



Archaeology, Environment and Human History: Examining the Spatial Links Between Human Settlements and Environmental Change in Iceland

School of GeoSciences

Dissertation
for the degree of

MSc in Geographical Information Science

Kaja Rønning

August 2020

Statement of Copyright

Copyright of this dissertation is retained by the author and the University of Edinburgh. Ideas contained in this dissertation remain the intellectual property of the author and their supervisors, except where explicitly otherwise referenced. All rights reserved. The use of any part of this dissertation reproduced, transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise or stored in a retrieval system without the prior written consent of the author and The University of Edinburgh (Institute of Geography) is not permitted.

I agree that this dissertation and associated electronic documents, web pages, data, files and computer programs can be retained by the University. YES

I agree that, with the permission of my supervisor(s) or the Programme Director, these materials be made available for the purposes of preparing a publication. YES

I agree that, with the permission of my supervisor(s) or the Programme Director, these materials can be used within the University of Edinburgh for continued research. YES

Statement of Originality

I declare that this dissertation represents my own work, and that where the work of others has been used it has been duly accredited. I further declare that the length of the components of this dissertation is 4946 words for the Research Paper and 7698 words for the Supporting Document.



KAJA RØNNING

12.08.2020

Acknowledgements

I would like to thank Dr Anthony Newton for being an excellent supervisor, and for his guidance and continuous support throughout. I further wish to thank Dr Rachel Opitz, who provided the data for the project and showed great interest in my work which has been very motivating. Additionally, Dr Emily Lethbridge, Dr Phil Buckland and Tom Ryan, who patiently answered all my questions regarding their research and datasets, and the rest of the dataARC team, for taking the time to provide feedback and discuss ideas.

Lastly, a very heartfelt thank you and “pat on the back” to all my course mates, for your encouragement and perseverance during this strange time. Completing a master’s dissertation in the midst a global pandemic is an impressive accomplishment, and I am very proud of all of us for managing to pull through.

PART ONE
Research Paper

Abstract

Research on how humans have interacted with a changing environment over time requires linking complex data and information from a range of disciplines and contextualise it in both time and space. In recent years such interdisciplinary research has become increasingly more frequent as a way of unveiling hidden patterns between data from a wide range of subjects. One such research initiative is dataARC, whose objective is to enable studies of human ecodynamics around the North Atlantic during the middle ages, using both archaeological, environmental and historical data.

This paper describes a project developed within the wider framework of dataARC and aims to help bridge the gap between research from multiple disciplines in a spatial context by implementing a multi-dimensional approach and produce a visualising tool which effectively combines cross-disciplinary datasets and appropriately map their connections in geographic space.

The study focuses on investigating spatial connections between literary, environmental and zooarchaeological data from Iceland, in order to create a visualisation prototype which can be implemented into the continuing work of dataARC. Using Self-Organising Maps (SOM), an unsupervised clustering technique, the study explores methods synthesising this information by identifying 10 cluster profiles with specific signatures related to the combination of sets of attributes or indicators. This analysis makes clear the considerable potential of a SOM approach in advancing pattern recognition between cross-disciplinary data.

Keywords: Self-Organising Maps, Cluster Analysis, High-Dimensional Data Visualisation, Environmental Archaeology, Human Ecodynamics, Cross-Disciplinary Research

1. Introduction

1.1. Humans, environment and the bridge between them

The history of the first human settlements in the North Atlantic region is one of constant change and adaption concurrently with rapid and dramatic environmental and climatic changes (McGovern et al. 2006). Archaeological evidence from islands in this region, such as Iceland, Orkney, Shetland, the Faroe Islands, Greenland, Ireland, and the Wester Isles, hold copious amounts of information on how human settlements have evolved interacted with their changing surroundings over time (Hartman et al. 2017). Palaeoenvironmental proxies have provided us with thorough insights into where and how climate and ecosystems were changing and their impacts on societies (Gupta et al. 2003).

The dynamic nature of humans, their societies and way of life, as well as the surrounding environment and ecosystems, force a sort of evolutionary symbiosis, where both dimensions are actively responding and adapting to changes with one another (Amorosi et al. 1997; Mairs et al. 2006; Dugmore et al. 2005). Linking human nature and history to changes in the environment has been studied extensively across different continents (McGovern et al. 1988; Haldon et al. 2018). Many of these studies focus on specific sites, events or identifying correlations between 2 variables, such as temperature changes and human migration (Fricke et al. 1995) or assess human responses to unpredictable changes in climate (Dugmore et al. 2007). However, there have been few attempts to create spatial links between archaeological and environmental datasets and to map these effectively.

Collaboration between researchers from different disciplines is crucial if we want to unveil how cultures and societies have co-existed and co-evolved with their surrounding environment and climate over time (Smiarowski et al. 2017). Inclusion and implementation of varied cross-disciplinary datasets into such analyses is important and has become increasingly more common over the past decades (Aagaard-Hansen, 2007). This shift in applied research practice opens a range of new challenges and obstacles which must be overcome in order to obtain findings and results that are both reliable and understandable (Butzer, 2008). Connectivity is especially important in archaeological thinking, as any archaeological site is embedded in an elaborate relational network with its surrounding environment (Pálsson, 2018). In order to understand relations and processes of adaptation, inclusion of varied and interdisciplinary datasets into the analysis is crucial.

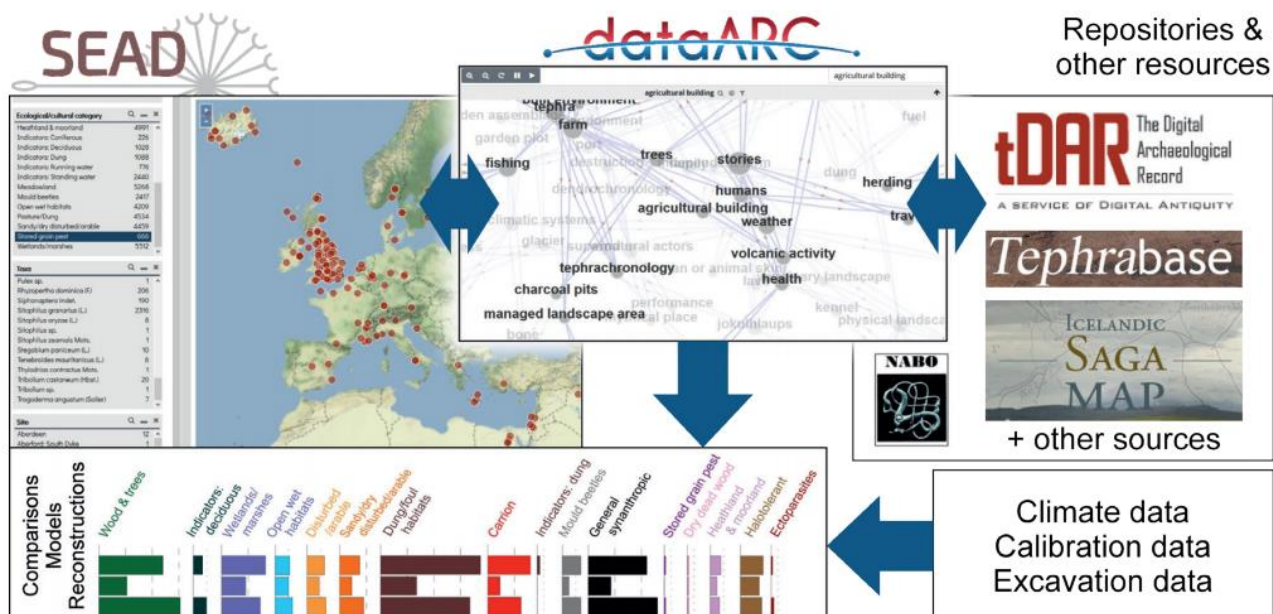
Regular spatial analysis methods have several limitations that make them unsuitable for or unable to explore large multidimensional datasets used in cross-disciplinary research (Gahegan et al. 2001). New approaches focus on improving pattern recognition within the datasets as a more effective type of data mining (Miller, 2010). Bringing together and comparing data from such a range of disciplines in a coherent and meaningful way is challenging for many reasons. Some of these issues include differences in data scales, values and level of detail, level of data complexity, and varying levels of objectivity vs subjectivity of the datasets (Aagaard-

Hansen, 2007). Within cross-disciplinary research one often tends to try to find a balance between including too much or too little detail from each discipline and dataset (Allard & Allard, 2009). Complexity must be reduced in order to produce an output that makes sense, however the risk of oversimplifying the data and thus losing a lot of dimension is very present.

Visualisation is an important part of data mining, especially in terms of combining several datasets because it can help with reduction of dimensions which presents the data in a way that is much more meaningful to a varied audience, from data providers to funders or the general public (Keim, 2002; Hofman & Chisholm, 2016). The concept is based on attempting to approach the data from different angles and perspectives to hopefully identify hidden patterns and connections (Lin et al. 2011). Issues arrive when attempting to spatially combine the datasets in a way which both preserves the integrity of the data and produces an output that is comprehensible to researchers from many backgrounds or disciplines.

1.2.dataARC- components and purpose

The dataARC project, funded by the National Science Foundation (NSF), aims to combine several environmental, archaeological and textual datasets from the North Atlantic region in order to encourage and aid interdisciplinary research (Figure 1). This type of data synthesising is important for several reasons, most of which ties back to the fact that humans, nature and environment need to be seen in context (McGovern, 2014). Although clearly important, finding an ideal way of spatially combining interdisciplinary datasets in the most efficient, accurate and least time consuming and error prone way has proven challenging.



The main aim for dataARC is to implement all datasets into one shared project which can be queried to study

Figure 1: Interdisciplinary research process using the dataARC concept map, which links data and information using information tags or “concepts” (from Kohler et al. 2018).

connections between all data points (dataARC, 2019).

The current dataARC model consist of two elements:

1. Map prototype. This is an interactive map of all the included study areas in the North Atlantic. The map can be queried by specific keywords, concepts, time period or place names, which help researchers look for and explore connections in the datasets
2. Concept map. This is a web map which aims to link datasets in the most coherent way through shared concepts and various data combinators (Figure 2).

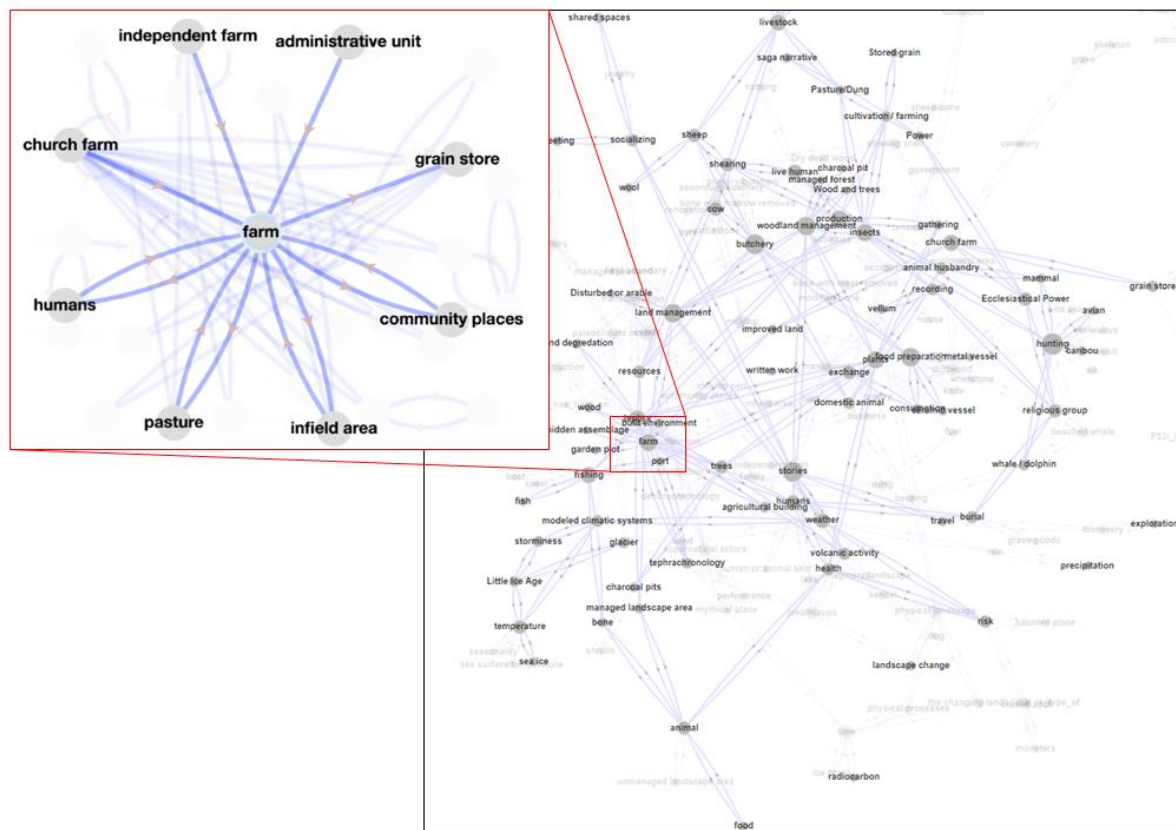


Figure 2: Snippet of the dataARC concept map. Datasets are connected by concepts and each data contributor will label their included datasets with a range of concepts. Close-up of example of the concept “farm” which is connected to 8 other concepts. Source: dataARC, 2019

Although highly useful for interdisciplinary research, both the shared map project and the concept map become increasingly more complex once you start to zoom out and study more than one or a few connections at a time. The connections between humans and their surroundings are inherently complex and thus difficult to visualize in a meaningful or comprehensive way. Concepts maps are intricate and offer intriguing insight into cross-disciplinary connections on a concept-by-concept scale (Kohler et al. 2018). Identifying overall spatial trends or connections in the data is however challenging using a concept map.

1.3. Research focus/aims

The final goal for the dataARC project is to build a model which help researchers from the different disciplines that have contributed data (palaeoenvironmental, archaeological etc.) and identify connections between their work and others, based on relationships linked to both time, space and concepts or topics. My study focuses exclusively on spatial visualisation methods. Through the application of unsupervised clustering we aim to both contribute to dataARC's further research on human ecodynamics in the North Atlantic, as well as assess the usefulness of this methodology on cross-disciplinary data mining. The results from my study should help to show relationships between multidisciplinary datasets in a spatial setting, which will aid further investigation of connections between data from a range of different disciplines.

This will be a novel approach into examining and combining historical interdisciplinary datasets using an unsupervised clustering technique. The primary aims for this project are as follows:

1. Build a model that links datasets from multiple disciplines (archaeological, palaeoenvironmental and textual) by identifying a series of clusters where the values for several or all of the included datasets are similar
2. Effectively map relationships and connections linking cross-disciplinary data contributed into the dataARC project by creating a visualised output of patterns within and between individual datasets

The end product of the project will be a series of identified clusters of spatial areas where the values for several or all of the included datasets are similar, which will be implemented into the continuing work of the dataARC team on relationship mapping between multiple disciplines.

1.4. Study Area

While the dataARC project include data from all over the North Atlantic, this study will focus primarily on Iceland. The data point density is currently the highest here, especially for textual data, which increases the likelihood of the analysis being able to find anything of interest (Figure 3). The overlap of data from multiple datasets and disciplines is also quite significant. Furthermore, because it is an island it is naturally geographically isolated. Modern Iceland is divided into municipalities, or *Sveitarfélög* (Sverrisson & Hannesson, 2014). Both the municipalities and general place names have been stable and subject to very little change over time (Lethbridge, 2016).

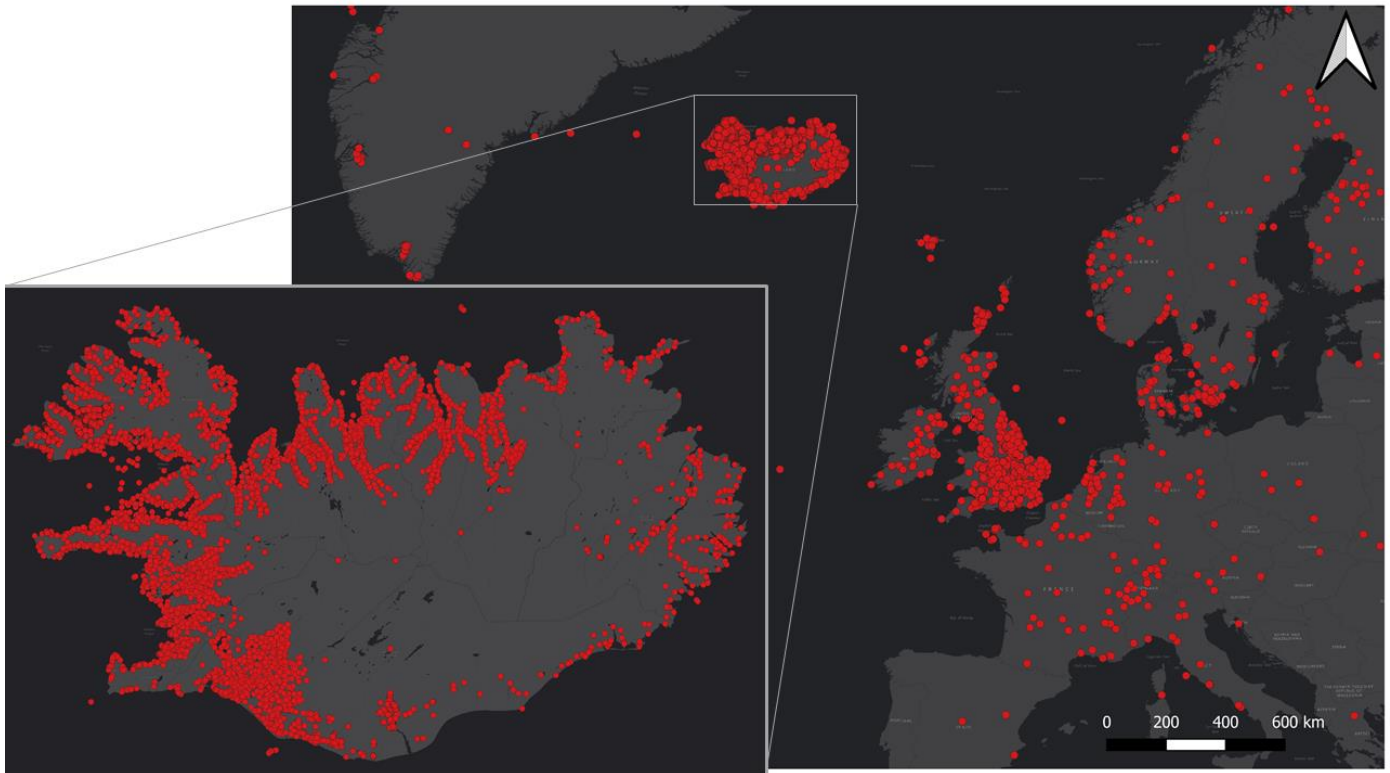


Figure 3: The spread and density of data points currently implemented into the dataARC project. Although data points are present across mainland Europe, dataARC focus primarily on islands and regions in the North Atlantic. This makes areas like England, which shows an equally high point density, unsuitable for this project. Out of the North Atlantic Islands, Iceland shows both the highest point density and the highest data overlap.

The history of the first Icelandic settlers has been preserved and transmitted in text as stories or sagas (Boulhosa, 2005; Orning, 2015). This puts the region in a special position in terms of our knowledge of the early settlements, where they lived and how they interacted with the landscape and their surroundings (Wyatt, 2004). The combination of well-preserved historical and literary documentation and data from extensive palaeoenvironmental and archaeological records that exist for Iceland makes this region ideal for developing our multidisciplinary data analysis model (Price & Gestsdóttir, 2006; McGovern et al. 2017).

2. Methodology

2.1. Workflow model

A conceptual model of the total workflow from start to finish is presented in Figure 4. I aspire to present visualising techniques which can be used for all components included in dataARC, however only a few selected datasets within a specified geographic range will be incorporated into this initial prototype. These are: Sagamap, the Strategic Environmental Archaeology Database (SEAD) and bone data from North Atlantic Biocultural Organization (NABOne).

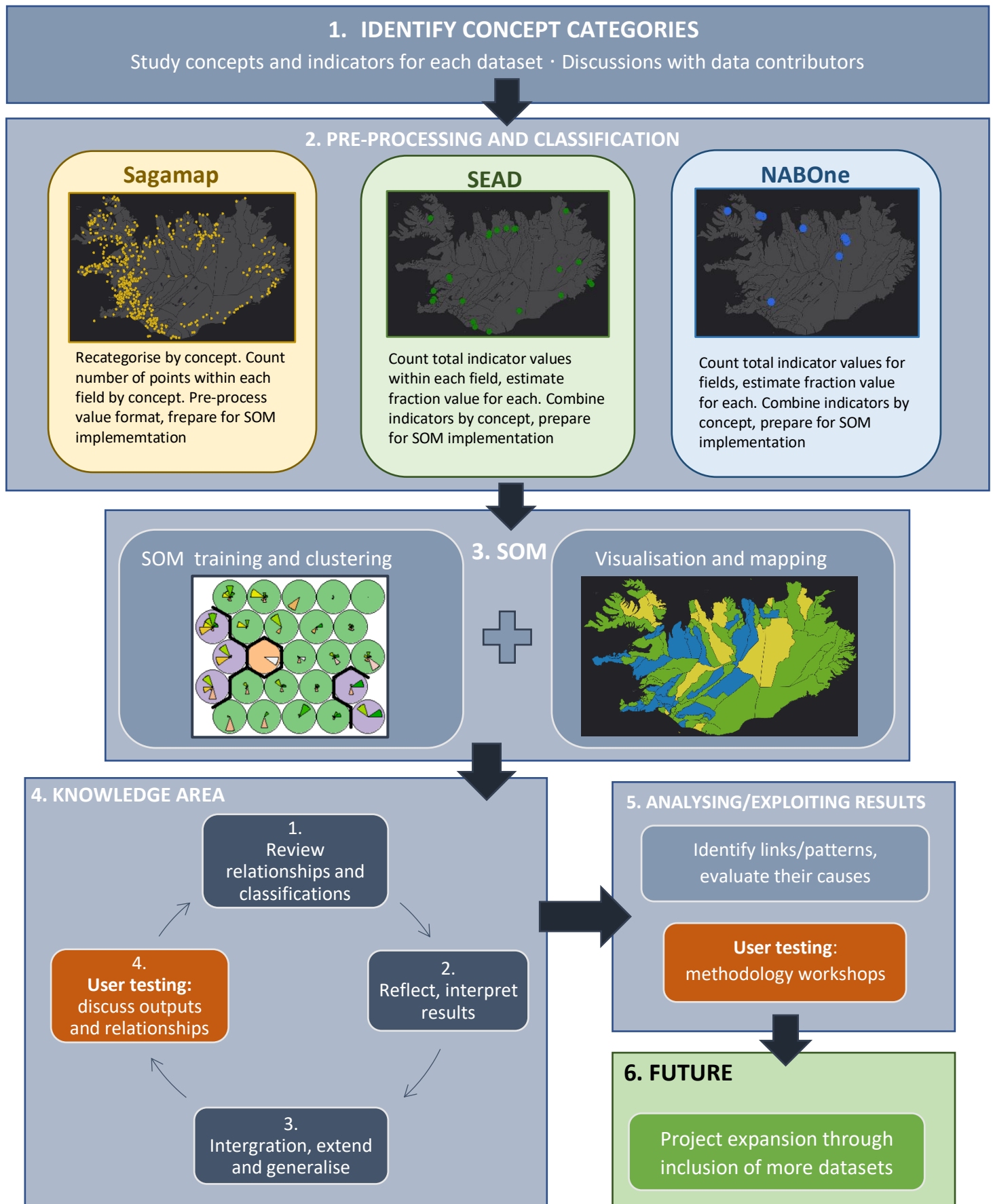


Figure 4: Workflow model explaining core methodology for this analysis, including flowchart for SOM computation analysis

2.2. Datasets and concept categories

3 datasets from the current dataARC project were selected for implementation into this initial analysis. The analysis itself is designed with every dataset currently included in dataARC in mind, as these will be implemented at a later stage by the research team. As the aim for this exercise is to build a visualising tool which can identify and present spatial connections between archaeological, environmental and historical data, one dataset from each discipline was selected based on size and spread of data points. Sagamap is a textual literary dataset, SEAD a palaeoenvironmental dataset constructed using a range of different proxies, and NABOne is a zooarchaeological dataset consisting of bone data. The 3 datasets are presented in Table 1.

Table 1: 3 included datasets, categorised and with additional information

Dataset	Category	# of entities	Additional information
SEAD	Environmental	Total: 9139 Iceland: 456	Strategic Environmental Archaeology Database. Mainly chemical, physical and biological proxy data derived from e.g. fossils, soil samples, geoarchaeological data and dendrochronological analyses. Also include insect/pollen/plant datasets that are used for palaeoenvironmental reconstruction (Buckland et al. 2018) Project website: https://www.sead.se/
Sagamap	Textual	Total: 4652 Iceland: 3869	Dataset containing places and locations mentioned in 42 Icelandic sagas. Each mention of a place is tagged with one or several concepts indicating or explaining what happens at the site or whether animals, buildings or items are found or seen there (Lethbridge, 2016) Project website: http://sagamap.hi.is/is/
NABOne data prepared for incorporation into SEAD	Zooarchaeological	Total: 928 Iceland: 928	Bone data collected and recorded by the North Atlantic Biocultural Organization (NABO) Zooarchaeology Working Group Data Records Project. NABO works to combine data from different disciplines in order to improve the research potential in the North Atlantic and, with the overarching aim being to reconstruct long term human ecodynamics by building and combining palaeoecological and geoarchaeological datasets. This particular dataset has been prepared for incorporation into the SEAD database (McGovern, 2014; Strawhacker et al. 2015). Project website: https://www.nabohome.org/

Data points from Sagamap, SEAD and NABOne are spread across Iceland, however the degree of spread differs quite significantly. Where Sagamap data is well represented across most of Iceland, there are currently only about 9 sites where zooarchaeological data has been collected and implemented into NABOne (Figure 5).

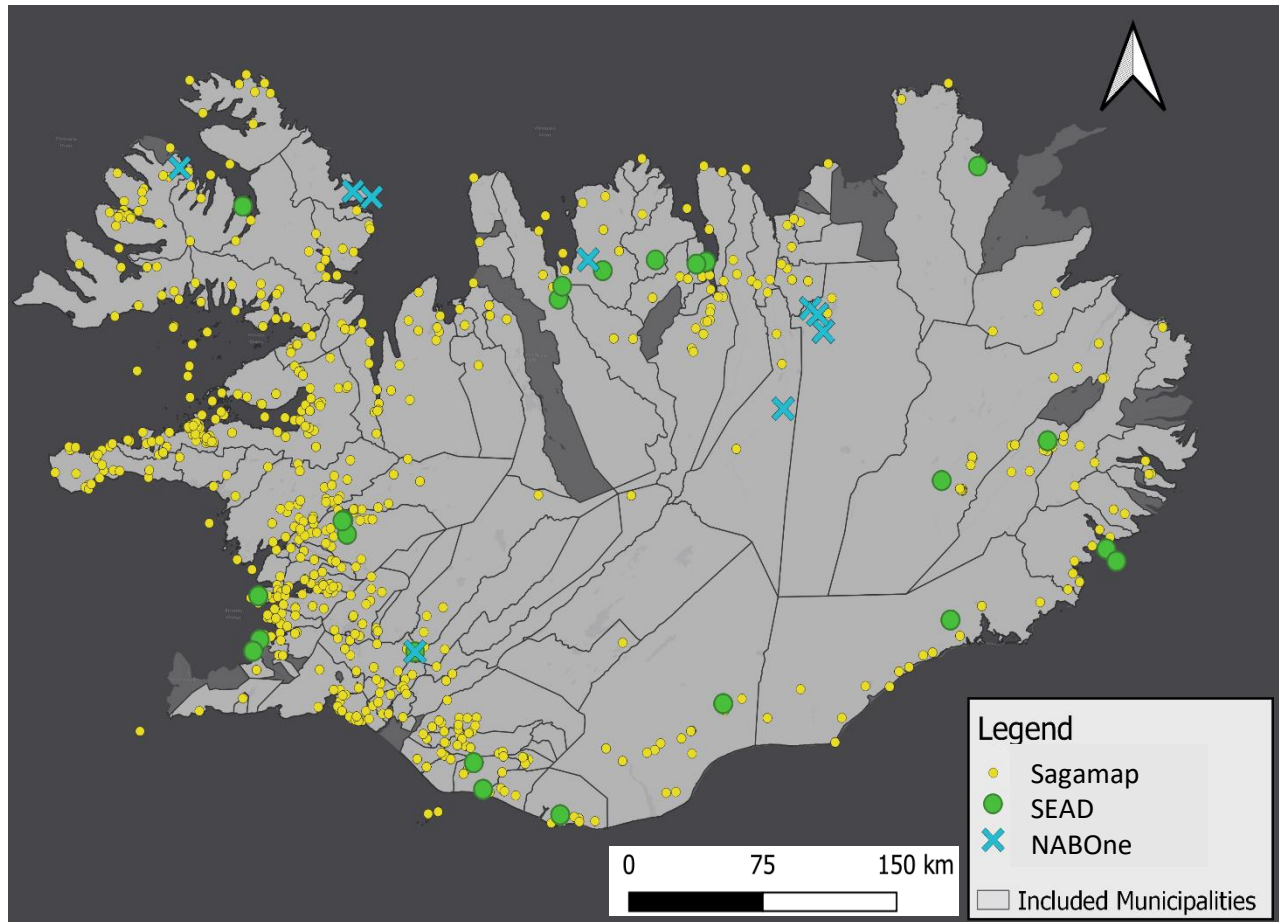


Figure 5: Geographic spread of data points for Sagamap (yellow), SEAD (blue) and NABOne (green). SEAD and NABOne points represent excavation sites; each site can represent hundreds of data points.

In addition to the datasets representing different disciplines, their varying data format was also accounted for. For these datasets to be accurate representations of the total dataARC data bank they need to reflect the differences in data formatting, which again reflects differences in data collection methods and the way scientists perceive and approach their data (Aagaard-Hansen, 2007).

In order to successfully combine datasets that have all been categorised on different scales with variable numbers of categories it is required that we develop a common scale of categories. This will also help downscale complexity within the individual datasets. From discussions with dataARC team members and literature research, 10 main concept categories have been selected to reflect the main topics and indicators each dataset represents. These concepts are presented in Table 2 along with a short description of each. For a more

detailed description of the concept categories and the individual categorisation of each dataset see Rønning 2020 section 5.

Table 2: *The 10 defined concept categories which are used to categorise each dataset onto a similar scale*

<i>Concept</i>	<i>Description</i>
<i>Activities</i>	Any mentions or evidence of human activities, apart from travelling or water related activities
<i>Buildings</i>	Any human-made buildings or construction, apart from cairns
<i>Managed</i>	Managed landscape, any evidence of land alteration or management by humans
<i>Domestic</i>	Domestic animals, livestock
<i>Natural</i>	Natural landscape, no or little human alteration
<i>Wild</i>	Wild animals, not managed by or living in relation to humans
<i>Water</i>	Water related activities, evidence or indicators of water bodies or of animals living in or near water
<i>Travel</i>	Any mentions or evidence of people travelling
<i>Weather</i>	Any weather observations
<i>Things</i>	Objects related to humans that are not buildings or animals

2.3. Self-Organising Maps: clustering and visualisation

This analysis applies Self-Organising Maps (SOMs) to the wide range of cross-disciplinary data as a way of combining and comparing them. SOMs are an increasingly popular data mining technique which applies computational clustering analysis to large and often heterogenous datasets in order to identify hidden patterns and connections within the data (Kohonen, 1989). The objective of a SOM is to project high-dimensional or multivariate data representing three or more independent parameters or features, onto a two-dimensional plane or grid without having to compromise on the complexity of the data (Whelan et al. 2010).

Using unsupervised training and neural network analysis, a SOM arrange areas or data into clusters based on shared characteristics, where the result is a grouping of areas with similarities in characteristics and data values

(Koua & Kraak, 2005; Skupin & Agarwal, 2008). The number of clusters are determined by the user depending on data structure or preferred output (Kohonen, 2013). Highly detailed or variable data usually requires a higher number of cluster types than more homogenous datasets (Brereton, 2012).

For this study a SOM map will be used to identify and present regions or areas in Iceland where data or information from several disciplines indicate similar conditions or the presence of specific features.

The SOM process itself can be divided into 2 main stages:

1. SOM training. Implement your standardised datasets into the SOM and train the model
2. Clustering and mapping of training results. Identify suitable number of clusters based on results from the training process (Figure 6). Visualise results spatially

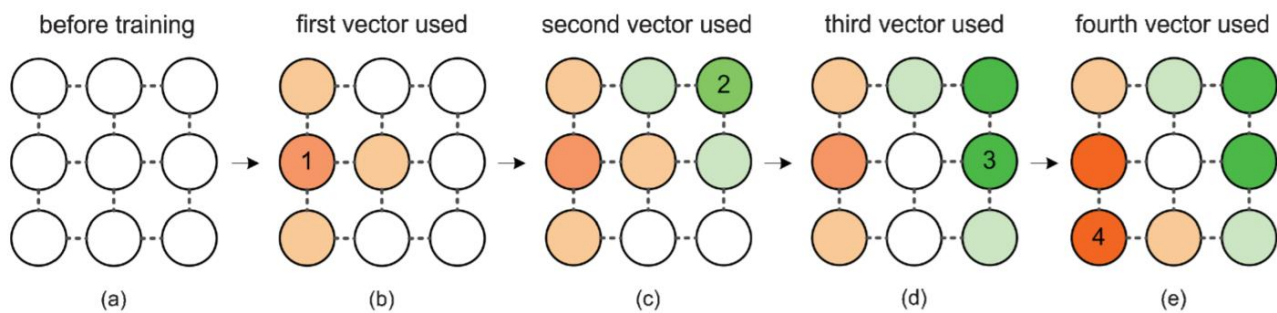


Figure 6: Process of SOM training using 3x3 grid and 4 vectors, or cluster types. As each vector finds their best matching unit, weights and surrounding nodes are being adjusted to match input vectors. Source: Skupin & Agarwal (2008).

The output of a SOM training process is presented as a pre-defined number of neurons presented on a lattice, where each neuron, or map unit, is attached to the input data (Koua & Kraak, 2005). During training the individual input data objects, in this case Icelandic municipalities, are presented to the SOM lattice one by one. The SOM is trained to produce “model units” which best represent the input data. One unit can represent one or several data objects. For a more extensive breakdown of the training process see Rønning (2020) section 6.2.2.

Units are partitioned into homogenous regions, or clusters, which helps reduce the high level of detail of the SOM output (Whelan et al. 2010). The structure and components of the SOM grid is described in Figure 7.

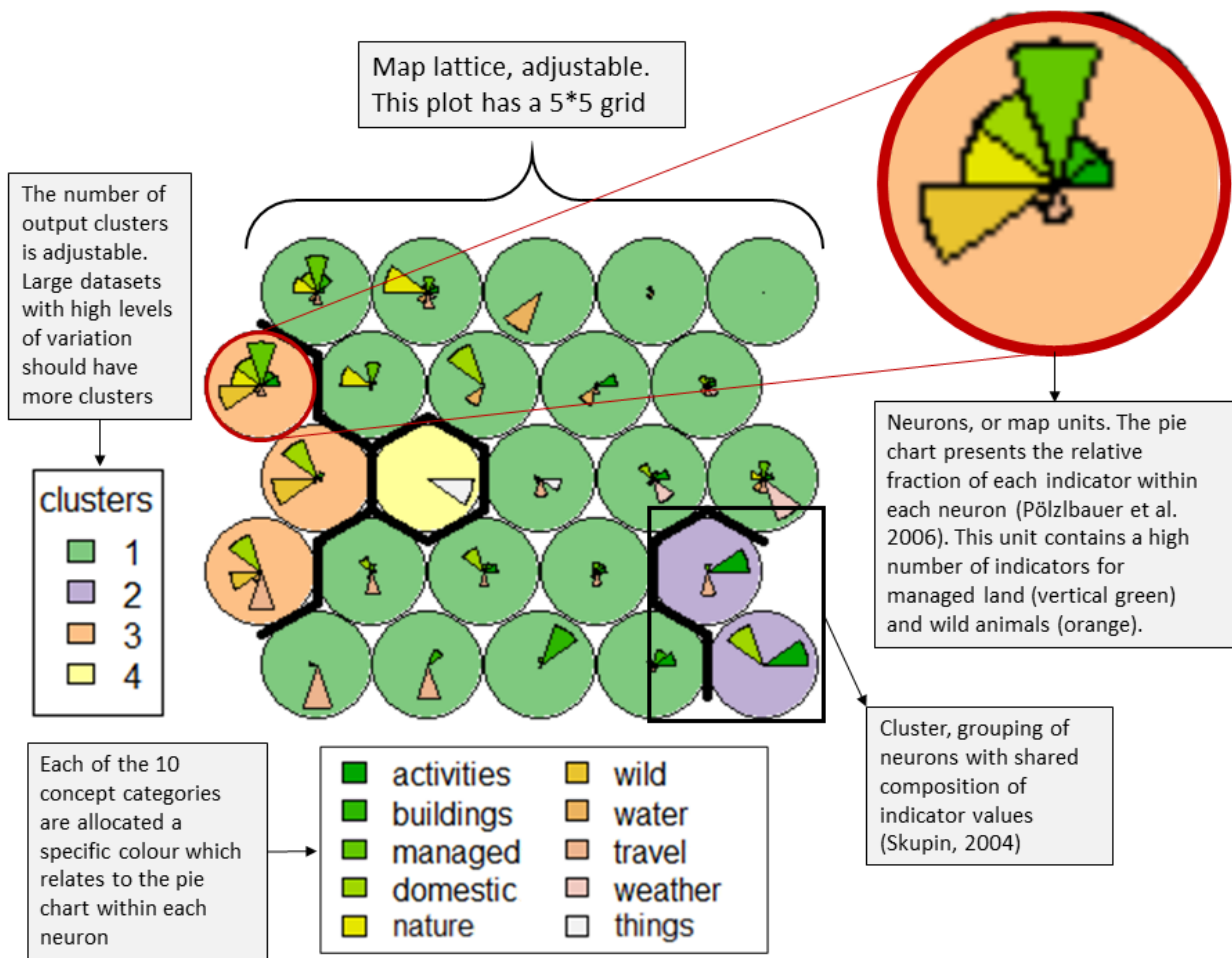


Figure 7: Components of a SOM training and clustering output explained, for a dataset with 4 defined clusters and 10 concept categories on a 5*5 grid.

Variables such as SOM grid parameters and number of clusters are defined based on the size of the included datasets (Kohonen, 2001). Larger datasets will, naturally, require larger grids, as is evident from Whelan et al.'s (2010) research on deprivation in Ireland. Ireland is divided into 119 municipalities, where 95 of them contain data points from the 3 included datasets. thus, an 8x8 grid is reasonable. Using a smaller grid risks over-generalising the data division, larger grids overcomplicate the data output as well as unnecessarily extending processing time (Kanevski et al. 2009).

The 10 concept categories and their abundances are represented as segments within each neuron (Figure 8) where the relative size of each segment represent the abundance of points and values correlated to this category within the area (Wehrens & Buydens, 2007).

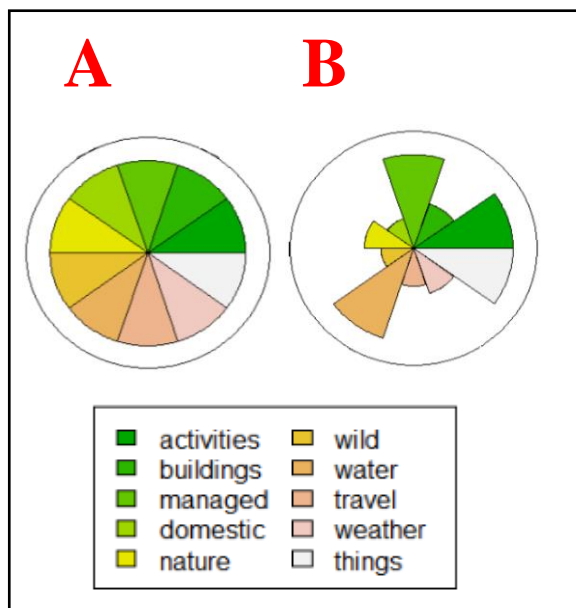


Figure 8: Example segments charts within each neuron representing the relative abundance/value for each concept category within that neuron. A represents a neuron where the values are the same for all concept categories, whereas B presents a more realistic output where the categories “activities”, “managed”, “wild” and “things” have the highest values.

Following SOM training and clustering the results can be portrayed visually by allocating symbols to points or polygons with a geographic location based on the cluster they belong to (Koua & Kraak, 2005).

Although this study focus on building a mapping tool where a range of cross-disciplinary data can be combined and visualised together, visualisations of the spread and different indicator combinations throughout Iceland should be made on an individual dataset scale as well, as a supplement to the main SOM output. This might help identify initial similarities between data values for the 3 datasets spatially across Iceland and be highly useful for individual data contributors or for anyone wanting to study intradisciplinary concept patterns or connections between only a few sets of data. Additionally, such maps will act as a form of validity test of the main map and present patterns which are not visible on a larger cross-disciplinary scale.

2.4 User testing

The overarching aim for this study is to produce a visualisation technique that can be implemented and used by the dataARC team. It is thus crucial to include the team in the development of the project. Both to make sure the end product is understandable to them, fulfils its purpose and fits in with other components of the dataARC project as a whole (Grudin, 2017). The technique is created with the data providers in mind and is designed for them to use and continue to implement their own data into on the future. Feedback from the dataARC team has also been necessary for developing the final list of 10 concept categories.

User testing has been conducted through both discussions and workshops. Discussions with the data providers Emily Lethbridge (Sagamap), Phil Buckland (SEAD) and Tom Ryan (NABOne) inspired the development of a list of concept categories. Because the structures of each dataset are complex and specific, continuous updates

and demonstrations of the analysis and model should be provided to the data contributors to make sure the integrity of their data is preserved in the final model output. Discussions with the wider dataARC team provide insight into the importance of their work as well as the main issues the team are facing in terms of spatially combining, analysing and querying cross-disciplinary data. These include:

- *Differing dataset structure, format and number of indicators for all datasets makes it difficult to combine them*
- *Significant variations in the geographic spread of each dataset*
- *Preservation of data complexity during dimension reduction*

Because this model is meant to be implemented into a data mining prototype consisting of various components, knowing and understanding the structure of these components is important. Team workshops were held to demonstrate the use of various parts of the dataARC prototype, such as the concept map. Additionally, the final model produced in this study was presented and during various user testing workshops. The workflow is demonstrated to the group, and any scripts are provided. This let the members attempt to run the analysis with their own data and provide feedback on efficiency and user-friendliness of workflow and code.

3. Results

3.1. SOM training and clustering

The output from the SOM training is presented on an 8x8 grid in Figure 9 below. The variation within and between the neurons can be seen as a measure of the vast differences in category compositions within individual municipalities.

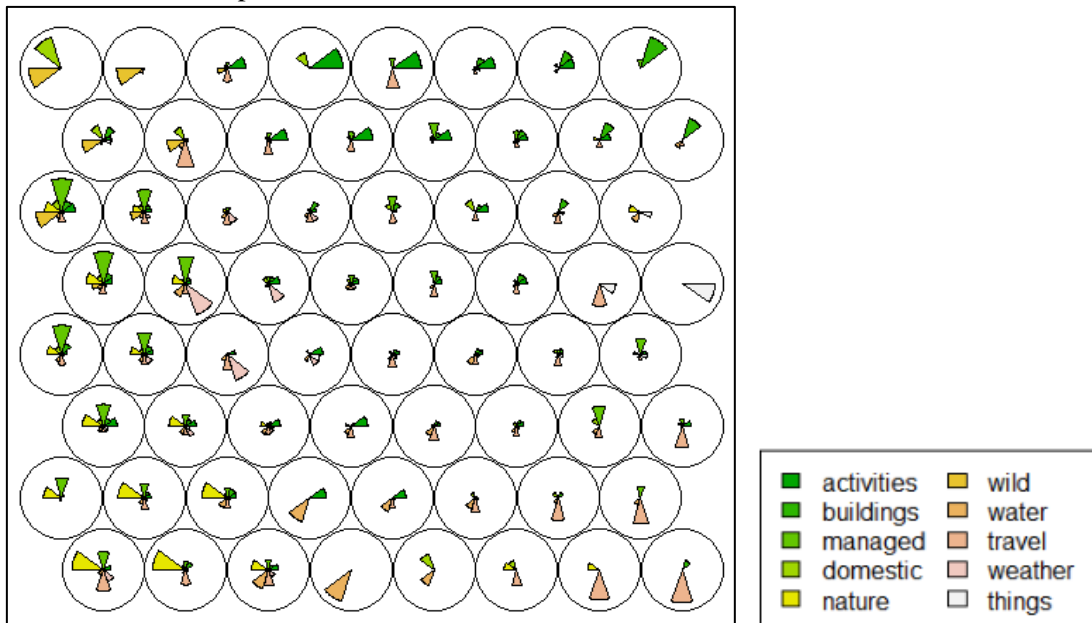


Figure 9: SOM training output, represented by segments plots within each unit.

From the scatter of data, grid size and number of concept categories included in the analysis, we identify ten clusters to be the most ideal number for the clustering of the combined datasets. The 10 clusters are projected in a lattice structure space shown in Figure 10.

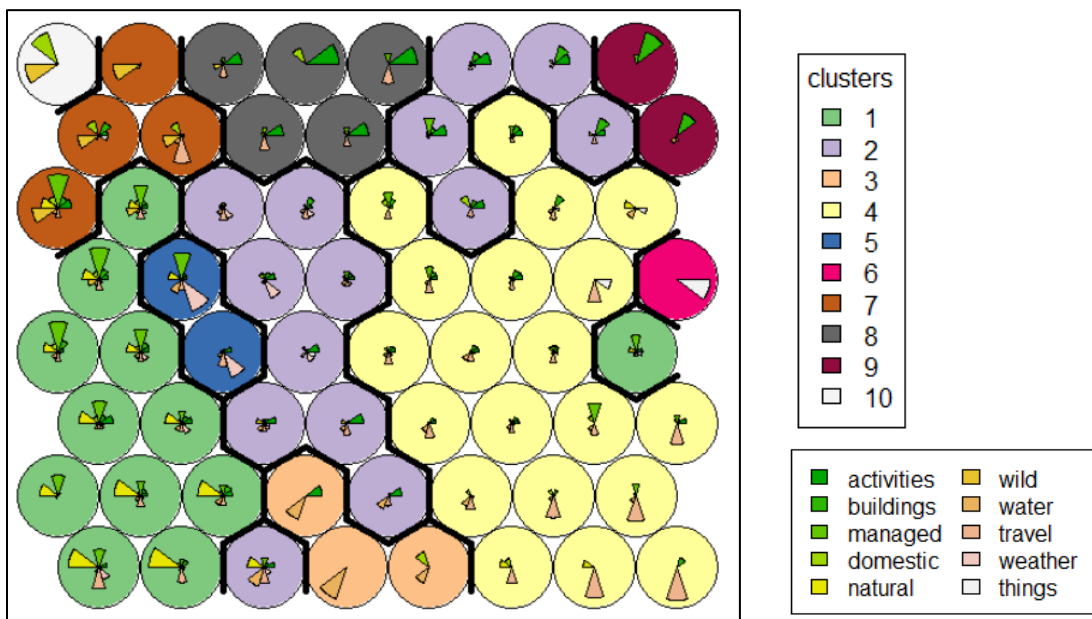


Figure 10: Final output of the SOM training for the 3 combined datasets, clustered into 10 respective clusters.

3.2. Visualisation and mapping

Mapping the cluster output aids the visual exploration part of this study by allowing us to get further insight into the spatial connections within our data (Keim, 2002). Following cluster identification each municipality in Iceland has been coloured to reflect their respective cluster (Figure 11).

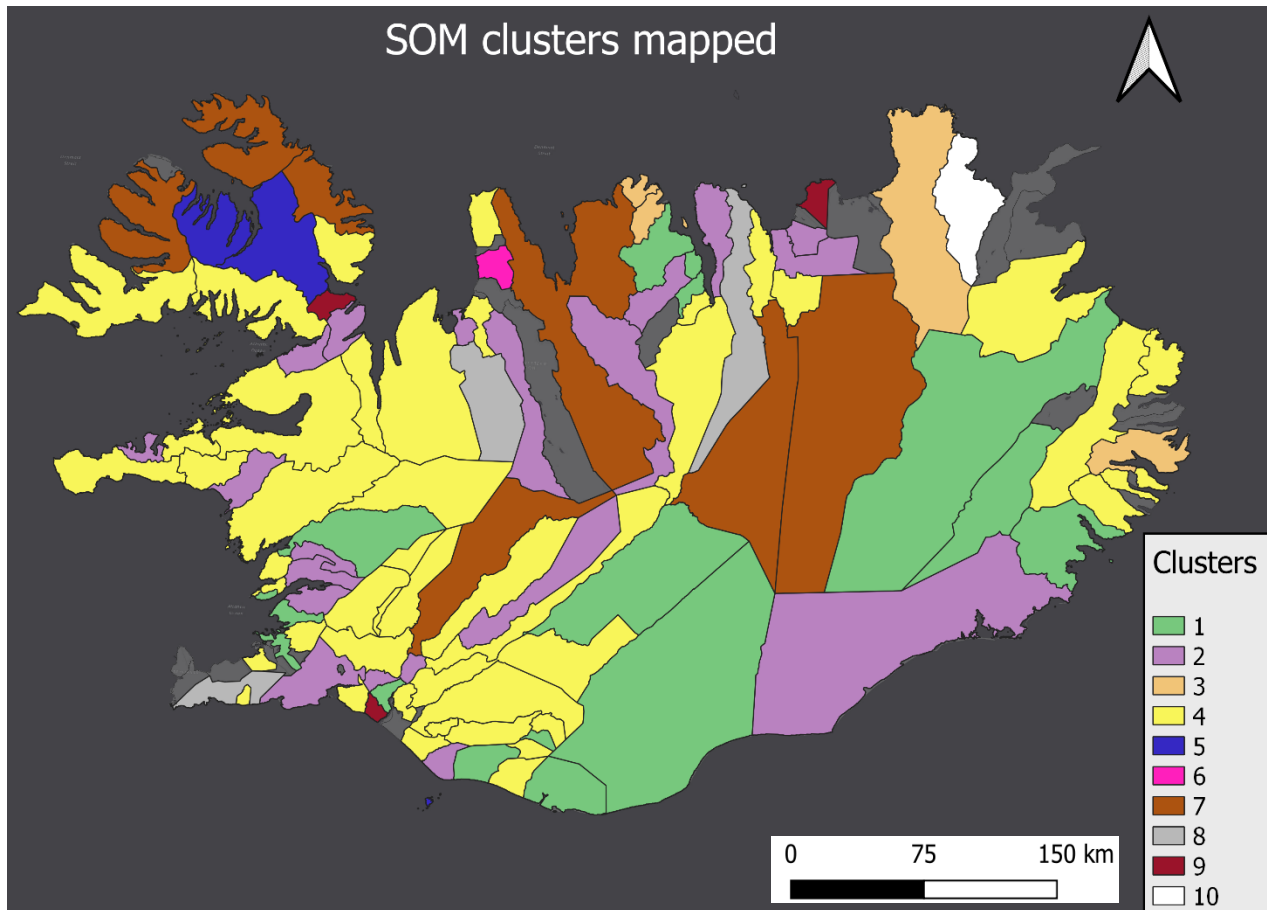



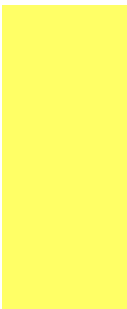



Figure 11: Visual representation of the clustered SOM results for all datasets combined. Municipalities containing no data have been excluded from the mapping. Numbering and colour coding of the clusters correspond with the training output shown in Figure 10.

By analysing the SOM grid, the unsupervised clustering and mapped output of this in unison, a brief interpretation of each of the 10 clusters are set out in Table 3.

Table 3: Descriptions of common features and concepts present within each cluster, with a short interpretation of what these findings suggest about the regions

Cluster	Colour	Description
1		Strong indications of both managed and natural landscape, as well as frequent mentions of travel. Landscape is clearly of importance in these regions, as mentions or indicators of both managed and unmanaged land are present. This cluster encompass regions made up by a mix of nature and cultivated land, areas that are clearly visited, travelled through s and changed by humans but not necessarily occupied by them.
2		Human activities are frequently mentioned, and evidence of buildings and constructions can be found both in text and in the archaeological record. There are also mentions of weather as well as indicators and mentions of water. More residential regions with little evidence of farming or cultivation of land
3		Very strong indicators of water, both in historical documents as well as in the zooarchaeological record. Some evidence of human activities and domestic animals as well. These areas could be used by humans for fishing and farming. There are no excavation sites positioned within these regions for neither SEAD nor NABOne.
4		Make up a large part of Iceland, particularly on the western side of the island. Travel and managed land are the categories most strongly represented here, we also see mentions and indicators of human belongings (things) and natural landscape. The majority of the municipalities that overall have the lowest number of indicators or mentions fall within this cluster. These are most likely regions used for farming and for people to travel through.
5		Quite few municipalities fall within this cluster, which suggest that the composition of concept categories here is vastly different compared to the rest of Iceland. Mentions of weather and indicators of managed landscapes are frequent here, especially weather mentions.

6

Another cluster containing only one municipality, named Vindhælishreppur. This region is dominated by mentions of things and belongings and contains no indicators of any other concept categories. There are no SEAD or NABOne excavation sites here, so the record is entirely made up of information from Icelandic sagas

7

This cluster represent municipalities where we see mostly evidence of wild animals, managed landscape and travelling. There are also a range of mentions of activities and buildings and some evidence of wild landscape but hardly any of water, weather or things and belongings. The only two regions which contain data from all 3 datasets, Skagafjörður and Biskupstungnahreppur, are categorised as cluster 7 (see section 4.2).

8

Very high number of mentions of human activities compare to the rest of the island, as well as some indicators of domestic animals and travelling. Again, this cluster seems to represent places used frequently by humans for both living, recreation, farming and travelling.

9

All records very strongly indicate the presence of buildings and constructions in these 3 municipalities, and not much else. Areas occupied by people, but of overall little significance. There is no evidence suggesting these areas were used for farming, activities or travelling. There are no SEAD or NABOne data present within these areas.

10

One single municipality is represented by cluster 10, which again symbolises the vast difference in concept category composition and values compared to the rest of Iceland. Here we only find evidence of animals, both wild and domestic. Data for this region is currently only available from the SEAD dataset.

3.3. Visualisation on a singular dataset scale

Results from individual visualisations of Sagamap, SEAD and NABOne are presented in Figures 12 and 13. The grid dimensions and number of defined clusters were customised for each dataset with respect to size and variation of data, as well as total number of municipalities within which data were present for each dataset (Wehrens & Buydens, 2007).

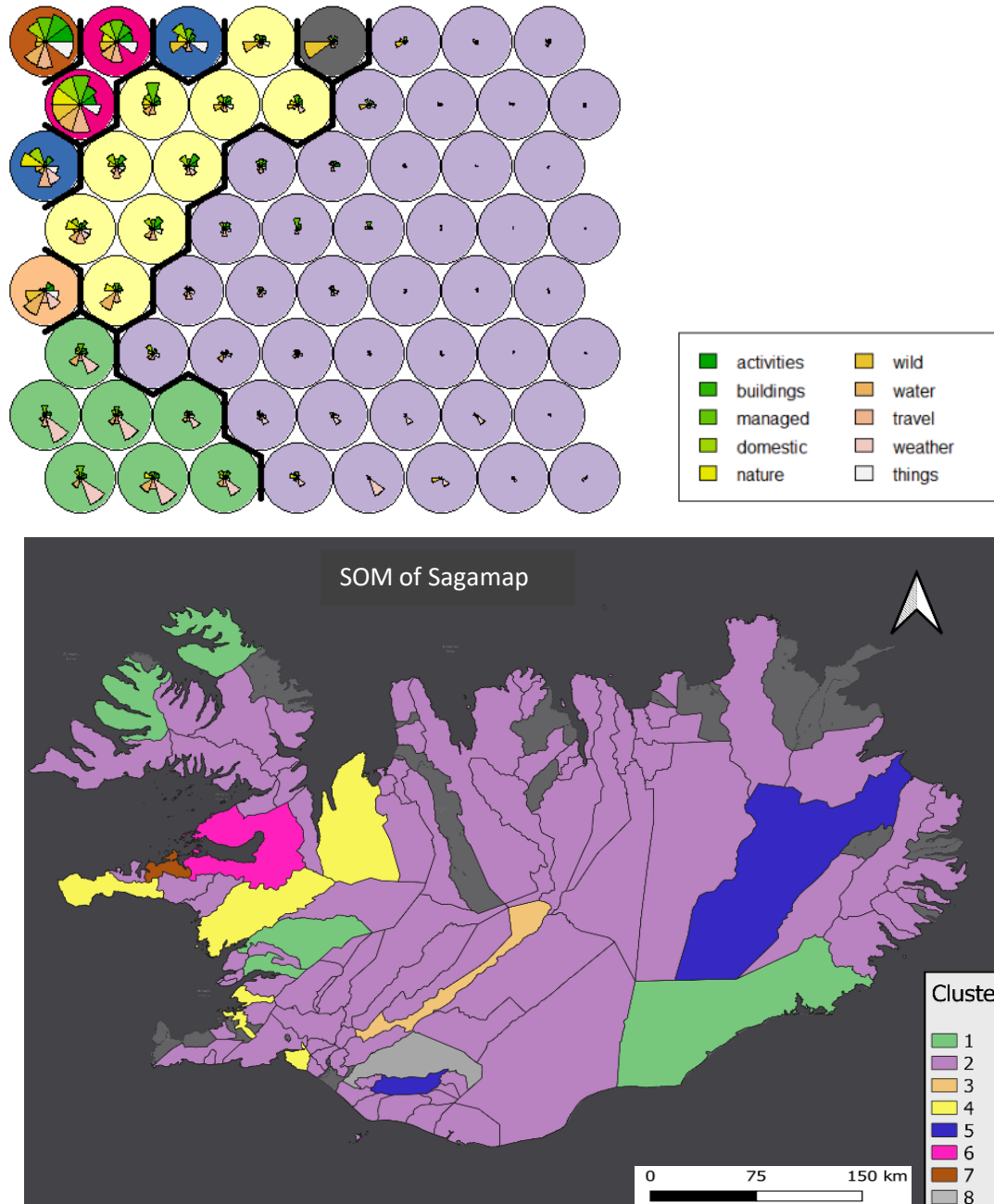


Figure 12: Results of SOM training and mapping of the Sagamap dataset. This dataset is spread over most of Iceland and is represented in every single concept category.

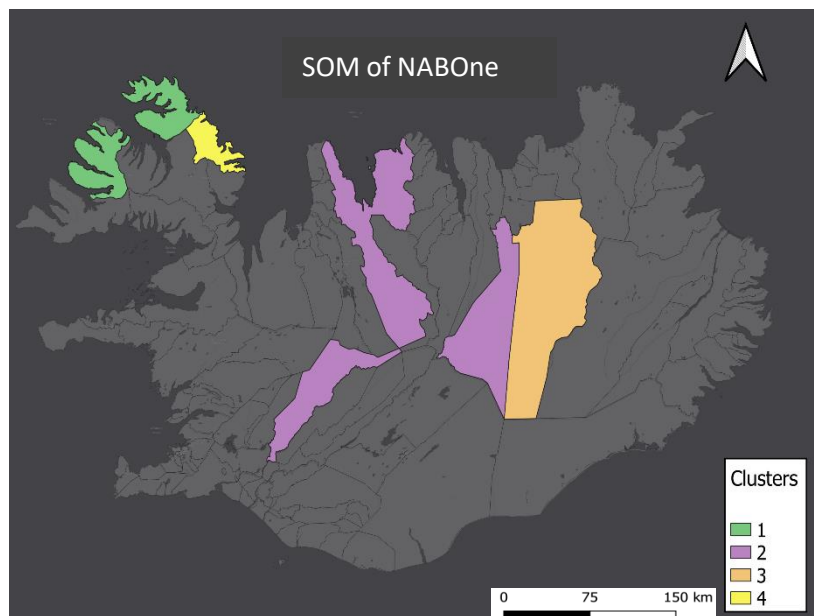
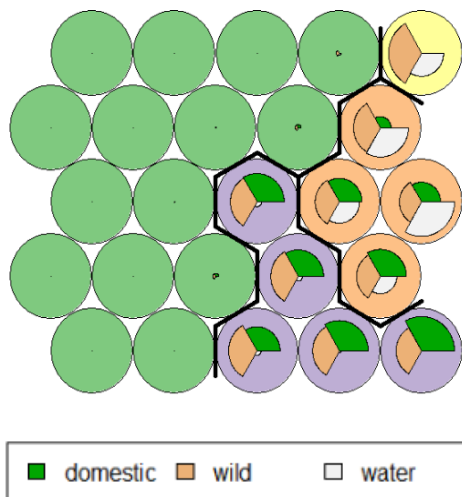
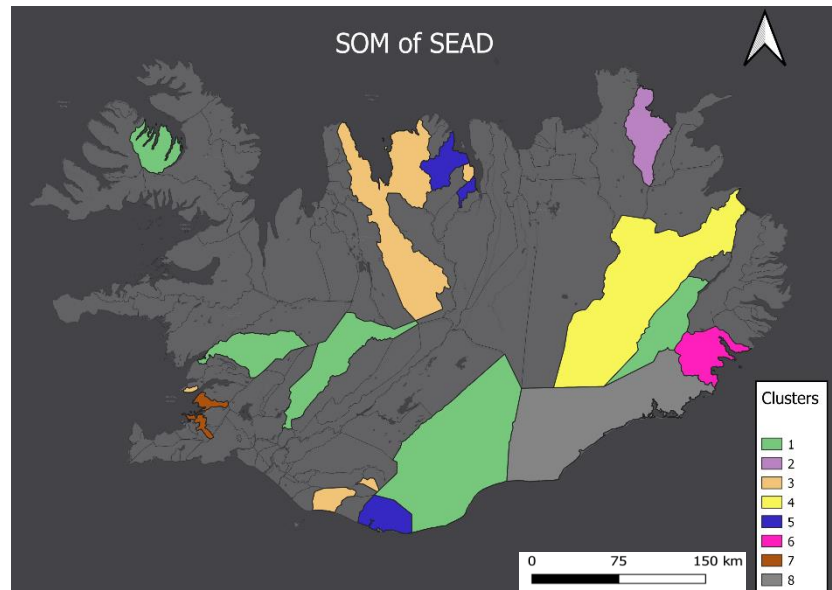
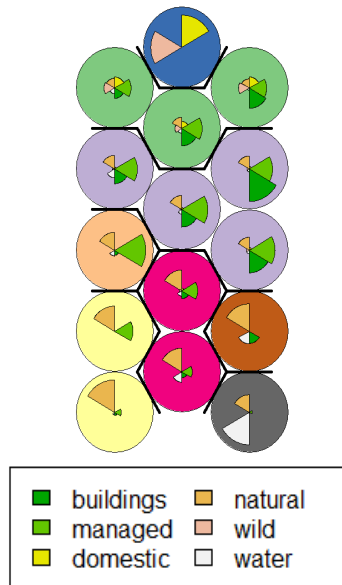


Figure 13: Results of SOM training and mapping of the SEAD and NABOne datasets respectively. Number of municipalities which are included for each, which are 17 and 7 for SEAD and NABOne respectively, indicate the current geographic spread of the two datasets in Iceland. Further, neither SEAD nor NABOne hold information related to every single concept category, which is evident from their neuron plots.

4. Discussion

4.1. Interpreting SOM results

From studying the raw output of the SOM training there seems to be very little evidence of any direct correlations between concept categories. Rather, the output suggests strong similarities between certain regions or municipalities in Iceland, as well as identifying areas that differ greatly from any other part of the island, such as Svalbarðshreppur (cluster 10) and Skagabyggð (cluster 6). The final output is to some extent affected by the lack of SEAD and NABOne data in certain areas, such as for cluster 3, 6 and 9. This is not so much a limitation as it is a call of attention to how the model is affected by geographic spread of data, and encouragement for more data to be implemented into the model.

Reykjavík, the current capital of Iceland, falls under cluster 1 (Figure 14). The first settlers on Iceland, led by Ingólfr Arnarson, settled in Reykjavík around 874 A.D. Given its long history of farming and cultivation up until the 18th century, it is sensible for the area near Reykjavík to be classified as cluster 1 (Róbertsdóttir, 2001). Þingvellir, the high seat of Iceland's first democratically elected parliament formed in the middle ages, around 930 A.D. (Bell, 2010), is situated within cluster 4. This cluster shows a high number of mentions of travel, managed land and human belongings, again a reasonable concept signature for politically important meeting spot like Þingvellir (Loftsdóttir & Lund, 2016).

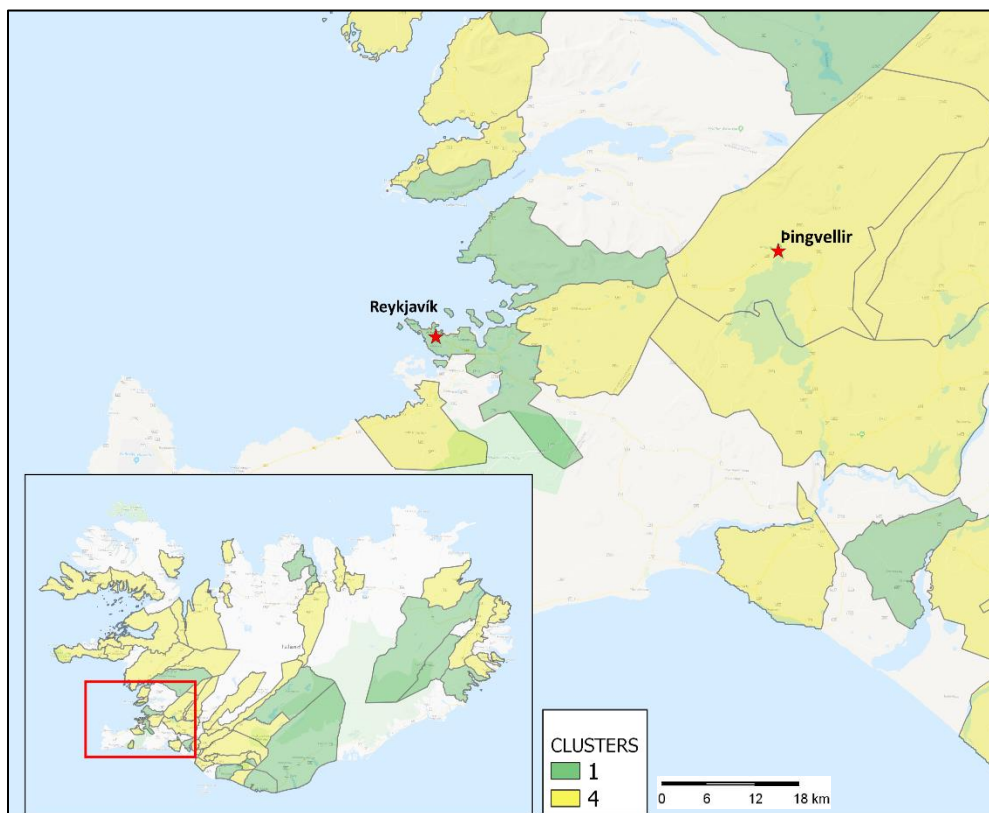


Figure 14: Locations of Reykjavík, the modern capital of Iceland, and Þingvellir, the former seat of parliament in the middle ages.

The clustering results successfully identified 10 clusters, all with quite distinct signatures and the mapped output seems reasonable given our knowledge about the history of people and their interaction with the environment in Iceland (McGovern et al. 2007; Rick et al. 2013). The signatures of the 4 largest clusters geographically (1, 2, 4 and 7) show interesting variations in concept category compositions. Where cluster 1 encompass what is most likely a mix of cultivated and natural landscapes with little evidence of human occupation (few mentions of buildings, activities or things but several of travelling), the municipalities falling under cluster 2 are dominated by mentions of activities and constructions, suggesting areas where people live, host meetings or other social activities.

Based on archaeological evidence indicating the spatial expansion of the first human settlements in Iceland (Smith, 1995), the patterns picked up by the SOM analysis can be considered reasonable as the mapped results from this study match settlement expansion findings (Hermanns-Audardóttir, 1991; Vésteinsson, 1998).

4.2. Analysis performance

The SOM training identified variances in the composition of concept category values between different regions in Iceland, which were then successfully visualised through clustering and mapping of these value variances. Here we will touch upon a range of common obstructions with regards to cross-disciplinary research identified both by data contributors within the dataARC team as well as by Aagaard-Hansen (2007), Shiu (2014) and Möbjork et al. (2020).

4.2.1. Differing dataset structure, format and number of indicators

The analysis deliberately incorporated three datasets with varying data structure to simulate the total format variation of the dataARC project database. The varying data formats and structure of information within each dataset made combining and analysing them in their original state, challenging. Introducing 10 concept categories which encompass the main pieces of information from all datasets, helped overcome two obstacles: 1) simplification and ordering of values within each respective dataset and 2) produce a common scale with which both archaeological, environmental and textual data can be reordered, combined and compared with each other.

4.2.2. Significant variations in the geographic spread of each dataset

In order to successfully conduct a SOM analysis, it is crucial to implement sufficient amounts of data from all included factors or indicators, and these must be available in all parts of the geographic region being investigated (Wehrens & Buydens, 2007). In another area of research, studies of multiple deprivation have implemented SOM approaches to investigate deprivation on a multi-dimensional scale. Whelan et al (2010) combined 5 socio-economic factors from the Irish EU-SILC, which allowed them to define deprivation as a combination of several factors, not just income. Such studies rely on sufficient information about each included factor for all sites.

In contrast to the example presented above, the three datasets included in this study show significant variations in data availability, density and spread throughout Iceland (Figure 5). One way of overcoming the issues this variation created is to primarily focus on municipalities containing data from 2 or more disciplines.

Results from these mappings identify 2 municipalities containing data from all three disciplines: Skagafjörður and Biskupstungnahreppur. Skagafjörður is categorised under cluster 7 for the combined SOM. Figure 15 presents the respective clusters this municipality is categorised under for each respective dataset.

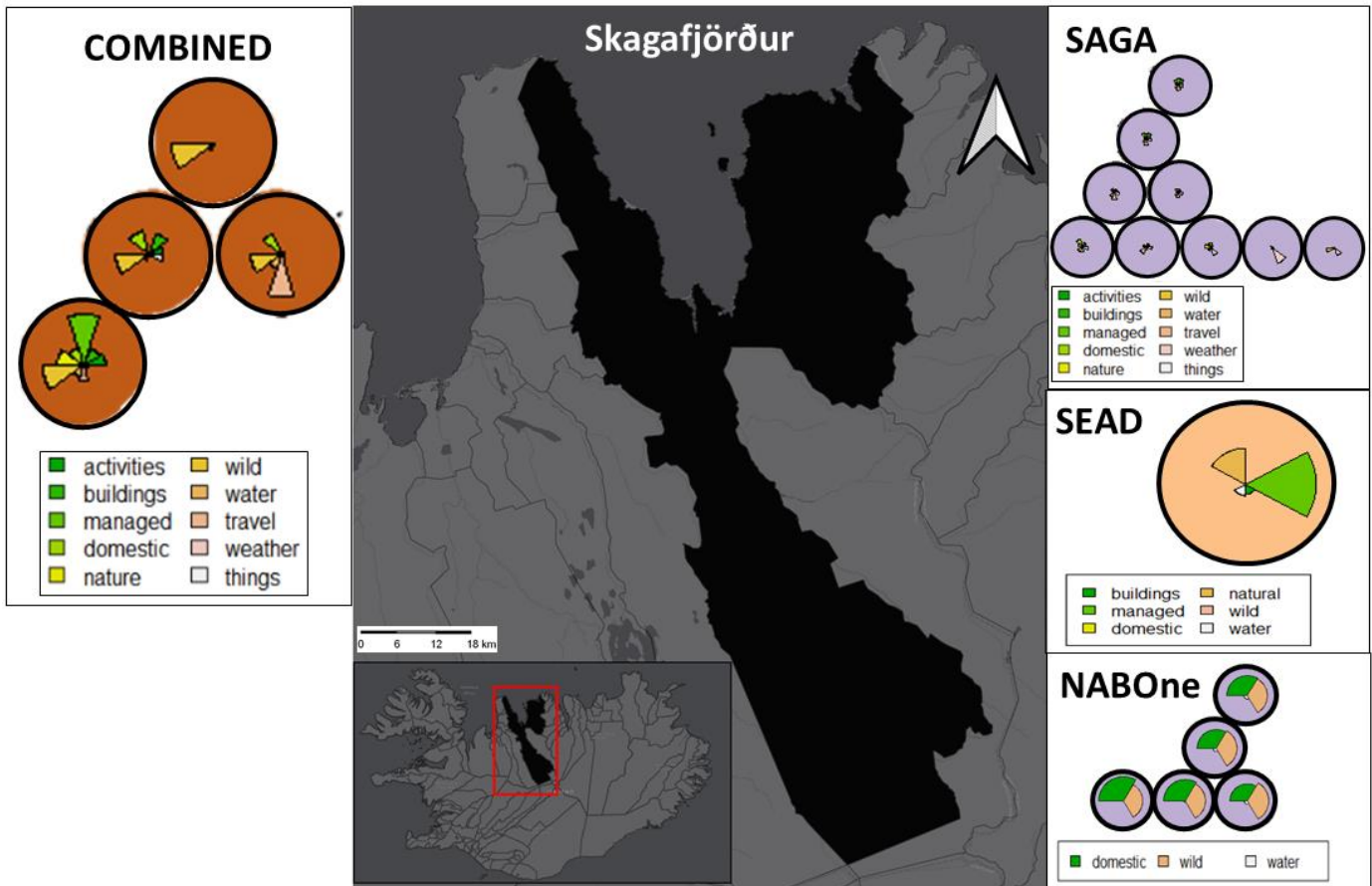


Figure 15: Skagafjörður, along with its respective clusters and cluster units for the COMBINED (cluster 7), Sagamap (cluster 2), SEAD (cluster 3) and NABOne (cluster 2) SOM analyses. Note that concept categories are represented by slightly different colours for each dataset (apart from COMBINED and Sagamap), and that only 9 of the total 41 units making up Sagamap cluster 2 are included.

Similar to the combined SOM, both SEAD and NABOne describe Skagafjörður as a region dominated by managed land and domestic animals, although evidence of more natural landforms is present in the record as well. Within the Sagamap categorisation Skagafjörður is categorised as cluster 2, which represents regions where data might be sparse and quite general, although there are frequent mentions of travelling (section 3.3, Figure 12). The region of Skagafjörður was colonised by early settlers around 900AD, who set up farmsteads and cultivated the land quite extensively (Steinberg et al. 2016). Not only do the results from the combined

SOM analysis correlate well with the known history of Skagafjörður and its early settlers, the comparison of the final clustering results from the individual SOM analyses with the combined analysis suggests that the combined analysis has been successful in preserving the integrity and level of detail which exist within each dataset.

4.2.3. Preservation of data complexity during dimension reduction

Every single dataset included in this study is complex; combining them as is increases the level of complexity by each new dataset included. Concept categories are a useful tool for downscaling data complexity. Transforming datasets from point to field data combats spatial complexity without affecting the integrity of the data (Kumar & Bangi, 2018).

Self-Organising Maps benefit from making minimal assumptions and aim to preserve the complexity of the input information (Whelan et al. 2010). This complexity is preserved by considering all included variables and indicators as being mutually dependent on each other. A number of studies applying multidimensional analysis methods to explore cross-disciplinary research questions have employed what is referred to as latent class analysis; a clustering analysis which assumes that each individual area can only be a member of one single cluster group (Moisio, 2004). Additionally, all included variables or indicators are considered as being mutually independent (Dewilde, 2007), an assumption that preserves and adds to redundant data complexity while at the same time not being necessarily true.

SOMs differ from other types of exploratory data analysis methods that apply dimension reduction, in that they aim to reproduce topology rather than distance (multidimensional scaling) (Cox & Cox, 2008). The SOM methods looks for similarities in the combination of dimensions in high-dimensional objects and map them as neighbours on a two-dimensional plane (Wehrens & Buydens, 2007). Details within the original data input are thus preserved, although dimensions are reduced, and data complexity will not increase.

4.3. User testing: comments and feedback

From discussions with Lethbridge, Buckland, Ryan and wider members of the dataARC team, several concerns regarding the combining of data and information from such a wide range of disciplines were identified. These concerns have been the predominant motivation and inspiration when developing the final model output, as overcoming them and thus producing an informative mapping output is the primary aim for the study.

Data providers, or researchers who will use the model to incorporate and visualise their shared data, are mostly concerned about the functionality of the model, whether it preserved the integrity of their data and whether the final model output is useful to them (Grudin, 2017). Gould & Lewis (1985) state the importance of iterative computer system design when designing for usability. This principle was integrated into the system design by

streamlining the overall methodology, constructing a web map and focussing on creating a program that is ready for incorporation into the dataARC user interface (Nielsen, 1994).

Several bugs and minor issues with the model were identified by various team members during user testing workshops, mainly concerning the concepts. Throughout the duration of the dataARC project one of the major struggles with combining these datasets in various ways have been to find a way to effectively query them so that a user is able to extract whatever information they desire. The concept map (section 1.2) can be queried down to specific concepts such as “glacier”, “church” or “fishing”. My project acts as a spatial counterpart to the concept map, where concepts can be queried in geographic space. Due to the inclusion of a spatial dimension, the level of detail within the concept dimension must be reduced in order to produce a meaningful queryable model. By combining the tools any users of the dataARC program should be able to identify the placement of any more general concept, such as “water” or “buildings”, then query the concept map to get more of an understanding of what this general concept comprises, and how it might be connected to other general concepts.

4.4. Potential improvements and future work

Although successful in capturing and visualising the general trends in connections between data and concepts in Iceland, there is a level of data generalisation which had to be introduced into the project given its temporal and financial frames. In Rønning (2020) section 7.2.4, the possibility of undertaking a SOM with point data rather than area data for NABOne is discussed in further detail. This approach should improve the spatial accuracy of SOM outputs for this dataset, but the possibility to combine them with other sets of data is lost. Even so, the process of running a SOM using a more segmented map should be explored further following the inclusion of more datasets.

Rønning (2020) section 7.3 explores the process of clustering concepts individually as a way of identifying sites where several datasets indicate the presence or occurrence of the same concept. This approach might also combat the high level of data generalisation. Due to the current inclusion of only three datasets, and because only three concepts are represented in all three datasets, the output of such an analysis at this time in the construction of this model is inadequate for its purpose.

As stated in the introduction, the goal for this study is to develop a prototype of a visualisation tool that helps researchers from cross-disciplinary fields to identify and analyse connections between data from the respective fields forming the dataARC project. The finalised model of this visualisation tool is set to include, in addition to its current components:

1. The total number of datasets making up the dataARC project, as well as a step-by-step instruction manual for the team to use when including their current and future data into the model
2. The full geographic spread of the dataARC project.

The inclusion of more datasets into the model will both increase the validity of the model output, as well as combat the issues related to variation in geographic spread of data between different disciplines (see section 4.2.2). Based on user feedback acquired from user testing sessions, a step-by-step methodology for data implementation is currently in the works and set to be ready before the final deadline for the dataARC project by the end of this year. The final version of this model will be presented as a web map tool that can be queried both spatially and by concept category.

5. Conclusion

Through this study I have aspired to contribute to the continuing efforts to appropriately combine and analyse archaeological, environmental and historical information on a multi-dimensional scale. The primary focus has been on developing visualisation tools which combines cross-disciplinary information and allows users to extract useful patterns within and between intricate data and display them using SOMs.

The study involves several data processing, mining and clustering stages. Firstly 10 concept categories were established, and all raw datasets had to be pre-processed and re-categorised into these categories. Then the SOM model was trained, which also involved identifying 10 cluster profiles showing varying combinations of the 10 concept categories. In the mapping stage I explored the spatial distribution of each of the cluster profiles, which allowed for discussion and interpretations regarding the clustering patterns.

This analysis strongly supports the view that the SOM approach has significant potential in improving our understanding of patterns between cross-disciplinary data, and to aid the ongoing work that is being done on revealing connections between the nature and continuous evolution of early human settlers and their surrounding environments and ecosystems in the North Atlantic island communities.

6. References

- Aagaard-Hansen, J., 2007. The challenges of cross-disciplinary research. *Social epistemology*, 21(4), pp.425-438.
- Allard, S. and Allard, G., 2009. Transdisciplinarity and information science in earth and environmental science research. *Proceedings of the American Society for Information Science and Technology*, 46(1), pp.1-9.
- Amorosi, T., Buckland, P., Dugmore, A., Ingimundarson, J.H. and McGovern, T.H., 1997. Raiding the landscape: human impact in the Scandinavian North Atlantic. *Human Ecology*, 25(3), pp.491-518.
- Bell, A., 2010. *Pingvellir: archaeology of the Althing* (Doctoral dissertation).
- Boulhosa, P.P., 2005. *Icelanders and the Kings of Norway: Mediaeval Sagas and Legal Texts*. Brill.
- Brereton, R.G., 2012. Self organising maps for visualising and modelling. *Chemistry Central Journal*, 6(2), pp.1-15.
- Buckland, P.I., Sjölander, M. and Eriksson, E.J., 2018. Strategic environmental archaeology database (SEAD).
- Butzer, K.W., 2008. Challenges for a cross-disciplinary geoarchaeology: the intersection between environmental history and geomorphology. *Geomorphology*, 101(1-2), pp.402-411.
- Cox, M.A. and Cox, T.F., 2008. Multidimensional scaling. In *Handbook of data visualization* (pp. 315-347). Springer, Berlin, Heidelberg.
- dataARC 2019, *Our Work*, viewed 22 April 2020, < <https://www.data-arc.org/> >
- Dewilde, C., 2008. Individual and institutional determinants of multidimensional poverty: A European comparison. *Social Indicators Research*, 86(2), pp.233-256.
- Dugmore, A.J., Borthwick, D.M., Church, M.J., Dawson, A., Edwards, K.J., Keller, C., Mayewski, P., McGovern, T.H., Mairs, K.A. and Sveinbjarnardóttir, G., 2007. The role of climate in settlement and landscape change in the North Atlantic islands: An assessment of cumulative deviations in high-resolution proxy climate records. *Human Ecology*, 35(2), pp.169-178.
- Dugmore, A.J., Church, M.J., Buckland, P.C., Edwards, K.J., Lawson, I.T., McGovern, T.H., Panagiotakopulu, E., Simpson, I.A., Skidmore, P. and Sveinbjarnardóttir, G., 2005. The Norse landnám on the North Atlantic islands: an environmental impact assessment. *Polar record.*, 41(1), pp.21-37.
- Fricke, H., O'Neil, J. and Lynnerup, N., 1995. SPECIAL REPORT: Oxygen isotope composition of human tooth enamel from medieval Greenland: Linking climate and society. *Geology*, 23(10), p.869.
- Gahegan, M., Wachowicz, M., Harrower, M. and Rhyne, T.M., 2001. The integration of geographic visualization with knowledge discovery in databases and geocomputation. *Cartography and Geographic Information Science*, 28(1), pp.29-44.

- Gould, J.D. and Lewis, C., 1985. Designing for usability: key principles and what designers think. *Communications of the ACM*, 28(3), pp.300-311.
- Grudin, J., 2017. From tool to partner: The evolution of human-computer interaction. *Synthesis Lectures on Human-Centered Interaction*, 10(1), pp.i-183.
- Gupta, A.K., Anderson, D.M. and Overpeck, J.T., 2003. Abrupt changes in the Asian southwest monsoon during the Holocene and their links to the North Atlantic Ocean. *Nature*, 421(6921), pp.354-357.
- Haldon, J., Mordechai, L., Newfield, T.P., Chase, A.F., Izdebski, A., Guzowski, P., Labuhn, I. and Roberts, N., 2018. History meets palaeoscience: Consilience and collaboration in studying past societal responses to environmental change. *Proceedings of the National Academy of Sciences*, 115(13), pp.3210-3218
- Hartman, S., Ogilvie, A.E.J., Ingimundarson, J.H., Dugmore, A.J., Hambrecht, G. and McGovern, T.H., 2017. Medieval Iceland, Greenland, and the new human condition: a case study in integrated environmental humanities. *Global and Planetary Change*, 156, pp.123-139.
- Hermanns-Audardóttir, M., 1991. The early settlement of Iceland. Results based on excavations of a Merovingian and Viking farm site at Herjólfssdalur in the westman islands, Iceland.
- Hofmann, M. and Chisholm, A. eds., 2016. *Text mining and visualization: case studies using open-source tools* (Vol. 40). CRC Press.
- Kanevski, M., Pozdnoukhov, A., Pozdnukhov, A. and Timonin, V., 2009. *Machine learning for spatial environmental data: theory, applications, and software*. EPFL press.
- Keim, D.A., 2002. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*, 8(1), pp.1-8.
- Kohler, T.A., Buckland, P.I., Kintigh, K.W., Bocinsky, R.K., Brin, A., Gillreath-Brown, A., Ludäscher, B., McPhillips, T.M., Opitz, R. and Terstriep, J., 2018. Paleodata for and from archaeology. *PAGES Magazine*, 26(2), pp.68-69.
- Kohonen, T., 1989. Self-organizing feature maps. In *Self-organization and associative memory* (pp. 119-157). Springer, Berlin, Heidelberg.
- Kohonen, T., 2001. *Self-Organizing Maps*. New York: Springer Series in Information Sciences.
- Kohonen, T., 2013. Essentials of the self-organizing map. *Neural networks*, 37, pp.52-65
- Koua, E.L. and Kraak, M.J., 2005. Integrating computational and visual analysis for the exploration of health statistics. In *Developments in Spatial Data Handling* (pp. 653-664). Springer, Berlin, Heidelberg.
- Kumar, G.N. and Bangi, M., 2018. An extension to winding number and point-in-polygon algorithm. *IFAC-PapersOnLine*, 51(1), pp.548-553.
- Lethbridge, E., 2016. The Icelandic sagas and saga landscapes. *Gripla*, 27, pp.51-92.

- Lin, Y.P., Chu, H.J., Wu, C.F., Chang, T.K. and Chen, C.Y., 2011. Hotspot analysis of spatial environmental pollutants using kernel density estimation and geostatistical techniques. *International journal of environmental research and public health*, 8(1), pp.75-88
- Loftsdóttir, K. and Lund, K.A., 2016. Þingvellir: Commodifying the “Heart” of Iceland. In *Postcolonial Perspectives on the European High North* (pp. 117-141). Palgrave Macmillan, London.
- Mairs, K.A., Church, M.J., Dugmore, A.J. and Sveinbjarnardóttir, G., 2006. Degrees of success: evaluating the environmental impacts of long term settlement in South Iceland. Aarhus University Press.
- McGovern, T., Smairowski, K., Hambrecht, G., Brewington, S., Harrison, R., Hicks, M., Feeley, F.J., Prehal, B. and Woollett, J., 2017. Zooarchaeology of the Scandinavian settlements in Iceland and Greenland: diverging pathways.
- McGovern, T.H., 2014. North Atlantic human ecodynamics research. *Human Ecodynamics in the North Atlantic: a Collaborative Model of Humans and Nature through Space and Time*. Lexington Books, Lanham, Maryland, pp.213-221.
- McGovern, T.H., Bigelow, G., Amorosi, T. and Russell, D., 1988. Northern islands, human error, and environmental degradation: A view of social and ecological change in the medieval North Atlantic. *Human Ecology*, 16(3), pp.225-270.
- McGovern, T.H., Perdikaris, S., Einarsson, A. and Sidell, J., 2006. Coastal connections, local fishing, and sustainable egg harvesting: patterns of Viking Age inland wild resource use in Mývatn district, Northern Iceland. *Environmental Archaeology*, 11(2), pp.187-205.
- McGovern, T.H., Vésteinsson, O., Fridriksson, A., Church, M., Lawson, I., Simpson, I.A., Einarsson, A., Dugmore, A., Cook, G., Perdikaris, S. and Edwards, K.J., 2007. Landscapes of settlement in northern Iceland: Historical ecology of human impact and climate fluctuation on the millennial scale. *American Anthropologist*, 109(1), pp.27-51.
- Miller, H.J., 2010. The data avalanche is here. Shouldn't we be digging?. *Journal of Regional Science*, 50(1), pp.181-201.
- Mobjörk, M., Berglund, C., Granberg, M. and Johansson, M., 2020. Sustainable development and cross-disciplinary research education: Challenges and opportunities for learning. *Högre utbildning*, 10(1), pp.76-89.
- Moisio, P., 2004. A latent class application to the multidimensional measurement of poverty. *Quality and Quantity*, 38(6), pp.703-717.
- Nielsen, J., 1994. Usability engineering. Morgan Kaufmann.
- Orning, H.J., 2015. Legendary sagas as historical sources. *Tabularia. Sources écrites des mondes normands médiévaux*.

- Pálsson, G., 2018. Storied Lines: Network Perspectives on Land Use in Early Modern Iceland. *Norwegian Archaeological Review*, 51(1-2), pp.112-141.
- Price, T.D. and Gestsdóttir, H., 2006. The first settlers of Iceland: an isotopic approach to colonisation. *antiquity*, 80(307), pp.130-144.
- Rick, T.C., Kirch, P.V., Erlandson, J.M. and Fitzpatrick, S.M., 2013. Archeology, deep history, and the human transformation of island ecosystems. *Anthropocene*, 4, pp.33-45.
- Róbertsdóttir, H (2001). “Hvaðan kemur nafnið „Innréttingarnar” á fyrirtækinu sem starfaði hér á 18. öld?” *Vísindavefurinn*, Available at: <http://visindavefur.is/svar.php?id=1752> (Accessed: 20 July, 2020)
- Rønning, K. (2020) *Archaeology, Environment and Human History: Examining the Spatial Links Between Human Settlements and Environmental Change in Iceland. MSc Dissertation [Technical Report]*. Edinburgh: University of Edinburgh
- Shiu, E. ed., 2014. *Creativity research: An inter-disciplinary and multi-disciplinary research handbook*. Routledge.
- Skupin, A., 2004. A picture from a thousand words [information visualization]. *Computing in Science & Engineering*, 6(5), pp.84-88.
- Smiarowski, K., Harrison, R., Brewington, S., Hicks, M., Feeley, F.J., Dupont-Hébert, C., Prehal, B., Hambrecht, G., Woollett, J. and McGovern, T.H., 2017. Zooarchaeology of the Scandinavian settlements in Iceland and Greenland. In *The Oxford Handbook of Zooarchaeology*
- Smith, K.P., 1995. Landnám: the settlement of Iceland in archaeological and historical perspective. *World Archaeology*, 26(3), pp.319-347.
- Steinberg, J.M., Bolender, D.J. and Damiata, B.N., 2016. The Viking Age settlement pattern of Langholt, North Iceland: results of the Skagafjörður archaeological settlement survey. *Journal of Field Archaeology*, 41(4), pp.389-412.
- Strawhacker, C., Buckland, P., Pálsson, G., Fridrikkson, A., Lethbridge, E., Brin, A., Opitz, R. and Dawson, T., 2015. Building cyberinfrastructure from the ground up for the North Atlantic Biocultural Organization introducing the cyberNABO Project. In *2015 Digital Heritage* (Vol. 2, pp. 457-460). IEEE.
- Sverrisson, Sigurdur, and Hannesson, Magnús Karel. 2014. *Local Governments in Iceland*. Reykjavik: Association of Local Authorities in Iceland
- Vésteinsson, O., 1998. Patterns of settlement in Iceland: a study in prehistory. *Saga book-Viking Society for Northern Research*, 25, pp.1-29.
- Wehrens, R. and Buydens, L.M., 2007. Self-and super-organizing maps in R: the Kohonen package. *Journal of Statistical Software*, 21(5), pp.1-19.

- Whelan, C.T., Lucchini, M., Pisati, M. and Maître, B., 2010. Understanding the socio-economic distribution of multiple deprivation: An application of self-organising maps. *Research in Social Stratification and Mobility*, 28(3), pp.325-342.
- Wyatt, I., 2004. The landscape of the Icelandic Sagas: text, place and national identity. *Landscapes*, 5(1), pp.55-73.

PART TWO

Technical Report

Table of Contents

1. Introduction and overview.....	6
2. Current data components of dataARC	6
3. Self-Organising Maps	9
3.1. Why is SOM applicable for this specific study?	9
4. Datasets and software	10
4.1. Sagamap.....	10
4.1.1. About.....	10
4.1.2. Structure.....	11
4.2. SEAD.....	13
4.2.1. About.....	13
4.2.2. Structure.....	13
4.3. NABOne	15
4.3.1. About.....	15
4.3.2. Structure.....	15
4.4. Software.....	17
5. Concept Categories	17
6. Methodology.....	19
6.1. Data preparation and pre-processing	19
6.1.1. Processing Sagamap.....	19
6.1.2. Processing SEAD and NABOne	21
6.2. SOM training and Clustering.....	26
6.2.1. Pre-training.....	26
6.2.2. SOM training.....	27
6.2.3. Clustering.....	30

6.3.	User Testing and analysis re-evaluation	33
7.	Mapping Results	33
7.1.	Combined SOM	33
7.2.	SOM of individual datasets.....	38
7.2.1.	<i>Sagamap</i>	38
7.2.2.	<i>SEAD</i>	40
7.2.3.	<i>NABOne</i>	41
7.2.4.	<i>NABOne – SOM of point outliers</i>	42
7.3.	SOM of individual concepts	44
7.3.1.	<i>Domestic animals</i>	44
7.3.2.	<i>Wild Animals</i>	45
7.3.3.	<i>Water</i>	45
8.	Answering the research questions	47
9.	Suggestions for further work.....	47
10.	References	48
Appendix	52	
sagas_concepts.py	52	
SEAD_indicators.py	56	
Nabonosead.py	58	
Nabonosead_outliers.py	60	
Saga_som.R.....	63	
Sead_som.R.....	66	
Nabone_som.R	69	
Combined_som.R	74	

List of Figures

Figure 1: Distribution of cairns in Iceland, according to the Cairns dataset. The geographic spread of this dataset is only representative for the area in which cairns were actually recorded.	8
Figure 2: Workflow model for the pre-processing of the Sagamap dataset	19
Figure 3: Workflow model for the pre-processing of the SEAD and NABOne dataset	22
Figure 4: SOM clustering of the SEAD dataset, unadjusted (left) and adjusted with respect to total indicator values within each municipality (right). Notice how the clustering of the unadjusted dataset is completely ruled by differences in total cluster values between the municipalities.....	24
Figure 5: SOM clustering of the NABOne dataset, unadjusted (left) and adjusted with respect to total indicator values within each municipality (right). Maps have been included to show the impact the adjustment of values have on the mapped output of the clustering.....	25
Figure 6: The relative abundance of each concept category within each municipality. The values represent a sum of the adjusted values for each dataset.	27
Figure 7: Training progress for the combined dataset.....	29
Figure 8: Node count plot (left) and mapping quality based on distance between objects and codebook neuro	29
Figure 9: SOM training output, represented by segments plots within each unit.....	30
Figure 10: WCSS of dataset, showing a distinct bend at 10 clusters	31
Figure 11: Clustered neuron grid with 8, 9, 10 and 11 defined clusters. The legend for the segment plots within each neuron is the same as for Figure 9.	32
Figure 12: Visual representation of the clustered SOM results for all datasets combined. Municipalities containing no data have been excluded from the mapping. Numbering and colour coding of the clusters correspond with the training output shown in Figure 9.	34
Figure 13: The 10 output clusters mapped individually (right), along with their corresponding neurons highlighted on the trained SOM matrix (left).....	38
Figure 14: Results of SOM training and mapping for the Sagamap dataset.....	38
Figure 15: Results of SOM training and mapping for the SEAD dataset.....	40

Figure 16: Results of SOM training and mapping for the NABOne dataset	41
Figure 17: Output of an isolation forest outlier detection method run in python.	42
Figure 18: NABOne outliers, trained, clustered and mapped by specific location or excavation site. .	43
Figure 19: Result of SOM clustering of the “domestic animals” concept category, where the values for each specific dataset is included as variable	44
Figure 20: Result of SOM clustering of the “wild animals” concept category. Note that Vestmannaeyjar, a small archipelago outside of the southern coast of Iceland, is the only municipality belonging to cluster 5 (dark blue).	45
Figure 21: Result of SOM clustering of the “wild animals” concept category	46

List of Tables

Table 1: Datasets included in the current dataARC prototype	7
Table 2: 10 concept categories, with full description and examples from all included datasets	18
Table 3: Examples of original concepts used to categorise the Sagamap dataset. Table presents the new concept categories into which these original concepts have been placed.	20
Table 4: The 22 indicators found within the SEAD dataset, described and recategorized into the 10 defined concept categories.....	23
Table 5: The 12 indicators found within the NABOne dataset, described and recategorized into the 10 defined concept categories.....	24

1. Introduction and overview

This report supports and supplements the associated research paper (Rønning, 2020) which presents a novel approach to investigate and spatially visualise connections between cross-disciplinary data using Self-Organising Maps.

The objectives for this report are to elaborate further on SOMs as a method, the structure of any included datasets, and to describe the steps taken to pre-process and combine the data to produce the final outcome. The main goal is to outline the analysis process for the dataARC team and other future researchers who would want to expand on the project further and include their own data in the study.

2. Current data components of dataARC

The current dataARC prototype consist of 16 datasets divided into 3 categories: Archaeological, environmental and textual (Table 1). Access to datasets were provided through GitHub by Dr Rachel Opitz, who co-directs the dataARC project. All datasets are stored in in JavaScript Object Notation (JSON or GEOJSON) format.

Archaeological datasets include excavation reports and zooarchaeological data from the NABOne project and identified cairn locations in the northwest of Iceland. Archaeological data helps with site locations and provides insight in evolving human ecodynamics (Haldon et al. 2018). Furthermore, zooarchaeological data provide insights in changing ecology and ecological zones over time (McGovern et al. 2006).

There are 2 textual datasets included in this project: Information on places mentioned in Icelandic sagas (Sagamap project) and historical documents from farms (Jardabok project). These datasets tell the story of a changing world and environment from a human perspective and provide historic and local knowledge and perceptions of environmental change in the past (Strawhacker et al. 2015).

The environmental category includes datasets holding information on paleoclimate from proxies such as tephra layers (Tephabase) and environmental archaeological proxies like insects, trees and ceramics (SEAD). Environmental datasets act as records for palaeoenvironmental and paleoclimatic conditions (Mann et al. 2009).

Table 1: Datasets included in the current dataARC prototype

Dataset	Category	Additional information
<i>SEAD</i>	environmental	Strategic Environmental Archaeology Database. Mainly chemical, physical and biological proxy data derived from e.g. fossils, soil samples, geoarchaeological data and dendrochronological analyses.
<i>Environmental threats</i>	environmental	Environmental threats to Icelandic archaeological sites
<i>Sagamap</i>	textual	Dataset containing places and locations mentioned in 42 Icelandic sagas. Each mention of a place is tagged with one or several concepts indicating or explaining what happens at the site or whether animals, buildings or items are found or seen there
<i>TephraBase</i>	environmental	Database of characteristics and distribution of Icelandic tephra layers. Aim to promote use of tephrochronology in palaeoenvironmental studies.
<i>Jardabok Icelandic farm data</i>	textual	Compilation of historical documents from Icelandic farms from late medieval to early modern times, around 1500-1860.
<i>Cairn Locations</i>	archaeological	Dataset of cairns identified in satellite imagery for NW Iceland.
<i>NABOne faunal data</i>	archaeological	North Atlantic Biocultural Organization. Combine data from different disciplines in order to improve the research potential in the North Atlantic. Aim to reconstruct long term human ecodynamics by building and combining palaeoecological and geoarchaeological datasets
<i>NABO excavation report topics</i>	archaeological	
<i>NABOne data prepared for incorporation into SEAD</i>	environmental	
<i>Eyrbyggja Saga</i>	textual	Mapped concepts from the Eyrbyggja Saga
<i>Whale Artefacts</i>	archaeological	Finds of whale bones and artefacts
<i>Orkney Faunal Database</i>	archaeological	Contains faunal data from the Orkney Islands, Scotland
<i>Excavated Archaeological Materials-grouped by context</i>	archaeological	Archaeological finds grouped by excavation unit. Correlates with other excavation data in the system
<i>Excavated Archaeological Materials-grouped by Find type</i>	archaeological	Finds from each group of stratigraphic units, correlates with other excavation data in the system
<i>Icelandic Excavated Archaeological Materials</i>	archaeological	Archaeological data on stratigraphic units and the key finds within them
<i>Archaeological Context Data</i>	archaeological	Descriptions of excavation units

The current variable datasets that have been included in this study and the dataARC project should be seen more as a representation of the distribution of a specific variable within a specific area where data is available, or research has been done, rather than as accurate representations of the full picture or reflections of real patterns. The cairns dataset for example, is only representative for the area which has been sampled for cairns (NW Iceland) and does not reflect the overall distribution of patterns in Iceland (Figure 1). The SEAD project is limited by the number of included archaeological sites, and several of the other datasets included are still being processed and expanded.



Figure 1: Distribution of cairns in Iceland, according to the Cairns dataset. The geographic spread of this dataset is only representative for the area in which cairns were actually recorded.

Additionally, it is important to keep in mind the subjectivity of textual and historical datasets, such as Sagamap. While zooarchaeological and environmental datasets are objective indicators of past environmental conditions, land cover and human impact on nature, historical notations have a much higher level of subjectivity and risk of not telling the full story (Pettersen, 2008). Although Icelandic sagas have been known to describe past events quite accurately and thus been given the status as historical documents, they are still stories and likely to include mentions of events that never happened or people that never existed (Lethbridge, 2016). In the context of this study however, where the sagas are used as indicators of people interacted with their surrounding environment, where they lived, cultivated the land or held meetings, rather than accurate descriptions of specific events, these stories fulfil their purpose effectively in spite of their subjective nature.

3. Self-Organising Maps

Attempting to tell a full story of real-life phenomena using only one dataset or one single factor or proxy is a crude oversimplification of natural systems, however including a wide range of datasets into one single analysis can very easily cause you to drown in information and make it increasingly harder to identify any patterns or connections (Aagaard-Hansen, 2007). Self-organising maps (SOMs) have provided great advances to not only data visualisation, but also for a variety of data handling and exploratory data analysis methods, especially in regard to multidimensional data analysis and interdisciplinary research (Pözlbauer et al. 2005; Príncipe & Miikkulainen, 2009). Being an unsupervised neural network model, it eliminates any issues related to human errors or biases, which makes it a favourable approach to other data handling and clustering methods as the output is completely unbiased (Mayer et al, 2007). As a visual analytics framework it benefits from being able to handle very large amounts of data with a range of different features and numerical values (Andrienko et al. 2010).

Self-organising maps use machine learning and unsupervised artificial neural networks to analyse and cluster data. The analysis method emphasize variation and bring out patterns in the datasets by eliminating dimensions (Kohonen, 1997) and visualise high-dimensional data on a two-dimensional grid, which makes the identification of patterns and connections easier (Yin, 2008). As a data analysis technique, it has both industrial as well as scientific applications (Laaksonen et al. 2001; Cracknell & Cowood. 2016).

The SOM technique cluster data based on relationships and similarities. Explained simply, similar factors for each dataset are clustered together. It identifies areas where values for each included dataset are similar and groups these areas together. SOMs also benefit from having the ability to adjust or determine a kernel smoothing parameter, meaning that a user can very easily determine the granularity levels for both number of clusters and on an individual cluster level (Pözlbauer et al. 2005). This enables users to extract information and investigate patterns at any level of detail they so desire (Moehrmann et al. 2011).

3.1. Why is SOM applicable for this specific study?

As this study includes large datasets from a variety of different disciplines, SOMs seems, from what has been discussed, like the ideal approach to investigate spatial connections and patterns within and between these datasets (Pölla et al. 2006).

Not only are SOMs good for data exploration and clustering analysis, it can also help combat limitations with the datasets and provide equal weightings to factors that might not be as well represented in the data as others but are still equally important. With the datasets implemented into dataARC, there is always the issue of the data distribution not being representative of the overall distribution of patterns in Iceland. The way the different

datasets are represented in space could make a geographic visualisation analysis challenging, as the spread of data strongly affects the outcome of the analysis. Thus, if the spread is an inaccurate representation of the truth, the output of a visualisation analysis will also be inaccurate. SOMs can help combat this to some extent. The overarching idea is for the analysis to identify areas which are statistically significant or similar, and cluster these together. If for example 4-5 of the Icelandic municipalities each contain hundreds of cairn locations whereas the rest of the island has none, it is likely that the 4-5 cairn municipalities will be clustered together unless other included factors indicate different patterns or are scattered differently (Kohonen, 1997).

There is also the risk of larger datasets being overrepresented which could potentially obfuscate or take away some of the focus and integrity of smaller datasets. Within the dataARC project the number of data points per dataset ranges from under 100 to nearly 10 000. SOMs provide equal weighting to each dataset and is concerned about common trends, both within the datasets themselves as well as the combination of all datasets.

4. Datasets and software

This section provides a more in-depth explanation of the 3 included datasets (Sagamap, SEAD and NABOne), their initial purpose and overall structure.

4.1. Sagamap

4.1.1. *About*

The Sagamap project aims to create a database where all mentions of place names in Icelandic sagas are geo-references with a description of the event that occurs there as well as a link to the saga and chapter it occurs in. The initiative was first started by Dr Emily Lethbridge at the University of Iceland in 2011.

Linking literary tales and stories with geography opens for the possibility to study the different functions of a landscape, and how it has changed with the changing history and societal evolution of Iceland. Further, studying the stories in a spatial context by mapping out where certain events, actions or interactions take place could help reveal and visualise any common patterns and identify important areas that had a specific use (meeting place, battlefield etc.) or common travel routes across Iceland. It is important to note that Sagas are not accurate historical documents and should not be seen as such. They can, however, give an indicator of where certain events usually took place, or placements of farmland and natural landscape forms (Ross, 1997).

Most if the Icelandic Sagas were written down during the 13th century and refers mostly to events that occurred from the late 9th century up until then. This was the age of the first settlers in Iceland and the stories mainly reflect the lives of the first generations of settlers. Sagas authors are usually not known or anonymous. Some

of the sagas are biographical and follow the lives and events of a specific person (Egill Skalla-Grímsson, Grettir Ásmundarson, Gísli Súrsson) whereas other focus on feuds or larger events (Eyrbyggje saga, Njáls saga etc.). Because sagas describe events, places and people in a very objective fashion, they can almost be said to be more like historical texts rather than novels (Ross, 1997; Lethbridge, 2010).

There are several logistical challenges linked to the Sagamap project, both in relation to the reliability of sagas as historical documents (whether they can be trusted as being historically accurate) and whether the place names mentioned in the sagas correlate to the place names of today (Lethbridge, 2020). This project is, like many other partner projects to dataARC (and dataARC itself) still under development and the dataset it thus not fully finished. This means that there might be errors in the dataset that have not been discovered yet.

4.1.2. Structure

The Sagamap data is saved as point data in a geojson file in the dataARC GitHub repository and currently consist of 4652 points, of which 3869 are located in Iceland. Each point has a point location and a feature which contains the information. The work that has been done with Sagamap is a text mining technique where events have been extracted from the text and given a spatial location as well as a feature definition explaining in one word or phrase what is going on (Dhillon & Modha, 2001).

Code example 1 presents the structure for one Sagamap data point. The “id” refers to place id, not point id. Several points can have the same id. These points will also have the same “name”. Each individual saga has a “sagaid” and “saganame”. The “concept” indicates in short what is happening in the text, followed by the actual text passing from the saga. This particular passing describes a meeting between a group of travellers who encounter a farmer herding his sheep at Búlandshöfði. A time frame for the event is also provided. Events in the texts can have multiple concepts. The id and paragraph itself will then be repeated in individual points.

```

    "type": "Feature",
    "properties": {
      "id": "1850",
      "name": "Búlandshöfði",
      "sagaid": 28,
      "saganame": "Eyrbyggja saga",
      "chapter": "18. kafli",
      "concept": "Activities: animal husbandry: herding",
      "text": "Þeir Þórarinn tóku hesta þeirra Þorbjarnar og ríða þeim
heim og sáu þeir þá hvar Nagli hljóp hið efra um hliðina. Og er þeir komu í
túnið sáu þeir að Nagli var kominn fram um garðinn og stefndi inn
til Búlandshöfða . Þar fann hann þræla Þórarins tvo er ráku sauði úr höfðanum.
Hann segir þeim fundinn og liðsmun hver var. Kallaðist hann víst vita að
Þórarinn og hans menn voru látnir og í því sáu þeir að menn riðu heiman eftir
vellinum. Þá tóku þeir Þórarinn að hleypa því að þeir vildu hjálpa Nagla að hann
hlypi eigi á sjó eða fyrir björg.",
      "action_start": "880",
      "action_end": "1031",
      "composition_start": "1240",
      "composition_end": "1310",
      "oldest_manuscript": "AM 162 e fol. ",
      "oldest_manuscript_start": "1290",
      "oldest_manuscript_end": "1310",
      "manuscript_link":
"https://handrit.is/en/manuscript/view/is/AM02-0162E"
    },
    "geometry": {
      "type": "Point",
      "coordinates": [
        -23.474655,
        64.940757
      ]
    }
  }

```

Code example 1: Point data example from the sagas.geojson file

4.2. SEAD

4.2.1. About

The Strategic Environmental Archaeology Database, or SEAD, focus on effectively store and analyse environmental data and data on how climate and environment have impacted humans in the past (Buckland et al. 2011).

The dataset consists of chemical, physical and biological proxy data derived from e.g. fossils, soil samples, geoarchaeological data and dendrochronological analyses, and is stored in a relational database. In addition, there are also some insect/pollen/plant datasets that are used for palaeoenvironmental reconstruction. The databases are linked temporally based on their dating. The time span goes from the Quaternary period (2.6 Ma) to today, with most of the data being from the Holocene (12-13 ka) where humans have started to have an impact on the planet. As of right now the project is focussed mainly in Scandinavia and northern Europe, but they are constantly growing and expanding as they get new partners (Buckland et al. 2018).

Some of the datasets included in SEAD are: MAL dataset consisting of geochemical and physical palaeodata, plant macrofossils and pollen (mainly in Sweden), BugsCEP fossil insect database, KFL ceramic thin section dataset, VDL pilot project dataset (dendrochronology, south Sweden only), AFL stable isotope and lipid dataset (test dataset, not yet incorporated into SEAD). The 560 SEAD data points located in Iceland are all from BugsCEP. This database was developed and is currently handled and supervised by Philip Buckland at the University of Umeå, Sweden (Buckland & Buckland, 2006; Buckland, 2007). The focus of the database is climatic and environmental reconstruction from coleoptera (beetle assemblages). Although primarily focussed in the North Atlantic, the dataset is currently being expanded to incorporate data from e.g. North and South America, Egypt and Japan (Buckland, 2014).

4.2.2. Structure

The structure follows archaeological data collection processes and methods. The result is a detailed database with a very intricate structure. The data is thoroughly tested prior to being implemented into the database to ensure data quality and reliability. Any inconsistencies are eliminated through thorough evaluation.

The SEAD dataset is stored as a .json file within the dataARC GitHub repository, see code example 2 for an example of point data structure. There is a total of 456 points of the dataset that are located in Iceland in 21 different locations. Each point has a dictionary of sample data information, such as site id and name, sample id, dating type, age (if available), and a dictionary of indicators. There are 22 indicators in total; all points have been given a value for each indicator, ranging from null or 1 to over 100. The BugsCEP database has been proven to produce habitat and climate reconstructions that are accurate and match other palaeoenvironmental studies (Buckland, 2007). All datapoints in Iceland date from within the Quaternary period (Buckland, 2010).


```

{
  "type": "FeatureCollection",
  "features": [
    {
      "type": "Feature",
      "id": 4214,
      "geometry": {
        "type": "Point",
        "coordinates": [
          -5.5633335,
          54.532776
        ]
      },
      "properties": {
        "id": 4214,
        "country": "Ireland",
        "sampleData": {
          "site_id": 590,
          "site_name": "Strangford Lough: Greyabbey Bay",
          "sample_name": "S2",
          "sample_group_id": 709,
          "dating_type": "Relative dates",
          "start": null,
          "end": null,
          "age_name": "Early Holocene",
          "age_abbreviation": "EH"
        },
        "indicators": {
          "Aquatics": 4,
          "Indicators: Standing water": 5,
          "Indicators: Running water": 5,
          "Pasture/Dung": 1,
          "Meadowland": 2,
          "Wood and trees": 10,
          "Indicators: Deciduous": 7,
          "Indicators: Coniferous": 5,
          "Wetlands/marshes": 4,
          "Open wet habitats": 4,
          "Disturbed/arable": 6,
          "Sandy/dry disturbed/arable": 1,
          "Dung/foul habitats": 4,
          "Carrion": 8,
          "Indicators: Dung": 1,
          "Mould beetles": 7,
          "General synanthropic": 2,
          "Stored grain pest": 6,
          "Dry dead wood": 1,
          "Heathland & moorland": 6,
          "Halotolerant": 2,
          "Ectoparasite": 1
        }
      }
    }
  ]
},

```

Code Example 2: Point data example from the *sead.json* file

Every point represents soil samples from different contexts at each of the 21 sites. The list of species, or indicators, depict different environmental signals. The number of points at each site reflects the level of detail in the environmental reconstruction possible for each site and the size of the excavation budget.

4.3. NABOne

4.3.1. *About*

The North Atlantic Biocultural Organisation (NABO) is a research cooperative founded in 1992 by a group of researchers from a wide range of disciplines, including archaeology, biology, climatology and history. NABO works to combine data from a range of disciplines to try and improve the research potential in the North Atlantic, with the primary aim being to reconstruct long term human ecodynamics by building and combining palaeoecological and geoarchaeological datasets (McGovern, 2014).

NABOne is a zooarchaeological database constructed by and within NABO. This database was founded by a working group in 1997 as a way off effectively compiling copious amounts of bone data from animals, birds and fish located in the North Atlantic region, and to come up with an all-encompassing way of describing and structuring them which would make both searching the database, comparing different finds and identifying new ones, easier (McGovern et al. 2017). The part of the NABOne dataset used in this study has been prepared for incorporation into the SEAD database by Thomas Ryan at the City University of New York.

4.3.2. *Structure*

The dataset consists of 928 points scattered over 9 different sites all located in Iceland. These 928 data points each represents an archaeological context, which relates to the position of an artefact or archaeological find in space and time (Schiffer, 1972). Large sites which have been subject to years of excavations, such as Hofstadir, will consist of larger numbers of contexts or points. Other sites are much smaller and holds a lower number of contexts. This specific part of the dataset has been prepared for implementation into SEAD, thus its structure resembles that of the SEAD dataset, see code example 3.

```

{
  "type": "Feature",
  "id": 7741,
  "geometry": {
    "type": "Point",
    "coordinates": [
      -17.163794,
      65.607995
    ]
  },
  "properties": {
    "id": 7741,
    "country": "Iceland",
    "sampleData": {
      "site_id": 1293,
      "site_name": "Hofstaðir",
      "sample_name": "5155",
      "sample_group_id": 1423,
      "dating_type": "Relative dates",
      "start": 650,
      "end": 100,
      "age_name": "POST 1300",
      "age_abbreviation": "HOF_Phase_POST 1300"
    },
    "indicators": {
      "domestic": 640,
      "wild": 1298,
      "Marine Mammal": 74,
      "Marine Fish": 11,
      "Freshwater Fish": 0,
      "terrestrial": 0,
      "aquatic": 0,
      "Sea Bird": 0,
      "Non Migratory Terrestrial": 0,
      "Fresh Water Migrant": 0,
      "On floe ice": 0,
      "On fast ice": 0
    }
  }
},

```

Code Example 3: Point data example from the Nabonosead.json file

Each point is related to a site id but has its own id and sample name, a start and end date and 12 different indicators which have been given a number based on abundance of records related to the specific indicator. One bone fragment is one record and these records are then lumped together and aggregated by the 928 contexts. The indicators are tagged to animals and fish species, the “domestic” indicator refers to domestic animals like sheep or cattle, “wild” to wild animals like wolves and so on.

4.4. Software

This study used a combination of python, R and qGIS, all on a Windows platform. Python was used primarily for pre-processing and recategorizing all datasets and prepare their structure to best fit the SOM training process. I applied a range of vector analysis tools in qGIS to transform the datasets from point to polygon data. qGIS was also used later in the process to produce final visualisations of the SOM and cluster output.

The SOM training and clustering was conducted in R using the “Kohonen” package (see Wehrens & Wehrens, 2019 for package documentation).

5. Concept Categories

Combining multidisciplinary data using self-organising maps has been proven effective and accurate through a multitude of different studies (Whelan et al, 2010; Tsademir & Merényi, 2012). The primary difference between these studies and the one conducted through this experiment, is that my project essentially is an attempt to combine several multidimensional datasets in a multidimensional study to look for connections between dimensions, rather than between scaled values

As the included datasets are annotated by concepts or indicators and have no common scale of values, a new categorisation system had to be developed which can be used on any included dataset. Discussions with data providers Emily Lethbridge (Sagamap) and Phil Buckland (SEAD) suggested 10 concept categories would be the most ideal, with concepts ranging from natural and cultivated landscape indicators, to mentions of travelling or human activity. Any more than 10 would cause unnecessary levels of detail whilst including less than 10 categories might severely oversimplify the categorisation of the data. The categories were also defined with all the datasets currently included in dataARC in mind.

The 10 concepts are explained in detail in Table 2, which also includes examples of data from each dataset for each individual category. There is the potential of this way of categorising data might cause some loss of integrity or information for specific datasets. However, as this analysis is an attempt to construct a framework for analysing connections between datasets, and not to map them all in full detail, dimension reduction and a certain level of dataset simplification is both expected and required (Yang et al. 2002; Akinduko et al. 2016).

Table 2: 10 concept categories, with full description and examples from all included datasets

Concept	Description	Data Examples	Appears in
Activities	Any mentions or evidence of human activities, apart from travelling or water related activities	<ul style="list-style-type: none"> • Mentions of meetings, saga narratives, exchanges or fights in Icelandic Sagas 	Sagamap
Buildings	Any human-made buildings or construction, apart from cairns	<ul style="list-style-type: none"> • Mentions of buildings in Icelandic sagas • Zooarchaeological evidence of stored grain pest or mould beetles in SEAD 	Sagamap SEAD
Managed	Managed landscape, any evidence of land alteration or management by humans	<ul style="list-style-type: none"> • Mentions of farming activities in Icelandic sagas, such as cultivation, grazing or pastureland • Zooarchaeological evidence of disturbed or arable landscape or synanthropic species in SEAD 	Sagamap SEAD
Domestic	Domestic animals, livestock	<ul style="list-style-type: none"> • Mentions of specific species or of livestock in Icelandic sagas • Zooarchaeological evidence of ectoparasites in SEAD • Indicators tagged to domestic animal species in NABOne 	Sagamap SEAD NABOne
Natural	Natural landscape, no or little human alteration	<ul style="list-style-type: none"> • Mentions of unmanaged forests, moorland or heathland in Icelandic sagas • Zooarchaeological evidence of meadowland, heathland or forest in SEAD 	Sagamap SEAD
Wild	Wild animals, not managed by or living in relation with humans	<ul style="list-style-type: none"> • Mentions of wild animals like bears, wolves or birds in Icelandic sagas • Zooarchaeological evidence of carrions in SEAD • Indicators tagged to wild animal species in NABOne 	Sagamap SEAD NABOne
Water	Water related activities, evidence or indicators of water bodies or of animals living in or near water	<ul style="list-style-type: none"> • Mentions of rivers, ocean, boats or fishing in Icelandic sagas • Zooarchaeological evidence of stagnant or running water or halotolerant species in SEAD • Indicators tagged to fish or marine mammals in NABOne 	Sagamap SEAD NABOne
Travel	Any mentions or evidence of people travelling	<ul style="list-style-type: none"> • Mentions of people travelling or of travel paths or cairns in Icelandic sagas 	Sagamap
Weather	Any weather observations	<ul style="list-style-type: none"> • Mentions or observations of rain, fog, wind or snow in Icelandic sagas 	Sagamap
Things	Objects related to humans that are not buildings or animals	<ul style="list-style-type: none"> • Mentions of objects such as weapons, money or food in Icelandic sagas 	Sagamap

6. Methodology

This section provides a step-by-step methodology of the data handling and mapping process, from pre-processing of the 3 included datasets, to training and clustering of the SOM model and mapping of the final results. All variables and parameters used are justified throughout.

6.1. Data preparation and pre-processing

Due to the differences in data structures, the pre-processing will be slightly different for nearly every single dataset included. The main aim for the pre-processing is to reshape the structure of the datasets in order to make them comparable, combinable and ready for implementation into the SOM training model.

6.1.1. Processing Sagamap

The workflow for pre-processing Sagamap is described in Figure 2 below, with the first step being extracting any data points located in Iceland, which adds up to 3869 points out of the total 4652.

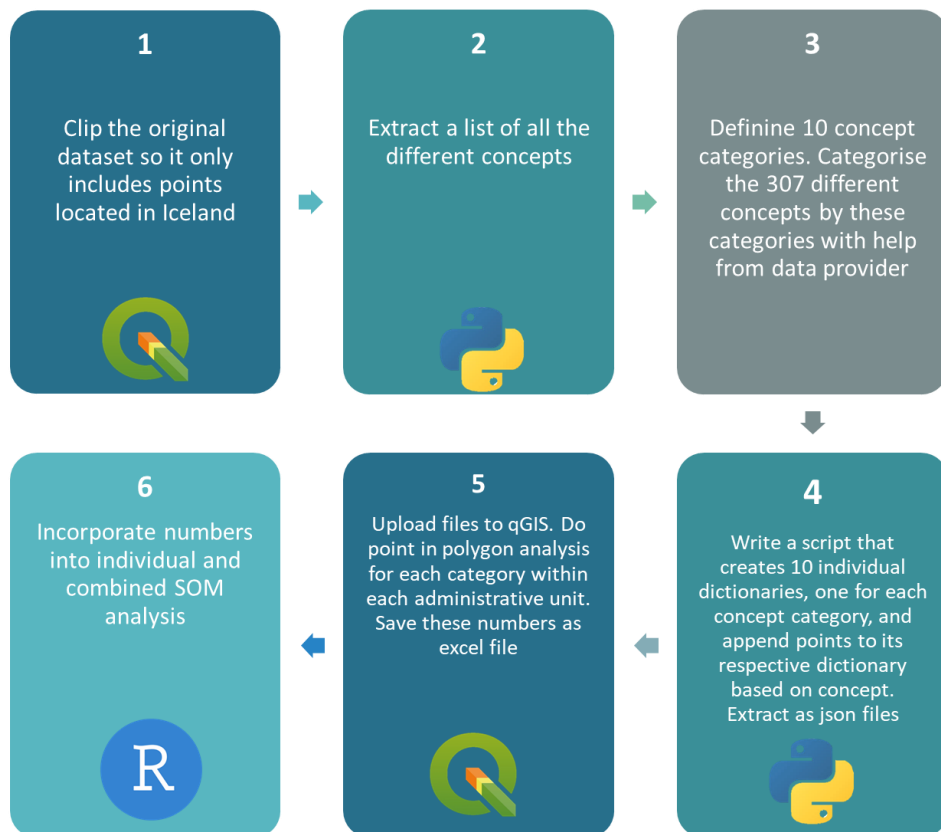


Figure 2: Workflow model for the pre-processing of the Sagamap dataset

Following the extraction of the data points located in Iceland only, the next part of the standardisation of the Sagas dataset is to filter it by feature. The idea behind this is to filter the data based on concept, so first all the different concepts must be extracted from the dataset, which was done in python. For the 3869 points located

in Iceland, there are a total of 307 different assigned concepts. These concepts were assigned into their respective concept categories by hand and validated several times by data provider Emily Lethbridge to make sure the concepts had been interpreted correctly.

Table 3 presents example concepts and the category to which these points have been appended based on their concept. The Table also indicates the total number of points appended to each concept:

Table 3: Examples of original concepts used to categorise the Sagamap dataset. The Table presents the new concept categories into which these original concepts have been placed.

Concept category	Sagamap concept examples	# of points
Activities	Activities: exchange Activities: socializing: performance Events: burial Ideas: power: obligation: social obligation	466
Buildings	Actors: institution: preChristian temple Physical Landscape: built environment: buildings: hall/house Physical Landscape: built environment: buildings	725
Managed	Activities: cultivation/farming Actors: plants: crops Physical Landscape: managed landscape area: field	190
Domestic	Actors: animals: dog Actors: animals: mammals: cow Animals: actors: mammals: ox	384
Natural	Actors: plants: tree Physical Landscape: ecological area: heathland Physical Landscape: ecological area: woods and trees	152
Wild	Actors: animals: mammals: bear Actors: animals: mammals: wolves Actors: animals: avian	20
Water	Activities: fishing Actor: thing: boat Physical Landscape: ecological area: moving water: river	699
Travel	Actors: things: bridge Activities: travels Physical Landscape: built environment: cairn	904
Weather	Physical processes: weather: fog/ rain Physical processes: weather: snow Physical Processes: weather: wind	35
Things	Actor: thing: spear Actors: things: bone Actors: things: food	294

Following concept categorisation, a script was written in python which categorised each point into its respective new concept category based on its original concept. 10 new geojson files, all with similar structures to the original Sagamap dataset, were written and imported into qGIS, where I estimated a count for each concept within each municipality by doing a point in polygon analysis for each respective concept category and saving the results in an excel format. This process fulfilled the goal of turning a nominal dataset into a numerical one that can later be used for further analysis or to be compared and analysed in unison with other numerical datasets.

The Sagamap dataset was not adjusted with respect to number of points within each area, which had to be done with the SEAD or NABOne dataset, see section 6.1.2. The overall spread of points in the Sagamap dataset hold very valuable information in terms of the importance of these areas to humans and the frequency of which they visited them, mentioned them or described them. Additionally, the number of points representing each concept category varies significantly, which is also believed to reflect the amount of mentions of each concept and thus also of importance.

6.1.2. Processing SEAD and NABOne

As mentioned in section 4.3.2, the structure of NABOne strongly resembles that of SEAD. Thus, the pre-processing methodology of these datasets is more or less identical, apart from the indicator reclassification. The full methodology of the pre-processing of SEAD and NABOne is presented in Figure 3 and was again done using a mixture of qGIS and python.

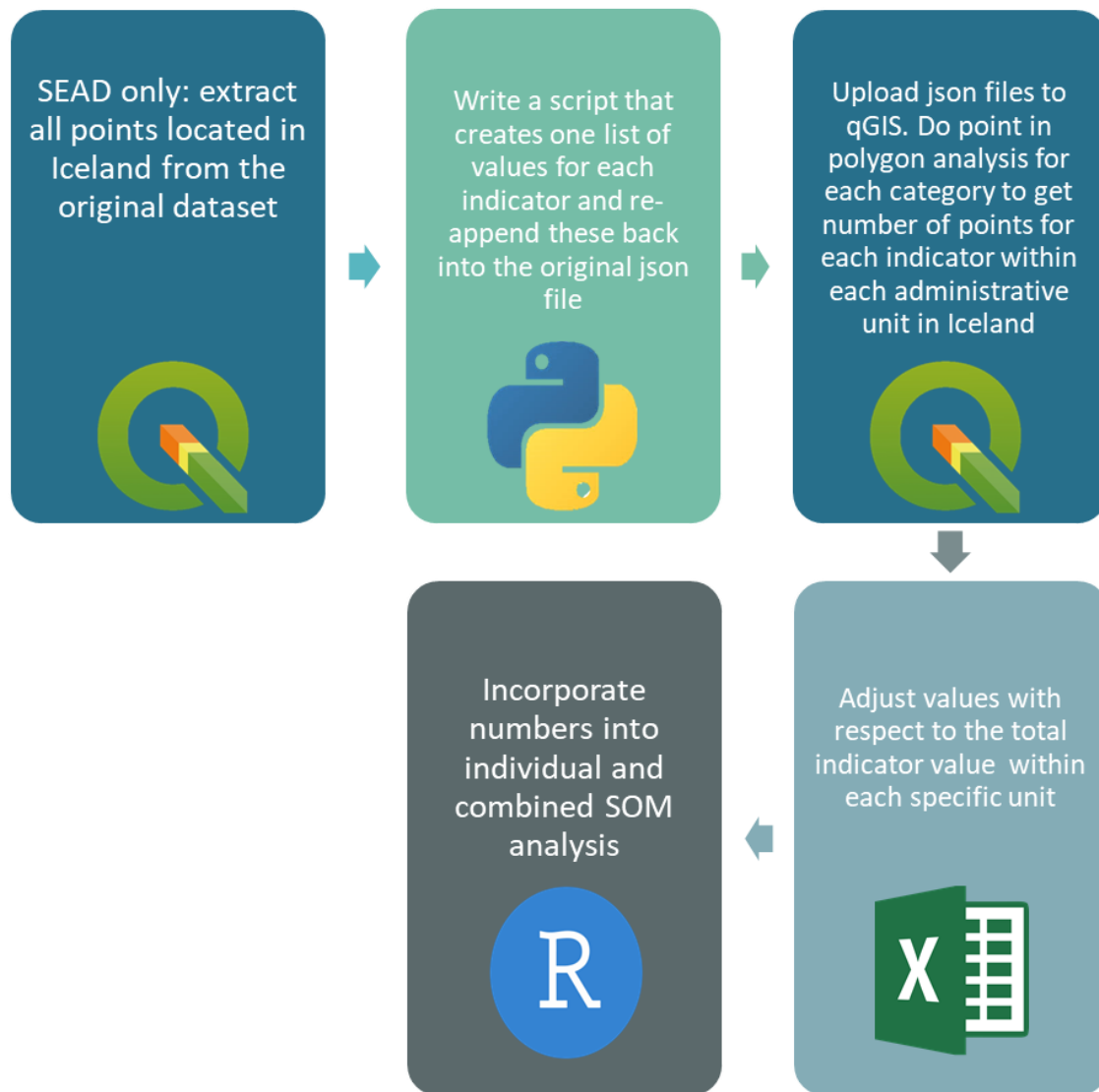


Figure 3: Workflow model for the pre-processing of the SEAD and NABOne dataset

Just like the Sagamap dataset, SEAD contains points located throughout most of the North Atlantic. Any data points located in Iceland were extracted prior to the processing and recategorization of the dataset. Table 4 presents the Concept category each SEAD indicator has been allocated to in this study. “Ectoparasite” and “carrion” indicators cannot be distinguished to represent either wild or domestic animals and have thus been set to represent both.

Table 4: The 22 indicators found within the SEAD dataset, described and recategorized into the 10 defined concept categories.

Indicator	Description	Concept category
<i>Aquatics</i>	Species adapted to living in water at any point during their life cycle	Water
<i>Indicators: Standing water</i>	Stagnant water bodies, lakes	Water
<i>Indicators: Running water</i>	Running water, rivers or streams	Water
<i>Pasture/Dung</i>	Land covered with grass/low plants, suitable for grazing	Managed
<i>Meadowland</i>	Indicators of open grassy habitats or meadows	Natural
<i>Wood and trees</i>	Indicators of habitats covered in trees and wood	Natural
<i>Indicators: Deciduous</i>	Deciduous trees, shed their leaves annually	Natural
<i>Indicators: Coniferous</i>	Coniferous trees with evergreen leaves, cone bearing	Natural
<i>Wetlands/marshes</i>	Marshes or swamps, saturated land	Natural
<i>Open wet habitats</i>	Indicators of wet open areas, like marshes or wet woodland	Natural
<i>Disturbed/arable</i>	Land used or suitable for growing crops	Managed
<i>Sandy/dry disturbed/arable</i>	Dry land used or suitable for growing crops	Managed
<i>Dung/foul habitats</i>	Suggest dirty or filthy conditions	Buildings
<i>Carrion</i>	Cadaver. Decaying flesh	Wild, domestic
<i>Indicators: dung</i>	Indicates animal faeces	Managed
<i>Mould beetles</i>	Suggest rotten or decaying stored food or mouldy grain	Buildings
<i>General synanthropic</i>	Species living in relation to or near humans and human-made artificial habitats	Managed
<i>Stored grain pest</i>	Indicators of harvested and stored grain	Buildings
<i>Dry dead wood</i>	Indicates trees and woods that have dried out and dies	Natural
<i>Heathland and moorland</i>	Wide open landscapes dominated by low plants like heather, indicates low soil fertility	Natural
<i>Halotolerant</i>	Tolerate conditions of high salinity, inland salt seas or springs	Water
<i>Ectoparasite</i>	Parasite that lives on the outside of its host	Wild, domestic

NABOne only contains points located in Iceland and did not have to be clipped or cropped in any way. For the current NABOne data none of the data points have any indicator values for 7 of the indicators. These indicators have thus been disregarded for our analysis. Table 5 below presents the 12 indicators for NABOne, which ones have been disregarded and the recategorization of the 5 indicators that have been kept.

Table 5: The 12 indicators found within the NABOne dataset, described and recategorized into the 10 defined concept categories.

Indicator	Description	Concept category
<i>domestic</i>	Domestic animals, such as horses, sheep or cattle	Domestic
<i>wild</i>	Wild animals, such as wolves or crows	Wild
<i>Marine Mammal</i>	Marine mammals, such as seals or whales	Water
<i>Marine Fish</i>	Saltwater fish such as cod, or mackerel	Water
<i>Freshwater Fish</i>	Freshwater fish, such as arctic char or mullet	Water
Not Included		
<i>terrestrial</i>		
<i>Aquatic</i>		
<i>Sea Bird</i>		
<i>Non Migratory Terrestrial</i>		
<i>Fresh Water Migrant</i>		
<i>On floe ice</i>		
<i>On fast ice</i>		

SEAD indicator values were adjusted by dividing the original value per indicator per polygon divided by the total indicator value within that polygon. This was done because the differences in values between polygons is so big that the final clustering is more a reflection of these differing values than of differences between the categories themselves if the raw values are implemented. An example is given in Figure 4 below, where the results of a SOM of the SEAD dataset has been conducted with adjusted and unadjusted values.

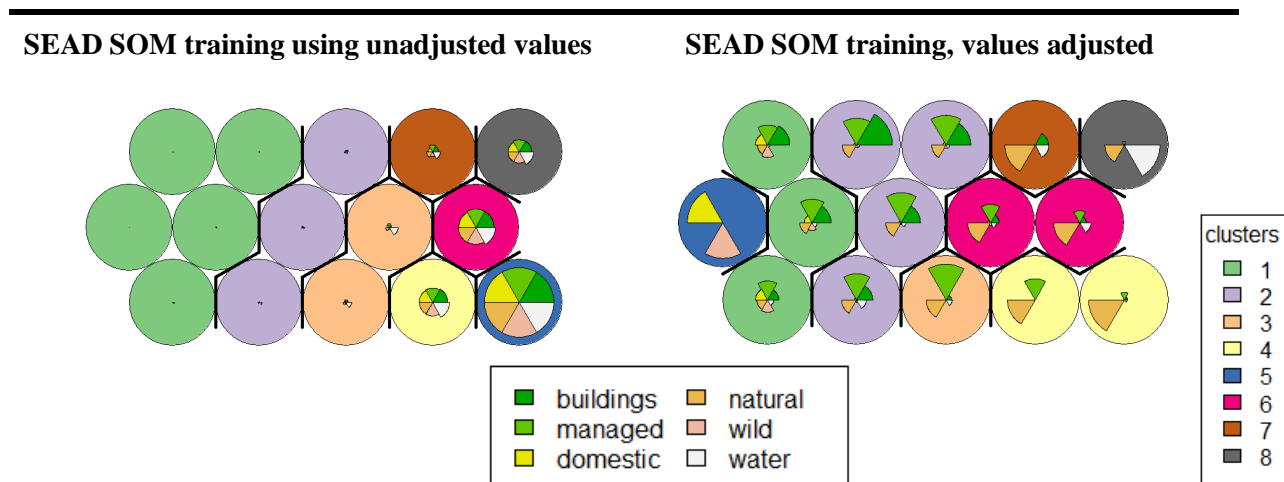


Figure 4: SOM clustering of the SEAD dataset, unadjusted (left) and adjusted with respect to total indicator values within each municipality (right). Notice how the clustering of the unadjusted dataset is completely ruled by differences in total cluster values between the municipalities

Similar to the SEAD indicator values, the NABOne dataset was adjusted to be fractions of each concept category within each municipality. The differences in indicator values differ wildly between sites; not adjusting values led to an exclusion of several sites (Figure 5).

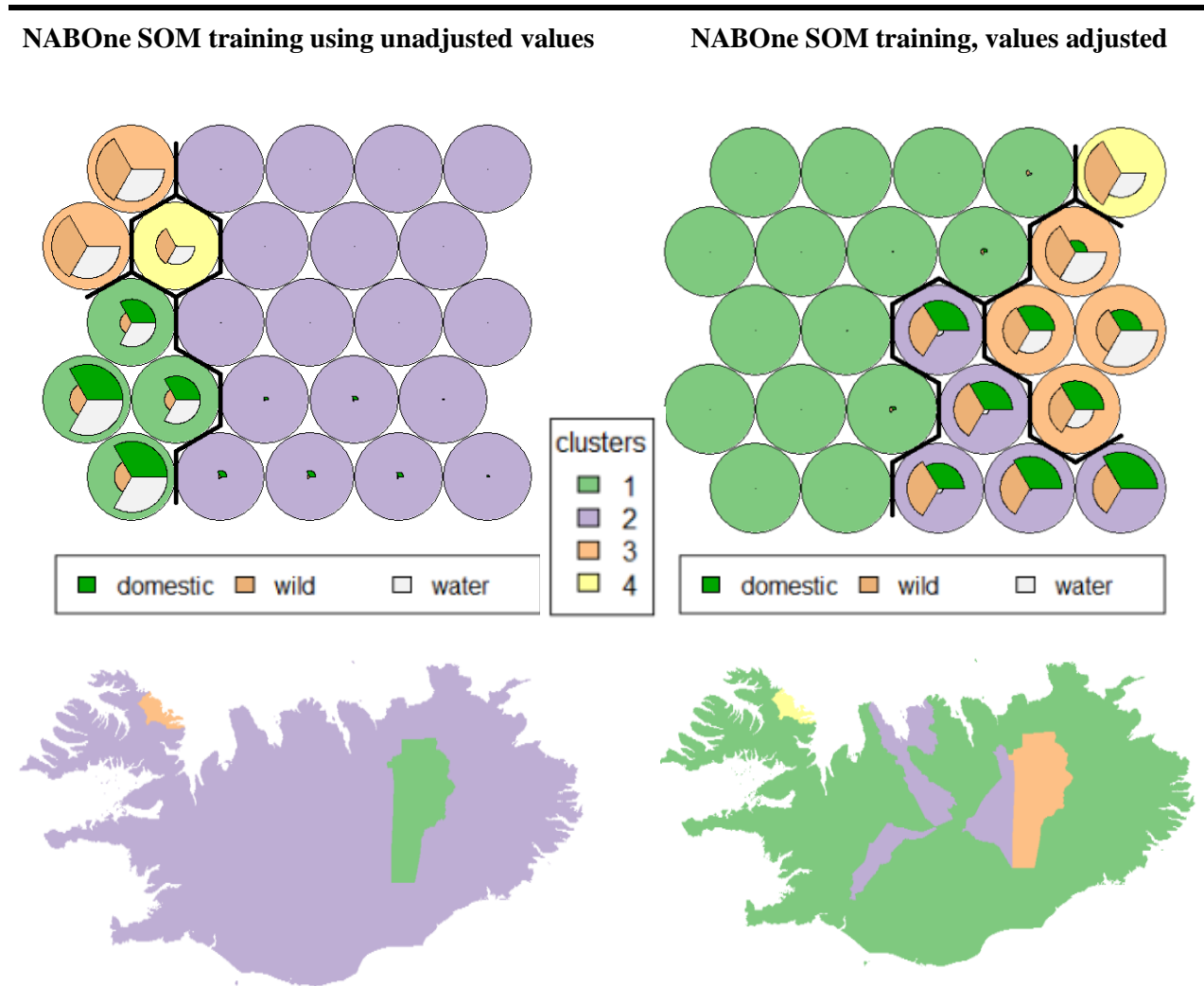


Figure 5: SOM clustering of the NABOne dataset, unadjusted (left) and adjusted with respect to total indicator values within each municipality (right). Maps have been included to show the impact the adjustment of values have on the mapped output of the clustering.

Adjusting values aims to preserve and present more of the information that is stored in the NABOne data points. This might cause some slight skewing of information, especially if the number of finds at a site had been a reflection of the actual abundance (or lack thereof) of fragments from animals or fish. From discussions with zooarchaeologists in the dataARC team became evident that these differences in numbers are more a

reflection of time, money and effort spent on certain excavation sites compared to others, not necessarily a reflection of any real trends or patterns. This idea is supported by the vast differences in number of points, or contexts, studied at each site.

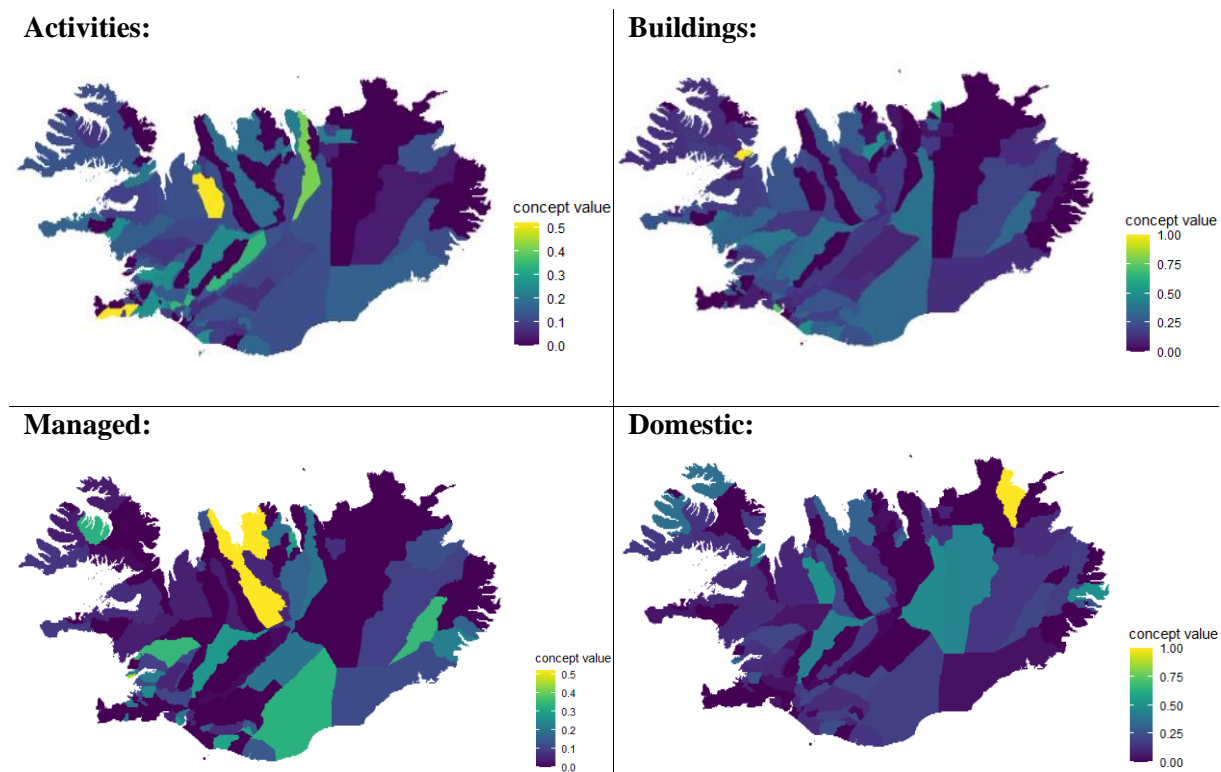
6.2. SOM training and Clustering

Following the standardising and pre-processing of each individual dataset, the prepared data is now ready to be implemented into the SOM training model.

6.2.1. Pre-training

Pre-processed and adjusted datasets were combined in an excel file, and any municipalities containing no data points were removed. Excel files can easily be imported into R using the “readxl” package.

Prior to the SOM training the datasets must be imported into R and merged with an Iceland shapefile for the SOM output to have a spatial reference which can be mapped. A shapefile of Iceland, with all municipalities mapped, was retrieved from <https://www.diva-gis.org/gdata> and merged with the prepared dataset by a common ID. Indicator values are now appended to each individual municipality and can be presented by abundance as shown in Figure 6 below.



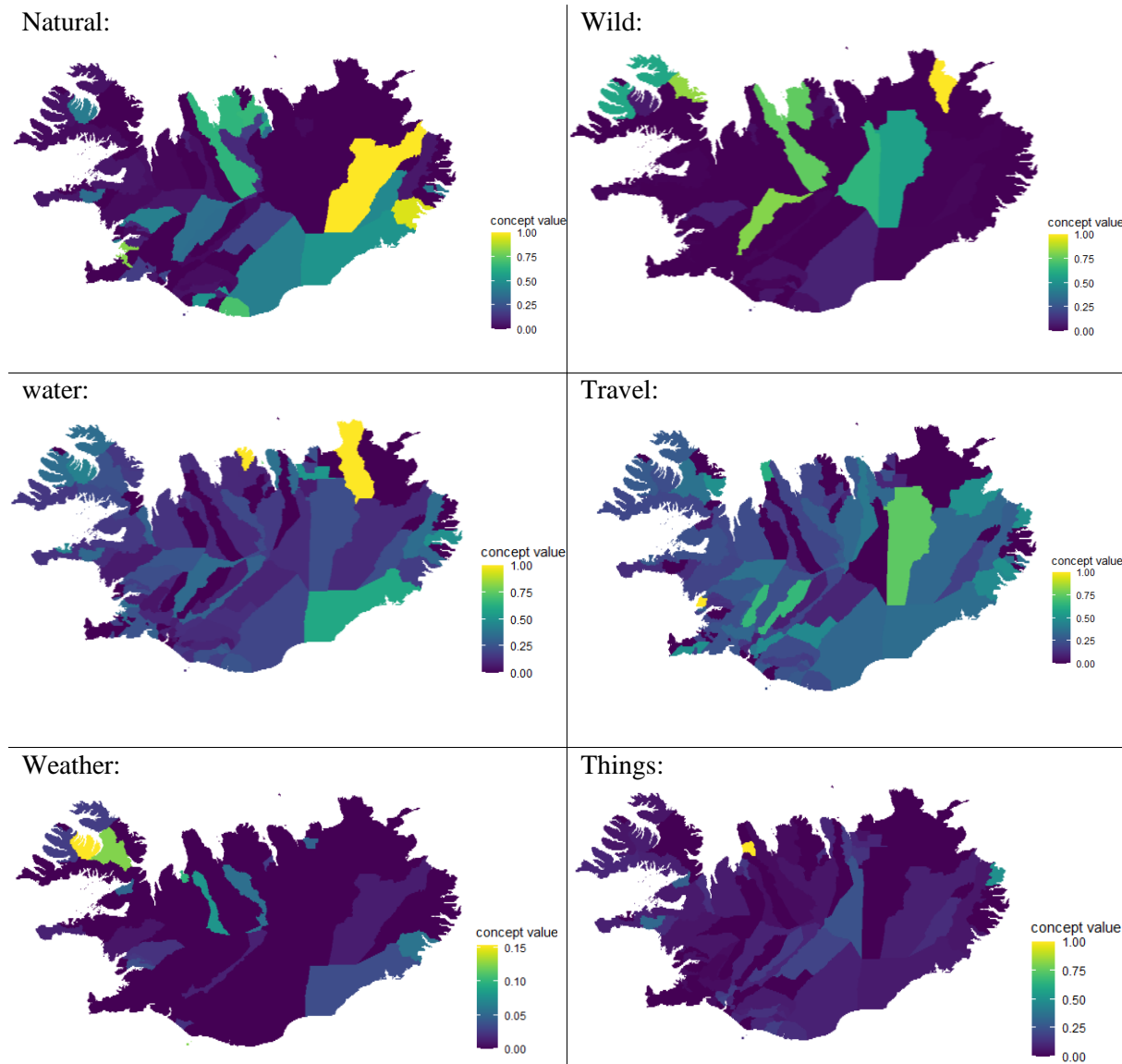


Figure 6: The relative abundance of each concept category within each municipality. The values represent a sum of the adjusted values for each dataset.

6.2.2.SOM training

The first steps of the training process are to convert the data frame to a matrix, or grid, and standardise it before implementing it into the model. The parameters of the SOM matrix can be adjusted based in size and complexity of the dataset (Kohonen, 2001). To avoid over-complicating our dataset and unnecessarily extending processing time, a relatively small grid of 8x8 seems the most fitting (Kanevski et al. 2009). A grid any smaller than this would struggle to present the topological relationships between the neurons (Skupin, 2004).

During training the N-dimensional dataset is presented to the bi-dimensional grid. This forms a layer where only one neuron on the grid responds to each input sample (or data point). Which neuron that would be is

determined by distance, where the neuron closest to the specific input sample “wins” that sample (Chicco et al. 2003). The winning neuron actively responds by being updated according to this relationship:

$$c_{new}^{(i)} = c_{old}^{(i)} + \eta(x^{(m)} - c_{old}^{(i)}) \quad (1)$$

η represents the pre-defined learning rate. During the training process the model updates both the weight of the winning unit as well as its neighbouring units, which helps preserve topology and means that any neurons that are spatially close on the map have corresponding patterns (Kohonen, 1989).

Code example 4 shows the specific R command that trains the model. “data_train_matrix” are the rescaled input variables which have been reshaped into a matrix, “som_grid” is a predefined grid with a size of 8x8 and a hexagonal topology. “rlen” is the iteration number and determines the number of times our dataset is presented to the network, and “alpha” refers to learning rate. Based on suggestions from the Kohonen package documentation (Wehrens & Wehrens, 2019), both “rlen” and “alpha” are kept as their default values.

```
som_model <- som(data_train_matrix,
  grid=som_grid,
  rlen=500,
  alpha=c(0.05, 0.1),
  keep.data = TRUE )
```

Code example 4: SOM training in R, displaying the actual values for the rlen and alpha variables

Figure 7 presents a plot of the SOM training process. During training, the codebook neurons are getting increasingly more similar to objects in the dataset. Visualising this process helps validate the chosen number of iterations and optimise training parameters (Wehrens & Buydens, 2007). The plot shows the average distance from all dataset objects to their closest codebook neuron.

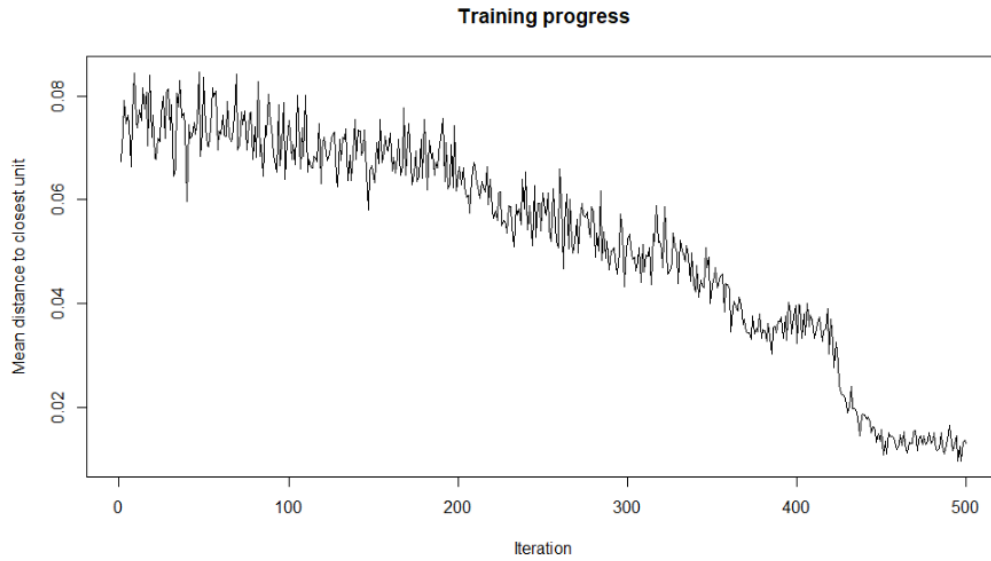


Figure 7: Training progress for the combined dataset.

In order to assess the quality and accuracy of the SOM training, and the fit of all variables, it can be useful to map the number of data objects mapped to each neuron and the distances between objects and their corresponding neurons (Wehrens & Buydens, 2007). This has been done in Figure 8 below. Empty units are depicted in grey. The plots show a reasonable spread of objects mapped to the codebook neurons, and the distances between neurons and objects are overall satisfactory (Hsu & Halgamuge, 2003).

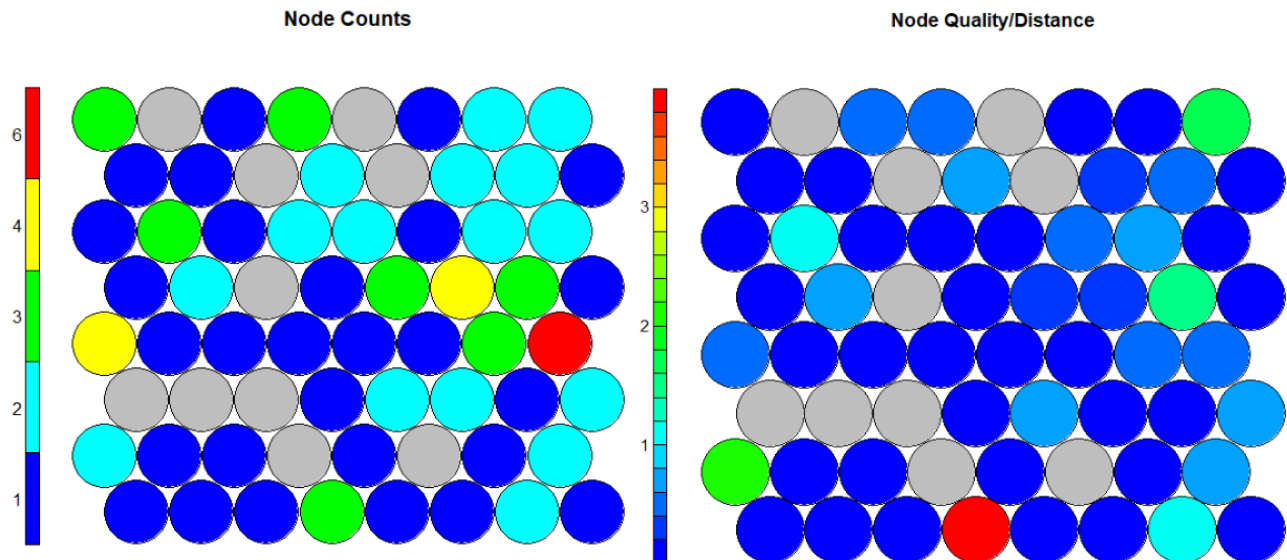


Figure 8: Node count plot (left) and mapping quality based on distance between objects and codebook neuro

Figure 9 shows the final codes spread, with the signature of each neuron presented using a segment charts corresponding to the combination of concept values.

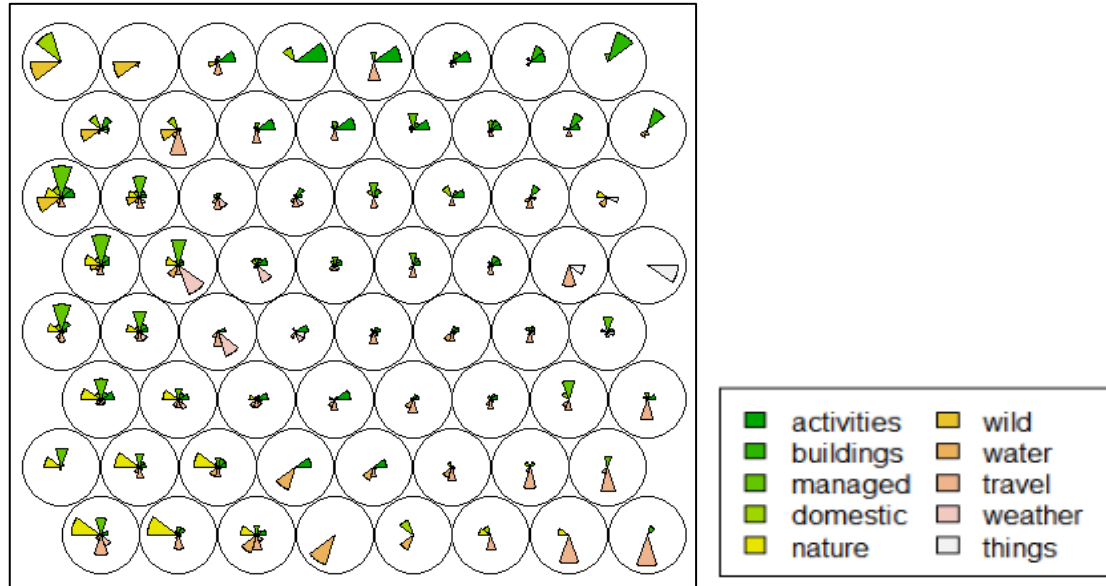


Figure 9: SOM training output, represented by segments plots within each unit.

From studying the codes spread alone one can start to perceive and form ideas about the general spread of data in Iceland and envisage where cluster boundaries might be placed based on the overall topology of the matrix and neurons. For a dataset presented in a smaller grid the clustering process could be carried out manually. However, in order to produce the most unbiased results for a grid of this particular size, the clusters must be identified under no human supervision (Malone et al. 2006).

6.2.3. Clustering

Clustering helps with identifying patterns in the trained matrix, or neighbourhoods of neurons with similar features. The process is again unsupervised, and clusters will consist of varying numbers of neurons depending on their similarities to other neurons (Chicco et al. 2003).

Identifying the ideal number of clusters for the specific dataset can be done in two ways: 1) Run the model several times with varying numbers of clusters to find the best fit or 2) Plot the Within Cluster Sum of Squares (WCSS), which will help indicate the most ideal number of clusters for the specific dataset (Hartigan, 1979). WCSS is an estimate of variability within a cluster, which ideally should be as low as possible in order to produce more uniform, or compact, clusters (Kaminka, 2016). For this analysis a combination of the 2 methods described above were used, by first plotting the WCSS and running a few clustering test-runs to identify the

most useful number of clusters. Then the clustering process will be executed a few times using some of the most promising number of clusters.

Figure 10 shows the result of the WCSS plotting, which presents the variances within the clusters. The internal variance decreases as the number of clusters increase but elbows at 10 clusters. This indicates that including more than 10 clusters will make little difference to the data (Pollard, 1981).

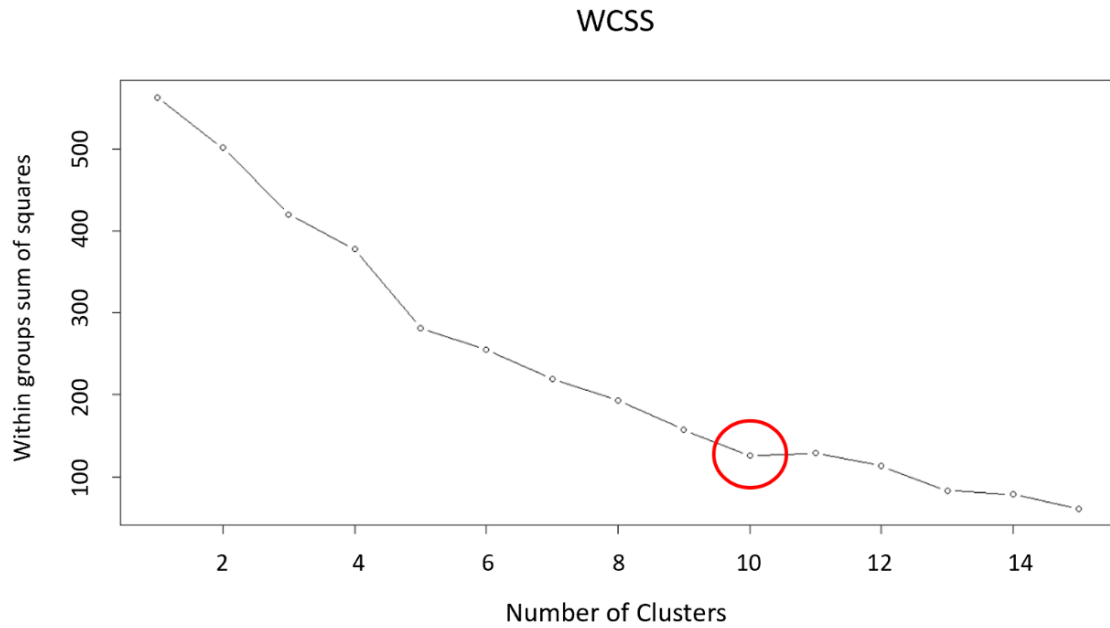


Figure 10: WCSS of dataset, showing a distinct bend at 10 clusters

Bearing in mind the results from the WCSS, the clustering process was run 4 times with 8, 9, 10 and 11 defined clusters (Figure 11). There is very little difference between 8 and 9 clusters, however the addition of a 10th cluster splits up one of the largest, but quite non-uniform, cluster. An additional 11th cluster creates unnecessary separation of neurons which are seemingly quite similar. Thus, proceeding with 10 defined clusters will give the best results in terms of both data preservation and appropriate separation of units.

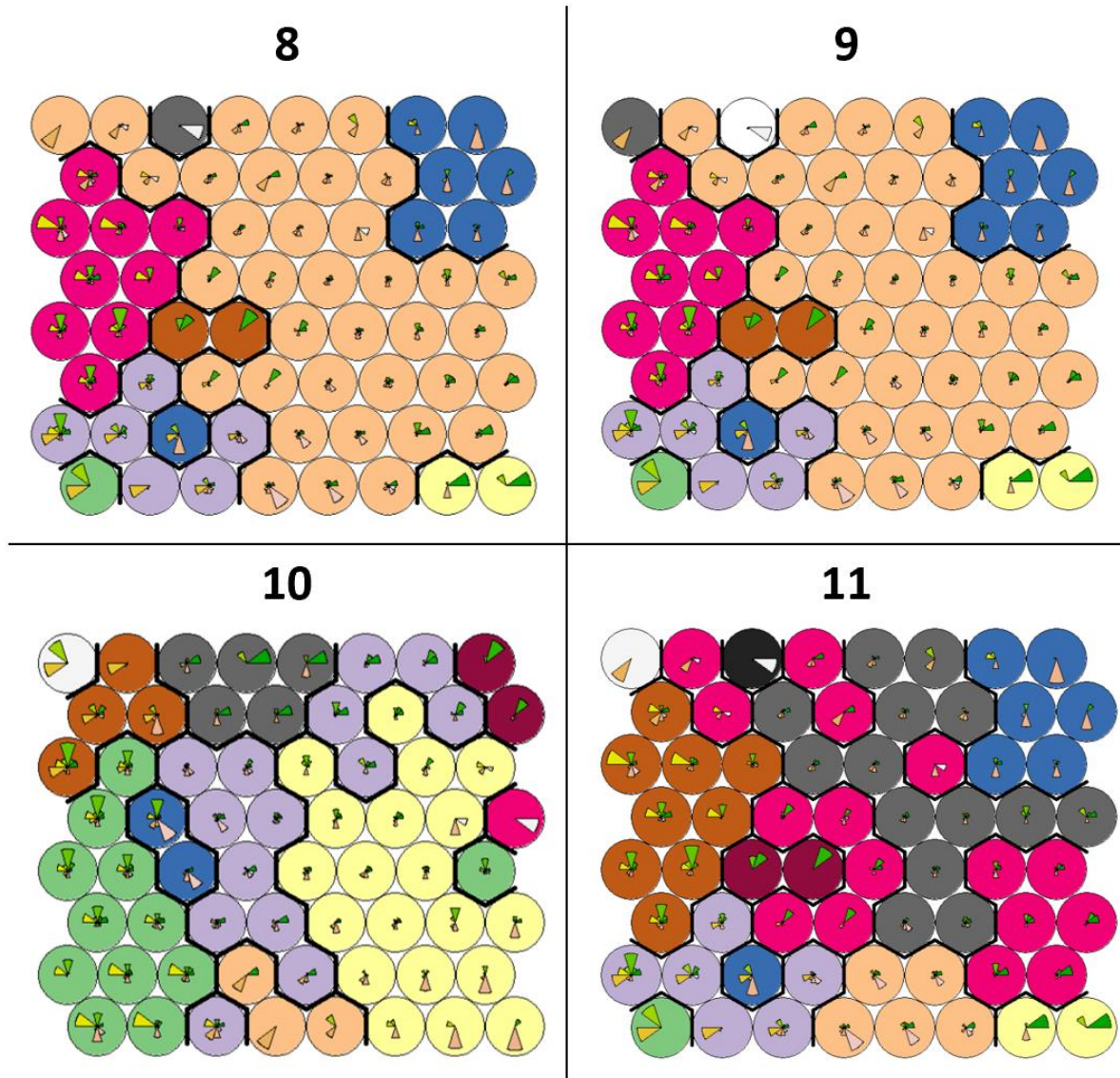


Figure 11: Clustered neuron grid with 8, 9, 10 and 11 defined clusters. The legend for the segment plots within each neuron is the same as for Figure 9.

6.2.4. Mapping of cluster results

In order to map the results of the clustering, the cluster units must be linked to their respective areas in a data frame and merged into the original spatial polygon data frame. The map can be exported from R into qGIS as a shapefile and mapped by colour corresponding to the 10 different clusters. The Self-organising map approach produce a choropleth map where each colour represents a different cluster “group” with its own set of distinguished features.

6.3. User Testing and analysis re-evaluation

As the main motivation for this project essentially evolves around designing a data mining tool for a research team, the inclusion of said team in the design project has been crucial in shaping the tool and provided great help and guidance during the pre-processing of individual datasets.

Through team meetings discussion about the most constructive way to categorise datasets led to the creation of the 10 established concept categories used for the final outputs of this study. Prior to making this decision, SOMs using both individual concept or indicator categories for all three datasets or using a higher or lower number of common categories were attempted, all with varying levels of success in terms of giving a meaningful output. Additionally, team members were worried about the potential complexity that would arise if tens of datasets, all with different concept or indicator categories, would be clustered together. As a response to this, and in order to produce the most meaningful and straightforward outputs possible from the analysis, the final 10 concept categories were established.

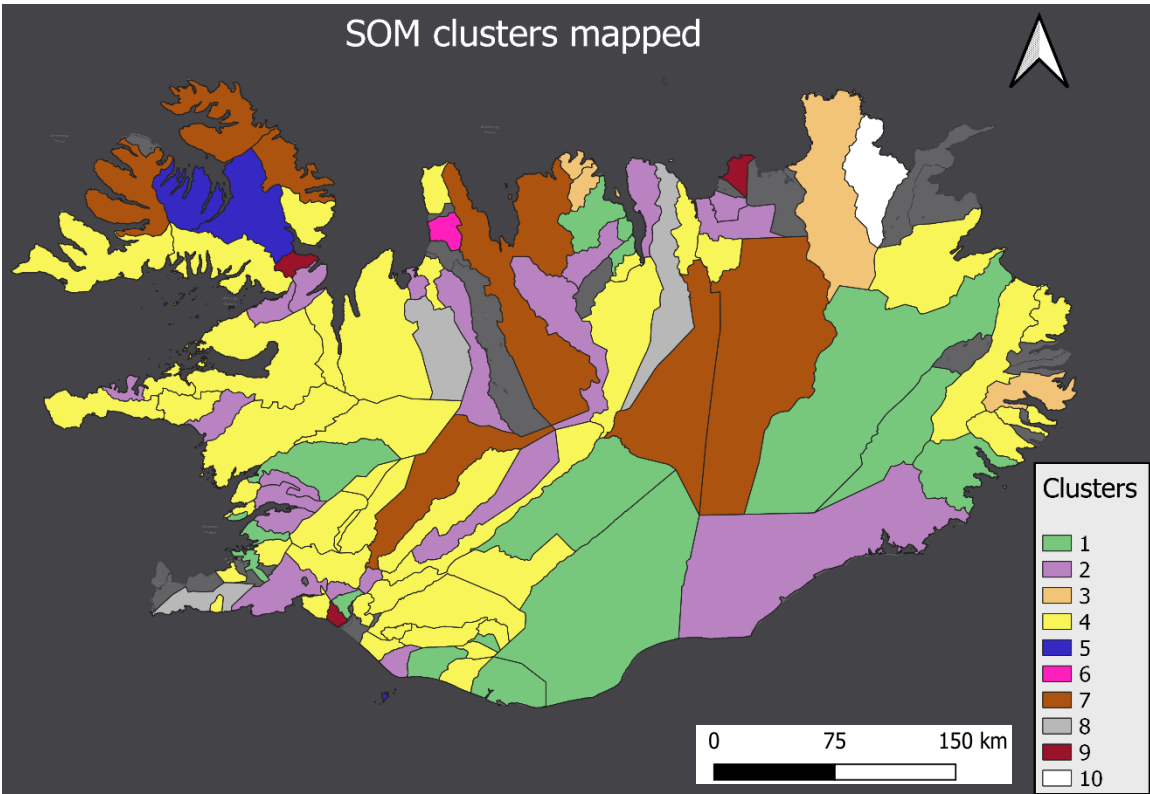
User testing sessions, where data providers are given a demonstration of the training and clustering process and trained in how to include their own datasets into the analysis, will be held following the finalisation of this mapping prototype.

7. Mapping Results

This section explores a set of different mapping methods on a combined and singular dataset scale. The results of some of the mapping procedures will also be examined in more detail and their importance as well as whether or how they might be used or implemented into the final product will be discussed.

7.1. Combined SOM

The clustered areas can be presented spatially on a choropleth map (Figure 12) where each colour corresponds to a specific cluster. Such maps are very useful for providing a full overview of the total combined dataset and how it maps visually. However, it might also hide certain patterns which will only be visible on a more detail or close-up level.

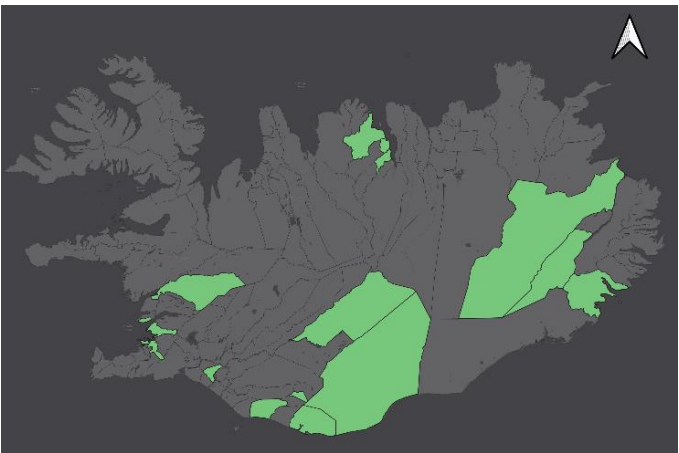
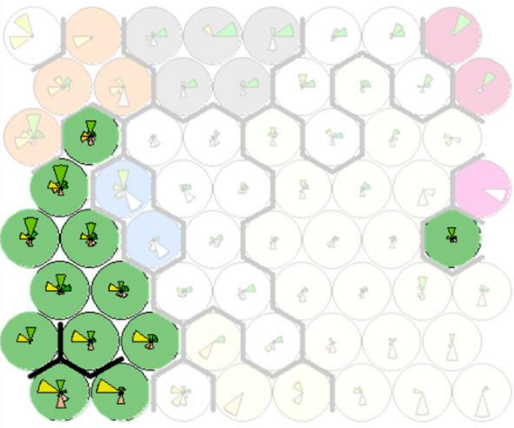


In

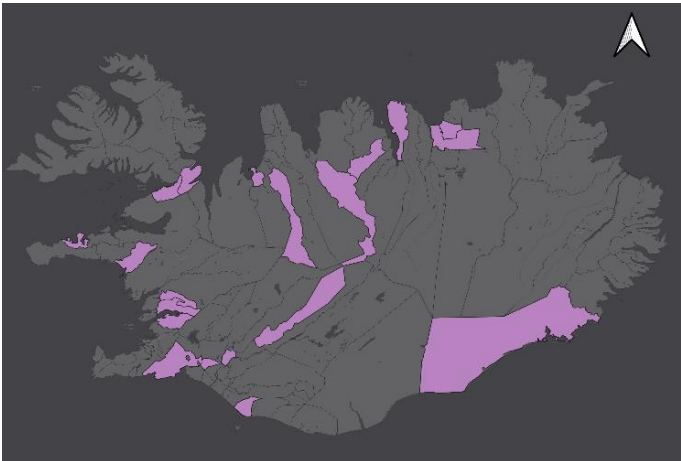
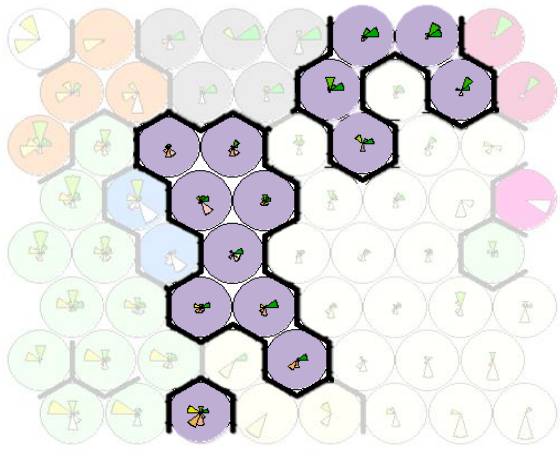
Figure 12: Visual representation of the clustered SOM results for all datasets combined. Municipalities containing no data have been excluded from the mapping. Numbering and colour coding of the clusters correspond with the training output shown in Figure 9.

order to better analyse and understand the composition of the respective clusters, the SOM clusters and corresponding municipalities can be mapped side by side (Figure 13). This makes analysing the nature of each cluster more straight forward as well as being a helpful visual tool

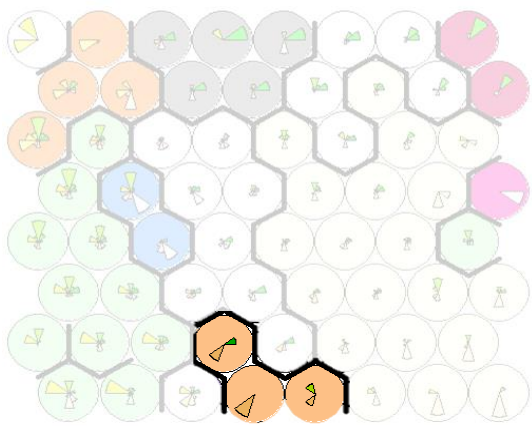
Cluster 1



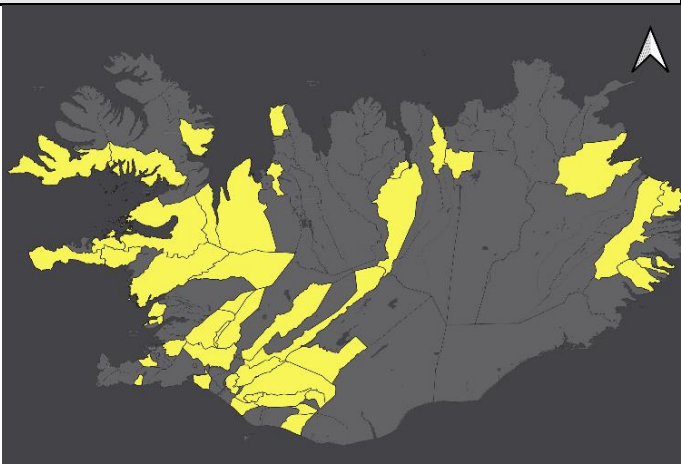
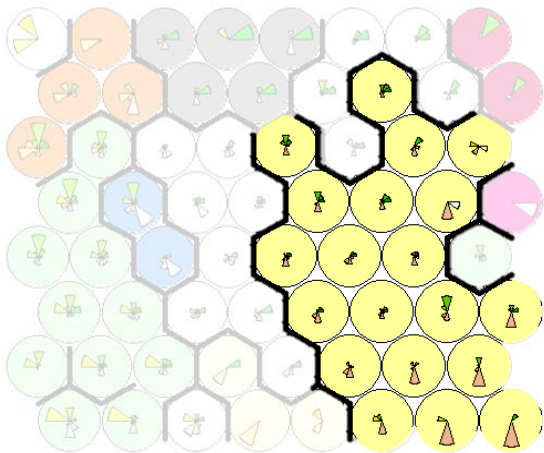
Cluster 2



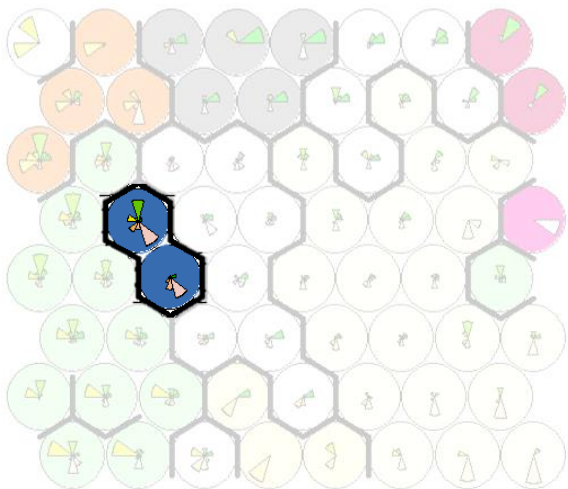
Cluster 3



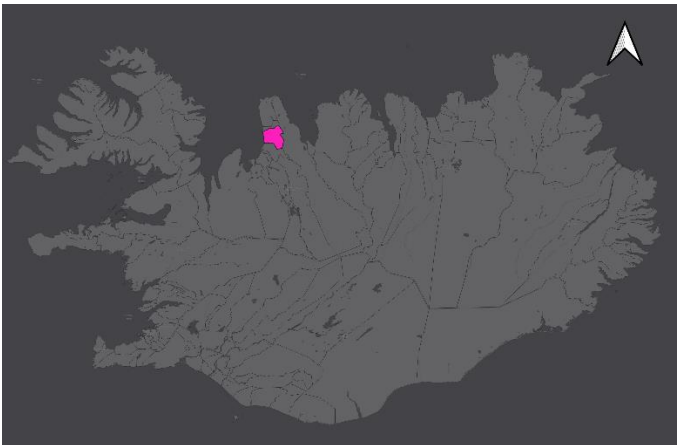
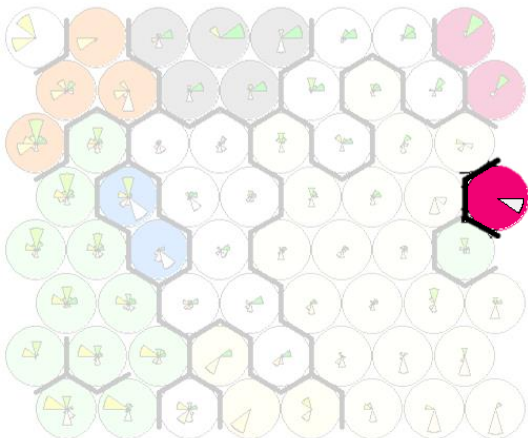
Cluster 4



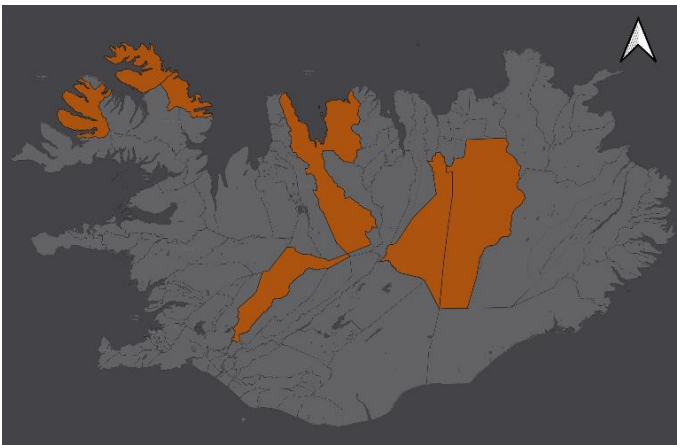
Cluster 5



Cluster 6



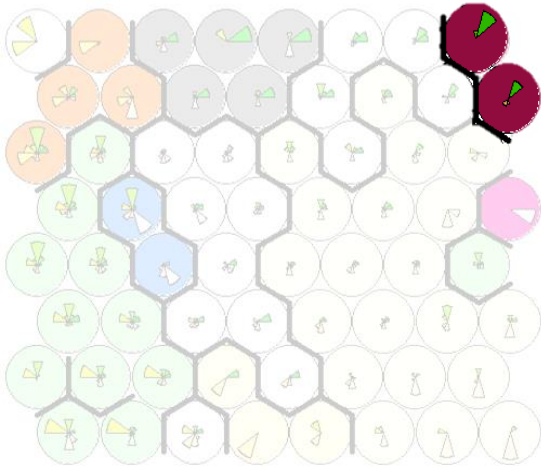
Cluster 7



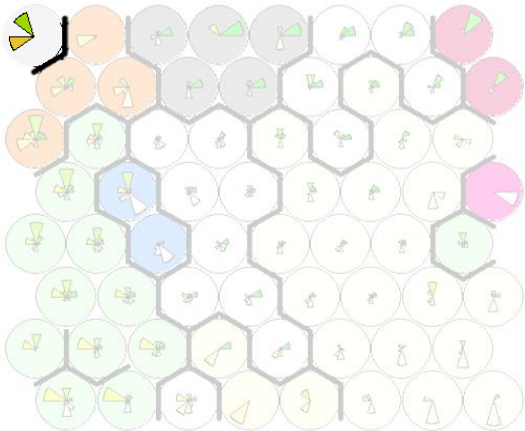
Cluster 8



Cluster 9



Cluster 10



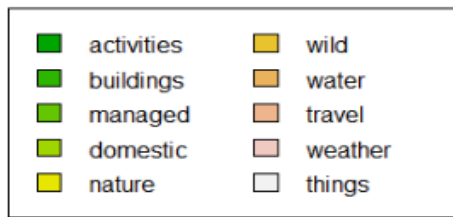


Figure 13: The 10 output clusters mapped individually (right), along with their corresponding neurons highlighted on the trained SOM matrix (left)

7.2. SOM of individual datasets

As an additional output feature to the combined SOM visualisation, the decision was made to produce individual SOMs for the 3 included datasets and map these spatially across Iceland as well. The idea behind this approach is to produce visualisations that can be used to validate the results of the final output map. Additionally, these maps might also help identify initial similarities between data values for the 3 datasets spatially across Iceland.

Trained SOM matrices for the 3 individual datasets are presented in Figures 14, 15 and 16 below, along with the mapped cluster results and descriptions for each cluster. As the number of municipalities and concept categories represented in the datasets differ greatly, different grid sizes and number of clusters were used to produce individual SOMs.

7.2.1. Sagamap

The Sagamap dataset contains data, which is represented in every single concept category, and data exist within 91 of the 119 municipalities.

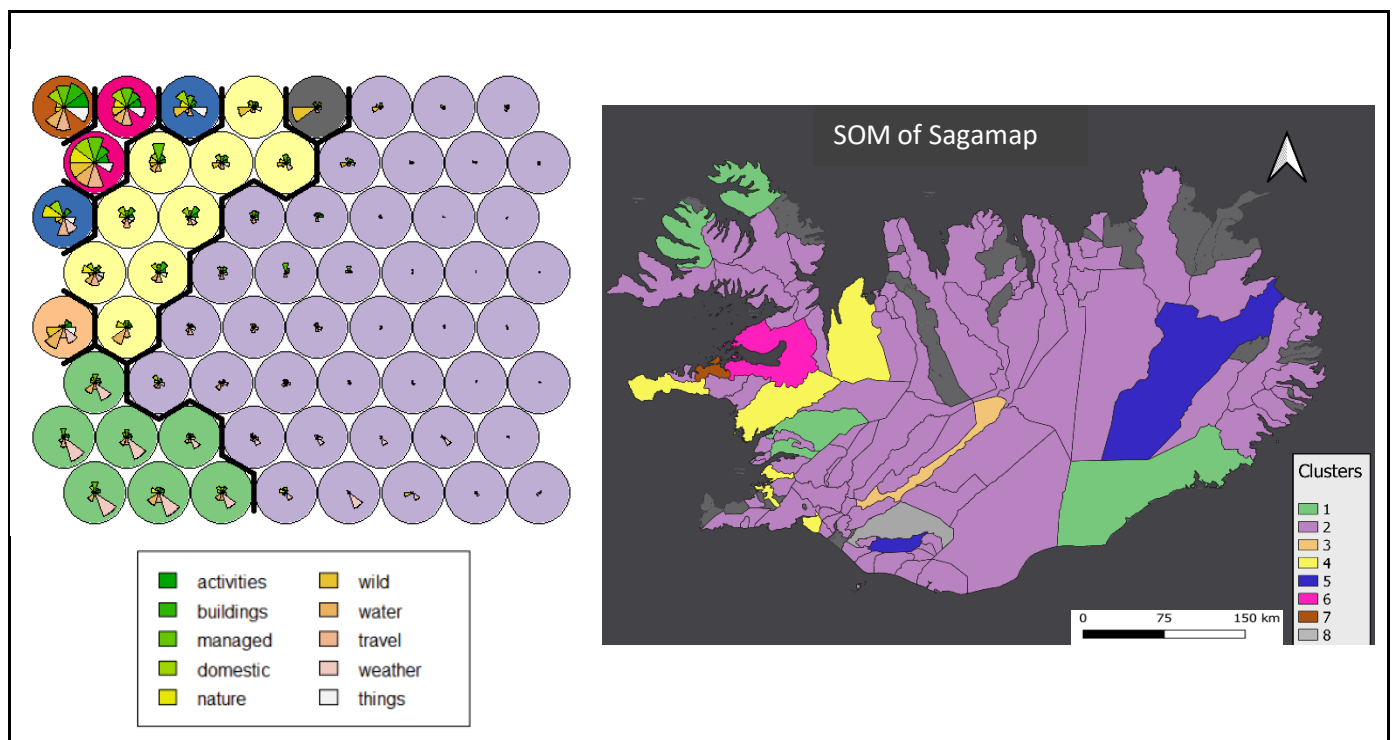


Figure 14: Results of SOM training and mapping for the Sagamap dataset

A cluster number of 8 was chosen both due to the slightly lower number of included municipalities, as well as the increased homogeneity within this dataset compared to the combined version. The 8 clusters are described in detail below.

CLUSTERS:

- 1. High number of mentions of weather conditions compared to the rest of Iceland, mostly due to the fact that there are very few mentions of weather in total.*
- 2. Represent most of Iceland, these are areas with overall few mentions in Icelandic Sagas. Some municipalities have some mentions of weather and some of wild animals.*
- 3. High number of wild animals and water, some of weather and things, otherwise very little. There are very few points located within this region overall. Little activity, whatever happens here is mostly related to water and wild animals*
- 4. Show a large number of managed land and water related mentions, as well as wild animals. People occupy these areas but seem to be in close contact with the wild*
- 5. Area dominated by objects, activities and events related to domestic animals, natural landscapes and travel*
- 6. High number of points, which mainly belong to managed and natural land, domestic and wild animals, water and travel. Certain similarities to cluster 7, although cluster 6 has a higher number of water-related mentions (which makes sense as this municipality is on the coast).*
- 7. This cluster only contains Helgafellssveit, which is the municipality with the highest number of points in total (383). High number of mentions of almost every concept apart from weather, wild animals and nature. Could be an important “high seat” for people, a place they live, meet up, travel to etc.*
- 8. Very high number of wild animals, few mentions of everything else. Lies close to areas that have been identified as more populated, so people probably roam in here and observe see wild animals.*

7.2.2. SEAD

The SEAD dataset is categorised using only 6 of the 10 defined concept categories and is represented in 17 of the 119 total municipalities. Even so, the variations within the dataset are significant enough that it seems sensible to arrange the results from the SOM training into 8 respective clusters.

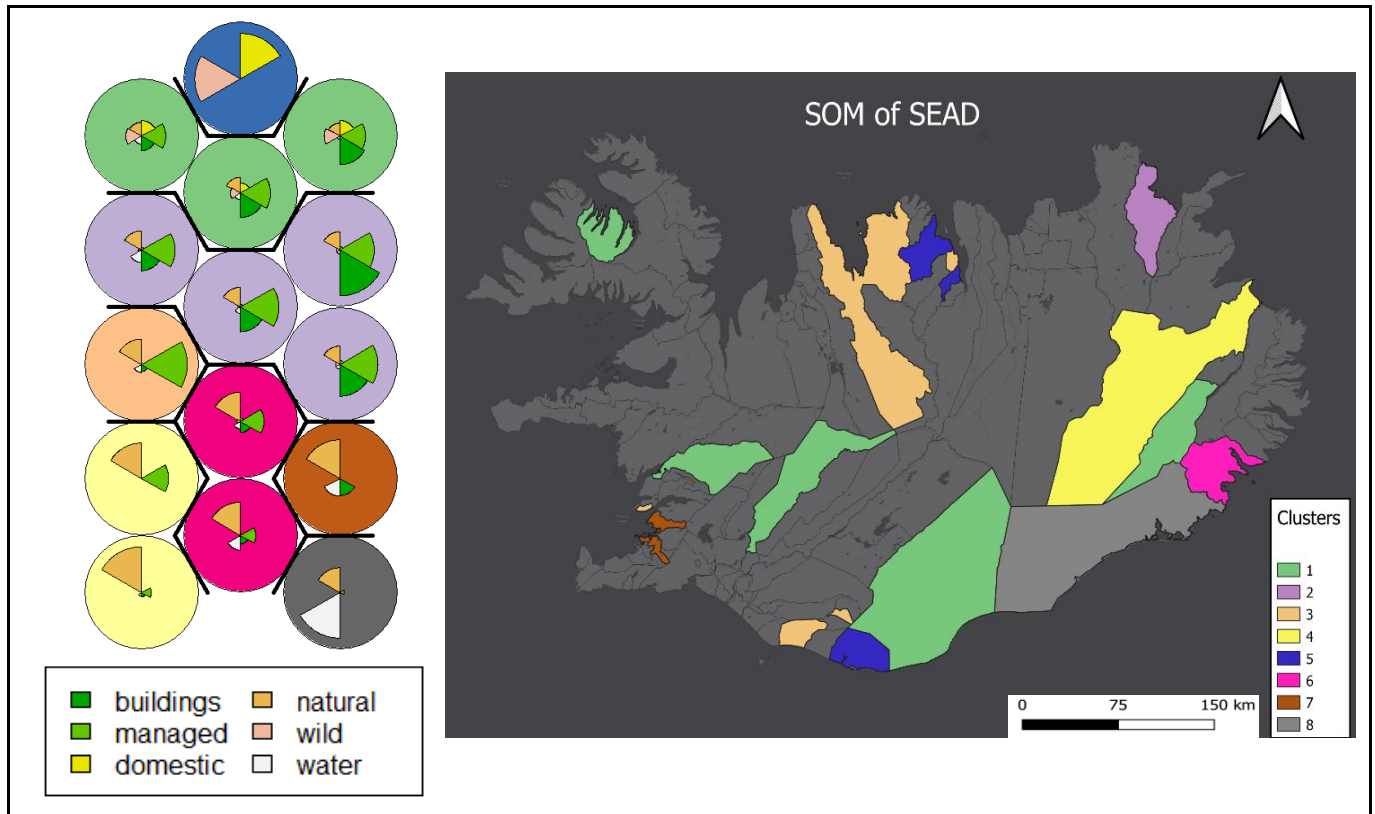


Figure 15: Results of SOM training and mapping for the SEAD dataset

CLUSTERS:

1. Buildings, managed land, some of domestic and wild animals and natural landscapes. Places where humans reside but where there's also an abundance of natural landscapes
2. Large numbers of indicators related to buildings and managed land, some linked to natural landscapes. Species indicating human activity are abundant in the zooarchaeological record
3. Indicators mainly related to managed landscape, some indicating natural landscape
4. Large number of indicators of natural landscape, some of managed land. Area mainly natural with some signs of human alteration
5. Record only show indicators related to animals, both wild and domestic
6. Mainly indicators of natural landscape, some signs of human indicators and species related to water
7. Mainly indicators of natural landscape, some indicators relating to buildings and water
8. Large number of indicators of water, some suggesting natural landscape. No sign of species suggesting human settlements or landscape alteration

7.2.3. NABOne

Although the NABOne dataset consist of a large number of data points, the spatial spread of the data is currently very limited. Data points from the 9 sites are located within 7 municipalities, which is too little to produce a meaningful SOM analysis. Thus, all 119 municipalities were included in the training and the 7 data holding areas were extracted from the final result and presented in R.

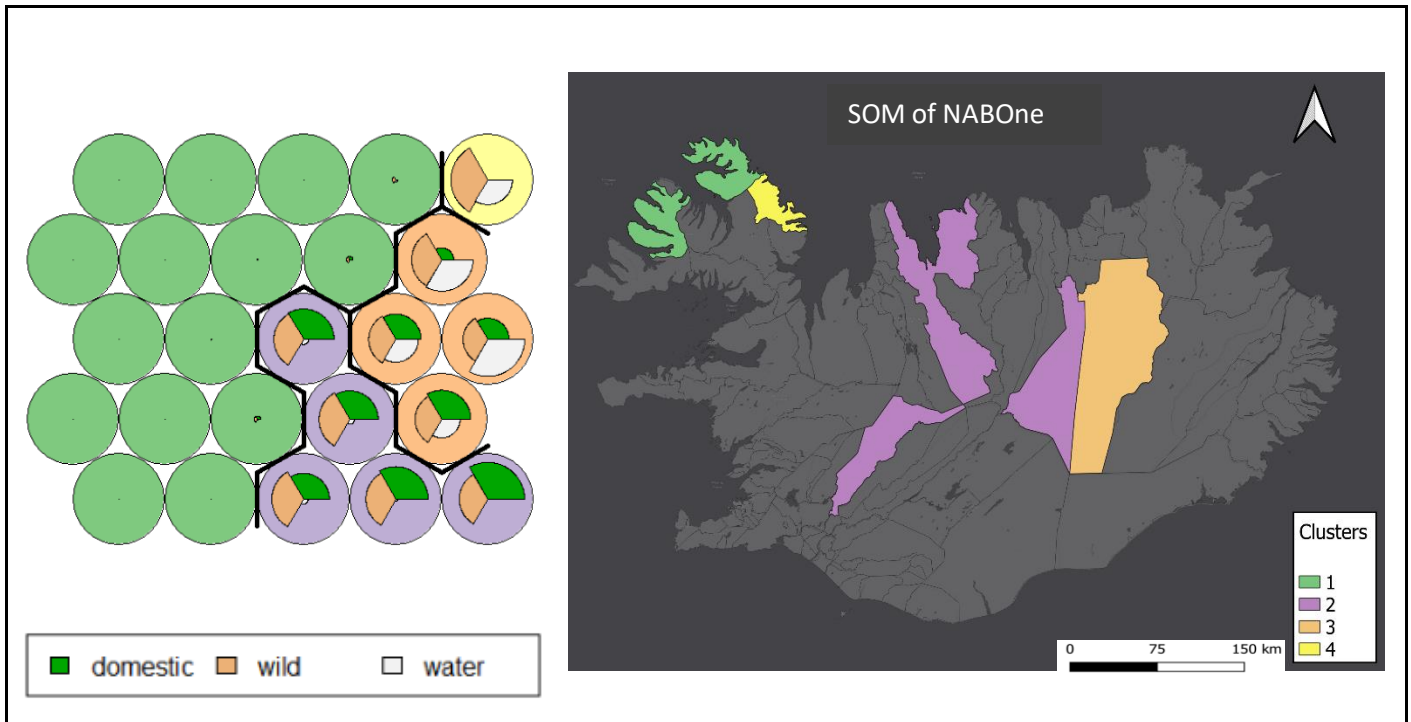


Figure 16: Results of SOM training and mapping for the NABOne dataset

A WCSS plot suggested 4 clusters to be the most ideal number for this dataset, these are described in more detail below:

CLUSTERS:

1. *Very few data points with low numbers of records within each context. Suggest mainly wild animals with some indicators of domestic animals*
2. *Primarily indicators of wild and domestic animals, more or less in equal quantities. Mainly inland, very little evidence of mammals or fish living in or near water in the zooarchaeological record*
3. *All categories well represented, this is the region with the highest number of contexts and records recorded. Show the overall highest abundance of species related to water*
4. *No signs of domestic animals in the zooarchaeological record, some of water related species and several records suggesting wild animals*

7.2.4. NABOne – SOM of point outliers

Because the NABOne dataset consist of a very high number of points scattered over a very limited number of sites or locations, an additional approach to the individual SOM training and mapping is to apply it directly to the point data as opposed to municipalities. This might help combat the oversimplification of the information contained in NABOne and better indicate the general variation in information and indicators that exist for individual sites.

Based on the values within the NABOne dataset, where most points have indicator values between 0 and 2 but some have disproportionally higher values, the most suitable way of standardising the dataset will be

through multivariate outlier identification (Bengal, 2005). This will eliminate a lot of unnecessary noise from the dataset and allows us to focus on data which identify true conditions.

An isolation forest multivariate outlier detection method was applied to the dataset (Liu et al. 2008). The isolation forest technique successfully identified 47 outliers out of the 928 total points and the indicator values for these points seemed reasonable for outliers (Figure 17).

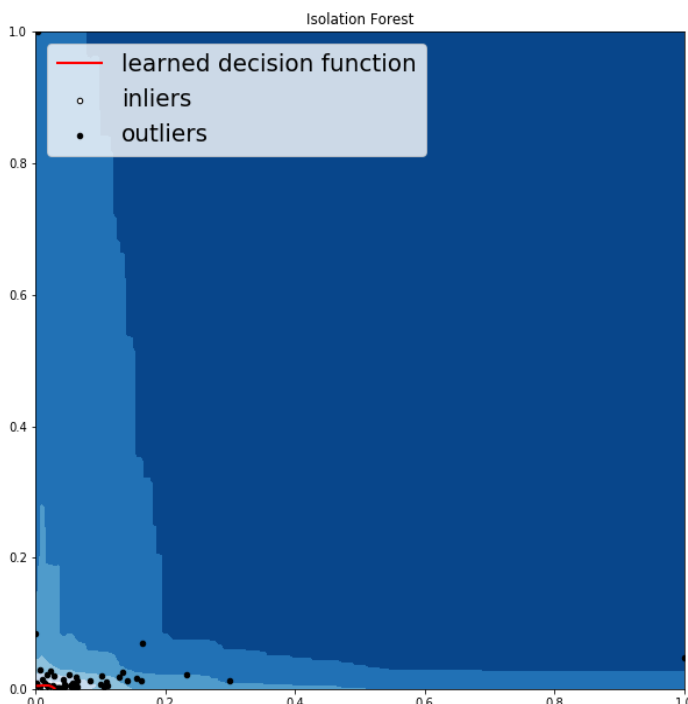


Figure 17: Output of an isolation forest outlier detection method run in python.

Following clustering, SOM training was applied to the 47 outliers and the results were mapped as 4 different clusters in GIS (Figure 18). The output of this point outlier clustering helps clarify certain characteristics with the NABOne dataset.

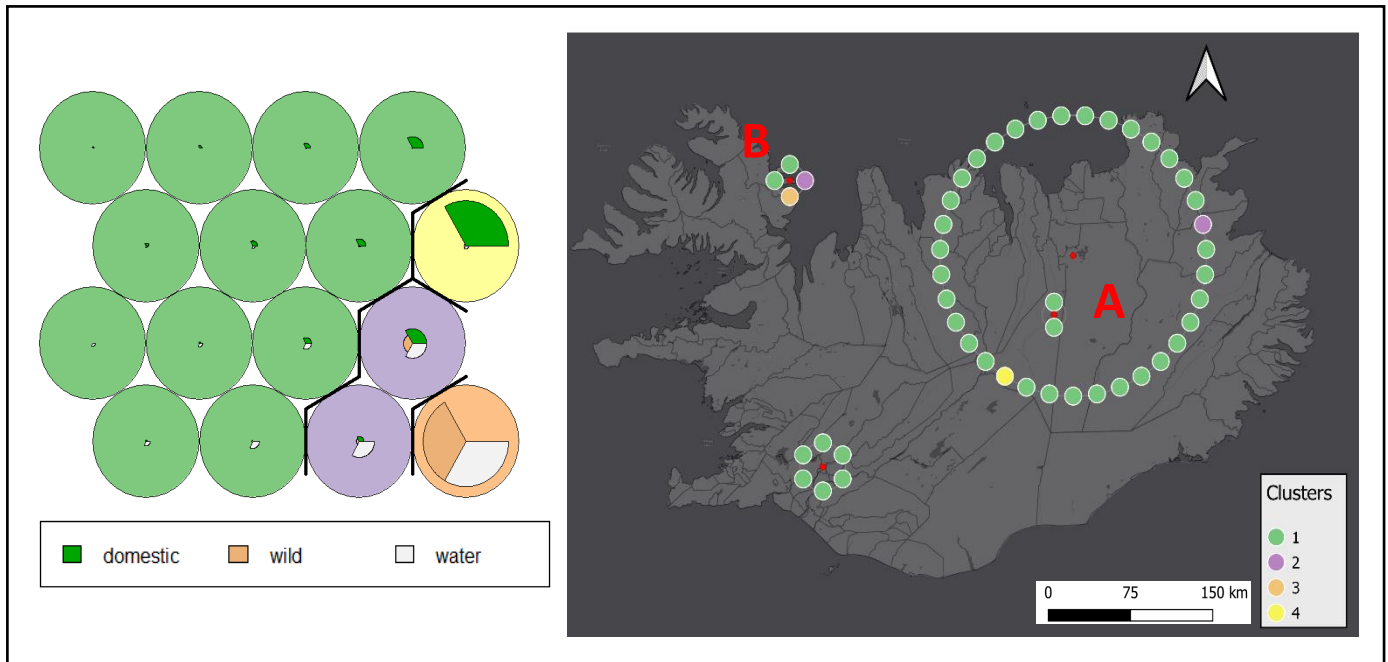


Figure 18: NABOne outliers, trained, clustered and mapped by specific location or excavation site.

Firstly, the spread of detected outliers suggests that the points making up cluster 1 and 2 in the initial clustered NABOne dataset have very low, uniform values (not outliers). There is significant variation within and between certain sites but for the most part the NABOne datasets proved to be rather uniform. The clustered outliers strengthen the results from the initial clustering; there are obvious similarities in trends between the results for both, with regions like Skútustaðahreppur (Figure 18.A) where indicators of domestic animals are frequent, whereas Árneshreppur (Figure.18.B) contains a higher number indicators for species related to water.

In order to create maps that enable effective comparison between datasets, the formatting and style of the maps should be identical. Although training and mapping NABOne data by point rather than area make for interesting visuals, it allows for a less straightforward comparison between this dataset and the others. It is impossible to train the Sagamap data in a similar manner, and the SEAD dataset contains few obvious outliers. The technique might be suitable for presenting smaller datasets or data which contains large amounts of information but poor geographic spread.

7.3. SOM of individual concepts

For anyone interested in investigating the correlations between spread of specific concept categories between the datasets, I propose an individual SOM clustering where each dataset represents a variable and the output maps shows patterns related to where each dataset indicates the chosen concept category.

There are only three categories which are represented in all 3 datasets: domestic animals, water, and wild animals. Thus, it is only possible to create individual cluster maps for these 3 concepts. The SOM training is performed as described for the combined map in section 6.2, with similar parameters and varying grid sizes and number of clusters depending on the size and variation for each dataset.

7.3.1. Domestic animals

Indicators of domestic animals are present in 65 out of 119 municipalities in Iceland. An applied SOM grid of 4x4 an cluster specification of 6 rendered the result presented in Figure 19. Sagamap clearly shows the highest spatial spread in indicators of domestic animals (green segment in the segments plot) whereas SEAD and NABOne are overrepresenting one or a few regions each.

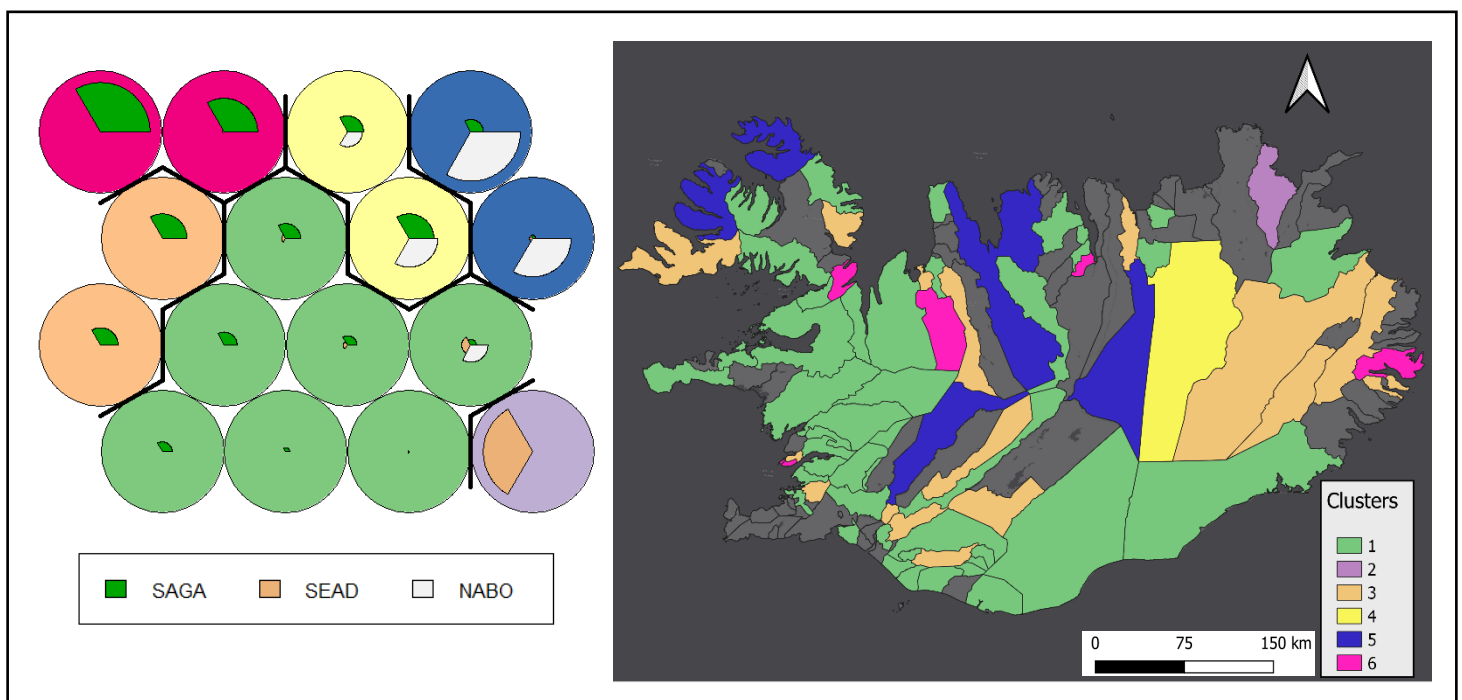


Figure 19: Result of SOM clustering of the “domestic animals” concept category, where the values for each specific dataset is included as variable

There is little evidence of a uniform or correlated spread of indicators related to domestic animals between the 3 datasets, although there are areas where more than one dataset strongly suggests the presence of domestic animals. In Skútustaðahreppur (cluster 3) indicators of domestic animals can be found within both Sagamap and NABOne data.

7.3.2. Wild Animals

Although indicators for wild animals can be found within all 3 datasets, these indicators are only present within 28 municipalities in Iceland. A smaller grid of 4x4 was applied, and the training successfully identified 5 specific clusters (Figure 20).

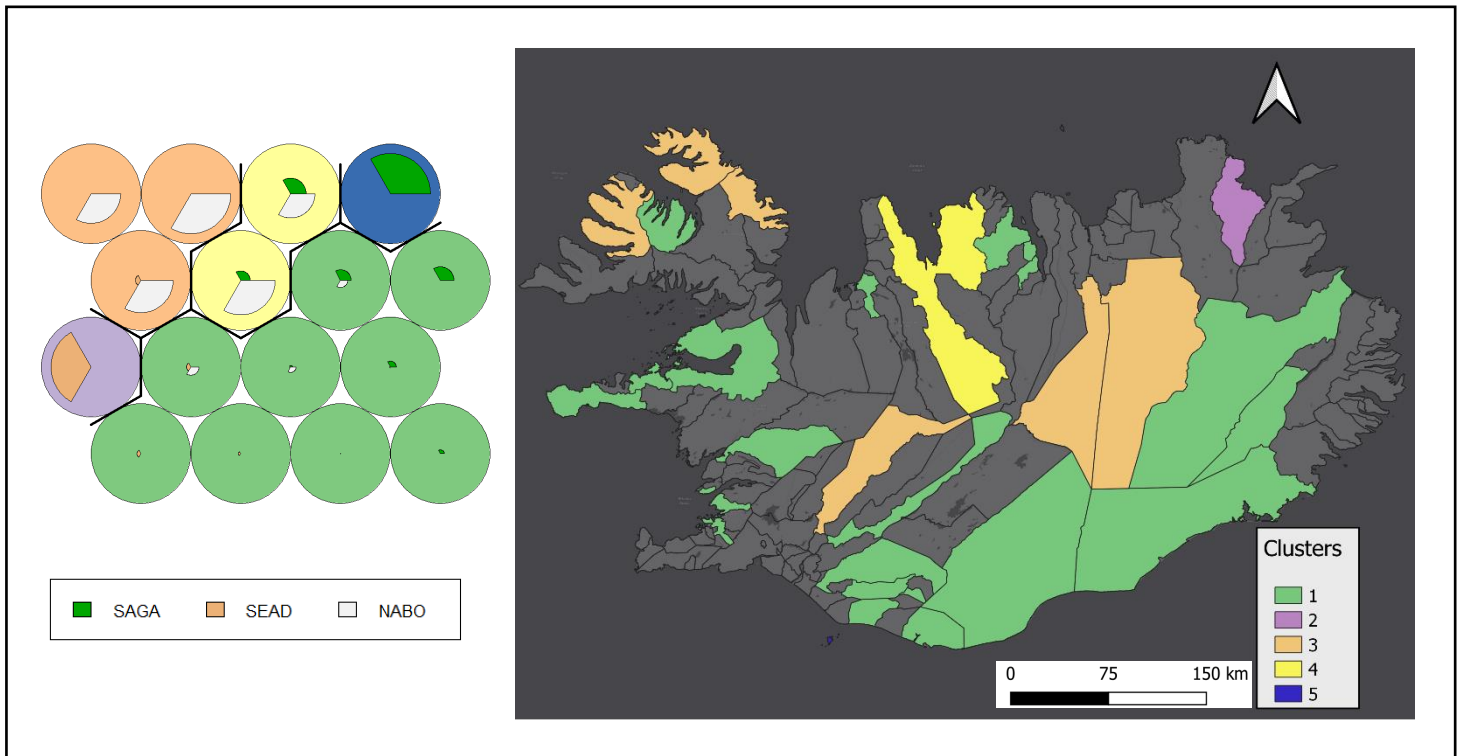


Figure 20: Result of SOM clustering of the “wild animals” concept category. Note that Vestmannaeyjar, a small archipelago outside of the southern coast of Iceland, is the only municipality belonging to cluster 5 (dark blue).

Cluster 4, which describes the conditions in Skagafjörður, indicates that within this region both Sagamap data and NABOne data mention or indicate the presence of wild animals. This trend is confirmed within the cluster categorisation and mapping of the three datasets combined (figure 12) where Skagafjörður belongs to cluster 7 where one can see good evidence for wild animals.

Indicators of wild animals do also seem to be present in areas clustered as 2, 3 and 5, although in these areas we generally see only one dataset indicating this. Within cluster one there are very few indicators of wild animals from either of the three datasets.

7.3.3. Water

Indicators of water or animals, objects or activities related to water are evident throughout most of Iceland; 76 municipalities contain data points or information indicating or related to water. Similar to the SOM training

for the domestic animals concept, a 5x5 grid was again applied and 6 unique clusters were identified. The results of the clustering are presented in Figure 21.

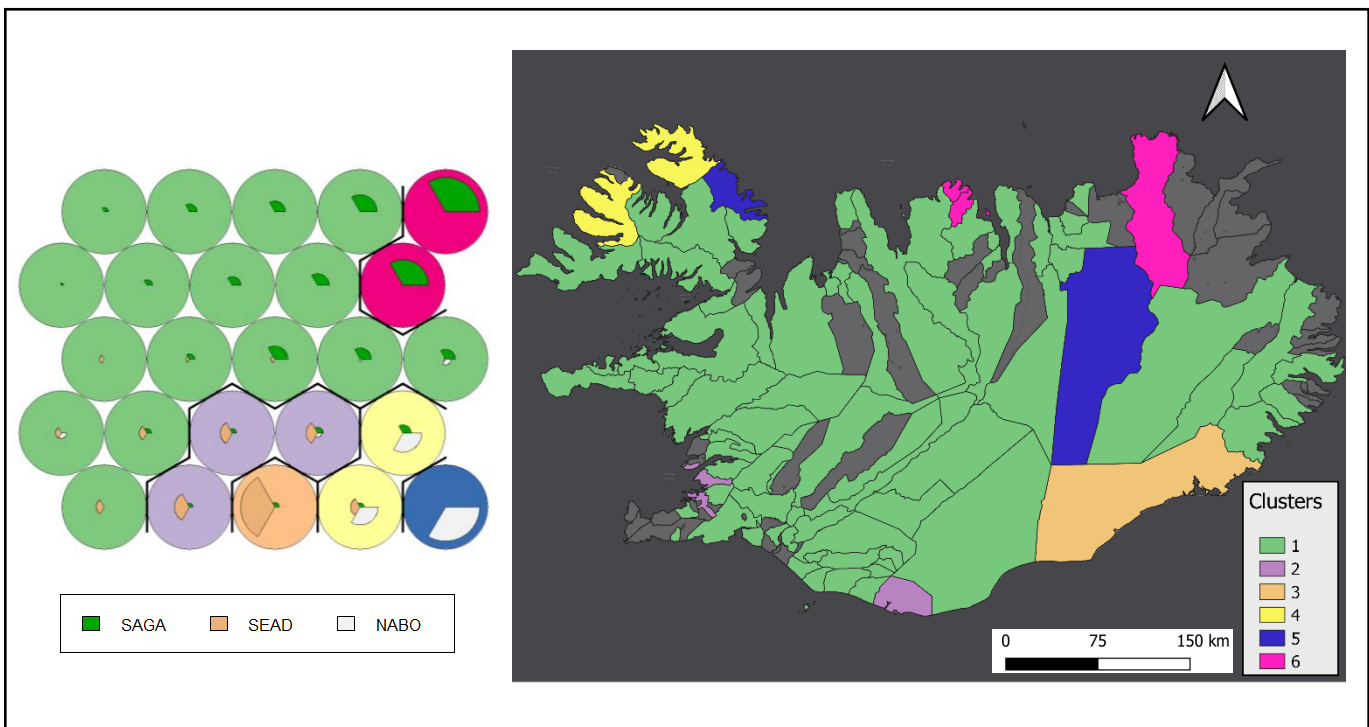


Figure 21: Result of SOM clustering of the “wild animals” concept category

Similarly to the plot for domestic animals, the clustered result of a trained SOM for the water concept category is strongly affected by the differences in spread of the datasets and whether data points from SEAD or NABO are present within an area or not, rather than showing any real patterns or spatial connections between the datasets. Clusters 2 and 4 do however represent areas where more than one dataset or discipline shows evidence of or indicate water.

Cluster plots for each specific concept category could potentially be of great help in identifying regions where more than one dataset or discipline strongly suggest the presence of or evidence for the chosen concept. However, due to the current inclusion of only 3 datasets in this analysis, 2 of which showing quite limited geographic spread, these concept cluster visualisations act more as a reflection of the regions in which data is available for each dataset, as oppose to where concept indicators are abundant. They are thus not included in the final presentation of major findings for this study, however it would be sensible to include SOMs by concept category into a finalised model, or present a user of the model with the option to visualise this type of clustering by concept, once the full range of data has been included. Hopefully then, more than three concept categories will be represented in three or more datasets or by data from several disciplines.

8. Answering the research questions

The research aims posed in Rønning (2020) will here be further expanded upon, which sets the scene for considering options for future work and suggest improvements. Throughout this project the research aims were the following:

Build a model that links datasets from multiple disciplines (archaeological, palaeoenvironmental and textual) by identifying a series of clusters where the values for several or all of the included datasets are similar

Through the application of an unsupervised data training and clustering method using self-organising maps, I successfully managed to compile multiple cross-disciplinary datasets within one single model. The datasets were recategorized using 10 established concept categories that managed to preserve the integrity of the included data while at the same time provide a common scale on which the datasets could be recategorized and combined. The model can be expanded to include more sets of data, which will further increase the validity of the output results.

Effectively map relationships and connections linking cross-disciplinary data contributed into the dataARC project by creating a visualised output of patterns within and between individual datasets

Visualising the final output of the clustered SOM results, both for individual datasets, individual concepts and of all datasets and concepts combined, was done through choropleth mapping of individual municipalities in Iceland. Through combining the clustered SOM output with a map with a similar colour scheme a reader or user is able to both get an idea of the composition of concept categories within a specific region, identify regions with similar concept compositions, and hopefully start to develop an idea or theory of why these patterns emerge.

9. Suggestions for further work

The develop visualisation model is still a prototype with room for further development and improvement. The inclusion of the full range of existing datasets, as well as a streamlined methodology for including more data in the future, is of the highest priority. In order to be implemented into the dataARC platform as a dynamic and useful tool, the model will have to be developed into a web-based interactive map, where users should be able to implement and combine whichever datasets they so desire. In combination with the already existing dataARC concept map, this visualised cluster map shows great potential as a highly useful tool within cross-disciplinary research on a spatial level, and will hopefully contribute to the important work that is being done on investigating human-environment co-evolution in the North Atlantic in the middle ages.

10. References

- Aagaard-Hansen, J., 2007. The challenges of cross-disciplinary research. *Social epistemology*, 21(4), pp.425-438.
- Akinduko, A.A., Mirkes, E.M. and Gorban, A.N., 2016. SOM: Stochastic initialization versus principal components. *Information Sciences*, 364, pp.213-221.
- Andrienko, G., Andrienko, N., Bak, P., Bremm, S., Keim, D., von Landesberger, T., Pölit, C. and Schreck, T., 2010. A framework for using self-organising maps to analyse spatio-temporal patterns, exemplified by analysis of mobile phone usage. *Journal of Location based services*, 4(3-4), pp.200-221.
- Ben-Gal, I., 2005. Outlier detection. In *Data mining and knowledge discovery handbook* (pp. 131-146). Springer, Boston, MA.
- Buckland, P., 2007. The development and implementation of software for palaeoenvironmental and palaeoclimatological research: the Bugs Coleopteran Ecology Package (BugsCEP) (Doctoral dissertation, Arkeologi och samiska studier).
- Buckland, P.I. and Buckland, P.C., 2006. BugsCEP: Coleopteran Ecology Package (software).
- Buckland, P.I., 2010. The Strategic Environmental Archaeology Database (SEAD): An International Research Cyber-Infrastructure for Studying Past Changes in Climate, Environment and Human Activities. *Journal of Northern Studies*, (1), pp.120-126.
- Buckland, P.I., 2014. The Bugs coleopteran ecology package (BugsCEP) database: 1000 sites and half a million fossils later. *Quaternary international*, 341, pp.272-282.
- Buckland, P.I., Eriksson, E., Linderholm, J., Viklund, K., Engelmark, R., Palm, F., Svensson, P., Buckland, P., Panagiotakopulu, E. and Olofsson, J., 2011. Integrating human dimensions of Arctic palaeoenvironmental science: SEAD—the strategic environmental archaeology database. *Journal of Archaeological Science*, 38(2), pp.345-351.
- Buckland, P.I., Sjölander, M. and Eriksson, E.J., 2018. Strategic environmental archaeology database (SEAD).
- Chicco, G., Napoli, R. and Piglion, F., 2003, June. Application of clustering algorithms and self organising maps to classify electricity customers. In *2003 IEEE Bologna Power Tech Conference Proceedings*, (Vol. 1, pp. 7-pp). IEEE.
- Cracknell, M.J. and Cowood, A.L., 2016. Construction and analysis of hydrogeological landscape units using self-organising maps. *Soil Research*, 54(3), pp.328-345.
- Dhillon, I.S. and Modha, D.S., 2001. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1-2), pp.143-175.

- Haldon, J., Mordechai, L., Newfield, T.P., Chase, A.F., Izdebski, A., Guzowski, P., Labuhn, I. and Roberts, N., 2018. History meets palaeoscience: Consilience and collaboration in studying past societal responses to environmental change. *Proceedings of the National Academy of Sciences*, 115(13), pp.3210-3218
- Hartigan, J.A. and Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1), pp.100-108.
- Hsu, A.L. and Halgamuge, S.K., 2003. Enhancement of topology preservation and hierarchical dynamic self-organising maps for data visualisation. *International Journal of Approximate Reasoning*, 32(2-3), pp.259-279.
- Kaminka, G.A., 2016, August. Repetitive branch-and-bound using constraint programming for constrained minimum sum-of-squares clustering. In *ECAI 2016: 22nd European Conference on Artificial Intelligence*, 29 August-2 September 2016, The Hague, The Netherlands-Including Prestigious Applications of Artificial Intelligence (PAIS 2016) (Vol. 285, p. 462). IOS Press.
- Kanevski, M., Pozdnoukhov, A., Pozdnukhov, A. and Timonin, V., 2009. *Machine learning for spatial environmental data: theory, applications, and software*. EPFL press.
- Kohonen, T., 1989. Self-organizing feature maps. In *Self-organization and associative memory* (pp. 119-157). Springer, Berlin, Heidelberg.
- Kohonen, T., 1997, June. Exploration of very large databases by self-organizing maps. In *Proceedings of international conference on neural networks (icnn'97)* (Vol. 1, pp. PL1-PL6). IEEE.
- Kohonen, T., 2001. *Self-Organizing Maps*. New York: Springer Series in Information Sciences.
- Laaksonen, J., Koskela, M., Laakso, S. and Oja, E., 2001. Self-organising maps as a relevance feedback technique in content-based image retrieval. *Pattern Analysis & Applications*, 4(2-3), pp.140-152.
- Lethbridge, E. 2010. 'Brief notes and sources of information', *The Saga-Steads of Iceland: A 21st-Century Pilgrimage*. Available at: <http://sagasteads.blogspot.com/p/what-are-medieval-icelandic-sagas.html> (Accessed: 16 May 2020)
- Lethbridge, E. 2020. Digital Mapping and the Narrative Stratigraphy of Iceland. In *Historical Geography, GIScience and Textual Analysis* (pp. 19-32). Springer, Cham.
- Lethbridge, E., 2016. The Icelandic sagas and saga landscapes. *Gripla*, 27, pp.51-92.
- Liu, F.T., Ting, K.M. and Zhou, Z.H., 2008, December. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 413-422). IEEE.
- Malone, J., McGarry, K., Wermter, S. and Bowerman, C., 2006. Data mining using rule extraction from Kohonen self-organising maps. *Neural Computing & Applications*, 15(1), pp.9-17.

- Mann, M.E., Zhang, Z., Rutherford, S., Bradley, R.S., Hughes, M.K., Shindell, D., Ammann, C., Faluvegi, G. and Ni, F., 2009. Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly. *Science*, 326(5957), pp.1256-1260.
- Mayer, R., Aziz, T.A. and Rauber, A., 2007, September. Visualising class distribution on self-organising maps. In *International Conference on Artificial Neural Networks* (pp. 359-368). Springer, Berlin, Heidelberg.
- McGovern, T.H., 2014. North Atlantic human ecodynamics research. *Human Ecodynamics in the North Atlantic: a Collaborative Model of Humans and Nature through Space and Time*. Lexington Books, Lanham, Maryland, pp.213-221.
- McGovern, T.H., Hambrecht, G., Brewington, S., Feeley, F., Harrison, R., Hicks, M., Smiarowski, K. and Woollett, J., 2017. Too many bones: data management and the NABONE experience. *The Wide Lens in Archaeology: Honoring Brian Hesse's Contributions to Anthropological Archaeology*, Lockwood Press, London, pp.29-42.
- McGovern, T.H., Perdikaris, S., Einarsson, A. and Sidell, J., 2006. Coastal connections, local fishing, and sustainable egg harvesting: patterns of Viking Age inland wild resource use in Mývatn district, Northern Iceland. *Environmental Archaeology*, 11(2), pp.187-205.
- Moehrmann, J., Burkovski, A., Baranovskiy, E., Heinze, G.A., Rapoport, A. and Heidemann, G., 2011, June. A discussion on visual interactive data exploration using self-organizing maps. In *International Workshop on Self-Organizing Maps* (pp. 178-187). Springer, Berlin, Heidelberg.
- Pettersson, J., 2008. Translators and Narrators. *The Translation of Subjectivity in Old Norse Literature*. Inbjuden talare pa konferensen Riddarasogur and the translation of court culture in 13th century Scandinavia, Universitetet i Oslo, Norge, pp.17-18.
- Pöllä, M., Honkela, T., Bruun, H. and Russell, A., 2006, October. Analysis of interdisciplinary text corpora. In *Proceedings of the 12th Finnish Artificial Intelligence Conference STeP* (pp. 26-27).
- Pollard, D., 1981. Strong consistency of k-means clustering. *The Annals of Statistics*, pp.135-140.
- Pözlbauer, G., Dittenbach, M. and Rauber, A., 2005, July. A visualization technique for self-organizing maps with vector fields to obtain the cluster structure at desired levels of detail. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005*. (Vol. 3, pp. 1558-1563). IEEE.
- Príncipe, J.C. and Miikkulainen, R., 2009. *Advances in Self-organising Maps*. Berlin, Germany: Springer.
- Rønning, K. (2020) *Archaeology, Environment and Human History: Examining the Spatial Links Between Human Settlements and Environmental Change in Iceland*. MSc Dissertation [Research Paper]. Edinburgh: University of Edinburgh
- Ross, M.C., 1997. The Intellectual Complexion of the Icelandic Middle Ages: Toward a New Profile of Old Icelandic Saga Literature. *Scandinavian Studies*, 69(4), pp.443-453.

- Schiffer, M.B., 1972. Archaeological context and systemic context. *American antiquity*, pp.156-165.
- Skupin, A., 2004. A picture from a thousand words [information visualization]. *Computing in Science & Engineering*, 6(5), pp.84-88.
- Strawhacker, C., Buckland, P., Palsson, G., Fridrikkson, A., Lethbridge, E., Brin, A., Opitz, R. and Dawson, T., 2015. Building cyberinfrastructure from the ground up for the North Atlantic Biocultural Organization introducing the cyberNABO Project. In *2015 Digital Heritage (Vol. 2, pp. 457-460)*. IEEE.
- Tasdemir, K. and Merényi, E., 2012. SOM-based topology visualisation for interactive analysis of high-dimensional large datasets. *Machine Learning Reports*, 1, pp.13-15.
- Wehrens, R. and Buydens, L.M., 2007. Self-and super-organizing maps in R: the Kohonen package. *Journal of Statistical Software*, 21(5), pp.1-19.
- Wehrens, R. and Wehrens, M.R., 2019. Package ‘kohonen’.
- Whelan, C.T., Lucchini, M., Pisati, M. and Maître, B., 2010. Understanding the socio-economic distribution of multiple deprivation: An application of self-organising maps. *Research in Social Stratification and Mobility*, 28(3), pp.325-342.
- Yang, J., Ward, M.O. and Rundensteiner, E.A., 2002. Visual hierarchical dimension reduction for exploration of high dimensional datasets.
- Yin, H., 2008. On multidimensional scaling and the embedding of self-organising maps. *Neural Networks*, 21(2-3), pp.160-169

Appendix

sagas_concepts.py

```
# -*- coding: utf-8 -*-
"""
Created on Tue May 12 15:47:27 2020

@author: kajar

This script takes the sagas datapoints from Iceland and separates them by the 10
established concept categories.
each concept category is given its own dictionary. we then have to save them all as
geojson files
so that we can export them to qgis and create counts for each based on area.
"""

!pip install geojson
import urllib
import geojson
import json

# upload the sagas.geojson file that only consist of points in Iceland
with open("sagas_ice.geojson", encoding='utf-8') as f:
    data_i = json.load(f)

# create list of all the different concepts
all_concepts = []
for feature in data_i['features']:
    all_concepts.append(feature['properties']['concept'])
# return list of concepts with no duplicates.
# we will use this to create the concept categories
concepts = list(dict.fromkeys(all_concepts))

"""
Now we are going to split the original .geojson file into 10 new json file,
one for each of the established concept categories
"""

# Create lists for the 10 'concept categories'

activities = []
buildings = []
managed = []
domestic = []
nature = []
wild = []
water = []
travel = []
weather = []
things = []
trash = []

# next we will append all points to their designated concept categories based on concept
for a in range(0, len(concepts)):
    if concepts[a][0:11] == 'Animals: ac' or concepts[a][0:18] == 'Actors: animals: d' or
concepts[a][0:9] == 'Actors: a' or concepts[a][0:8] == 'Actor: a' or concepts[a][0:18] ==
'Actors: animals: h' or concepts[a][0:28] == 'Actors: animals: mammals: bu' or
```

```

concepts[a][0:27] == 'Actors: animals: mammals: c' or concepts[a][0:27] == 'Actors:
animals: mammals: d' or concepts[a][0:27] == 'Actors: animals: mammals: g' or
concepts[a][0:27] == 'Actors: animals: mammals: h' or concepts[a][0:27] == 'Actors:
animals: mammals: o' or concepts[a][0:28] == 'Actors: animals: mammals: pi' or
concepts[a][0:28] == 'Actors: animals: mammals: sh' or concepts[a][0:22] == 'Actors:
animals: mamml' or concepts[a][0:17] == 'Actors: animals:m' or concepts[a][0:10] ==
'Actors: ta':
    domestic.append(concepts[a]) #append to list "domestic"

    elif concepts[a][-4:] == 'snow' or concepts[a][-4:] == 'calm' or concepts[a][-4:] ==
'rain' or concepts[a][-4:] == 'wind' or concepts[a][-3:] == 'fog' or concepts[a][-5:] ==
'frost' or concepts[a][-5:] == 'calm ' or concepts[a][-5:] == 'rain ':
        weather.append(concepts[a]) #append to list "weather"

    elif concepts[a][-6:] == 'bridge' or concepts[a][-7:] == 'bridge ' or concepts[a][-
6:] == 'travel' or concepts[a][-7:] == 'travel ' or concepts[a][-7:] == 'travels' or
concepts[a] == 'Actors: animals: ' or concepts[a][-4:] == 'airn':
        travel.append(concepts[a]) #append to list "travel"

    elif concepts[a][0:18] == 'Actors: animals: a' or concepts[a][-4:] == 'bear' or
concepts[a][-3:] == 'fox' or concepts[a][-4:] == 'wolf' or concepts[a][-6:] == 'wolves':
        wild.append(concepts[a]) #append to list "wild"

    elif concepts[a][-5:] == 'spear' or concepts[a][-5:] == ' wood' or concepts[a][-5:]
== 'money' or concepts[a][-7:] == 'jewelry' or concepts[a][-4:] == 'bone' or
concepts[a][-4:] == 'food' or concepts[a][-4:] == 'horn' or concepts[a][-6:] == 'spears'
or concepts[a][-8:] == '(timber)' or concepts[a][-8:] == 'jewelry ' or concepts[a][0:9]
== 'Actors: h':
        things.append(concepts[a]) #append to list "things"

    elif concepts[a][-4:] == 'boat' or concepts[a][-5:] == 'boats' or concepts[a][0:14]
== 'Activities: fi' or concepts[a][-4:] == 'ship' or concepts[a][0:18] == 'Actors:
animals: f' or concepts[a][-6:] == 't shed' or concepts[a][-4:] == 'boag' or
concepts[a][-5:] == 'beach' or concepts[a][-10:] == '/building ' or concepts[a][-8:] ==
'boatshed' or concepts[a][-4:] == 'seal' or concepts[a][-5:] == 'whale' or concepts[a][-
6:] == 'walrus' or concepts[a][-5:] == "river" or concepts[a][-6:] == "stream" or
concepts[a][-7:] == "streams" or concepts[a][-5:] == "water" or concepts[a][-6:] ==
"spring" or concepts[a][-4:] == "pool" or concepts[a][-5:] == "ocean" or concepts[a][-4:]
== "lake" or concepts[a][-5:] == "lakes" or concepts[a][-9:] == 'warehouse':
        water.append(concepts[a]) #append to list "water"

    elif concepts[a][0:17] == 'Actors: plants: g' or concepts[a][-9:] == 'al change' or
concepts[a][-4:] == 'land' or concepts[a][-9:] == 'brushwood' or concepts[a][-4:] ==
'tree' or concepts[a][-7:] == 'glacier' or concepts[a][-8:] == 'glacier ' or
concepts[a][-6:] == 'forest' or concepts[a][-9:] == 'stability' or concepts[a][-5:] ==
'scape' or concepts[a][-5:] == 'heath' or concepts[a][-4:] == 'lava' or concepts[a][-6:]
== 'heaths' or concepts[a][-5:] == 'woods' or concepts[a][-5:] == 'trees':
        nature.append(concepts[a]) #append to list "nature"

    elif concepts[a][0:13] == 'Activities: c' or concepts[a][-11:] == 'cultivation' or
concepts[a][-10:] == 'management' or concepts[a][-11:] == 'management ' or concepts[a][-
9:] == 'managment' or concepts[a][0:17] == 'Actors: plants: c' or concepts[a][-7:] ==
'herding' or concepts[a][-7:] == 'grazing' or concepts[a][-5:] == 'field' or
concepts[a][-6:] == 'field ' or concepts[a][-9:] == 'ed change' or concepts[a][-7:] ==
'pe area' or concepts[a][-7:] == 'pasture' or concepts[a][-5:] == '/wall' or
concepts[a][-4:] == 'yard':
        managed.append(concepts[a]) #append to list "managed"

```



```

    elif concepts[a][0:6] == 'Events' or concepts[a][0:5] == 'Ideas' or concepts[a][0:4]
== 'Imag' or concepts[a][0:9] == 'Actors: s' or concepts[a][-7:] == '(dream)' or
concepts[a] == 'Actors: ' or concepts[a][-8:] == 'butchery' or concepts[a][-8:] ==
'butchery' or concepts[a][-9:] == 'butchery' or concepts[a][0:5] == 'Actir' or
concepts[a][0:10] == 'Activitie:' or concepts[a][-7:] == 'milking' or concepts[a][0:13]
== 'Activities: s' or concepts[a][0:13] == 'Activities: e' or concepts[a][0:13] ==
'Activities: n' or concepts[a][0:13] == 'Activities: p' or concepts[a][0:10] ==
'Activitise':
        activities.append(concepts[a]) #append to list "activities"

    elif concepts[a][-3:] == 'pen' or concepts[a][-8:] == 'shieling' or concepts[a][-9:]
== 'shieling ' or concepts[a][-4:] == 'byre' or concepts[a][-9:] == 'buildings' or
concepts[a][-10:] == 'buildings ' or concepts[a][-4:] == 'barn' or concepts[a][-6:] == '-
house' or concepts[a][-5:] == 'fence' or concepts[a][-6:] == ': wall' or concepts[a][-6:]
== 'temple' or concepts[a][-6:] == 'church' or concepts[a][-5:] == 'booth' or
concepts[a][-4:] == 'hall' or concepts[a][0:6] == 'Commun' or concepts[a][-6:] ==
'/house' or concepts[a] == 'Physical Landscape: built environment: building ' or
concepts[a][-3:] == 'pit' or concepts[a][-4:] == 'fold' or concepts[a][-12:] == 'out
building' or concepts[a][-11:] == 'outbuilding' or concepts[a][-7:] == ': house' or
concepts[a][-9:] == 'out house' or concepts[a][-7:] == 'h house' or concepts[a][-10:] ==
'longhouse ' or concepts[a][-6:] == 'booths' or concepts[a][-6:] == ': shed' or
concepts[a][-7:] == 'w shed ' or concepts[a][-5:] == '-shed' or concepts[a][-6:] == 'p
shed' or concepts[a][-5:] == '/shed':
        buildings.append(concepts[a]) #append to list "buildings"

    else:
        trash.append(concepts[a]) #useful for picking up concepts which haven't been
appended into any categories. Ideally, this list should be empty

#Create dictionaries for each category and append the lists created above
activitiesd = {}
buildingsd = {}
managedd = {}
domesticd = {}
natured = {}
wildd = {}
waterd = {}
traveld = {}
weatherd = {}
thingsd = {}

def sagaconcepts(concd, conc):
    concd['features'] = [] #create feature list and type list in each dictionary
    concd['type'] = data_i['type']
    #append list to 'feature' list within the dictionary
    for a in range(0, len(data_i['features'])): #numbers each point, 3868
        if data_i['features'][a]['properties']['concept'] in conc: #if the concept for
point number a is similar to any point in "data-i":
            concd['features'].append(data_i['features'][a]) #append the information from
that point into the feature list for the dictionary for the fitting concept category
            print (len(concd['features']))

sagaconcepts(activitiesd, activities)
sagaconcepts(buildingsd, buildings)
sagaconcepts(managedd, managed)
sagaconcepts(domesticd, domestic)
sagaconcepts(natured, nature)
sagaconcepts(wildd, wild)
sagaconcepts(waterd, water)
sagaconcepts(traveld, travel)

```

```
sagaconcepts(weatherd, weather)
sagaconcepts(thingsd, things)
```

```
#save the dictionaries as json files that can be uploaded into qgis
with open('s_activities.json', 'w') as fp:
    json.dump(activitiesd, fp)
with open('s_buildings.json', 'w') as fp:
    json.dump(buildingsd, fp)
with open('s_managed.json', 'w') as fp:
    json.dump(managedd, fp)
with open('s_domestic.json', 'w') as fp:
    json.dump(domesticd, fp)
with open('s_nature.json', 'w') as fp:
    json.dump(natured, fp)
with open('s_wild.json', 'w') as fp:
    json.dump(wildd, fp)
with open('s_water.json', 'w') as fp:
    json.dump(waterd, fp)
with open('s_travel.json', 'w') as fp:
    json.dump(traveld, fp)
with open('s_weather.json', 'w') as fp:
    json.dump(weatherd, fp)
with open('s_things.json', 'w') as fp:
    json.dump(thingsd, fp)
```

SEAD_indicators.py

```

# -*- coding: utf-8 -*-
"""
Created on Tue May 12 15:47:27 2020

@author: kajar

This script takes the SEAD datapoints from Iceland and separates the indicator values
within each data point
"""
!pip install geojson
import urllib
import geojson
import json

#open SEAD file, filtered in qgis to only contain points located within Iceland
with open("sead_ip.geojson", encoding='utf-8') as f:
    sead_p = json.load(f)

'''
EXTRACTING INDICATOR VALUES

extracting values from the 'indicators'dictionary, making a list for each and placing
them in the 'properties' dictionary,
makes it so that we can add them together in qGIS through point in polygon analysis
'''
aquatics = []
swater = [] #stagnant water
rwater = [] #running water
pasturedung = []
meadow = []
wood = []
deciduous = [] #evergreen
coniferous = [] #drop leaves seasonally
wetland = [] #marshes or swamps, saturated land
openwet = []
arable = [] #land suitable for growing crops
dryarable = []
foul = [] #assume filthy or dirty
carrion = [] #decaying flesh or dead animals
dung = []
mould = []
synantropic = [] #in relation to humans
storedgrain = []
deadwood = []
heathland = [] #myr og lynghei
halotolerant = [] #tolerate conditions of high salinity, inland salt seas or springs
ectoparasite = [] #parasite that lives on the outside of its host

#write a function that append indicator values to the lists above and places the lists
within the 'properties' dictionary
def allindicators(ind, ind2, ogind):
    for feature in sead_p['features']:
        ind.append(feature['properties'][ogind])
    for g in range(0, len(sead_p['features'])):
        sead_p['features'][g]['properties'][ind2] = ()
        sead_p['features'][g]['properties'][ind2] = ind[g]

allindicators(aquatics, 'aquatics', 'Aquatics')
allindicators(swater, 'swater', 'Indicators: Standing water')

```

```

allindicators(rwater, 'rwater', 'Indicators: Running water')
allindicators(pasturedung, 'pasturedung', 'Pasture/Dung')
allindicators(meadow, 'meadow', 'Meadowland')
allindicators(wood, 'wood', 'Wood and trees')
allindicators(deciduous, 'deciduous', 'Indicators: Deciduous')
allindicators(coniferous, 'coniferous', 'Indicators: Coniferous')
allindicators(wetland, 'wetland', 'Wetlands/marshes')
allindicators(openwet, 'openwet', 'Open wet habitats')
allindicators(arable, 'arable', 'Disturbed/arable')
allindicators(dryarable, 'dryarable', 'Sandy/dry disturbed/arable')
allindicators(foul, 'foul', 'Dung/foul habitats')
allindicators(carrion, 'carrion', 'Carrion')
allindicators(dung, 'dung', 'Indicators: Dung')
allindicators(mould, 'mould', 'Mould beetles')
allindicators(synanthropic, 'synanthropic', 'General synanthropic')
allindicators(storedgrain, 'storedgrain', 'Stored grain pest')
allindicators(deadwood, 'deadwood', 'Dry dead wood')
allindicators(heathland, 'heathland', 'Heathland & moorland')
allindicators(halotolerant, 'halotolerant', 'Halotolerant')
allindicators(ectoparasite, 'ectoparasite', 'Ectoparasite')

```

```

#save sead_i as new json file
with open('SEAD_indp.json', 'w') as fp:
    json.dump(sead_p, fp)

```

```

#make a list of the point ids, need these for the spreadsheet
ids = []
for b in sead_p['features']:
    ids.append(b['properties']['id'])

```

```

#Make an excel file and append point id and indicator columns
import xlswriter
workbook = xlswriter.Workbook('seadindp.xlsx')
worksheet1 = workbook.add_worksheet()

```

```

worksheet1.write_column('A2', ids)
worksheet1.write_column('B2', aquatics)
worksheet1.write_column('C2', swater)
worksheet1.write_column('D2', rwater)
worksheet1.write_column('E2', pasturedung)
worksheet1.write_column('F2', meadow)
worksheet1.write_column('G2', wood)
worksheet1.write_column('H2', deciduous)
worksheet1.write_column('I2', coniferous)
worksheet1.write_column('J2', wetland)
worksheet1.write_column('K2', openwet)
worksheet1.write_column('L2', arable)
worksheet1.write_column('M2', dryarable)
worksheet1.write_column('N2', foul)
worksheet1.write_column('O2', carrion)
worksheet1.write_column('P2', dung)
worksheet1.write_column('Q2', mould)
worksheet1.write_column('R2', synanthropic)
worksheet1.write_column('S2', storedgrain)
worksheet1.write_column('T2', deadwood)
worksheet1.write_column('U2', heathland)
worksheet1.write_column('V2', halotolerant)
worksheet1.write_column('W2', ectoparasite)
workbook.close()

```

Nabonosead.py

```
# -*- coding: utf-8 -*-
"""
Created on Wed Jun  3 15:05:05 2020

@author: kajar

This script takes the NABONOSEAD datapoints from Iceland and separates the indicator
values within each data point
"""

!pip install geojson
import urllib
import geojson
import json

# upload nabonosead.json file that only consist of points in Iceland
with open("nabonosead_data.json", encoding='utf-8-sig') as f:
    nsead = json.load(f)

'''
Tests to make sure dataset is fine and only contains points that we want
'''

#make sure all points are in Iceland
for p in range(0, len(nsead['features'])):
    print (nsead['features'][p]['properties']['country'])

'''
EXTRACTING INDICATOR VALUES
'''

#make lists of values for each indicator that contains any values
domestic = []
wild = []
marinem = []
marinef = []
freshf = []

#append indicator values tp lists
def indlists(ind, ind2): #ind is the name of the list we append values to, ind2 is the
name of the indicator list within the nsead dataset
    for feature in nsead['features']:
        ind.append(feature['properties']['indicators'][ind2])

indlists(domestic, 'domestic')
indlists(wild, 'wild')
indlists(marinem, 'Marine Mammal')
indlists(marinef, 'Marine Fish')
indlists(freshf, 'Freshwater Fish')

#write this data into excel columns, which we can import into qGIS and join with the
original dataset

#fist have to make a list of point ids
pointids = []
for feature in nsead['features']:
    pointids.append(feature['id'])

#then make the excel file and append the columns
import xlswriter
```

```
workbook = xlsxwriter.Workbook('naboind.xlsx')
worksheet1 = workbook.add_worksheet()

worksheet1.write_column('A2', pointids)
worksheet1.write_column('B2', domestic)
worksheet1.write_column('C2', wild)
worksheet1.write_column('D2', marinem)
worksheet1.write_column('E2', marinef)
worksheet1.write_column('F2', freshf)
workbook.close()
#in excel gace each column a name (same as list name) and saved it as csv file
```

Nabonosead_outliers.py

```
# -*- coding: utf-8 -*-
"""
Created on Wed Jun  3 15:05:05 2020

@author: kajar

This script identifies multivariate outliers within the NABONOSEAD datasets through the
application of a Isolation Forest methodology
"""

!pip install geojson
import urllib
import geojson
import json
import copy
from geojson import Point, Feature, FeatureCollection, dump
import statistics

# upload nabonosead.json file that only consist of points in Iceland
with open("nabonosead_data.json", encoding='utf-8-sig') as f:
    nsead = json.load(f)

'''
Identify multivariate outliers
Code snippets from https://www.analyticsvidhya.com/blog/2019/02/outlier-detection-python-pyod/
we will attempt 3 different methods for muotidimensional outlier idenification
'''

#we need lists of value for each indicator
domestic = []
wild = []
marinem = []
marinef = []
freshf = []

def indlists(ind, ind2):    #ind is the name of the list we append values to, ind2 is the
    #name of the indicator list within the nsead dataset
    for feature in nsead['features']:
        ind.append(feature['properties']['indicators'][ind2])

indlists(domestic, 'domestic')
indlists(wild, 'wild')
indlists(marinem, 'Marine Mammal')
indlists(marinef, 'Marine Fish')
indlists(freshf, 'Freshwater Fish')

#write this data into excel columns

#fist have to make a list of point ids
pointids = []
for feature in nsead['features']:
    pointids.append(feature['id'])

#then make the excel file and append the columns
import xlswriter
workbook = xlswriter.Workbook('naboind.xlsx')
worksheet1 = workbook.add_worksheet()
```

```

worksheet1.write_column('A2', pointids)
worksheet1.write_column('B2', domestic)
worksheet1.write_column('C2', wild)
worksheet1.write_column('D2', marinem)
worksheet1.write_column('E2', marinef)
worksheet1.write_column('F2', freshf)
workbook.close()
#in excel gace each column a name (same as list name) and saved it as csv file

#import necessary packages and models
import pandas as pd
import numpy as np
from scipy import stats
!pip install pyod
!pip install --upgrade pyod # to make sure that the latest version is installed!
!pip install --upgrade pip
# Import models
from pyod.models.abod import ABOD
from pyod.models.hbos import HBOS
from pyod.models.iforest import IForest
#read the nsead file into a pandas dataframe constructor
df = pd.read_csv('naboind.csv', sep = ";")
df.plot.scatter("domestic", "freshf") #just to test that it's been read in correclty

# create a meshgrid, don't really know what this does....
xx , yy = np.meshgrid(np.linspace(-1000, 1000, 20000), np.linspace(-1000, 1000, 20000))

# scatter plot
plt.scatter(domestic,wild)
plt.xlabel('domestic')
plt.ylabel('wild')

# Define 3 outlier detection tools to be compared
random_state = np.random.RandomState(42)
outliers_fraction = 0.05 #percentage of observations you want to detect that are not
similar to the rest of the data

classifiers = {
    #'Angle-based Outlier Detector (ABOD)': ABOD(contamination=outliers_fraction),
    #'Histogram-base Outlier Detection (HBOS)':
    HBOS(contamination=outliers_fraction),
    'Isolation Forest':
    IForest(contamination=outliers_fraction,random_state=random_state)
}

#scale all values down to a range between 0 and 1
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler(feature_range=(0, 1))
df[['domestic', 'wild']] = scaler.fit_transform(df[['domestic', 'wild']])
df[['domestic', 'wild']].head()

#store these values in numpy arrays (?)
X1 = df['domestic'].values.reshape(-1,1)
X2 = df['wild'].values.reshape(-1,1)
#X3 = df['marinem'].values.reshape(-1,1)
#X4 = df['marinef'].values.reshape(-1,1)
#X5 = df['freshf'].values.reshape(-1,1)
X = np.concatenate((X1,X2),axis=1)

```



```

xx , yy = np.meshgrid(np.linspace(0,1 , 200), np.linspace(0, 1, 200))

for i, (clf_name, clf) in enumerate(classifiers.items()):
    clf.fit(X)
    # predict raw anomaly score
    scores_pred = clf.decision_function(X) * -1

    # prediction of a datapoint category outlier or inlier
    y_pred = clf.predict(X)
    n_inliers = len(y_pred) - np.count_nonzero(y_pred)
    n_outliers = np.count_nonzero(y_pred == 1)
    plt.figure(figsize=(10, 10))

    # copy of dataframe
    dfx = df
    dfx['outlier'] = y_pred.tolist()

    # IX1 - inlier feature 1, IX2 - inlier feature 2
    IX1 = np.array(dfx['domestic'][dfx['outlier'] == 0]).reshape(-1,1)
    IX2 = np.array(dfx['wild'][dfx['outlier'] == 0]).reshape(-1,1)

    # OX1 - outlier feature 1, OX2 - outlier feature 2
    OX1 = dfx['domestic'][dfx['outlier'] == 1].values.reshape(-1,1)
    OX2 = dfx['wild'][dfx['outlier'] == 1].values.reshape(-1,1)

    print('OUTLIERS : ',n_outliers,'INLIERS : ',n_inliers, clf_name)

    # threshold value to consider a datapoint inlier or outlier
    threshold = stats.scoreatpercentile(scores_pred,100 * outliers_fraction)

    # decision function calculates the raw anomaly score for every point
    Z = clf.decision_function(np.c_[xx.ravel(), yy.ravel()]) * -1
    Z = Z.reshape(xx.shape)

    # fill blue map colormap from minimum anomaly score to threshold value
    plt.contourf(xx, yy, Z, levels=np.linspace(Z.min(), threshold,
7), cmap=plt.cm.Blues_r)

    # draw red contour line where anomaly score is equal to threshold
    a = plt.contour(xx, yy, Z, levels=[threshold],linewidths=2, colors='red')

    # fill orange contour lines where range of anomaly score is from threshold to maximum
    anomaly score
    plt.contourf(xx, yy, Z, levels=[threshold, Z.max()],colors='orange')

    b = plt.scatter(IX1,IX2, c='white',s=20, edgecolor='k')

    c = plt.scatter(OX1,OX2, c='black',s=20, edgecolor='k')

    plt.axis('tight')

    # loc=2 is used for the top left corner
    plt.legend(
        [a.collections[0], b,c],
        ['learned decision function', 'inliers','outliers'],
        prop=matplotlib.font_manager.FontProperties(size=20),
        loc=2)

    plt.xlim((0, 1))
    plt.ylim((0, 1))
    plt.title(clf_name)
    plt.savefig('testthey.png')
    plt.show()

#all outlier points have been given a value of 1 in the df "file", 47 in total using
isolation forest. The excel file can be uploaded into qGIS and joined with the original
NABONOSEAD dataset.

```

Saga_som.R

Script that trains and clusters the Sagamap dataset

```
#set working directory
setwd("~/GIS_pg/dissertation/data")

#install packages and libraries.
install.packages('kohonen')
install.packages('ggplot2')
install.packages('rgdal')
install.packages('gridExtra')
install.packages('grid')
install.packages('viridis')
install.packages('dplyr')
install.packages('maptools')
install.packages('gpclib')
install.packages('devtools')
install.packages('readxl')
library(kohonen)
library(ggplot2)
library(rgdal)
library(gridExtra)
library(grid)
library(viridis)
library(dplyr)
library(maptools)
library(gpclib)
library(readxl)

#read in the data. This data has been categorised into 10 concept categories
#any polygons containing zero Saga data points have been removed
saga_data <- read_excel("sagas_pp.xlsx", sheet = "nozero")

#read in boundary data for Iceland, which has been matched up with the above data (by ID)
iceland_map <- readOGR("../isl/ISL_adm2.shp", stringsAsFactors =
  FALSE)

#convert map into latitude and longitude, easier to implement into ggmap and plot it.
plot to check it's fine.
iceland_map <- spTransform(iceland_map, CRS("+proj=longlat +ellps=WGS84
+datum=WGS84 +no_defs"))
plot(iceland_map)

#combine your dataset with the map of Iceland
#first check you gpclib permit. if true, proceed
gpclibPermit()
#convert spatial polygon to dataframe including columns of spatial information
iceland_fort <- fortify(iceland_map, region= "ID_2")

#merge the new dataframe with the imported dataset using their shared column (id)
iceland_fort <- merge(iceland_fort, saga_data, by.x="id", by.y="ID_2")

#Test to see that this has worked by creating a plot of whatever you want
ggplot(data=iceland_fort, aes(x=long, y=lat, fill=activities, #this plots the spatial
spread of the concept categy "activities"
  group=group)) +
```

```

    scale_fill_viridis(name = "some rate")+
    geom_polygon(colour=NA)+
    theme_void() +
    coord_equal()

#SOM training

#Select variables used to train SOM by subsetting the 'data' dataframe
data_train <- select(saga_data, activities, buildings, managed, domestic, nature, wild,
water, travel, weather, things)

#standardise the data and convert to a matrix
# first we resacle varaibles and create a matrix
data_train_matrix <- as.matrix(scale(data_train))
#keep the column names of data_train as names in our new matrix
names(data_train_matrix) <- names(data_train)

#define the size and topology of the som grid
som_grid <- somgrid(xdim = 8, ydim = 8, topo="hexagonal")

# Train the SOM model
som_model <- som(data_train_matrix,
                grid=som_grid,
                rlen=500,
                alpha=c(0.05,0.1),
                keep.data = TRUE,)
# Plot SOM training progress - how the node distances have stabilised over time.
plot(som_model, type = "changes")

#load custom palette, make som visualisations more appealing
source('coolBlueHotRed.R')

#counts within nodes - how many counts/points exist within each node
plot(som_model, type = "counts", main="Node Counts",
     palette.name=coolBlueHotRed)

#map quality as node distance
plot(som_model, type = "quality", main="Node Quality/Distance",
     palette.name=coolBlueHotRed)

#neighbour distances
plot(som_model, type="dist.neighbours", main = "SOM neighbour distances",
     palette.name=grey.colors)

#code spread, this creates the codes plot that helps determine clusters
plot(som_model, type = "codes")

#Clustering of SOM results

# show the WCSS metric for kmeans for different clustering sizes.
# Use this to indicate the ideal number of clusters
mydata <- getCodes(som_model)
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(mydata, centers=i)$withinss)
#Plot WCSS
plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares", main="WCSS")

# Form clusters on grid

```

```

#define number of clusters
som_cluster <- cutree(hclust(dist(getCodes(som_model))), 8)

# Colour palette definition. Download the colour palette package "RColorBrewer" and
specifying palette name ("Accent")
#as well as number of cluster colours
library("RColorBrewer")
cbp <- brewer.pal(n = 8, name = "Accent")

#Plot som_model data, adding colour and cluster boundaries
plot(som_model, type="codes", bgcol = cbp[som_cluster], main =
      "Clusters")
add.cluster.boundaries(som_model, som_cluster)
bgcol = cbPalette[som_cluster]
legend("right", title="clusters", legend = c("1", "2", "3", "4", "5", "6", "7", "8"),
fill = cbp)

#Mapping cluster results

#create dataframe of the small area id and of the cluster unit
cluster_details <- data.frame(id=saga_data$ID_2,
cluster=som_cluster[som_model$unit.classif])

#merge our cluster details onto the fortified spatial polygon dataframe we created
earlier
mappoints <- merge(iceland_fort, cluster_details, by="id")

# Map the areas and colour by cluster
ggplot(data=mappoints, aes(x=long, y=lat, group=group, fill=factor(cluster)))+
  geom_polygon(colour="transparent") +
  theme_void() +
  coord_equal() +
  scale_fill_manual(name = "Clusters", values = cbp)

#combine the cluster data onto our original spatial polygons
iceland_map <- merge(iceland_map, saga_data, by.x="ID_2", by.y="ID_2")
iceland_map <- merge(iceland_map, cluster_details, by.x="ID_2", by.y="id")
#write the edinburgh_map as a shapefile
writeOGR(obj=iceland_map,
        dsn="saga_10",
        layer="saga_10",
        driver="ESRI Shapefile")

```

Sead_som.R

Script that trains and clusters the SEAD dataset

```
#set working directory
setwd("~/GIS_pg/dissertation/data")

#install packages and libraries.
install.packages('kohonen')
install.packages('ggplot2')
install.packages('rgdal')
install.packages('gridExtra')
install.packages('grid')
install.packages('viridis')
install.packages('dplyr')
install.packages('maptools')
install.packages('gpclib')
install.packages('devtools')
install.packages('readxl')
library(kohonen)
library(ggplot2)
library(rgdal)
library(gridExtra)
library(grid)
library(viridis)
library(dplyr)
library(maptools)
library(gpclib)
library(readxl)

#read in the data from the excel workbook, this data has been categorised into 10 concept
categories and adjusted wrt total point value within each polygon
#any polygons containing zero SEAD data points have been removed
sead_data <- read_excel("SEAD_points_in_polygons.xlsx", sheet = "adjustedph2")

#read in boundary data for Iceland, which has been matched up with the above data (by ID)
iceland_map <- readOGR("../isl/ISL_adm2.shp", stringsAsFactors =
  FALSE)

#convert map into latitude and longitude, easier to implement into ggmap. plot to check
it's fine
iceland_map <- spTransform(iceland_map, CRS("+proj=longlat +ellps=WGS84
+datum=WGS84 +no_defs"))
plot(iceland_map)

#combine your dataset with the map of Iceland
#first check you gpclib permit. if true, proceed
gpclibPermit()
#convert spatial polygon to dataframe including columns of spatial information
iceland_fort <- fortify(iceland_map, region= "ID_2")

#merge the new dataframe with the imported dataset using their shared column (id)
iceland_fort <- merge(iceland_fort, sead_data, by.x="id", by.y="ID_2")

#Test that this has worked by creating a plot of whatever you want
ggplot(data=iceland_fort, aes(x=long, y=lat, fill=buildings, #this plots the spatial
spread of the concept categ "buildings"
```

```

    group=group)) +
  scale_fill_viridis(name = "some rate")+
  geom_polygon(colour=NA)+
  theme_void() +
  coord_equal()

#SOM training

#Select variables used to train SOM by subsetting the 'data' dataframe
data_train <- select(sead_data, buildings, managed, domestic, natural, wild, water)

#standardise the data and convert to a matrix
# first we resacle varaibles and create a matrix
data_train_matrix <- as.matrix(scale(data_train))
#keep the column names of data_train as names in our new matrix
names(data_train_matrix) <- names(data_train)

#define the size and topology of the som grid. must be 5*3 because Iceland only 17
municipalities have data in them
som_grid <- somgrid(xdim = 3, ydim=5, topo="hexagonal")

# Train the SOM model!
som_model <- som(data_train_matrix,
  grid=som_grid,
  rlen=500,
  alpha=c(0.05,0.1),
  keep.data = TRUE)
# Plot SOM training progress - how the node distances have stabilised over time.
plot(som_model, type = "changes")

#load custom palette, make som visualisations more appealing
source('coolBlueHotRed.R')

#counts within nodes - how many "counts"/points are within each node
plot(som_model, type = "counts", main="Node Counts",
  palette.name=coolBlueHotRed)

#map quality as node distance
plot(som_model, type = "quality", main="Node Quality/Distance",
  palette.name=coolBlueHotRed)

#neighbour distances
plot(som_model, type="dist.neighbours", main = "SOM neighbour distances",
  palette.name=grey.colors)

#code spread, this creates the codes plot that helps determine clusters
plot(som_model, type = "codes")

#Clustering of SOM results

# show the WCSS metric for kmeans for different clustering sizes.
# Use this to indicate the ideal number of clusters
mydata <- getCodes(som_model)
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 1:10) wss[i] <- sum(kmeans(mydata, centers=i)$withinss)
#Plot WCSS
plot(1:10, wss, type="b", xlab="Number of Clusters",

```

```

ylab="Within groups sum of squares", main="WCSS")

# Form clusters on grid

#define number of clusters
som_cluster <- cutree(hclust(dist(getCodes(som_model))), 8)

# Colour palette definition. Download the colour palette package "RColorBrewer" and
specifying palette name ("Accent")
library("RColorBrewer")
cbp <- brewer.pal(n = 8, name = "Accent")

#Plot som_model data, adding colour and cluster boundaries
plot(som_model, type="codes", bgcol = cbp[som_cluster], main =
      "Clusters")
add.cluster.boundaries(som_model, som_cluster)
legend("right", title="clusters", legend = c("1", "2", "3", "4", "5", "6", "7", "8"),
fill = cbp)

#Mapping cluster results

#create dataframe of the small area id and of the cluster unit
cluster_details <- data.frame(id=sead_data$ID_2,
cluster=som_cluster[som_model$unit.classif])

#we can just merge our cluster details onto the fortified spatial polygon dataframe we
created earlier
mappoints <- merge(iceland_fort, cluster_details, by="id")

# Map the areas and colour by cluster
ggplot(data=mappoints, aes(x=long, y=lat, group=group, fill=factor(cluster)))+
  geom_polygon(colour="transparent") +
  theme_void() +
  coord_equal() +
  scale_fill_manual(name = "Clusters", values = cbp)

#combine the cluster data onto our original spatial polygons
iceland_map <- merge(iceland_map, sead_data, by.x="ID_2", by.y="ID_2")
iceland_map <- merge(iceland_map, cluster_details, by.x="ID_2", by.y="id")
#write the edinburgh_map as a shapefile
writeOGR(obj=iceland_map,
        dsn="sead_10",
        layer="sead_10",
        driver="ESRI Shapefile")

```

Nabone_som.R

Script that trains and clusters the NABOne dataset, both as areas and data points

```
#set working directory
setwd("~/GIS_pg/dissertation/data")

#install packages and libraries.
install.packages('kohonen')
install.packages('ggplot2')
install.packages('rgdal')
install.packages('gridExtra')
install.packages('grid')
install.packages('viridis')
install.packages('dplyr')
install.packages('maptools')
install.packages('gpclib')
install.packages('devtools')
install.packages('writexl')
install.packages('readxl')
library(kohonen)
library(ggplot2)
library(rgdal)
library(gridExtra)
library(grid)
library(viridis)
library(dplyr)
library(maptools)
library(gpclib)
library(readxl)
library(writexl)

#----1. SOM OF ALL NABO DATA -----

"
DATA INPUT, PREPARE FOR TRAINING
"

#read in the data from the excel workbook
nabo_data <- read_excel("nabo_indicators.xlsx", sheet = "combined")

#read in boundary data for Iceland, which has been matched up with the above data (by ID)
iceland_map <- readOGR("../isl/ISL_adm2.shp", stringsAsFactors =
                      FALSE)

#convert map into latitude and longitude, easier to implement into ggmap. plot to check
it's fine.
iceland_map <- spTransform(iceland_map, CRS("+proj=longlat +ellps=WGS84
+datum=WGS84 +no_defs"))
plot(iceland_map)
```



```
#combine your dataset with the map of Iceland
#first check you gpclib permit. if true, proceed
gpclibPermit()
#convert spatial polygon to dataframe including columns of spatial information
iceland_fort <- fortify(iceland_map, region= "ID_2")

#merge the new dataframe with the census data using their shared column (id)
iceland_fort <- merge(iceland_fort, nabo_data, by.x="id", by.y="ID_2")

#Test to see that this has worked by creating a plot of whatever you want
ggplot(data=iceland_fort, aes(x=long, y=lat, fill=wild,
                             group=group)) +
  scale_fill_viridis(name = "some rate")+
  geom_polygon(colour=NA)+
  theme_void() +
  coord_equal()

#SOM training

#Select variables used to train SOM by subsetting the 'data' dataframe
data_train <- select(nabo_data, domesticaval, wildaval, wateraval)

#standardise the data and convert to a matrix
# first we resacle varaibles and create a matrix
data_train_matrix <- as.matrix(scale(data_train))
#keep the column names of data_train as names in our new matrix
names(data_train_matrix) <- names(data_train)

#define the size and topology of the som grid.
som_grid <- somgrid(xdim = 5, ydim=5, topo="hexagonal")

# Train the SOM model
som_model <- som(data_train_matrix,
                grid=som_grid,
                rlen=300,
                alpha=c(0.05,0.1),
                keep.data = TRUE)
# Plot SOM training progress - how the node distances have stabilised over time.
plot(som_model, type = "changes")

#load custom palette, make som visualisations more appealing
source('coolBlueHotRed.R')

#counts within nodes - how many "counts"/points are within each node?
plot(som_model, type = "counts", main="Node Counts",
     palette.name=coolBlueHotRed)

#map quality
plot(som_model, type = "quality", main="Node Quality/Distance",
     palette.name=coolBlueHotRed)

#neighbour distances
plot(som_model, type="dist.neighbours", main = "SOM neighbour distances",
     palette.name=grey.colors)

#code spread
plot(som_model, type = "codes")
```

```

#Clustering of SOM results

# show the WCSS metric for kmeans for different clustering sizes.
# Use this to indicate the ideal number of clusters
mydata <- getCodes(som_model)
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(mydata, centers=i)$withinss)
#Plot WCSS
plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares", main="WCSS")

# Form clusters on grid

#define number of clusters
som_cluster <- cutree(hclust(dist(getCodes(som_model))), 4)

# Colour palette definition. Download the colour palette package "RColorBrewer" and
specifying palette name ("Accent")
library("RColorBrewer")
cbp <- brewer.pal(n = 4, name = "Accent")
#6 might seem like too many but for now we have to keep it the same as for the point data
so the colours refer to the same cluster number

#Plot som_model data, adding colour and cluster boundaries
plot(som_model, type="codes", bgcol = cbp[som_cluster], main =
     "Clusters")
add.cluster.boundaries(som_model, som_cluster)
legend("right", title="clusters", legend = c("1", "2", "3", "4"), fill = cbp)

#Mapping cluster results

#create dataframe of the small area id and of the cluster unit
cluster_details <- data.frame(id=nabo_data$ID_2,
cluster=som_cluster[som_model$unit.classif])

#we can just merge our cluster details onto the fortified spatial polygon dataframe we
created earlier
mappoints <- merge(iceland_fort, cluster_details, by="id")

# Map the areas and colour by cluster
ggplot(data=mappoints, aes(x=long, y=lat, group=group, fill=factor(cluster)))+
  geom_polygon(colour="transparent") +
  theme_void() +
  coord_equal() +
  scale_fill_manual(name = "Clusters", values = cbp)

"
Export cluster results to shapefile so it can be mapped in qGIS
"

#combine the cluster data onto our original spatial polygons
iceland_map <- merge(iceland_map, nabo_data, by.x="ID_2", by.y="ID_2")
iceland_map <- merge(iceland_map, cluster_details, by.x="ID_2", by.y="id")
#write the Iceland_map as a shapefile
writeOGR(obj=iceland_map,
        dsn="nabo_10_2",
        layer="nabo_10_2",
        driver="ESRI Shapefile")

```

```

#----2. SOM OF NABO OUTLIERS, POINTS-----

"
DATA INPUT, PREPARE FOR TRAINING
"

#read in the data from the excel workbook
nabo_data <- read_excel("naboind_o.xlsx", sheet = "naboind")

#SOM training

#Select variables used to train SOM by subsetting the 'data' dataframe
data_train <- select(nabo_data, domestic, wild, water)

#standardise the data and convert to a matrix
# first we resacle varaibles and create a matrix
data_train_matrix <- as.matrix(scale(data_train))
#keep the column names of data_train as names in our new matrix
names(data_train_matrix) <- names(data_train)

#define the size and topology of the som grid.
som_grid <- somgrid(xdim = 4, ydim=4, topo="hexagonal")

# Train the SOM model!
som_model <- som(data_train_matrix,
                 grid=som_grid,
                 rlen=200,
                 alpha=c(0.05,0.1),
                 keep.data = TRUE)
# Plot SOM training progress - how the node distances have stabilised over time.
plot(som_model, type = "changes")

#load custom palette, make som visualisations more appealing
source('coolBlueHotRed.R')

#counts within nodes - how many "counts"/points are within each node
plot(som_model, type = "counts", main="Node Counts",
     palette.name=coolBlueHotRed)
#map quality
plot(som_model, type = "quality", main="Node Quality/Distance",
     palette.name=coolBlueHotRed)
#neighbour distances
plot(som_model, type="dist.neighbours", main = "SOM neighbour distances",
     palette.name=grey.colors)
#code spread
plot(som_model, type = "codes")

#Clustering of SOM results

# show the WCSS metric for kmeans for different clustering sizes.
# Use this to indicate the ideal number of clusters
mydata <- getCodes(som_model)
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(mydata, centers=i)$withinss)
#Plot WCSS
plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares", main="WCSS")

```

```
# Form clusters on grid

som_cluster <- cutree(hclust(dist(getCodes(som_model))), 4)

# Colour palette definition. Download the colour palette package "RColorBrewer" and
specifying palette name ("Accent")
library("RColorBrewer")
cbp <- brewer.pal(n = 4, name = "Accent")

#Plot som_model data, adding colour and cluster boundaries
plot(som_model, type="codes", bgcol = cbp[som_cluster], main =
      "Clusters")
add.cluster.boundaries(som_model, som_cluster)
legend("right", title="clusters", legend = c("1", "2", "3", "4"), fill = cbp)

#create dataframe of the cluster "number" for each point
cluster_details <- data.frame(id=nabo_data$pointids,
cluster=som_cluster[som_model$unit.classif])

write_xlsx(cluster_details, path = "nabo_o2.xlsx")
#this spreadsheet will be imported into qGIS, joined spatially with the original NABO
json point file and displayed visually
```

Combined_som.R

Script that trains and clusters the 3 combined datasets, as well as individually training datasets for the concept categories “domestic”, “water” and “wild”

```
#set working directory
setwd("~/GIS_pg/dissertation/data")

#install packages and libraries. Copied from SOM assessment, you probably don't
need all of these
install.packages('kohonen')
install.packages('ggplot2')
install.packages('rgdal')
install.packages('gridExtra')
install.packages('grid')
install.packages('viridis')
install.packages('dplyr')
install.packages('maptools')
install.packages('gpclib')
install.packages('devtools')
install.packages('writexl')
install.packages('readxl')
library(kohonen)
library(ggplot2)
library(rgdal)
library(gridExtra)
library(grid)
library(viridis)
library(dplyr)
library(maptools)
library(gpclib)
library(readxl)
library(writexl)

#----SOM OF ALL DATASETS COMBINED, EMPTY FIELDS REMOVED ----

"
DATA INPUT, PREPARE FOR TRAINING
"

#read in the data from the excel workbook
all_data <- read_excel("everything_combined.xlsx", sheet = "test3")

#read in boundary data for Iceland, which has been matched up with the above
data (by ID)
iceland_map <- readOGR("../isl/ISL_adm2.shp", stringsAsFactors =
                        FALSE)
```

```
#convert map into latitude and longitude, easier to implement into ggmap. plot
to check it's fine.
iceland_map <- spTransform(iceland_map, CRS("+proj=longlat +ellps=WGS84
+datum=WGS84 +no_defs"))
plot(iceland_map)

#combine your dataset with the map of Iceland
#first check you gpclib permit. if true, proceed
gpclibPermit()
#convert spatial polygon to dataframe including columns of spatial information
iceland_fort <- fortify(iceland_map, region= "ID_2")

#merge the new dataframe with the census data using their shared column (id)
iceland_fort <- merge(iceland_fort, all_data, by.x="id", by.y="ID_2")

#SOM training

#Select variables used to train SOM by subsetting the 'data' dataframe
data_train <- select(all_data, activities, buildings, managed, domestic,
natural, wild, water, travel, weather, things)
#standardise the data and convert to a matrix
# first we resacle varaibles and create a matrix
data_train_matrix <- as.matrix(scale(data_train))
#keep the column names of data_train as names in our new matrix
names(data_train_matrix) <- names(data_train)

#define the size and topology of the som grid
som_grid <- somgrid(xdim = 8, ydim=8, topo="hexagonal")

# Train the SOM model!
som_model <- som(data_train_matrix,
                grid=som_grid,
                rlen=500,
                alpha=c(0.05,0.1), #0.05,0.1
                keep.data = TRUE,)
# Plot SOM training progress - how the node distances have stabilised over time.
plot(som_model, type = "changes")

#load custom palette, make som visualisations more appealing
source('coolBlueHotRed.R')

#counts within nodes - how many "counts"/points are within each node?
plot(som_model, type = "counts", main="Node Counts",
     palette.name=coolBlueHotRed)
#map quality
plot(som_model, type = "quality", main="Node Quality/Distance",
     palette.name=coolBlueHotRed)
#neighbour distances
plot(som_model, type="dist.neighbours", main = "SOM neighbour distances",
     palette.name=coolBlueHotRed)
#code spread
plot(som_model, type = "codes")
```

```

#Clustering of SOM results

# show the WCSS metric for kmeans for different clustering sizes.
# Use this to indicate the ideal number of clusters
mydata <- getCodes(som_model)
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(mydata, centers=i)$withinss)
#Plot WCSS
plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares", main="WCSS")

# Form clusters on grid

#define number of clusters
som_cluster <- cutree(hclust(dist(getCodes(som_model))), 10)

# Colour palette definition.
cbp <- c("#7FC97F", "#BEAED4", "#FDC086", "#FFFF99", "#386CB0", "#F00274",
"#BF5B17", "#666666", "#900C3F", "#F5F5F5" )

#Plot som_model data, adding colour and cluster boundaries
plot(som_model, type="codes", bgcol = cbp[som_cluster], main =
     "Clusters")
add.cluster.boundaries(som_model, som_cluster)
legend("right", title="clusters", legend = c("1", "2", "3", "4", "5", "6", "7",
"8", "9", "10"), fill = cbp)

#Mapping cluster results

#create dataframe of the area id and of the cluster unit
cluster_details <- data.frame(id=all_data$ID_2,
cluster=som_cluster[som_model$unit.classif])

#we can just merge our cluster details onto the fortified spatial polygon
dataframe we created earlier
mappoints <- merge(iceland_fort, cluster_details, by="id")

# Map the areas and colour by cluster
ggplot(data=mappoints, aes(x=long, y=lat, group=group, fill=factor(cluster)))+
  geom_polygon(colour="transparent") +
  theme_void() +
  coord_equal() +
  scale_fill_manual(name = "Clusters", values = cbp)

#combine the cluster data onto our original spatial polygons
iceland_map <- merge(iceland_map, all_data, by.x="ID_2", by.y="ID_2")
iceland_map <- merge(iceland_map, cluster_details, by.x="ID_2", by.y="id")
class(iceland_map)

```

```
#write the map as a shapefile
writeOGR(obj=iceland_map,
        dsn="combined3",
        layer="combined3",
        driver="ESRI Shapefile")

#----- SOM OF INDIVIDUAL CATEGORIES: DOMESTIC-----

#read in the data from the excel workbook
all_data <- read_excel("everything_combined.xlsx", sheet = "domestic")

#read in boundary data for Iceland, which has been matched up with the above
data (by ID)
iceland_map <- readOGR("../isl/ISL_adm2.shp", stringsAsFactors =
                      FALSE)

#convert map into latitude and longitude, easier to implement into ggmap. plot
to check it's fine
iceland_map <- spTransform(iceland_map, CRS("+proj=longlat +ellps=WGS84
+datum=WGS84 +no_defs"))

#create map by first checking you gpclib permit. if true, proceed
gpclibPermit()
#convert spatial polygon to dataframe including columns of spatial information
iceland_fort <- fortify(iceland_map, region= "ID_2")

#merge the new dataframe with the census data using their shared column (id)
iceland_fort <- merge(iceland_fort, all_data, by.x="id", by.y="ID_2")

#SOM training

#Select variables used to train SOM by subsetting the 'data' dataframe
data_train <- select(all_data, SAGA, SEAD, NABO)

#standardise the data and convert to a matrix
# first we resacle varaibles and create a matrix
data_train_matrix <- as.matrix(scale(data_train))
#keep the column names of data_train as names in our new matrix
names(data_train_matrix) <- names(data_train)

#define the size and topology of the som grid
som_grid <- somgrid(xdim = 4, ydim=4, topo="hexagonal")

# Train the SOM model!
som_model <- som(data_train_matrix,
                grid=som_grid,
                rlen=500,
                alpha=c(0.05,0.1), #0.05,0.1
                keep.data = TRUE,)
```



```

# Plot SOM training progress - how the node distances have stabilised over time.
plot(som_model, type = "changes")

#load custom palette, make som visualisations more appealing
source('coolBlueHotRed.R')

#counts within nodes - how many "counts"/points are within each node?
plot(som_model, type = "counts", main="Node Counts",
      palette.name=coolBlueHotRed)
#map quality
plot(som_model, type = "quality", main="Node Quality/Distance",
      palette.name=coolBlueHotRed)
#neighbour distances
plot(som_model, type="dist.neighbours", main = "SOM neighbour distances",
      palette.name=grey.colors)
#code spread
plot(som_model, type = "codes")

#Clustering of SOM results

# show the WCSS metric for kmeans for different clustering sizes.
# Use this to indicate the ideal number of clusters
mydata <- getCodes(som_model)
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(mydata, centers=i)$withinss)
#Plot WCSS
plot(1:15, wss, type="b", xlab="Number of Clusters",
      ylab="Within groups sum of squares", main="WCSS")

# Form clusters on grid

som_cluster <- cutree(hclust(dist(getCodes(som_model))),6)

# Colour palette definition. Download the colour palette package "RColorBrewer"
and specifying palette name ("Accent")
library("RColorBrewer")
cbp <- brewer.pal(n = 6, name = "Accent")

#Plot som_model data, adding colour and cluster boundaries
plot(som_model, type="codes", bgcol = cbp[som_cluster], main =
      "Clusters")
add.cluster.boundaries(som_model, som_cluster)
legend("right", title="clusters", legend = c("1", "2", "3", "4", "5", "6"), fill
= cbp)

#Mapping cluster results

#create dataframe of the small area id and of the cluster unit
cluster_details <- data.frame(id=all_data$ID_2,
cluster=som_cluster[som_model$unit.classif])

```

```

#we can just merge our cluster details onto the fortified spatial polygon
dataframe we created earlier
mappoints <- merge(iceland_fort, cluster_details, by="id")

# Map the areas and colour by cluster
ggplot(data=mappoints, aes(x=long, y=lat, group=group, fill=factor(cluster)))+
  geom_polygon(colour="transparent") +
  theme_void() +
  coord_equal() +
  scale_fill_manual(name = "Clusters", values = cbp)

#combine the cluster data onto our original spatial polygons
iceland_map <- merge(iceland_map, all_data, by.x="ID_2", by.y="ID_2")
iceland_map <- merge(iceland_map, cluster_details, by.x="ID_2", by.y="id")
class(iceland_map)

#write the map as a shapefile
writeOGR(obj=iceland_map,
        dsn="som_domestic3",
        layer="som_domestic3",
        driver="ESRI Shapefile")

#-----SOM OF INDIVIDUAL CATEGORIES: wild-----

#read in the data from the excel workbook
all_data <- read_excel("everything_combined.xlsx", sheet = "wild")

#read in boundary data for Iceland, which has been matched up with the above
data (by ID)
iceland_map <- readOGR("../isl/ISL_adm2.shp", stringsAsFactors =
                      FALSE)

#convert map into latitude and longitude, easier to implement into ggmap. plot
to check it's fine
iceland_map <- spTransform(iceland_map, CRS("+proj=longlat +ellps=WGS84
+datum=WGS84 +no_defs"))

#create map by first checking you gpclib permit. if true, proceed
gpclibPermit()
#convert spatial polygon to dataframe including columns of spatial information
iceland_fort <- fortify(iceland_map, region= "ID_2")

#merge the new dataframe with the census data using their shared column (id)
iceland_fort <- merge(iceland_fort, all_data, by.x="id", by.y="ID_2")

#SOM training

#Select variables used to train SOM by subsetting the 'data' dataframe
data_train <- select(all_data, SAGA, SEAD, NABO)

```

```
#standardise the data and convert to a matrix
# first we resacle varaibles and create a matrix
data_train_matrix <- as.matrix(scale(data_train))
#keep the column names of data_train as names in our new matrix
names(data_train_matrix) <- names(data_train)

#define the size and topology of the som grid
som_grid <- somgrid(xdim = 4, ydim= 4, topo="hexagonal")

# Train the SOM model!
som_model <- som(data_train_matrix,
                grid=som_grid,
                rlen=500,
                alpha=c(0.05,0.1), #0.05,0.1
                keep.data = TRUE,)
# Plot SOM training progress - how the node distances have stabilised over time.
plot(som_model, type = "changes")

#load custom palette, make som visualisations more appealing
source('coolBlueHotRed.R')

#counts within nodes - how many "counts"/points are within each node?
plot(som_model, type = "counts", main="Node Counts",
     palette.name=coolBlueHotRed)
#map quality
plot(som_model, type = "quality", main="Node Quality/Distance",
     palette.name=coolBlueHotRed)
#neighbour distances
plot(som_model, type="dist.neighbours", main = "SOM neighbour distances",
     palette.name=grey.colors)
#code spread
plot(som_model, type = "codes")

#Clustering of SOM results

# show the WCSS metric for kmeans for different clustering sizes.
# Use this to indicate the ideal number of clusters
mydata <- getCodes(som_model)
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(mydata, centers=i)$withinss)
#Plot WCSS
plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares", main="WCSS")

# Form clusters on grid

som_cluster <- cutree(hclust(dist(getCodes(som_model))),5)

# Colour palette definition. Download the colour palette package "RColorBrewer"
and specifying palette name ("Accent")
library("RColorBrewer")
cbp <- brewer.pal(n = 5, name = "Accent")
```

```

#Plot som_model data, adding colour and cluster boundaries
plot(som_model, type="codes", bgcol = cbp[som_cluster], main =
      "Clusters")
add.cluster.boundaries(som_model, som_cluster)
legend("right", title="clusters", legend = c("1", "2", "3", "4", "5"), fill =
cbp)

#Mapping cluster results

#create dataframe of the small area id and of the cluster unit
cluster_details <- data.frame(id=all_data$ID_2,
cluster=som_cluster[som_model$unit.classif])

#we can just merge our cluster details onto the fortified spatial polygon
dataframe we created earlier
mappoints <- merge(iceland_fort, cluster_details, by="id")

# Map the areas and colour by cluster
ggplot(data=mappoints, aes(x=long, y=lat, group=group, fill=factor(cluster)))+
  geom_polygon(colour="transparent") +
  theme_void() +
  coord_equal() +
  scale_fill_manual(name = "Clusters", values = cbp)

#combine the cluster data onto our original spatial polygons
iceland_map <- merge(iceland_map, all_data, by.x="ID_2", by.y="ID_2")
iceland_map <- merge(iceland_map, cluster_details, by.x="ID_2", by.y="id")
class(iceland_map)

#write the map as a shapefile
writeOGR(obj=iceland_map,
        dsn="som_wild",
        layer="som_wild",
        driver="ESRI Shapefile")

#-----SOM OF INDIVIDUAL CATEGORIES: water-----

#read in the data from the excel workbook
all_data <- read_excel("everything_combined.xlsx", sheet = "water")

#read in boundary data for Iceland, which has been matched up with the above
data (by ID)
iceland_map <- readOGR("../isl/ISL_adm2.shp", stringsAsFactors =
FALSE)

#convert map into latitude and longitude, easier to implement into ggmap. plot
to check it's fine
iceland_map <- spTransform(iceland_map, CRS("+proj=longlat +ellps=WGS84
+datum=WGS84 +no_defs"))

```

```
#create map by first checking you gpclib permit. if true, proceed
gpclibPermit()
#convert spatial polygon to dataframe including columns of spatial information
iceland_fort <- fortify(iceland_map, region= "ID_2")

#merge the new dataframe with the census data using their shared column (id)
iceland_fort <- merge(iceland_fort, all_data, by.x="id", by.y="ID_2")

#SOM training

#Select variables used to train SOM by subsetting the 'data' dataframe
data_train <- select(all_data, SAGA, SEAD, NABO)

#standardise the data and convert to a matrix
# first we resacle variables and create a matrix
data_train_matrix <- as.matrix(scale(data_train))
#keep the column names of data_train as names in our new matrix
names(data_train_matrix) <- names(data_train)

#define the size and topology of the som grid
som_grid <- somgrid(xdim = 5, ydim=5, topo="hexagonal")

# Train the SOM model!
som_model <- som(data_train_matrix,
                 grid=som_grid,
                 rlen=500,
                 alpha=c(0.05,0.1), #0.05,0.1
                 keep.data = TRUE,)
# Plot SOM training progress - how the node distances have stabilised over time.
plot(som_model, type = "changes")

#load custom palette, make som visualisations more appealing
source('coolBlueHotRed.R')

#counts within nodes - how many "counts"/points are within each node?
plot(som_model, type = "counts", main="Node Counts",
     palette.name=coolBlueHotRed)
#map quality
plot(som_model, type = "quality", main="Node Quality/Distance",
     palette.name=coolBlueHotRed)
#neighbour distances
plot(som_model, type="dist.neighbours", main = "SOM neighbour distances",
     palette.name=grey.colors)
#code spread
plot(som_model, type = "codes")

#Clustering of SOM results

# show the WCSS metric for kmeans for different clustering sizes.
# Use this to indicate the ideal number of clusters
mydata <- getCodes(som_model)
```

```

wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(mydata, centers=i)$withinss)
#Plot WCSS
plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares", main="WCSS")

# Form clusters on grid

som_cluster <- cutree(hclust(dist(getCodes(som_model))), 6)

# Colour palette definition. Download the colour palette package "RColorBrewer"
and specifying palette name ("Accent")
library("RColorBrewer")
cbp <- brewer.pal(n = 6, name = "Accent")

#Plot som_model data, adding colour and cluster boundaries
plot(som_model, type="codes", bgcol = cbp[som_cluster], main =
     "Clusters")
add.cluster.boundaries(som_model, som_cluster)
legend("right", title="clusters", legend = c("1", "2", "3", "4", "5", "6"), fill
= cbp)

#Mapping cluster results

#create dataframe of the small area id and of the cluster unit
cluster_details <- data.frame(id=all_data$ID_2,
cluster=som_cluster[som_model$unit.classif])

#we can just merge our cluster details onto the fortified spatial polygon
dataframe we created earlier
mappoints <- merge(iceland_fort, cluster_details, by="id")

# Map the areas and colour by cluster
ggplot(data=mappoints, aes(x=long, y=lat, group=group, fill=factor(cluster)))+
  geom_polygon(colour="transparent") +
  theme_void() +
  coord_equal() +
  scale_fill_manual(name = "Clusters", values = cbp)

#combine the cluster data onto our original spatial polygons
iceland_map <- merge(iceland_map, all_data, by.x="ID_2", by.y="ID_2")
iceland_map <- merge(iceland_map, cluster_details, by.x="ID_2", by.y="id")
class(iceland_map)

#write the map as a shapefile
writeOGR(obj=iceland_map,
        dsn="som_water",
        layer="som_water",
        driver="ESRI Shapefile")

```