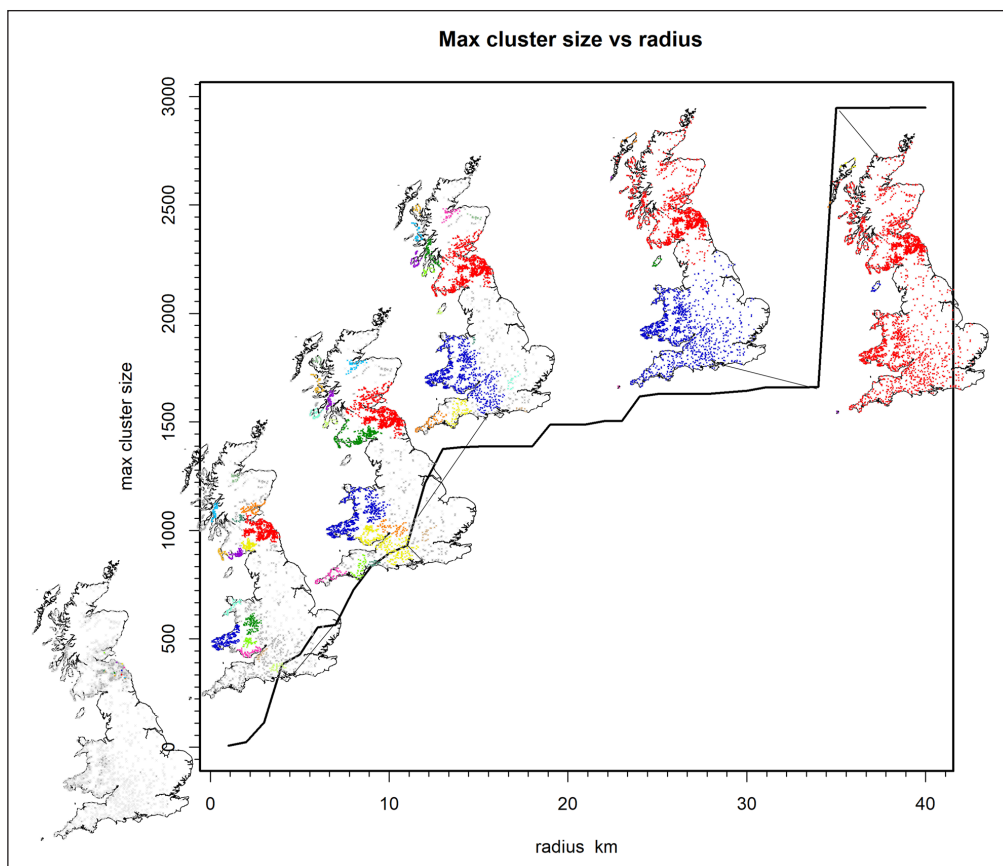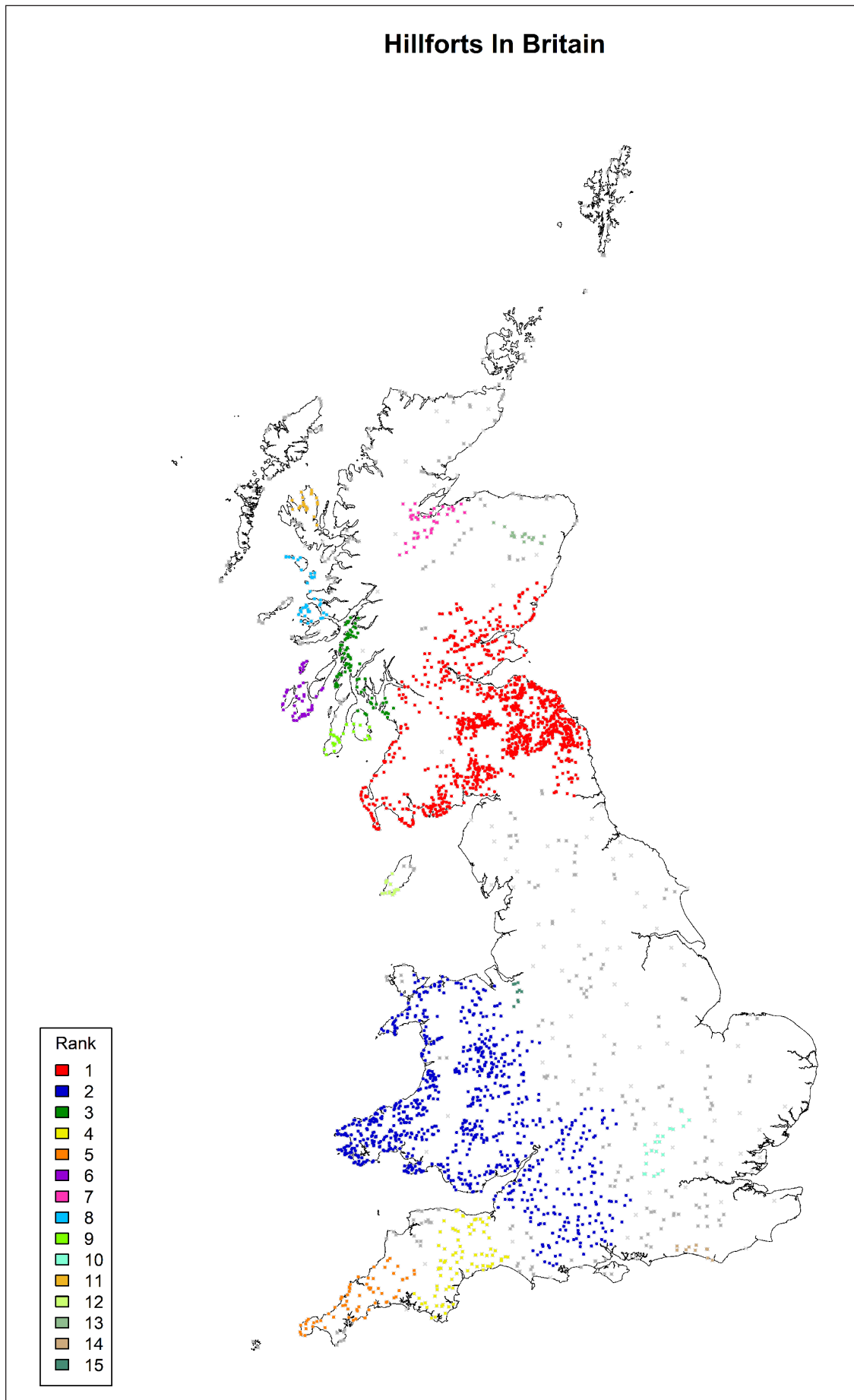

Figure 2: Cluster transition diagram, showing the maximum cluster size (normalised) vs. percolation radius, for hillforts in Britain.
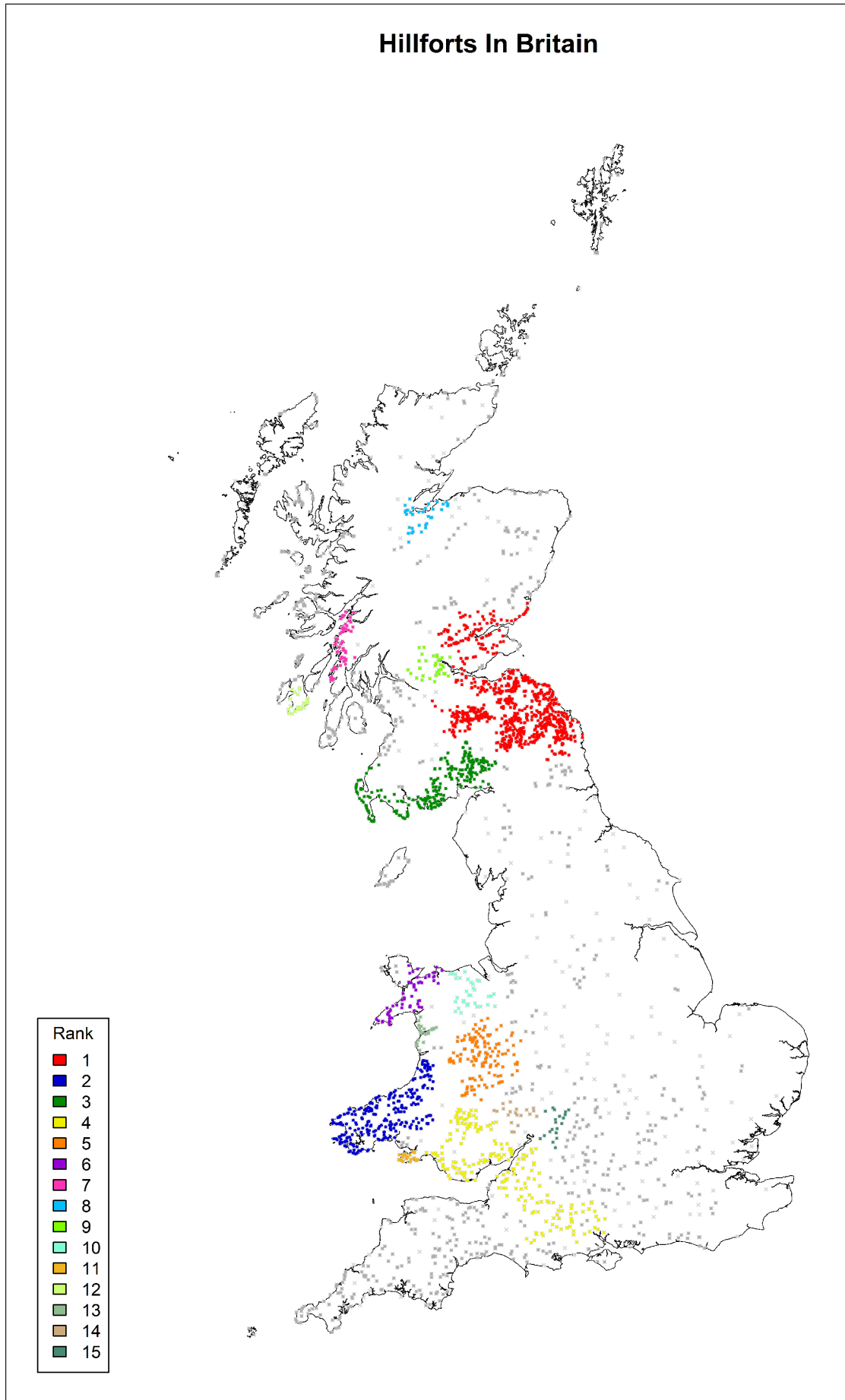
Hillforts In Britain
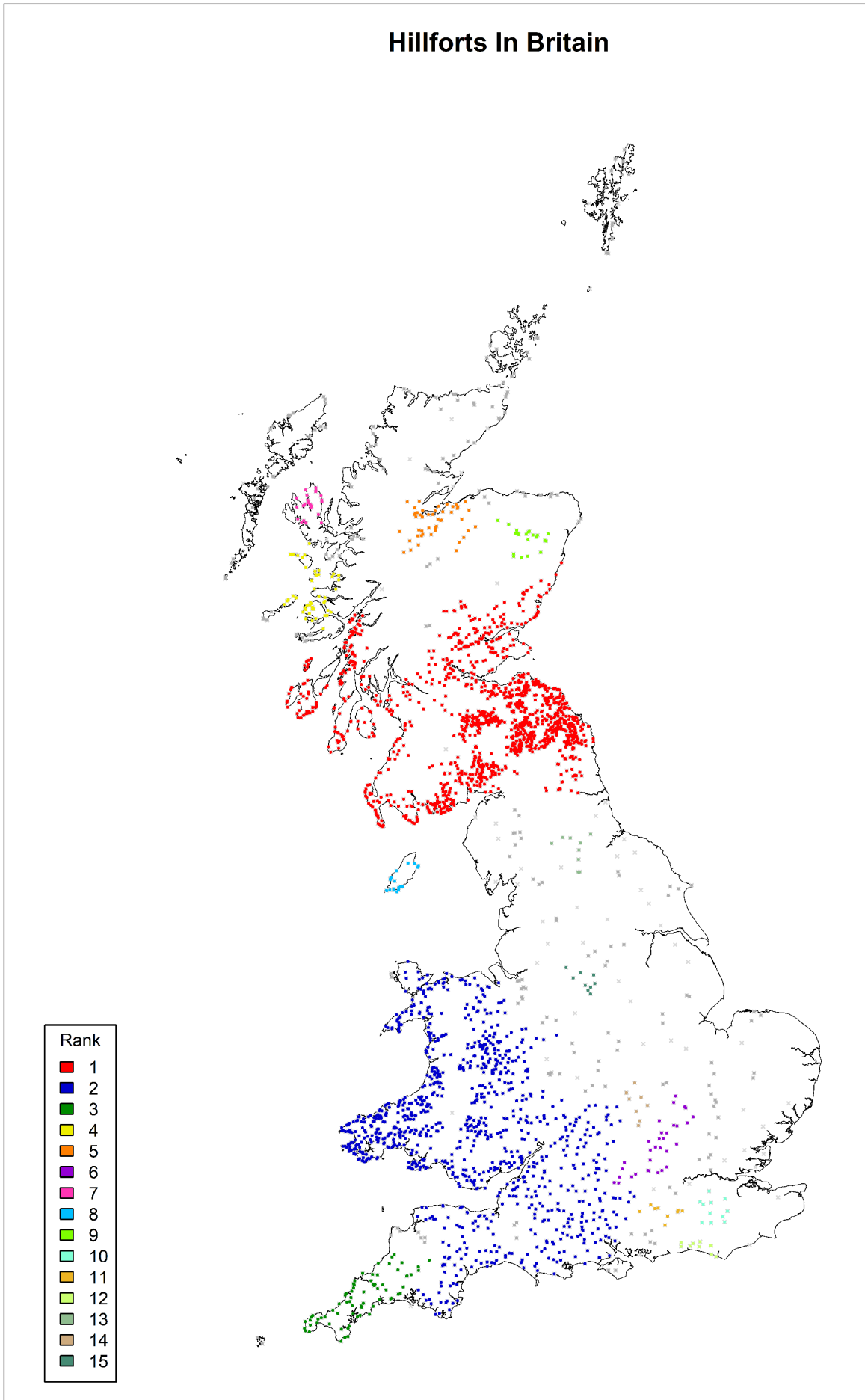
Figure 3: Percolation clusters of hillforts in Britain at 14 km radius.
Source: Atlas_of_Hillforts_GB_IoM_coords. 170725.csv; percolation distance: 14 km.

**Figure 4:** Percolation clusters of hillforts in Britain at 12 km radius.
Source: Atlas_of_Hillforts_GB_IoM_coords. 170725.csv; percolation distance: 12 km.

Figure 5: Percolation clusters of hillforts in Britain at 9 km radius.
Source: Atlas_of_Hillforts_GB_IoM_coords. 170725.csv; percolation distance: 9 km.

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. To illustrate, three clusters are shown plotted on a terrain model, which suggests a regional coherence and strong correlation with the local topography. The cluster in Central Wales at 6 km radius (**Figure 6**) shows hillforts on the hills above the valley of the river Severn and its tributaries as it comes down to the Shropshire plain. The Cornwall cluster at 12 km (**Figure 7**) shows sites in the peninsula, clearly
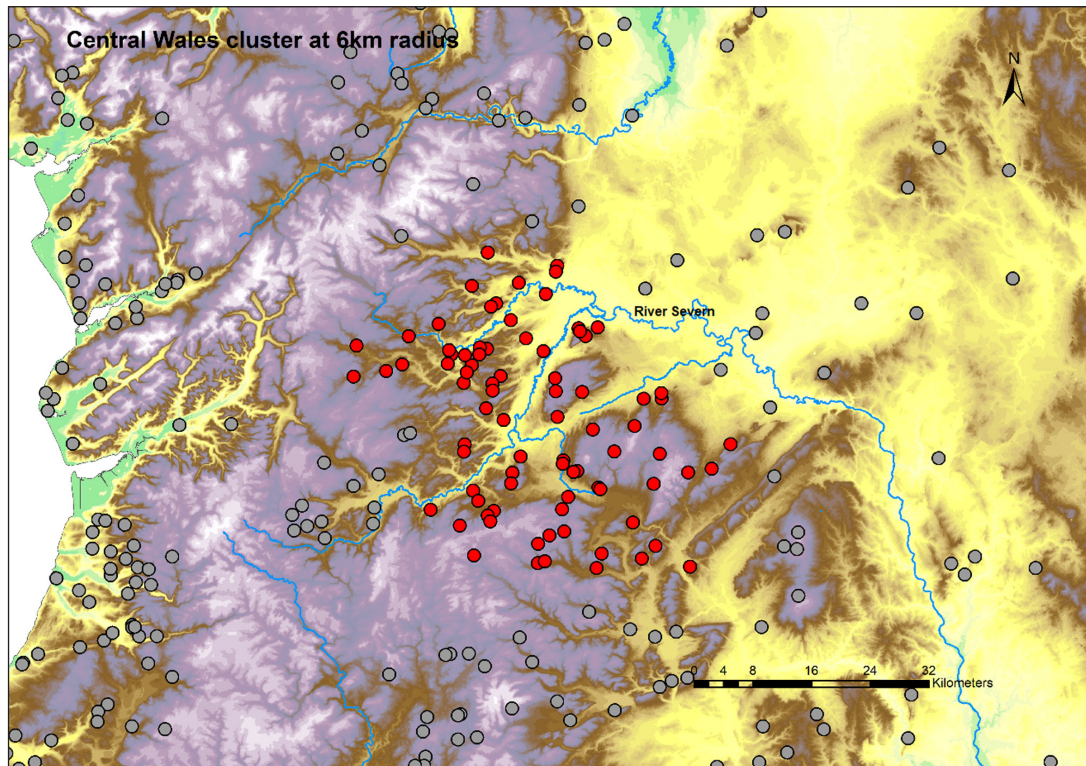
Figure 6: Central Wales hillfort cluster at 6km radius overlaid on terrain model with the River Severn and its main tributaries.
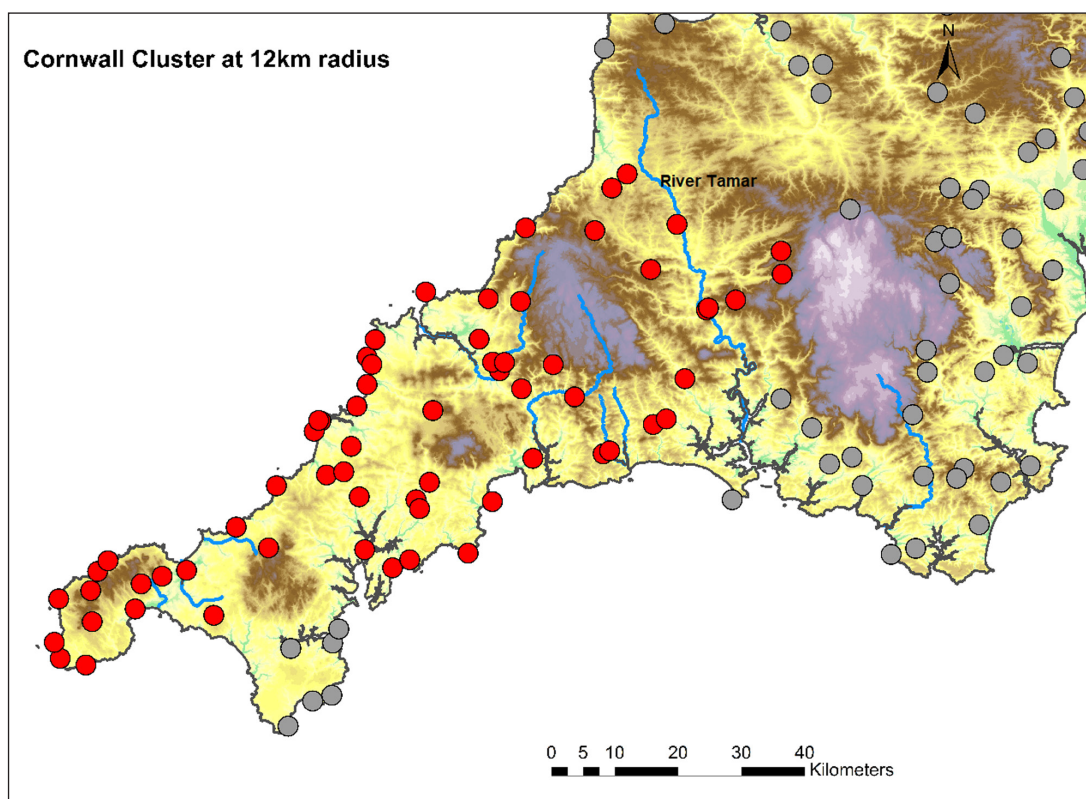

Figure 7: The Cornwall hillfort cluster at 12 km radius, overlaid on a terrain model, with principal rivers.

bounded by Dartmoor to the east and the river Tamar. The Cotswold cluster at 6 km (**Figure 8**) shows the hillforts along the escarpment overlooking the Severn valley with some sites close to the river, particularly where it joins the river Avon. These three examples are on quite distinct geographical terrains; they illustrate the potential for identifying clusters for further detailed regional study utilising other sources of data. This has been started (Maddison, 2019), and is an ongoing project.

It is clear from the above discussion that the formation of clusters is very dependent upon the density of sites. In high density areas, clusters form at lower percolation radius values as compared with low density areas where the clusters form at higher radius values. In areas of high site density such as south-west Wales and south-east Scotland, then a finer grained analysis with radius increments of 0.1 km reveals more localised cluster patterns in the range 3–7 km radius values, for example see **Figures**
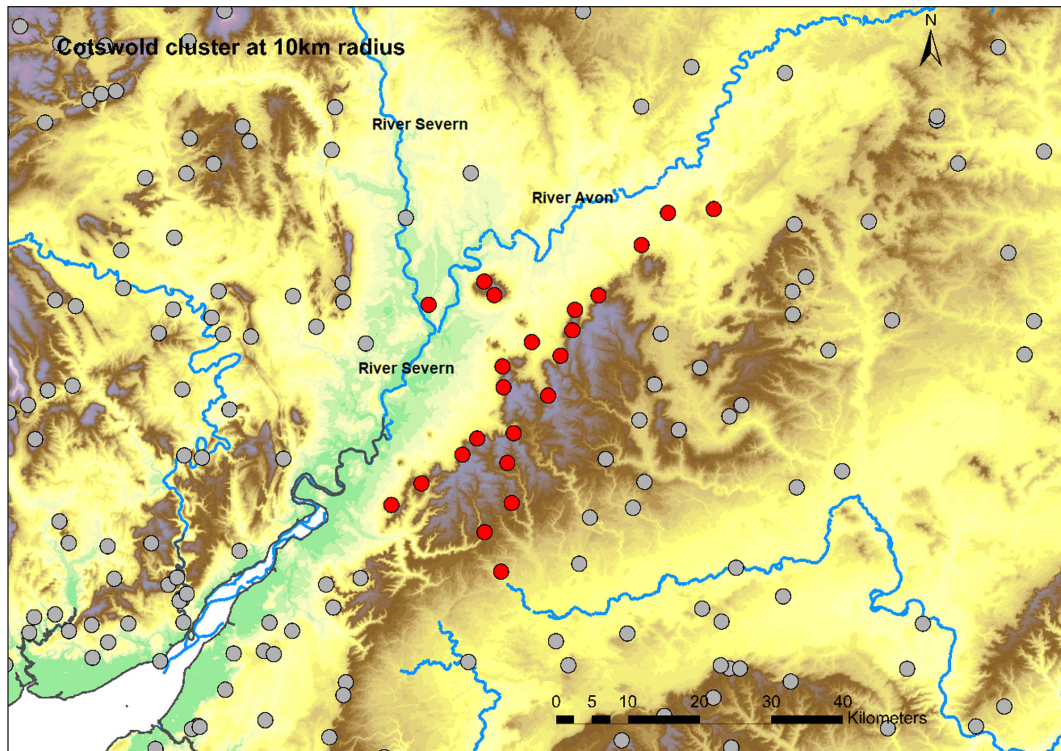
Figure 8: The Cotswolds hillfort cluster at 6 km radius, overlaid on a terrain model, with the rivers Wye, Severn, Avon and Thames.
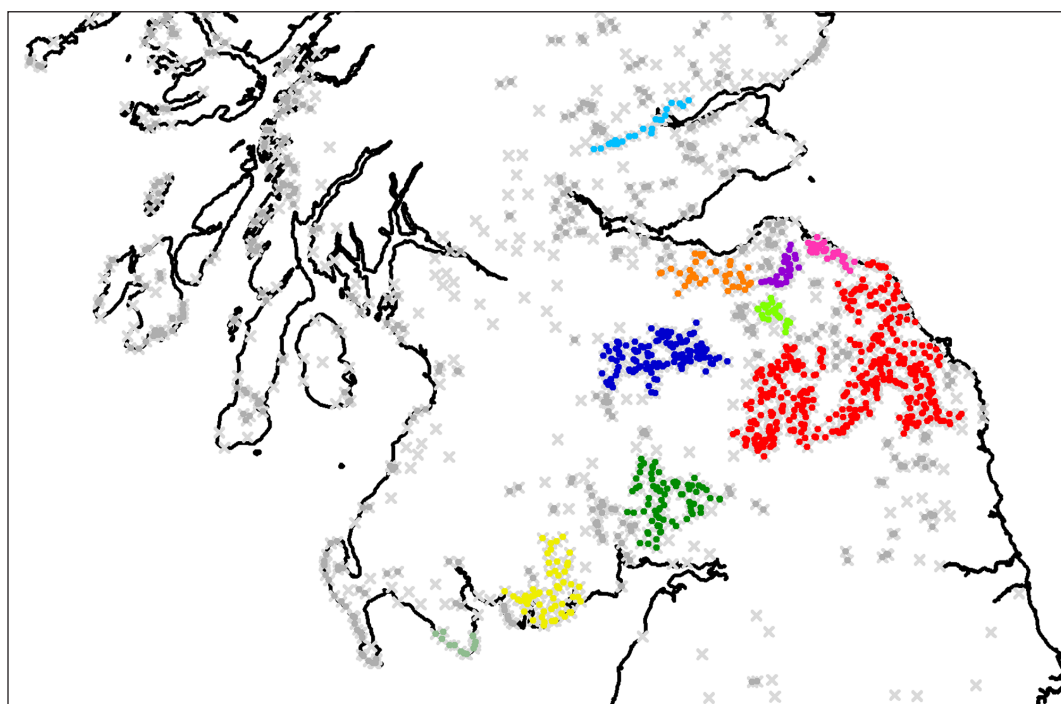
Figure 9: Southern Scotland hillfort clusters at 3.7 km radius.
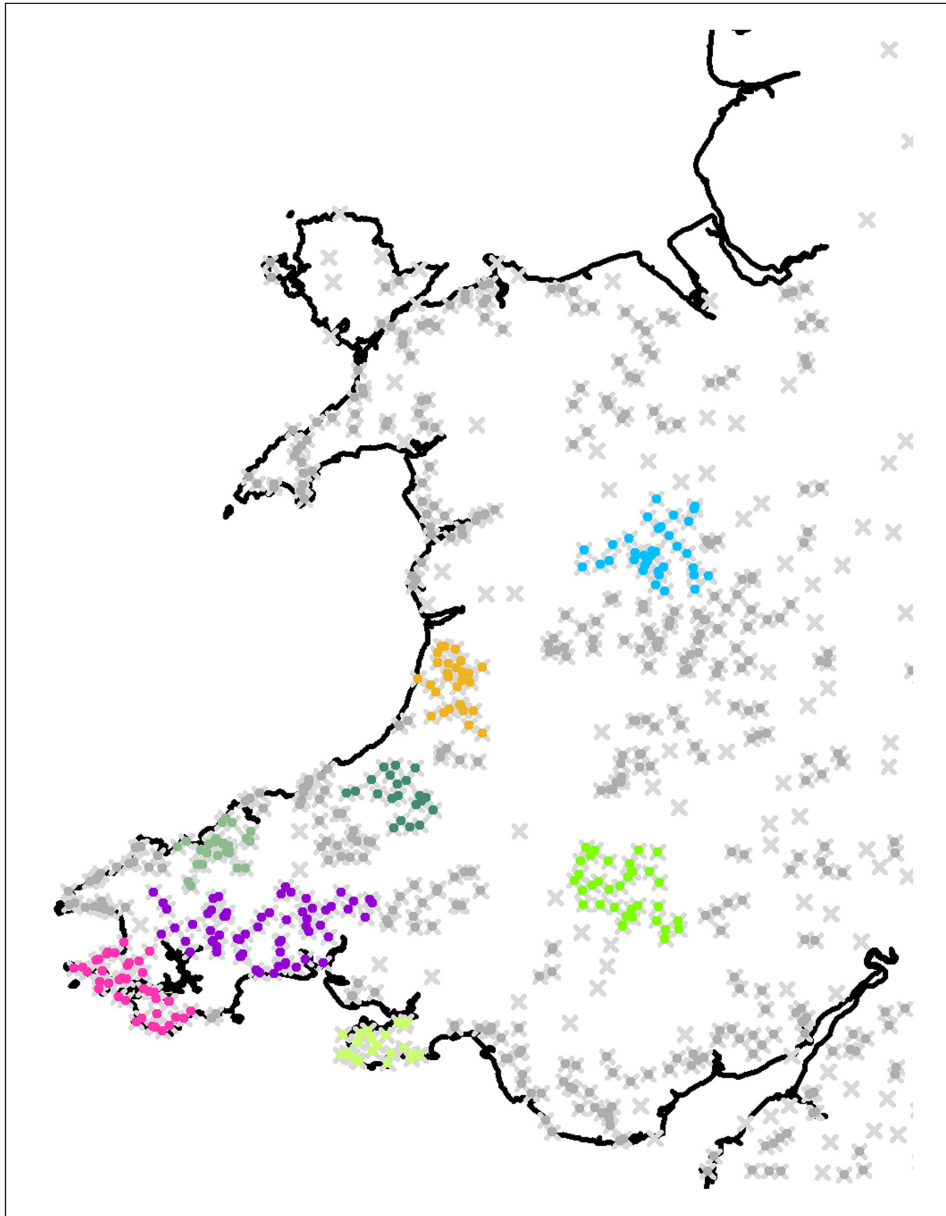
Figure 10: Wales hillfort clusters at 5.1 km radius.

9 and **10**. In both these regions the size of hillforts is much smaller than in less dense areas such as Wessex, so may reflect different social and political structures and hillfort roles. Further study is required to establish what significance these clusters might have.

3.2. Strip excavation in Germany

The second case study is located in Saxony-Anhalt, central Germany. During the early 2000s the 'Bundesstraße 6' road was built near the town of Köthen. The area of the main road, approach roads, exits and retention reservoirs were machine test-trenched and excavated by Heritage Management Saxony-Anhalt to find archaeological sites (see **Figure 11**). Along the ca. 13 km long and mostly 40 m wide location of the road, over 6000 features were uncovered (Fahr in preparation), but this paper will focus on features from the Late Bronze Age ('Saalemündungs-gruppe'). For this period the aim is to find the clusters which delineate a site with the help of percolation analy-

sis. The radius will estimate the distance between features that belong to a single site.

There are 1006 settlement features catalogued as belonging to this period (see **Figure 12**).

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 centre of a lively debate in survey archaeology over the last 40 years. Until the 1970s it was mostly defined as a place (Dunnell 1992: 23–24) – a find spot, which is similar to the German definition, where the term 'Fundplatz' means exactly that. In German archaeological inventories, though, a site may contain several culturally distinct loci, 'Fundstellen' (Dauber 1950: 96), which are usually catalogued as a settlement, a burial ground, a hoard or as a single find spot of a particular
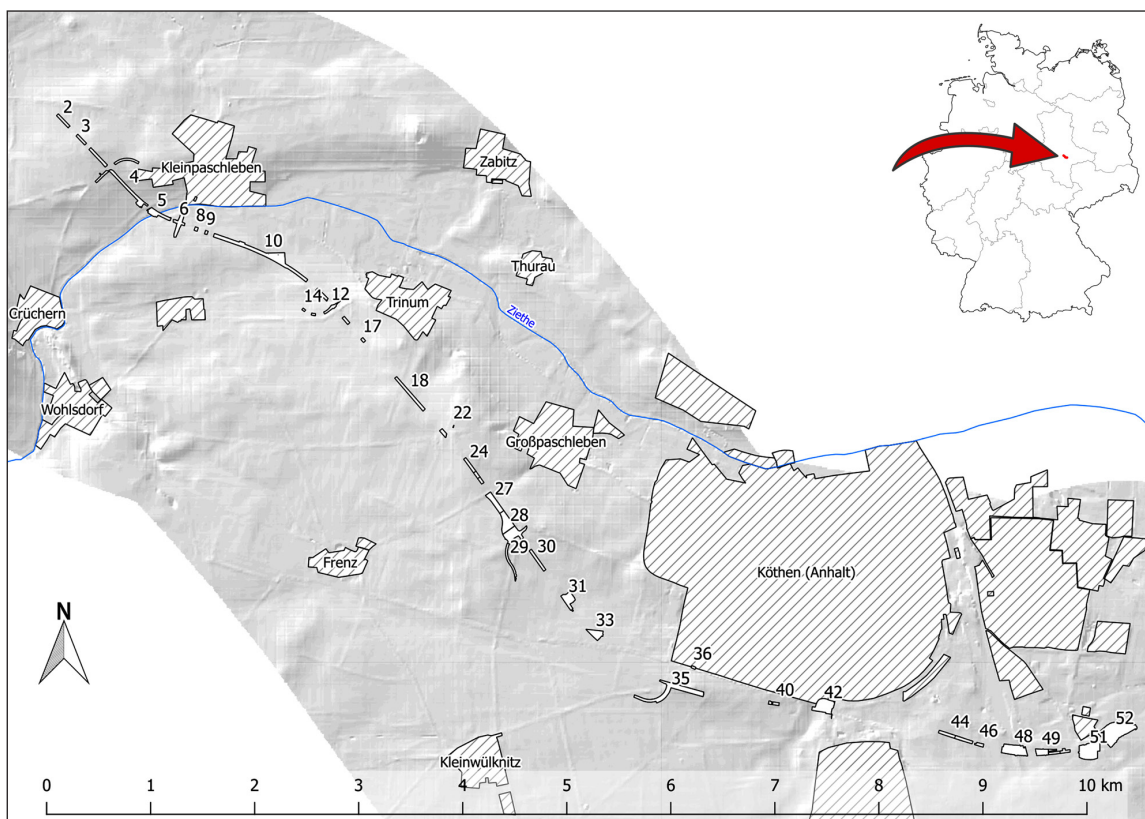
Figure 11: Road excavation, sites delimited by the Heritage Management Saxony-Anhalt. Residential areas marked by stripes. Background: DEM by State Agency for Surveying and Geoinformation Saxony-Anhalt.
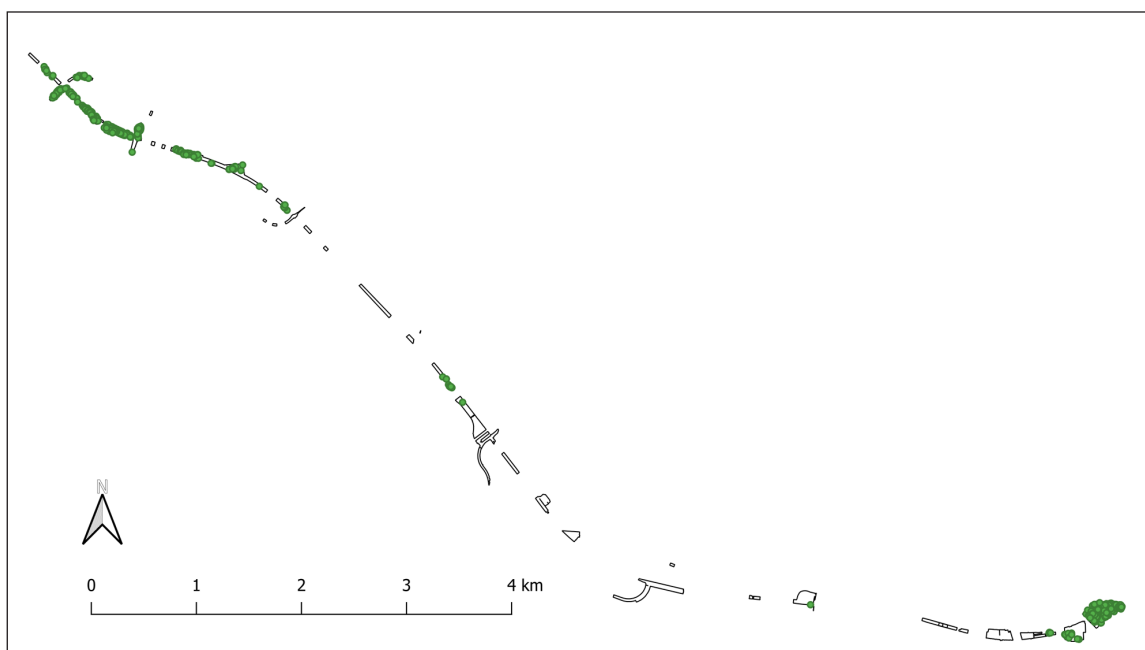
Figure 12: Excavation sites delimited by the Heritage Management Saxony-Anhalt, green points denoting Late Bronze Age features in the transect (overlapping).

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, how far apart two 'find spots' are supposed to be. If two sherds of the same time period are found 50 m apart – are they from two sites or do they belong to one site? In every larger settlement or landscape archaeological study this topic is being discussed and in the end a decision reached, which the authors agree is arbitrary and should actually rely on empirical studies (Malmer 1962: 258; Mischka 2007: 49–50). This is the question this case study aims to answer.

Since the middle of the 1970s there has been discussion of a non- or off-site approach in anglophone archaeology. The terms were introduced by Thomas (1975) and Foley (1981) and describe the analysis of finds 'between sites'. This approach interprets the archaeological record as a continual distribution in space, with spots of higher density, the so-called 'sites', and areas of lower density (Wobst 1983: 39).

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

Assuming a connection between sub-surface features and surface finds enables us to associate the distances between features to those of finds; this can inform landscape archaeological surveys on the decision to attribute single finds to an already known site or registering a new site. Of course, settlement features only comprise the 'built space' of a settlement. Although features in-between sites may exist, such as traps or temporary storage facilities, places such as working areas or fields which may be detected by off-site surface find distributions and will not be included.

3.2.1. Percolation analysis

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 1 m. As the size of a prehistoric village in central Germany cannot be assumed to be larger than a kilometre in diameter, the largest radius used was 500 m (see following **Figure 13**).

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 (as an example of this, see **Figure 14** for a percolation radius of 5 m). Longer stretches of flattening might be interpreted as the algorithm running out of features to add to the clusters. If we define a settlement as a cluster of features, we can take the values before a flattening to show at which radius no more features are added to a cluster and thereby suggest the edges of settlements. The step rises at higher percolation radii also need explanation. To provide these, the results will be compared to a large-scale excavation of a Late Bronze Age site in central Germany.

As in the example above, the maps generated show only the 15 largest clusters, marked in colour. In **Figure 14** the clusters that develop at a percolation radius of 5 m are depicted for the north-western part of the excavation. There the densest concentration of Late Bronze Age features has been uncovered. In **Figure 14** one can see a number of small clusters and a number of points not yet belonging in any cluster (marked with a cross). Note that the size of the cluster is determined by the numbers of features inside the cluster, not by its geographical spread.

The most distinctive steps of the continuous percolation analysis in the distribution are highlighted in **Figure 13**. We can see the radii after which the algorithm does not change the cluster size for some time, showing that stretches of space between features exist here. The turning point of the mean cluster sizes is at the radius of 68 m. In **Figure 15** the north-western part of the transect is shown with the results of different percolation analyses. At the radii of 5, 10 and 20 m the clusters grow considerably in size, without necessarily merging. This illustrates the steep rise of the curve in **Figure 13** at low percolation values. The differences between the clusters at radius values of 46 m, 68 m and
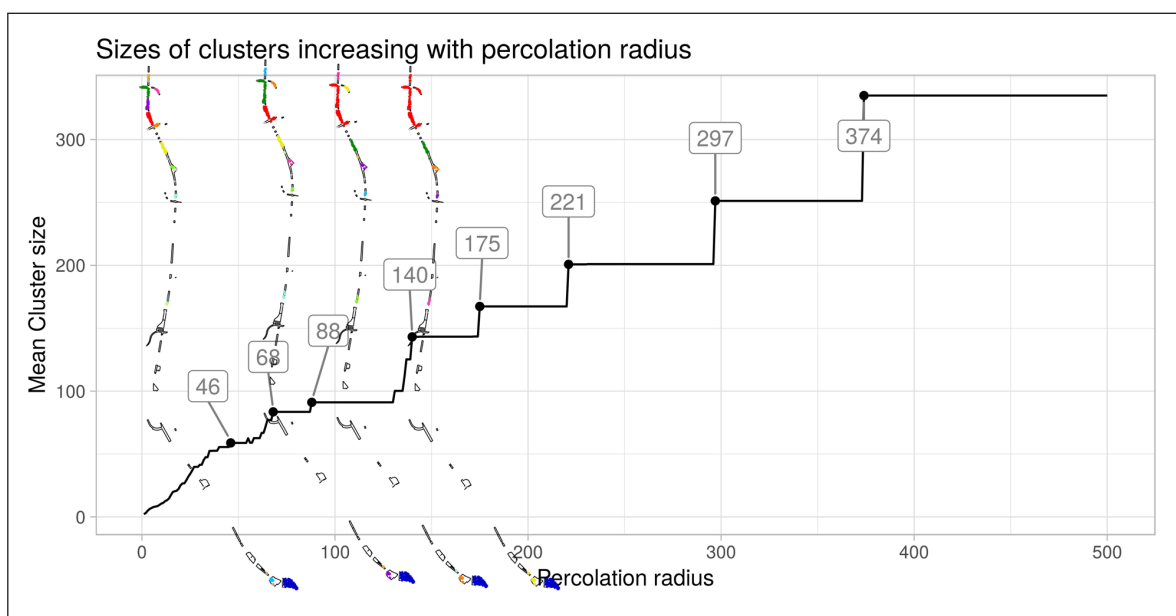
Figure 13: Percolation radius and mean cluster size of the transect excavation at Köthen, central Germany. Clustered strip excavation plots in the background correspond to the marked percolation radii.
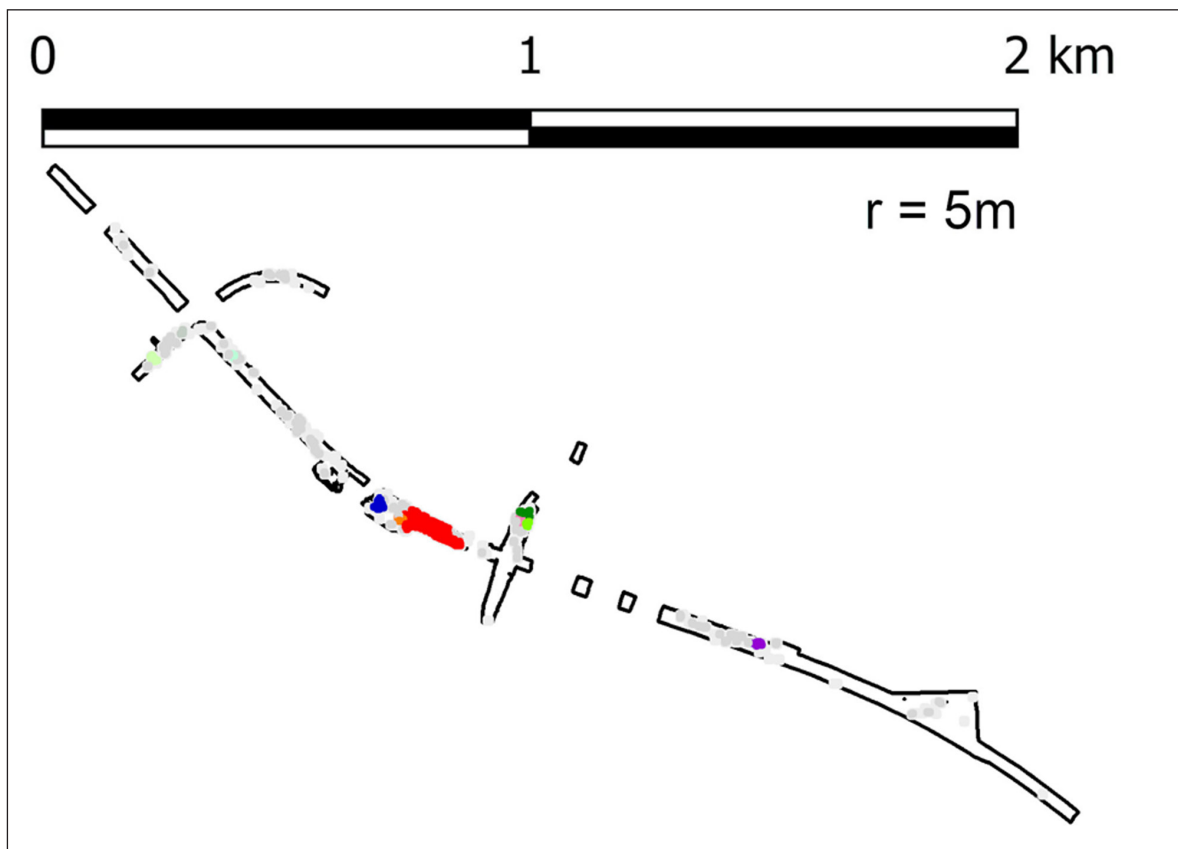
Figure 14: North-western end of the strip excavation at Köthen. Clusters that develop at a 5 m radius are shown. The points for the Bronze Age features are overlapping.

88 m are not as significant. At these percolation radii, two larger pre-existing clusters are merged, which leads to an increase of the mean cluster size value as depicted in **Figure 13**.

3.2.2. Interpretation

The Late Bronze Age features are mostly concentrated in the western half and on the eastern-most part of the transect, which is why the north-western part of the transect has been chosen for illustration (**Figures 14** and **15**). As can be seen by the rapidly rising graph describing the mean cluster size in relation to the percolation radius in **Figure 13**, the features cluster closely. The 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 do not grow with increasing percolation radii. As mentioned above, the mean width of the excavated strip was 40 m. 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 central Germany is 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 × 825 m (Huth and Stäuble 1998: 194–213). The farmsteads in Zwenkau do not overlap (Huth and Stäuble 1998: 214) and the distances between them are quite diverse. Nonetheless a number of them seem to scatter at intervals of ca. 65–75

m. (cf. Huth and Stäuble 1998: 216, fig. 6). The farmsteads themselves are not all the same size, but the longest sides are about 50–85 m long (cumulating to sizes of 2000–2500 m²) (Huth and Stäuble 1998: 213). It is assumed that not all of them existed at the same time, but that there were always several isolated farmsteads in a scattered distribution relative to each other (Huth and 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 (see Schmidt 2019). We can therefore assume that at least in part the high density of settlement features belonging to the Late Bronze Age might be interpreted as evidence of several phases or local re-settlement. Nevertheless, we can note that the turning points of the radius values at 68 m falls well within the range of the farmstead sizes of Zwenkau-Eythra and might be a possible farmstead site delimiter revealed by the percolation algorithm.

Two points can be noted here. Firstly, even though the transect only reveals a "slim" window into the archaeological record, the percolation analysis reveals a pattern, which is similar to that recorded on large-scale excavations. It therefore shows this to be a useful approach.

Secondly, taking the measure from percolation as a guide, we suggest that as long as the next feature is within approximately 68 m, features seem to belong to the same cluster. We can therefore assume that Late Bronze Age
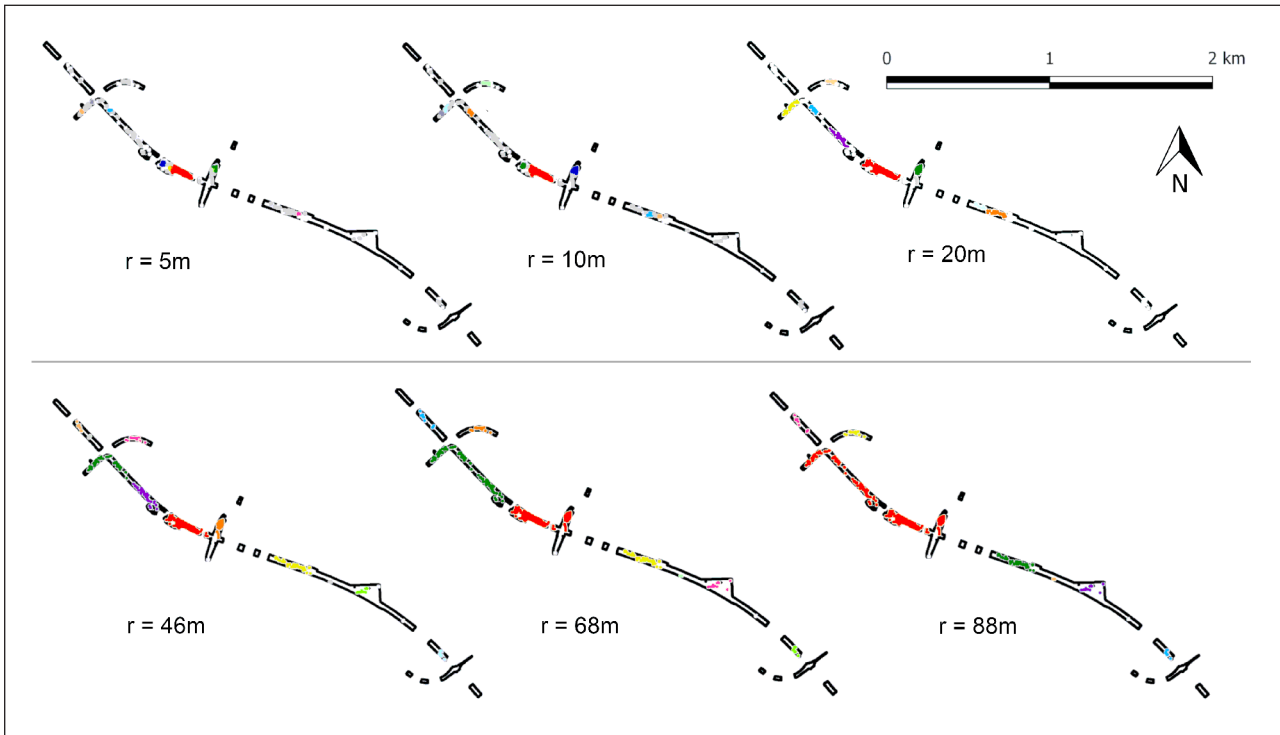
Figure 15: Details of percolation analyses of the north-western end of the transect at different radii. The scale given is relevant for all six maps.

finds within such a distance to each other belong to the same site. This is not to be understood as a general rule for all Late Bronze Age settlements everywhere, but rather as a rule of thumb for survey archaeology in this particular region.

Of course, if a survey misses a find and the find density is so low that no other was found in proximity, two findspots belonging to the same site might not be identified as such. This problem cannot be mitigated by a cluster algorithm, as it is a data issue.

Mapping the clusters (see **Figures 14** and **15**) reveals a more complete picture: As the percolation algorithm moves from node to node, the cluster groups are much larger than the radii. The sites therefore do not consist of single farmsteads, but are conglomerations of these.

In this way, the question posed by Malmer (1962) and Mischka (2007) on how to define the distance between two findspots before delimiting a new site has been answered in an empirical fashion. A similar analysis applied to different and larger data sets might shed further light on the stability of this interpretation.

4. Conclusion

In this paper percolation analysis was used for two very different archaeological applications and goals at widely different spatial scales. It was applied in Great Britain to investigate the distribution of hillforts, using data at a national, pan-regional scale. The objective was to start with the data, namely the spatial dataset of hillforts, and identify possible territorial groupings purely based on their spatial distribution. No assumptions were made as to modern or historical administrative or political boundaries, so the technique was applied very much in an explor-

atory way. Clear groupings have been identified, and some of these have been selected for further detailed analysis, for comparison with other sources of data (such as historical county boundaries for example), and utilising other attributes such as architecture, size and dating. Where these different sources support each other, it strengthens the case for their indicating prehistoric socio-political territories.

In contrast, the second application was at a sub-regional level. Features excavated along a transect were clustered to find relevant sizes of settlements of the Late Bronze Age in central Germany. These may be used as indicators of how far apart two findspots should be to delimit two different sites. It is suggested here that two finds should have, as a guideline, a distance of at least 68 m between each other for them not to belong to the same site.

In conclusion, percolation analysis offers a fruitful and easily approachable cluster algorithm, which lends itself well to archaeological middle range theory. It is applicable at different scales and may answer a variety of research questions. As it is exploratory in nature, it demands almost no assumptions, just the range of radii that are to be explored needs to be determined beforehand. Also, while Euclidian distance has been utilised in the examples shown here, the application to least-cost research questions is possible as well. With regards to Ducke's desiderata for cluster algorithms in archaeology, percolation analysis as implemented in our R-package offers an easy to understand alternative to more complex algorithms, such as DBscan and local Ripley's K. The assumptions are minimal, and the algorithm dependably identifies clusters and noise at different scales. The form of the clusters is not relevant, which sets it off positively

against e.g. the k-means algorithm which tends to create round clusters.

It is our belief that percolation analysis should be a standard part of the spatial analysis toolkit for archaeologists (see for example Gillings, Hacıgüzeller and Lock 2020) and that the examples given here are an effective demonstration of its value and flexibility.

Percolation Analysis: detailed description of the R code

The program suite, along with a sample program and test data can be found and downloaded in Open Access from:

DOI 10.17605/OSF.IO/7EXTC.

The code, written in R, applies a spatial clustering process to a spatial point data set.

The **percolate()** function processes this data set to generate the clusters for a range of percolation radius values. It also generates and stores a set of data tables for use by the other functions, as well as optional external processing, as described below.

mapClusters() prints maps of the generated clusters, overlaid on a given shape file.

plotClustFreq() plots analysis results for the clusters showing the key characteristics of the clusters generated for a given point data set.

A working directory path needs to be defined before running the functions. The functions will create the following sub-directories off this path:

/working_data – where the generated data tables are stored

/analysis_results – where the analysis results and the plots are stored

/maps – where the generated output maps are stored

It is suggested that */source_data* and */shape_files* directories are also created.

Graphical outputs are generated as png files with an A4 format.

Percolate

percolate(data,distance_table,upper_radius, lower_radius, step_value, limit, radius_unit)

The **data** needs to be provided as a dataframe in the format below:

PlcIndex,Easting,Northing

The *PlcIndex* field is a provided identity for each point, which is specified by the coordinates. The resolution of the coordinates is assumed to be in metres.

In addition to the spatial data, parameters are passed to the percolate function as below:

distance_table is normally set to NULL. This dataframe needs to be provided if you wish to use the analysis with your own inter-point distances, for example least-cost path or weighted distances.

upper_radius is the upper value of the radius range to be used, scale is m * radius unit.

lower_radius is the lower value of the radius range to be used, scale is m * radius unit.

step_value is the step value to be used between these two values, scale is m * radius unit (NOTE: this value can be decimal, e.g. 0.2).

limit is the value above which distances between sites will not be stored, scale is m * radius unit. This is typically above the value at which all points are within a single cluster.

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.

As an example, the upper_radius is 40, the lower_radius is 2, the step value is 1, the limit is 50 and the radius unit is 1000, i.e. it sets all the above values to km. 'data' is a file containing the point data and identifier.

The first function to be called, with example parameters, is:

"percolate(data, , 40, 2, 1, 50, 1000)"

The percolate program first computes an inter-point distance matrix. Given the dataset of points and XY coordinates, it computes a partial matrix of inter-point Euclidean distances using Pythagoras' theorem. The *limit* parameter sets the maximum value stored, in this case to 50 km. For example, with a spatial dataset covering the whole of Britain, there is little point in storing the distance between points at the extremes of the country. This reduces the data file sizes and speeds up data file handling. The distance for each pair of points is computed and stored only once.

The inter-point distance table comprises a data frame consisting of: first point identity; second point identity; the distance between them, all rounded to two decimal places of the unit value. This is stored as a space delimited text file called 'nodes_list_d.txt'. The format is comma separated with columns:

ID1,ID2,d12 – the identifiers for point 1, point 2 and the distance between them.

Duplicate and null entries are identified, and if present these data points are stored in space delimited text files, 'duplicate_entries.txt' and 'null_entries.txt'.

The inter-point distance data is used to generate the clusters. However, if so desired, a distance matrix may be provided to the program. So, for example, weighted values may be used rather than the Euclidean distance between points, or the computed Euclidean distances may be modified in some way (e.g. applying a factor or non-linear scaling). In this case, the program will skip the computation, and use the given data frame, passed as a parameter. For example: "percolate(data, distance_table, 40, 2, 1, , 1000)"

Note that the various other parameters are still needed for the cluster computation, with the exception of the limit value, which is irrelevant.

The program now generates the clusters for each radius value required, starting at the largest radius value. The radius values to be used are computed from the upper and lower radius values, decremented by the given step value. In this example, the values start at 40 and decrement to 2, with a step value of 1; all values are in km. This data is stored for use by the mapping function, in the working directory. For each radius value, data is extracted from the distance table, including only those point-pairs where the inter-point distance <= the current radius value, creating a sub-matrix of the original point data set. This sub-matrix can be considered as an edge list for a graph.

The following R functions from library 'igraph' are then applied to this sub-matrix as follows:

graph_edgelist() – creates a graph from the sub-matrix, with directed=TRUE to count the edges only once.

clusters() – identifies the clusters in the graph and generates a unique identity for each. This identity is then bound to each point (i.e. graph vertex) using the **V()** function.

A table is created indexed by the given point identity (PlcIndex), with a column for each computed radius value, containing the cluster identity for that radius value (assuming it is a member of a cluster, else it is NULL). As the computation progresses, columns are added to this table. This table is stored as "member_cluster_by_radius.csv".

This table is then processed to generate analytics, so that for each radius value, the number of clusters, the maximum, mean and median number of points per cluster are all computed and stored in a file called "analysis_by_radius.csv".

These tables are stored in the analysis results directory, for use by the other routines as well as for any desired additional processing that the user may undertake.

If required the clusters can be mapped using the **mapClusters()** function.

Mapping Clusters
mapClusters(shape, map_name, source_file_name)
shape – an imported shape file, read using **readOGR()** – note this requires the rgdal library, e.g. "Namibia.shp".
map_name – to be printed at the head of the maps, e.g. "Archaeological sites in Namibia".
source_file_name – to be printed at the bottom of the map as a reference to the source data used to generate the cluster maps, e.g. "point_data_Namibia.v4".
dpi – the resolution for the maps, defaults to 300. Higher resolution may be required for publications, for example, or lower values for exploratory plots.

Using the example parameters, the call would be: "mapClusters(shape,"Archaeological sites in Namibia", "point_data_Namibia.v4").

The radius values are taken from the file generated by the percolate() function, to ensure consistency.

The map projection is extracted from the shape file and applied to the maps generated. This means the input point data needs to be in the same coordinate and projection system as the background map. A base map is first generated with all the points overlaying the shape file, plotting all points as a basic check. Then, for each radius value a cluster map is generated; all points are plotted as a small cross in pale grey, and then overlaid with the points coloured according to the rank of their cluster. The largest 15 clusters are coloured, red for the largest, blue for the next and so on. For clusters outside this grouping a mid-grey is used. Noise, that is points that are not within a cluster, are not overlaid and remain as a small grey cross. All the maps are stored in the /map directory. png is used as the map format, being highly portable and without using compression.

The final step, if required, is to plot the statistics for the data.

Cluster Analytics
plotClustFreq(source_file_name)
For example:
plotClustFreq("point_data_Namibia.v4")

All of the data required is taken from the file "analysis_by_radius.csv" in the */analysis results* directory. The source file name is passed as a parameter for inclusion in the plots. These plots, similar to that in Figure 2, are generated as png files and stored in the */analysis_results* directory.

The three functions are provided separately, to more quickly identify issues with data, maps and formats, without having to run the entire suite each time.

Competing Interests
The authors have no competing interests to declare.

References
Arcaute, E, Brookes, S, Brown, T, Lake, M and **Reynolds, A.** forthcoming. Case studies in percolation analysis: the distribution of English settlement in the 11th

and 19th centuries compared. *Journal of Archaeological Science.*

Arcaute, E, Molinero, C, Hatna, E, Murcio, R, Vargas-Ruiz, C, Masucci, AP and **Batty, M.** 2016. Cities and regions in Britain through hierarchical percolation. *Royal Society Open Science,* 3(4): 150691. DOI: https://doi.org/10.1098/rsos.150691

Argote-Espino, D, Solé, J, López-García, P and **Sterpone, O.** 2012. Obsidian Subsource Identification in the Sierra de Pachuca and Otumba Volcanic Regions, Central Mexico, by ICP-MS and DBSCAN Statistical Analysis. *Geoarchaeology,* 27(1): 48–62. DOI: https://doi.org/10.1002/gea.21389

Baddeley, A, Rubak, E and **Turner, R.** 2015. *Spatial Point Patterns: Methodology and Applications with R.* New York: Chapman and Hall/CRC Press. DOI: https://doi.org/10.1201/b19708

Baxter, MJ. 2015a. Notes on Quantitative Archaeology and R. Available at https://www.researchgate.net/profile/Mike_Baxter/publication/277931925_Notes_on_Quantitative_Archaeology_and_R/links/5576bf9708aeacff1ffe7fea.pdf Last accessed 2016/03/24/13:14:57.

Baxter, MJ. 2015b. *Exploratory multivariate analysis in archaeology, with a new introduction by the author.* Clinton Corners, NY: Percheron Press, a division of Eliot Werner Publications, Inc.

Baxter, MJ. 2015c. Spatial k-means clustering in archaeology – variations on a theme. Available at https://www.academia.edu/18142974/Spatial_k-means_clustering_in_archaeology_--_variations_on_a_theme Last accessed 2019/05/23/12:16:25.

Bevan, A and **Conolly, J.** 2006. Multiscalar Approaches to Settlement Pattern Analysis. In: Lock, G and Molyneaux, BL (eds.), *Confronting Scale in Archaeology: Issues of Theory and Practice.* New York: Springer.

Dauber, A. 1950. Der Forschungsstand als innere Gültigkeitsgrenze der Fundkarte. In: Kirchner, H (ed.), *Ur-und Frühgeschichte als historische Wissenschaft: Festschrift zum 60. Geburtstag von Ernst Wahle.* Heidelberg: Winter.

Digimap Resource Centre. Available at http://digimap.edina.ac.uk/webhelp/resources/ Last accessed 2016.

Ducke, B. 2015. Spatial Cluster Detection in Archaeology: Current Theory and Practice. In: Barcelo, JA and Bogdanovic, I (eds.), *Mathematics and Archaeology.* Boca Raton: Taylor & Francis. DOI: https://doi.org/10.1201/b18530-22

Dunnell, RC. 1992. The Notion Site. In: Rossignol, J and Wandsnider, L (eds.), *Space, Time, and Archaeological Landscapes.* New York: Springer Science & Business Media.

Eggert, MKH. 2005. *Prähistorische Archäologie. Konzepte und Methoden.* Tübingen und Basel: A. Francke Verlag.

Enloe, JG, David, F and **Hare, TS.** 1994. Patterns of Faunal Processing at Section 27 of Pincevent: The Use of Spatial Analysis and Ethnoarchaeological Data in the Interpretation of Archaeological Site Structure. *Journal of Anthropological Archaeology,* 13: 105–124. DOI: https://doi.org/10.1006/jaar.1994.1007

Ester, M, Kriegel, H-P, Sander, J and **Xu, X.** A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis, E, Han, J and Fayyad, UM (eds.), *Second International Conference on Knowledge Discovery and Data Mining (KDD-96),* Portland, OR. 1996. Palo Alto, CA: AAAI Press.

Fahr, J. in preparation. Lange Streifen und große Entfernungen – Die archäologischen Untersuchungen in Vorbereitung der flächigen Ausgrabungen im PA 16 der B 6n.

Foley, R. 1981. Off-site archaeology: an alternative approach for the short-sited. In: Hodder, I, Isaac, G and Hammond, N (eds.), *Pattern of the past. Studies in honour of David Clarke.* Cambridge: Cambridge University Press.

Frisch, HL and **Hammersley, JM.** 1963. Percolation Processes and Related Topics. *Journal of the Society for Industrial and Applied Mathematics,* 11(4): 894–918. DOI: https://doi.org/10.1137/0111066

Geofabrik, OpenStreetMap. Available at https://download.geofabrik.de/europe/germany/sachsen-anhalt.html Last accessed 2016.

Getis, A and **Franklin, J.** 1987. Second-Order Neighborhood Analysis of Mapped Point Patterns. *Ecology,* 68(3): 473–477. DOI: https://doi.org/10.2307/1938452

Gillings, M, Hacıgüzeller, P and **Lock, G.** (eds.) 2020. *Archaeological Spatial Analysis: A Methodological Guide.* London and New York: Routledge. DOI: https://doi.org/10.4324/9781351243858

Hey, G. 2006. Scale and Archaeological Evaluation: What are We Looking For? In: Lock, G and Molyneaux, BL (eds.), *Confronting Scale in Archaeology. Issues of Theory and Practice.* New York: Springer Science+Business Media, LLC.

Huth, C and **Stäuble, H.** 1998. Ländliche Siedlungen der Bronzezeit und der älteren Eisenzeit. Ein Zwischenbericht aus Zwenkau. In: Küster, H, Lang, A and Schauer, A (eds.), *Archäologische Forschungen in urgeschichtlichen Siedlungslandschaften. Festschr. G. Kossack.* Regensburg: Universitätsverlag Regensburg.

Jendryke, M and **Caspari, G.** 2017. Archsphere – A cluster algorithm for archaeological applications. *Journal of Archaeological Science,* 14: 181–188. DOI: https://doi.org/10.1016/j.jasrep.2017.05.052

Kintigh, KW. 1990. Intrasite spatial analysis: A commentary on major methods. *Mathematics and Information Science in Archaeology: A Flexible Framework.* Bonn: Holos.

Kintigh, KW and **Ammerman, AJ.** 1982. Heuristic Approaches to Spatial Analysis in Archaeology. *American Antiquity,* 47(1): 31–63. DOI: https://doi.org/10.2307/280052

Koetje, TA. 1994. Intrasite Spatial Structure in the European Upper Palaeolithic: Evidence and Patterning from the SW of France. *Journal of Anthropological Archaeology,* 13: 161–169. DOI: https://doi.org/10.1006/jaar.1994.1011

Ladefoged, TN and **Pearson, R.** 2000. Fortified castles on Okinawa Island during the Gusuku Period, AD 1200–1600. *Antiquity,* 74(284): 404–412. DOI: https://doi.org/10.1017/S0003598X00059494

Lemke, AK. 2013. Cutmark systematics: Analyzing morphometrics and spatial patterning at Palangana. *Journal of Anthropological Archaeology,* 32(1): 16–27. DOI: https://doi.org/10.1016/j.jaa.2012.08.001

Linke, W. 1976. *Frühes Bauerntum und geographische Umwelt. Eine historisch-geographische Untersuchung westfälischer und nordhessischer Bördenlandschaften.* Paderborn: Schöningh.

Lock, G and **Ralston, I.** 2017. Atlas of Hillforts of Britain and Ireland. Available at https://hillforts.arch.ox.ac.uk/ Last accessed 2017–2019.

Maddison, S. 2019. Using Atlas data: the distribution of Hillforts in Britain and Ireland. In: Lock, G and Ralston, I (eds.), *Hillforts: Britain, Ireland and the Nearer Continent: Papers from the Atlas of Hillforts of Britain and Ireland Conference, June 2017.* Oxford: Archaeopress Archaeology. DOI: https://doi.org/10.2307/j.ctvnb7r0b.15

Maddison, S. 2020. Percolation Analysis. In: Gillings, M, Haciguzeller, P and Lock, GR (eds.), *Archaeological Spatial Analysis: A Methodological Guide.* London and New York: Routledge. DOI: https://doi.org/10.4324/9781351243858-5

Malmer, MP. 1962. *Jungneolithische Studien.* Bonn: Rudolf Habelt.

Mischka, D. 2007. *Methodische Aspekte zur Rekonstruktion Prähistorischer Siedlungsmuster. Landschaftsgenese vom Ende des Neolithikums bis zur Eisenzeit im Gebiet des südlichen Oberrheins.* Rahden/Westf.: Marie Leidorf GmbH.

Nakoinz, O and **Knitter, D.** 2016. *Modelling Human Behaviour in Landscapes. Basic Concepts and Modelling Element.* Switzerland: Springer International. DOI: https://doi.org/10.1007/978-3-319-29538-1

Neethu, CV and **Surendran, S.** 2012. Review of Spatial Clustering Methods. *International Journal of Information Technology Infrastructure,* 2(3): 15–24. Available at http://warse.org/pdfs/2013/ijiti01232013.pdf

OS Open Rivers. Available at www.os.uk/opendata Last accessed August–September 2016.

Palmisano, A. 2013. Zooming Patterns Among the Scales: a Statistics Technique to Detect Spatial Patterns Among Settlements. In: Earl, G, Sly, T, Chrysanthi, A, Murietta-Flores, P, Papadopoulos, C, Romanowska, I and Wheatley, D (eds.), *CAA 2012: Proceedings of the 40th Annual Conference of Computer Applications and Quantitative Methods in Archaeology (CAA), Southampton, United Kingdom, 26–30 March 2012.* Amsterdam: Amsterdam University Press.

Pillot, L and **Saligny, L.** 2013. The Evolution of Territorial Occupation: Exploratory Spatial Data Analysis. Uncertainty and Heterogeneity of Data in Archaeology. In: Earl, G, Sly, T, Chrysanthi, A, Murietta-Flores, P, Papadopoulos, C, Romanowska, I and Wheatley, D (eds.), *CAA 2012: Proceedings of the 40th Annual Conference of Computer Applications and Quantitative Methods in Archaeology (CAA), Southampton, United Kingdom, 26–30 March 2012.* Amsterdam: Amsterdam University Press.

Ripley, BD. 1976. The second-order analysis of stationary point processes. *Journal of Applied Probability,* 13(2): 255–266. DOI: https://doi.org/10.2307/3212829

Rozenfeld, HD, Rybski, D, Andrade, JS, Batty, M, Stanley, HE and **Makse, HA.** 2008. Laws of population growth. *Proceedings of the National Academy of Sciences of the United States of America,* 105(48): 18702. DOI: https://doi.org/10.1073/pnas.0807435105

Rozenfeld, HD, Rybski, D, Gabaix, X and **Makse, HA.** 2011. The Area and Population of Cities: New Insights from a Different Perspective on Cities. *American Economic Review* 101(5): 2205–25. DOI: https://doi.org/10.1257/aer.101.5.2205

Safadi, C. 2013. The Beqa'a Valley During the Early Bronze Age: A GIS Approach to Settlement Patterns. Unpublished M.Sc. thesis, University of Southampton.

Savage, SH. 1997. Descent Group Competition and Economic Strategies in Predynastic Egypt. *Journal of Anthropological Archaeology,* 16: 226–268. DOI: https://doi.org/10.1006/jaar.1997.0312

Sayer, D and **Wienhold, M.** 2013. A GIS-Investigation of Four Early Anglo-Saxon Cemeteries: Ripley's K-function Analysis of Spatial Groupings Amongst Graves. *Social Science Computer Review,* 31(1): 71–89. DOI: https://doi.org/10.1177/0894439312453276

Schirren, CM. 1997. *Studien zur Trichterbecherkultur in Südostholstein.* Bonn: Rudolf Habelt.

Schmidt, SC. 2019. Siedlungsgrößen und Abstände zwischen Siedlungsstandorten – eine geostatistische Analyse von Transektdaten. *Prähistorische Zeitschrift,* 94(2). DOI: https://doi.org/10.1515/pz-2019-0015

Smith, TE. 2016. *K-Function Analysis of Point Patterns,* [online]: http://www.seas.upenn.edu/~ese502/#notebook.

Stauffer, D and **Aharony, A.** 1991. *Introduction to Percolation Theory.* Boca Raton: CRC Press.

Thomas, DH. 1975. Nonsite Sampling in Archaeology: Up the Creek Without a Site? In: Mueller, JW (ed.), *Sampling in Archaeology.* Tucson: University of Arizona Press.

Varghese, B, Unnikrishnan, A and **Jacob, K.** 2014. Spatial Clustering Algorithms- An Overview. *Asian Journal of Computer Science And Information Technology,* 3. Available at https://pdfs.semanticscholar.org/0ea3/078ff0cac5e4385339d84d90f9d099875658.pdf

Wobst, HM. 1983. We Can't See the Forest for the Trees: Sampling and the Shapes of Archaeological Distributions. In: Moore, JA and Keene, AS (eds.), *Archaeological Hammers and Theories.* New York: Academic Press. DOI: https://doi.org/10.1016/B978-0-12-505980-0.50010-0