THE UNIVERSITY
of EDINBURGH

Spatio-Temporal Clustering in Application

Antonia Gieschen

Doctor of Philosophy
University of Edinburgh
2021

While each doctoral thesis presents the outcome of independent and at times lonely work, I am sure that I speak for myself and my peers when saying that academia is more often than not a real team effort. I could not have done it without the people behind, in front of and next to me, who supported, guided and walked with me this journey.

To my primary supervisor, Belen Martin-Barragan, who with much patience and good humour helped me stay on a much more structured and purposeful path than I could have done on my own, and who taught me so much.
Thank you for your kindness and knowledge.

To my second supervisor, Jake Ansell, whose both academic and mental support was much needed and received most gratefully.
Thank you for your wisdom and humour.

To all my academic colleagues and co-authors, and most importantly to Ewelina Lacka and Raffaella Calabrese.
Thank you for your trust and belief in me.

To my parents, who accepted my choice of pursuing an academic career without hesitation and whose love and encouragement were invaluable. Thank you for your patience and understanding.

Finally, and most importantly, to my partner, Ben Moews, without whom I would not be the person I am today.
My endless love and gratitude.

And now, to new adventures.

Lay summary

Machine learning (ML) has become an important approach to how data in businesses, organisations and research is analysed. The array of methods allows us to better understand, visualise, forecast and structure data, and as such machine learning finds application in a number of areas. One differentiates between two approaches to machine learning methods: supervised and unsupervised. Supervised ML uses a part of the data set, which is called training data, to derive rules from and learn how to handle new and unseen data in the future. In contrast, unsupervised learning does not need any pre-labelling to structure data and instead it independently finds interesting structure and groups within the data set. This thesis will focus on the second type of ML approaches and specifically on a group of algorithms that is called clustering. Clustering looks for groups of data points which are as similar to each other as possible and as different from other groups as possible.

Data can be of many different types and this affects the method which can be used to analyse them. Spatial data has a geographic component and can experience dependency between locations, which is often assumed to be stronger between points closer to each other compared to those further away. Another perspective on spatial data which is of high relevance in clustering and this thesis is the notion of density. As clusters are defined as groups of similar data points, in spatial contexts this similarity is described through closeness and density. The second type of data this thesis will focus on is time series data. This type of data is measured through time and as such forms a sequential array. Similarly to spatial data, temporal dependency is a concept often integrated in time series analysis. It describes how a value measured at one point in time is dependent on other values observed at a point close in time, rather than values observed much earlier or later. Another challenge that temporal data brings is how to measure the similarity of multiple time series. Different approaches have been suggested in the literature, which compare series one point in time after the other or methods which take into account the overall shape of the two time series. Finally, spatio-temporal data denotes the combination of these two data types. Data is being observed both from a spatial and a temporal perspective, which makes analysis complex as concepts from both dimensions are being brought together. Clearly, spatial and temporal data bring unique challenges to data analysis in general and clustering in particular. As such, the first aspect that will be demonstrated through this thesis is how clustering algorithms have to be methodologically adapted or data has to be handled if data is spatial and/or temporal in nature.

The second contribution of this thesis will come through its application. Through application to three different areas, health, finance and marketing, this work will demonstrate the flexibility of clustering and how clustering algorithms can be used to derive insights from data. Chapter 3 explores the use of spatio-temporal clustering in the analysis of drug prescriptions in Scotland. By using clustering, groups of GPs are formed which behave similar to each other over time and

are geographically close to each other. This allows not only for insights into the landscape of Scottish prescription behaviour, but also uncovers extreme behaviour of very high or low prescription levels. Individual GPs can compare themselves to their peers and policy makers can detect potential areas at risk of overprescription, the relevance of which will be demonstrated through the use-case of antibiotic prescription. From a methodological perspective, Scotland presents a challenging case due to the uneven population distribution. As described earlier, spatial clustering uses notions of density to detect clusters. The spatio-temporal clustering algorithm ST-DBSCAN, which will be used in Chapter 3, suffers from the known drawback of not being able to handle more spatially dense and less spatially dense geographic regions simultaneously. The Chapter thus will introduce a novel approach to using a density factor to make such areas more comparable and improve the performance of the algorithm.

Chapter 4 demonstrates the usefulness of clustering in the analysis of access to finance for small and medium-sized enterprises (SMEs). Understanding factors which impact the ability of SMEs to access external financing, in this case in the form of bank loans and overdrafts, is important because financial constraints can impact the companies' ability to grow and these businesses form an integral part of many economies. SMEs also experience geographical connections and spillover effects between each other, likely through formal and informal networks between them. Methodologically speaking, this Chapter will use a novel combination of supervised and unsupervised ML in explaining access to finance: In order to account for spatial spillover effects between companies, clustering will be used. Clustering will form groups of SMEs which are both geographically close and similar to each other regarding their financial situation and operating sector. In a second step, a spatial regression model will be used to detect which SME characteristics have an impact on the success of bank loan and overdraft applications. The regression model will use the cluster groups to account for effects only spilling over between SMEs which are connected through their spatial closeness and similarity. The Chapter will demonstrate that such spillover effects between such companies exist and can explain which companies are able to access finance. This carries important implications for policy makers in an effort to formulate more regional efforts to support SMEs struggling to secure financing.

Finally, in Chapter 5 the use of clustering in the context of tourism will be presented. The COVID-19 pandemic has hit the tourism sector especially hard due to travel restrictions and uncertainty among travellers. For businesses trying to recover, understanding from which countries tourists are the most likely to visit is important because it allows them to formulate marketing strategies which target those specific markets first. The Chapter will present the use of clustering to form groups of countries which are similar in terms of tourists' interest, affordability of their travel, and reachability of Edinburgh and Scotland as a location. Interest will be taken into account in the form of Google Trend search data over time and affordability will be implemented through the Consumer Confidence Index. A novel approach to including reachability will be introduced in this Chapter. In our modern connected world, geographic distance can not always explain how easy to reach a location is for a tourist. Instead, the presence of established travel routes can be considered more relevant to a traveller. The Chapter therefore introduces a reachability factor based on the number and directness of flight connections between the tourists' country and Edinburgh. Three groups of countries are formed through the use of clustering. These results offer actionable insights for businesses in their pandemic recovery efforts. Marketing strategies can be formulated based on the likeliness of visitors from those locations and in other cases address the reason for not visiting due to the combination of the three factors interest, affordability and reachability in the analysis.

Two contributions are being made through this thesis: methodologically, challenges from spatial, temporal and spatio-temporal data in clustering will be tackled. From an application perspective, the insights which clustering can deliver to diverse contexts like health, finance and marketing, are demonstrated through real-world applications. On the backdrop of the increasing importance of machine learning in data analysis of businesses and research alike, this work will thus form an important perspective and bridge between theory and practice. While challenges arise through complex data types, the insights that can be derived through clustering offer valuable information for businesses and policy makers alike.

Abstract

The importance of machine learning methods in the data analysis of both academic research and industry applications has advanced rapidly in recent years. This thesis will investigate how a method of unsupervised machine learning known as clustering can be employed to analyse spatial and spatio-temporal data from different fields of application. Spatio-temporal data present a particular challenge. In spatial contexts, the notion of dependency among geographically close elements needs to be considered when analysing the geographic distance as well as other spatial components. The temporal dimension of the data makes traditional dissimilarity metrics unsuitable due to the sequential ordering of data points. For this reason, this thesis will present ways of overcoming the shortcomings in existing methodologies when applied to these data types. By doing so, it will contribute to the literature on clustering through innovative extensions, adaptations, and considerations. The flexibility of clustering will be demonstrated in three different application contexts in health, finance, and marketing. As such, this thesis will also contribute to the academic literature in these areas and offer valuable insights into applicable machine learning methodology for practitioners.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and motivation

I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to

speak of machines thinking without expecting to be contradicted."

- Alan Turing

Almost 70 years after the death of Alan Turing and 21 years after the referred-to date for prominence of unanimous conversations about "machines thinking", the prophecy above seems to have come to fruition. Artificial intelligence (AI) has become an element of the digital age which has encompassed almost all aspects of our daily lives. While discussions about what constitutes intelligence or "thinking" in the context of machines remain heated, the reliance of humans on programmes to learn, adapt, make decisions, communicate (with each other or indeed us), answer questions, drive cars, or water our potted plants[1] seems undeniable.

In Scotland, Scotland's AI Strategy was launched in March 2021 as an indicator that governments are responding to the increasing interest and opportunity arising in both the private and public sector (Scotland's AI Strategy, 2021). Overseen by the Scottish Government and The

[1] Halifax's Deprolabs Use AI to Water Plants, `https://www.borndigital.com/2017/10/23/halifaxs-deprolabs-use-ai-to-water-plants`

Data Lab[2], it maps out how AI and machine learning are expected to shape Scotland's future and the principles and guidance it should follow within that. A key mission that the strategy highlights is that they expect that "Scotland will become a leader in the development and use of trustworthy, ethical and inclusive AI" (The Scottish Government, 2021, p.6). While AI encompasses many different forms of machine intelligence, such as natural language processing and some areas of robotics, machine learning (ML) has become a prominent sub-category within it. Put simply, ML describes computer programmes (algorithms) learning through data. The importance of machine learning for decision making in organisations has been on the increase for the past years with no sign of slowing down. Closely associated areas such as data science have seen a similar movement towards becoming an integral part of organisations.

Public services like the Scottish Government and the National Health Service (NHS) Scotland show an increasing involvement in the research, development and integration of AI and general data science. One example for such an involvement of the Scottish Government is in their presentation of the Scotland's AI Strategy above, another in the founding of Research Data Scotland[3] as a collaboration of Scottish Government, Scottish universities and other public bodies. NHS Scotland is demonstrating their engagement through initiatives like Health Data Research UK[4] which has placed one of its sites in Scotland. The main purpose of this programme is to bring together capabilities and resources for research employing computational methods including AI. The connection between academia and industry is also at the focus of the Data-Driven Innovation (DDI)[5] cluster. As an organisation sitting between the Edinburgh universities and industry partners, their aim is to bring together resources to solve real-world problems with data. The access to data is a key factor for researchers. Organisations like the UK Data Archive[6] provide researchers with such access to data, enabling them to conduct breakthrough research. Similar trends as in the public sector can be seen in smaller companies and enterprises. Reporting on a recent study conducted by Algorithmia in their *2021 Enterprise Trends in Machine Learning* survey, Louis Columbus writes for Forbes that 76% of the Enterprises interviewed would pri-

[2] https://www.thedatalab.com/
[3] https://www.researchdata.scot/
[4] https://www.hdruk.ac.uk/
[5] https://ddi.ac.uk/
[6] https://www.data-archive.ac.uk/

oritise AI and ML in 2021 (Columbus, 2021). In Scotland, tech startups such as Wallscope[7] and incubators like CodeBase[8] create an atmosphere of excitement and creativity around AI in businesses and foster relationships between entrepreneurs and investors. In the Scottish tourism sector, data-driven insights have become more important too, as reflected in the founding of the Tourism, Technology & Data research cluster of the University of Edinburgh Centre for Data, Culture & Society[9].

This thesis very much builds on and sits within this ecosystem of initiatives and networks. The concrete objective of the presented thesis thus lies in introducing methodological advancements in the context of their contributions to practice. This is made possible through collaborative efforts with local organisations and by using local real life data. After an introduction (Chapter 1) and a review of the background (Chapter 2), the first core chapter (Chapter 3) is based on a research project conducted in collaboration with local startup Wallscope[10] and funded by The Data Lab. Utilising public open data by NHS Scotland in the form of GP prescription data, we demonstrate that clustering can be used to form peer groups of GPs based on their spatio-temporal antibiotic prescription behaviour. These peer groups can be used by GPs themselves and by health authorities for information and training purposes, uncovering instances of overprescribing of antibiotics as an indicator for risk within specific spatial regions. The results were presented to the NHS and the Scottish Government at a public event series to showcase the usability of Open Data in health contexts[11]. From a methodological point of view, a weakness in the used algorithm had to be managed by introducing a way of weighing spatial distances based on local point density. The chapter thus demonstrates both the applicability of clustering for public health uses and the need for making methodological changes due to the nature of real life data.

Chapter 4 employs data from the UK Data Archive on small and medium enterprise (SME) financing. Access to finance is an important factor for SMEs and as those businesses are a crucial part of many economies including the UK, understanding and responding to their needs is a part

[7] https://www.wallscope.co.uk/

[8] https://www.thisiscodebase.com/

[9] https://www.cdcs.ed.ac.uk/research-clusters/tourism-technology-data

[10] https://www.wallscope.co.uk/

[11] https://medium.com/wallscope/open-data-mixer-2019-94a7a39c0192

of policy making in many regions. By combining clustering and logistic regression we demonstrate which factors are impacting whether an SME can access external financing in the form of loans or overdrafts, as well as delivering evidence for the existence of spatially bound contagion effects between them. These contagion effects are shown to exist between companies which are both geographically close to each other and similar in terms of turnover and sector. As such, our findings carry important information for policy makers in the way that they should approach access to finance as a local issue instead of a national one. Methodologically, we propose a way of using clustering to create a sparse $\mathbf{W}$ matrix in a spatial logistic model, which allows for spatial effects in the model to take place between similar companies only.

In Chapter 5 we introduce work based on a project in collaboration with the Edinburgh Tourism Action Group (ETAG)[12] which was funded by DDI. Tourism is an important part of the Scottish economy but the COVID-19 pandemic has impacted the sector dramatically. One objective of the project was to aid recovery efforts through data driven insights into which tourism countries are most likely to visit Edinburgh. Chapter 5 builds on the initial findings of this DDI funded project, and demonstrates how the combination of three factors, namely interest, affordability and reachability, can be used to cluster countries by their likeliness to visit. Data from different sources, including search data, consumer confidence indices and flight connections, are used to cluster countries into groups which are then interpreted. Methodologically, we introduce a reachability factor which is not based on geographic distance but instead uses flight connections as a measure of closeness. In a modern connected world the geographic distance that has to be overcome is not the same as the perceived ease of visiting thanks to modern travel routes via planes. This is especially true during and after a pandemic, when flight connections are impacted by local restrictions. The empirical findings carry important implications for businesses as well as policy makers in the economic recovery of the country, because they can be used as a starting point to develop a marketing strategy by giving a classification and implied timeline for targeting countries through marketing efforts.

[12] https://www.etag.org.uk/

Overall, this thesis will explore the use of clustering in different contexts of businesses and organisations. Methodological developments will be presented which make the applied methods more applicable to the given circumstances. In doing so, this work will contribute not only to the existing literature on machine learning from a methodology perspective, but also offer insights into the flexibility of cluster analysis and how organisations from backgrounds as diverse as public health, SME financing and tourism can benefit from it. Generally speaking, clustering is presented as a valuable method for both data exploration and in the context of data modelling. The applications of clustering should not be seen as limited to the presented cases, but instead those contexts should be seen as case studies for the variability of the chosen method. A focus is put on creative ways of adapting and combining existing methodologies, such as using clustering in a two-step approach to logistic regression or overcoming challenges stemming from real world data. These approaches can be used in other contexts, making the findings of this thesis generalisable. Each Chapter will therefore include a section in its concluding remarks which discusses avenues for future research and generalisation of clustering as used in that case scenario. The challenge of combining data sources is another important issue addressed by this thesis. The thesis will contribute to the knowledge on the combination of data of different formats, such as time series and geospatial information in Chapter 3, and the combination of data from different sources such as search engine data, country characteristics and flight information in Chapter 5. This is considered a crucial step, given the increasing amount of data available while also taking into account the challenges of different data formats and standards this brings with it. This thesis thus aims to develop and present the use of clustering in such a way that is useful for both real life practitioners and researchers working in close-to-practice areas. The relevance of the methodological advancements made within this thesis is always seen in their usefulness to real world scenarios. This bridging between theory and practice and the call for data driven and practice lead research is a key overarching objective of the thesis.

We will follow the following structure: The now following Chapter 2 will explore the background literature on clustering as a technique with a focus on its application to spatial, temporal and spatio-temporal data. Then, Chapters 3, 4 and 5 will present the application cases to public health, SME financing and tourism respectively. Each chapter can be seen as an independent piece of work as well as one building block of this thesis, forming a coherent picture of the ways clustering can contribute to decision making and explorative data analysis. Lastly, we will conclude with Chapter 6 which summarises the findings and puts them into perspective based on this motivational introduction.

By bringing together methodologies and empirical findings from three different areas that use the same underlying methodological foundation of clustering, this thesis aims to contribute to the movement towards utilising machine learning capabilities in organisations. At the same time, we believe that real world applications add an important dimension to methodological thinking which take into consideration whether a method works in practice or which steps have to be undertaken to make it do so. The conversation about machines thinking might be ongoing, but equally important might be the conversation about how we think about machines and their role in contributing to our knowledge development.

A note on notation

The three main chapters, 3, 4 and 5, should be seen as independent parts of this work that work together to support the overall thesis notion. Due to that nature, while efforts have been undertaken to ensure notation consistency, that might at some points not be possible resulting in slight deviations. These instances are indicated as such and variables are clearly defined. In this sense, the notation will apply throughout the whole thesis unless noted that it refers only to a specific chapter.

Chapter 2

Background

2.1 A brief overview of machine learning

Murphy (2012, p.1) define machine learning as “a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty”. As such, machine learning can be seen as a large and growing collection of different computational approaches to data analysis, differing in their objective and how they approach a problem. Generally, we distinguish between supervised and unsupervised learning. While supervised machine learning relies on a labeled data set to learn structures from, unsupervised methods automatically detect structures which are deemed interesting. In the following sections some examples for common methods within supervised and unsupervised classification are introduced, before this thesis focuses on a subset of methods within unsupervised machine learning, clustering.

2.1.1 Supervised classification

Supervised learning utilises a labeled data set called training data to learn patterns from and then applies them to new or unlabeled data. Common problems within this category are classification ones, where based on a feature set the categorical class of a new data point is predicted. We distinguish between different kinds of classification depending on the input and output data, such a binary classification where the predicted class is binary and multiclass classification which allows for more than two possible outcomes (Murphy, 2012, p.3). Examples for algorithms include discriminant functions, which assign an input vector $\mathbf{x} = x_1, ..., x_N$ to one of k classes C_k (Bishop, 2006, p.181). A linear discriminant function for binary classification with $k = 2$ takes the form

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 \tag{2.1}$$

with weight vector $\mathbf{w} = w_1, ..., w_N$ and bias w_0. Input vector $\mathbf{x}$ is then assigned to class C_1 if $y(\mathbf{x}) \geq 0$ and class C_2 in all other cases. Extending the linear discriminant function to $k > 2$ classes gives us a generalisation of Equation 2.1 as follows:

$$y_k(\mathbf{x}) = \mathbf{w}_k^T\mathbf{x} + w_{k0} \tag{2.2}$$

which assigns $\mathbf{x}$ to a class C_k with $y_k(\mathbf{x}) > y_j$ for $j \neq k$ (Bishop, 2006, p.183).

In classification problems the input data can be numerical, but they can also be categorical labels, images, text or time series. Another example for supervised learning is regression, where the outcome value is continuous (Murphy, 2012, p.2). In its simplest form a linear regression is a linear combination of input variables for the prediction of a continuous output variable as shown in Equation 2.3 for input variables $\mathbf{x}$ and parameters $\mathbf{w}$:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1x_1 + ... + w_Nx_N \tag{2.3}$$

Multiple variations on this basis formula are possible which then allow for example for non-linear combinations (Bishop, 2006, p.138). The main difference between the linear regression in Equation 2.3 and the linear discriminant function in Equation 2.1 is in their output variable. Linear regressions predict a continuous value while discriminant functions focus on classification problems and therefore predict membership in a categorical class (Bishop, 2006, p.180).

Logistic regression is an approach which utilises approaches from both classification and regression, but is generally regarded as a classification algorithm in the machine learning literature

(Bishop, 2006, p.205; Murphy, 2012, p.247). In machine learning it can be interpreted as a probabilistic generative model for a two-class classification problem where we describe the posterior probability of class membership in C_1 for a feature vector $\mathbf{x}$ and parameter vector $\mathbf{w}$. For this purpose, logistic regression uses a sigmoid function, in the following denoted as logit(s), to act on the linear function of the feature vector. This allows logistic regression to map the output value to a probability, which can then be used to predict one of two (or in some cases multiple) classes. The model takes the following form for class C_1:

$$p(C_1|\mathbf{x}) = y(\mathbf{x}) = \text{logit}(\mathbf{w}^T\mathbf{x}) \tag{2.4}$$

and the equivalent for class C_2 as

$$p(C_2) = 1 - p(C_1|\mathbf{x}). \tag{2.5}$$

logit(s) in Equation 2.4 is the logistic sigmoid function which is defined as

$$\text{logit}(s) = \frac{1}{1 + \exp(-s)} \tag{2.6}$$

(Bishop, 2006, p.197-205).

The parameters in the model shown in Equation 2.4 can be estimated for example using maximum likelihood. A very similar approach to logistic regression is a probit model. The main difference between the two is in the link function, where the logistic regression model uses the logit transform and the probit model utilises the inverse normal cumulative distribution function, but outcomes are often indistinguishable. The preference of one over the other in many cases seems to comes down to personal preference within an area of research, with some exceptions when analysing extreme behaviour. The difference between logistic regression and probit models will be touched upon again in Chapter 4 where a spatial regression model will be used for for predicting access to finance of companies.

2.1.2 Unsupervised classification

In contrast to such supervised learning approaches, which uses training data to learn patterns from, unsupervised learning does not require pre-labelled training data. Instead, patterns are recognised which are deemed "interesting" (Murphy, 2012, p.2). Unsupervised methods include self-organising maps (SOMs). This method, also called a Kohonen map after its original author (Kohonen, 1982, 1990), maps high dimensional vectors onto a two-dimensional space in such a way that close locations of nodes in the two dimensions indicate closeness of vectors (Bishop, 2006, p.598). Bishop (2006) note, however, that as SOMs do not optimise a cost function, their parameters are difficult to set and their convergence is difficult to assess, with the self-organisation of the SOMs depending on the former.

Another group of algorithms within unsupervised learning is called cluster analysis. In this thesis, we will focus on this specific group of methods and demonstrate their application in three different contexts. The following chapter will introduce the reader to the existing methods in clustering through a literature review in Section 2.2. A special focus will be put on the additional challenges posed by spatial and temporal data in Sections 2.3 and 2.4 respectively, as well as their combination in spatio-temporal data in Section 2.5.

2.2 Cluster analysis

Clustering is an unsupervised machine learning technique for grouping data points which finds and groups data structures according to their (dis)similarity. Its nature as an unsupervised method means the algorithms do not require pre-labelled training data for identifying these groups. This makes clustering especially interesting for contexts in which little is known about an existing group structure of the data.

2.2.1 Dissimilarity measures

As the membership in clusters is dependent on the definition of dissimilarity, a brief introduction to ways of defining it is necessary in order to understand how different algorithms operate. The interested reader is recommended the extensive chapter on proximity measures given in Xu and Wunsch (2008) for a more detailed review. Given are two data points i and j which are defined by a set of n features each. These features can be measured on different scales which will impact the way we define the dissimilarity of the two sample points. Generally, a feature can be defined as binary, discrete, or continuous, where a binary feature has two possible outcomes only, a discrete feature has a finite or at the most a countable infinite number of possible outcomes, and a continuous feature has an infinite and uncountable number of possible values. The values can also be defined by their comparability to each other, with nominal or categorical values having no mathematical interpretation or possibility of comparison, ordinal values only carrying a rank order but no definition of the size of the difference, interval enabling some degree of comparison but no absolute zero, and ratio which allows for the ratio of two numbers to be taken. Furthermore, the n features can be all of the same type or mixed, and they can be dependent or independent from another (Xu & Wunsch, 2008).

For binary features the dissimilarity can be calculated using a matching coefficient which counts matching occurrences of the features for the two data points. If both an occurrence and a non-occurrence are considered equally important then the Simple Matching Coefficient for our example with two data points i and j is defined as

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{n_{11} + n00}{n11 + n00 + n10 + n01} \tag{2.7}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are n-dimensional vectors. n_{11} and n_{00} are the numbers of simultaneous occurrence or non-occurrence for i and j, and n_{10} and n_{01} are mismatches. If only occurrence is considered an important factor for similarity, the Jaccard coefficient can be calculated as

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{n_{11}}{n11 + n10 + n01} \tag{2.8}$$

where simultaneous non-occurrence is not considered.

For multiple discrete features the Simple Matching Coefficient can be generalised to

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{n} \sum_{l=1}^{n} S_{ijl} \tag{2.9}$$

where $S_{ijl} = 1$ if i and j match in the l-th feature and $S_{ijl} = 0$ if not.

For continuous variables the most well known dissimilarity measure is the Euclidean distance and its generalisation the Minkowski distance. For our two data points i and j the Minkowski distance D is defined as

$$D(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^{n} |x_{il} - x_{jl}|^{1/p} \right)^{p}. \tag{2.10}$$

The Euclidean distance presents the case in which $p = 2$, while $p = 1$ gives us the Manhattan distance.

The similarity of two points with mixed data types can be calculated for example using the Gower distance (Gower, 1971). This distance measure will be used in Chapter 4 for calculating the similarity of SMEs based on mixed type characteristics. We assume that we have n attributes of mixed type. The Gower distance for points i and j is then defined as

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{l=1}^{n} \delta_{ijl} s_{ijl}}{\sum_{l=1}^{n} \delta_{ijl}} \tag{2.11}$$

where s_{ijl} is the similarity at the lth feature between objects i and j. δ_{ijl} is a binary coefficient which is 1 if the lth feature is nonmissing for both x_i and x_j and it is zero otherwise.

For discrete and binary variables, s_{ijl} is then defined as

$$s_{ijl} = \begin{cases} 1, & \text{if } x_{il} = x_{jl} \\ 0, & \text{if } x_{il} \neq x_{jl}. \end{cases}$$

And for continuous variables, s_{ijl} is given as

$$s_{ijl} = 1 - \frac{|x_{il} - x_{jl}|}{R_l} \tag{2.12}$$

with R_l is the range of variable l.

For all cases above the dissimilarity is then typically defined as $D(\mathbf{x}_i, \mathbf{x}_j) = 1 - S(\mathbf{x}_i, \mathbf{x}_j)$.

All similarity and distance measures discussed above give the same weight to every attribute, that is, they do not assign a higher importance to one feature over the over. Weighted dissimilarity measures exist also, for example in the form of a weighted Euclidean distance, and they allow for features to be weighted based on pre-existing domain knowledge.

Another note should be made here about the importance of scaling and normalising continuous variables. If variables are measured at different scales, distance measures such as the Euclidean distance will not be able to accurately capture similarity between points. For example, in Chapter 5 three different dimensions of similarity will be considered for categorising countries. With interest on a scale of 0 to 100, affordabiltiy being reported as varying around a value of 100, and reachability measured on a scale of 0 to 9999, min-max normalisation will be employed to make these values comparable.

2.2.2 Clustering algorithms

The choice of algorithm depends on the type and amount of data, assumptions made about the shape and nature of the clusters, as well as on the objective. One way of classifying clustering algorithms is depending on the way they assign the data points to their groups. Common categories of algorithms for this include partitioning, hierarchical, probabilistic, graph-based and density methods. This list is by no means exclusive and recent developments have seen for example the application of neural networks to cluster problems. Members of clusters can belong exclusively to one group, known as hard clustering, or be assigned to multiple groups expressed in degrees or probabilities, an approach known as fuzzy clustering (Xu & Wunsch, 2008).

Partitioning approaches

One of the most well-known approaches is the partitioning algorithm k-means which aims to partition the data into k groups. Its popularity largely stems from its computational efficiency and ease of implementation (Fränti & Sieranoja, 2019). It operates by minimising an objective function calculating the intra-cluster distances as the sum of squared errors (SSE) as shown in Equation 2.13 for a data set of size N. SSE is then defined as

$$SSE = \sum_{i=1}^{N} ||x_i - c_j||^2 \tag{2.13}$$

where x_i is one data point and c_j is its nearest centre point.

In doing so the algorithm aims to form clusters of data points which are very similar to other points within their cluster, but very different to those outside. This approach means that k-means operates best on data sets with well separated clusters (Fränti & Sieranoja, 2019).

When initialising the algorithm one has to choose the optimal number of clusters k. There are multiple approaches to this described in the literature which usually involve some kind of iterative calculation of results with multiple values of k and comparing the outcome regarding some measure of cluster quality. Three common approaches will now be explored in more detail: the Elbow criterion, the average silhouette measure, and the Gap statistic.

The Elbow criterion utilises the objective function in Equation 2.13 for optimising the number of clusters in such a way that adding another cluster does not remarkably reduce the SSE value. Plotting the SSE value for multiple values of k creates a lineplot as shown in Figure 2.1. The value of k where this plot has an “Elbow”, a point from which on the added benefit diminishes, is the optimal number of clusters for this dataset. In the case of the example shown in Figure 2.1, this would be $k = 4$.

The Elbow criterion has an element of subjectivity to it as it relies on the researcher making a choice as to where the Elbow lies. This can in some cases be challenging to determine. As an alternative to this, the average Silhouette value can be used. Silhouettes are a measure for

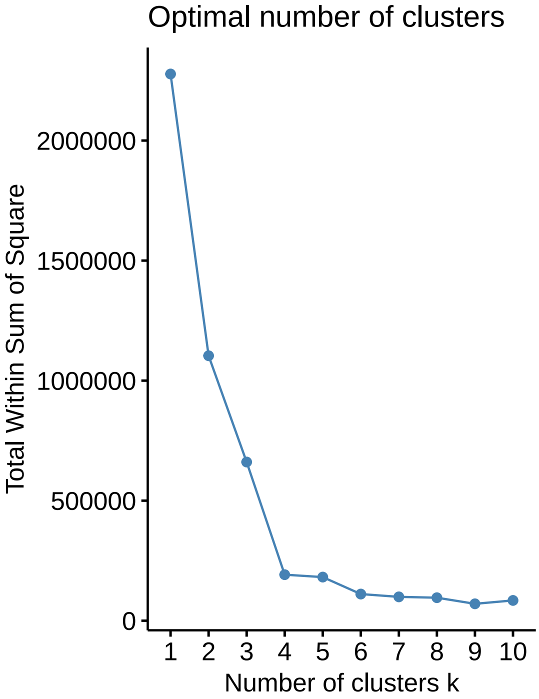

Figure 2.1: Example for an Elbow criterion plot produced with the `fviz_nbclust` function from package `NbClust` (Charrad et al., 2014) using the simulated dataset with four clusters introduced in Chapter 3. The x-axis depicts the number of clusters k with the y-axis describing the total within cluster sum of squares (SSE).

cluster quality first introduced by Rousseeuw (1987). They use a balance for tightness within and separation between clusters, which describes how well a point lies within a cluster. For an observation i which is member of cluster A the silhouette measure is defined as

$$s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]} \tag{2.14}$$

where $a(i)$ is the average dissimilarity of object i to all other members of its cluster A and $b(i)$ is the minimum dissimilarity of object i to a member of neighbouring cluster B. The range of $s(i)$ is given as $-1 \leq s(i) \leq 1$ where a good value indicating well separated clusters with small inter- and large intra-cluster dissimilarities would be close to 1 and and a worse value closer to -1. Plotting all individual silhouette values for a clustering gives the researcher a good idea of which specific observations are driving the quality of an outcome, but we can also use the average value as an overall measure of quality. This is how the silhouette measure is used for choosing the optimal number of clusters k. Figure 2.2 shows a plot of average Silhouette values for different values of k. As described above, the largest possible value is the optimal number for k.

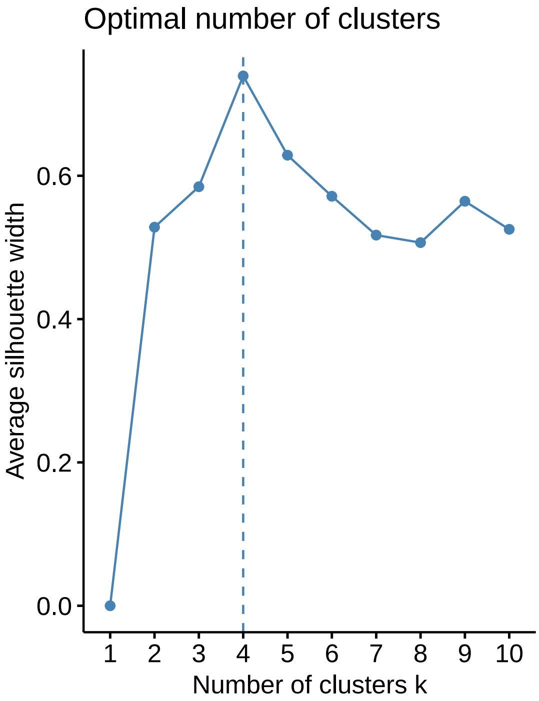

Figure 2.2: Example for an average Silhouette measure plot produced with the `fviz_nbclust` function from package `NbClust` (Charrad et al., 2014) using the simulated dataset with four clusters introduced in Chapter 3. The x-axis depicts the number of clusters k with the y-axis describing the average Silhouette width.

In 2001, Tibshirani et al. introduce what they describe as a statistic to formalise the heuristic that is the Elbow criterion. The Gap statistic for a number of clusters k is defined as

$$\mathrm{Gap}_n(k) = E_n^*\{\log(W_k)\} - \log(W_k) \tag{2.15}$$

where W_k is the within cluster sum of squares and E_n^* is the expectation value of W_k under a null reference distribution of the data. The value of k for which the Gap statistic is maximised is then chosen as the optimal number of clusters. An example for this is shown visually in Figure 2.3, where the optimal number of k is given as 5. This also illustrates a common problem in clustering where different methods for choosing k yield different results for the same data set.

The choice therefore continues to be somewhat subjective, even with objective measures such as the Gap statistic.

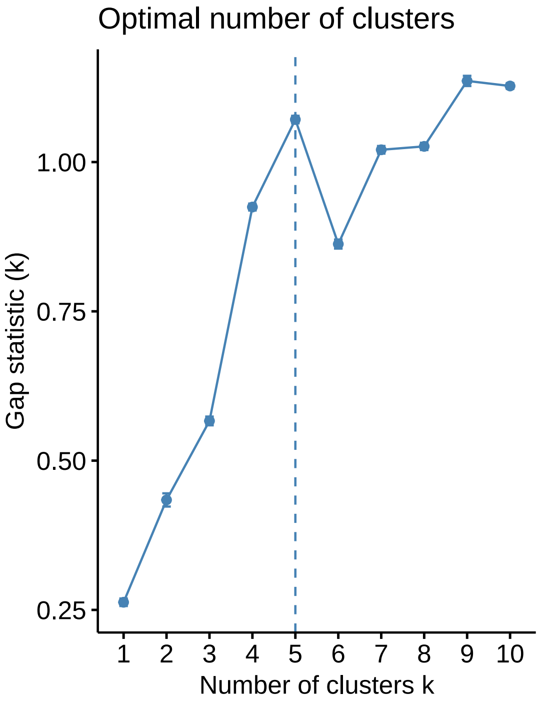

Figure 2.3: Example for a Gap statistic plot produced with the `fviz_nbclust` function from package `NbClust` (Charrad et al., 2014) using the simulated dataset with four clusters introduced in Chapter 3. The x-axis depicts the number of clusters k with the y-axis describing the Gap statistic values.

After choosing the number of k the algorithm sets these initial random cluster centres, often uniformly random across the data space, and assigns each data point to that centre which is closest to it. The original centre points are then recalculated and updated as the mean of the newly formed cluster. This process is repeated a fixed number of times or until the algorithm converges, meaning that the changes to the position of the centres are sufficiently small (Fränti & Sieranoja, 2019). Clearly, the position of the original cluster centres has a high impact on the found solution, meaning that this algorithm can suffer from inaccuracy (Celebi, Kingravi, & Vela, 2013). Through the updating step the algorithm is able to finetune the position of the clusters locally but can not do so globally over the whole data space (Fränti & Sieranoja, 2019). Thus choosing the right initial centres is a crucial step. Extensive reviews and comparisons of initialisation methods for k-means, for example by Celebi et al. (2013), demonstrates the high interest in overcoming this issue. One proposed solution is the $k++$-means or k-means++ algorithm introduced by Arthur and Vassilvitskii (2006) which, when choosing the centres at random, takes into consideration the shortest distance of the data point to an already chosen centre through a weighting system. This approach has become a popular adaptation to k-means however it suffers from scalability issues, prompting further developments into a scalable version employing Markov chain Monte Carlo (MCMC) (Bachem, Lucic, Hassani, & Krause, 2016a) and an assumption-free MCMC which overcomes potential issues around assumptions made when employing MCMC (Bachem, Lucic, Hassani, & Krause, 2016b). Other common approaches include the Maxmin approach by Kaufman and Rousseeuw (2009) or running the algorithm multiple times with random initial centres and choosing the best outcome (Fränti & Sieranoja, 2019).

Another popular partitioning algorithm is the Partitioning Around Medoids (PAM) (Kaufman & Rousseeuw, 2009). First presented at a conference in 1987 (Kaufman & Rousseeuw, 1987) and described in more detail in their book in 1990, this algorithm searches for k "representative objects" which function as the clusters centroids which are called medoids in the case of PAM. This stands in contrast to k-means, where centroids are artificial centre points which move when iteratively recalculating the new centres. For PAM, in a first step the algorithm selects these k medoids from among the data points and assigns each surrounding point to its nearest medoid. Then it calculates the average dissimilarity within the clusters and repeats the process with a different set of k. This average dissimilarity describes the quality of the clustering result as it calculates how similar the objects within each cluster are to each other. By selecting the solution with the lowest dissimilarity value PAM is able to find the set of k which optimises the within cluster quality. As an algorithm PAM accepts not only a set of data points to calculate dissimilarities between, but also dissimilarity matrices as input objects. As such, PAM can be used in cases in which dissimilarity matrices are constructed in a different manner than calculating direct distances between data points using for example Euclidean distances as is the case for k-means. This makes the algorithm more flexible regarding the data format, accepting for example binary, ordinal or mixed type data without issues. This is also the reasoning why we will use this particular algorithm in Chapter 5, where similarity between countries will be calculated by combining three features. Another advantage of the method comes with its functionality of identifying medoids as representative objects for each cluster, which can be a useful feature for describing the clusters.

Hierarchical approaches

Instead of finding one outcome to a clustering problem, hierarchical approaches produce tree-like structures called dendrograms (Han, Kamber, & Pei, 2011). There are two general approaches to this called agglomerative or divisive methods. A common algorithm for agglomerative clustering is AGNES, the Agglomerative Nesting algorithm, introduced by Kaufman and Rousseeuw (2009) in their book "Finding Groups in Data". It starts by assigning each data point to its own cluster and then subsequently merges the closest clusters in a step-wise process until all data points are in one cluster. The less common opposite approach of divisive hierarchical clustering, for example using DIANA, the Divisive Analysis algorithm (Kaufman & Rousseeuw, 2009), operates by starting with one cluster containing all data points and then step-wise dividing those points which are most dissimilar until each data point is again in its own individual cluster. The reason why divisive algorithms are far less common outside of the theoretical literature is the fact that the number of partitions which have to be considered is extremely large. While Kaufman and Rousseeuw (2009) refer to some approaches on how a limited number of divisions can be considered to restrict these computational expenses using heuristics, agglomerative approaches remain the norm in hierarchical clustering. For making a decision on which clusters to merge or divide the dissimilarity between groups of points has to be calculated using a so-called linkage criterion. The three most common linkage criteria are single, complete, centroid and average linkage, as well as Ward's criterion. Each of these criteria carries different advantages and disadvantages, making the choice similar to the "right" choice of a dissimilarity measure a matter of application context and result sought. For two clusters A and B the distance d is defined as follows:

- Single linkage calculates the distance between clusters as the minimum distance between two points of the two clusters. $d(A, B) = \min d(i, j)$ for $i \in A$ and $j \in B$ This approach tends to form elongated clusters.

- Complete linkage does the very opposite by defining the distance of the two clusters as the distance of the two furthest points, thus forming small and compact clusters. $d(A, B) = \max d(i, j)$ for $i \in A$ and $j \in B$.

- Centroid linkage defines the distance of clusters A and B as the Euclidean distance between their respective centroids. $d(A, B) = ||\bar{x}(A) - \bar{x}(B)||$. This linkage is not suitable for complex data where Euclidean distance can not be used.

- Weighted and unweighted average linkage (WPGMA and UPGMA (Sokal & Michener, 1958)) calculate the averages of the linkages between all points of the two clusters. For the unweighted variant and $|A|$ and $|B|$ the cardinality of clusters A and B, the distance is $d(A, B) = \frac{1}{|A| \cdot |B|} \sum_{i \in A} \sum_{j \in B} d(i, j)$.

- Ward's criterion uses a similar approach as the centroid linkage but includes another factor which takes into account cluster size. The result leads to this criterion using variance minimization as an indicator for cluster merging or dividing. $d^2(A, B) = \frac{2|A||B|}{|A|+|B|}||\bar{x}(A) - \bar{x}(B)||^2$.

Hierarchical clustering using the Ward criterion is also simply called "Ward's method". It has been used for example by Van Puyenbroeck, Montalto, and Saisana (2021) in order to cluster cities in the context of urban culture, indicating the usefulness of clustering in aiding policy making. Hierarchical approaches can also be used to decide the number of clusters for subsequent partitioning methods. Romano, Cambini, Fumagalli, and Rondi (2021) do so in the form of a **hybrid clustering procedure** which combines partitioning and hierarchical methods. Specifically, they begin by using Ward's method to cluster the Italian gas distribution network and the number of clusters is chosen using a stopping criterion. They then apply k-means using this number for k. Romano et al. (2021) do acknowledge, however, that this hybrid method suffers from sensitivity to outliers.

Probabilistic approaches

Mixture models can provide a framework for modelling complex probability distributions (Bishop, 2006, p.423). For example, a mixture of Gaussians can be used to model data for which a single Gaussian is not sufficient for capturing all structures. Such a Gaussian mixture is defined as

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2.16}$$

with K Gaussian densities, each with their own mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$. The parameters π_k are called mixing coefficients with

$$\sum_{k=1}^{K} \pi_k = 1 \tag{2.17}$$

and

$$0 \leq \pi_k \leq 1 \tag{2.18}$$

which means that they fulfill the requirements for being probabilities (Bishop, 2006, p.111).

Mixture models can now also be used for clustering data by assuming that each cluster is one of a mixture of subpopulations (Liao, 2005; Mai, Fry, & Ohlmann, 2018). This is done by extending the model in Equation 2.16 with a K dimensional binary random variable $\mathbf{z}$ where $z_k \in \{0,1\}$ and $\sum_k z_k = 1$. One element of z_k is equal to 1 and all other elements are 0 (1-of-K representation) (Bishop, 2006, p.430). With $p(z_k = 1) = \pi_k$ and $p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$ we can then write the conditional distribution of our data $\mathbf{x}$ given a particular value of $\mathbf{z}$ as

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k}{}^{z_k}) \tag{2.19}$$

while the marginal distribution is the same as shown in Equation 2.16. Clustering with mixture models now uses the conditional probability $p(z_k = 1|\mathbf{x})$ for grouping data: we interpret this conditional probability as the "responsibility that component k takes for explaining observation

x" (Bishop, 2006, p.432). In other words, the conditional probability gives us information about to which Gaussian component (i.e. cluster) a specific observation belongs.

One problem with mixture models lies in the estimation of all unknown distribution parameters at the same time. A commonly used solution for this is the expectation maximisation (EM) algorithm. Initial values for means, covariances and mixing coefficients of the Gaussian mixture are chosen and then two steps, expectation (E) and maximisation (M) are repeated iteratively. For the expectation step the posterior probabilities are calculated and evaluated with the current parameter configurations. The maximisation step then re-estimates the parameters. The two steps are repeated until convergence is reached, that is, the change in log likelihood or parameters falls below a threshold (Bishop, 2006, p.437). One weakness of the EM algorithm is the selection of initialisation values. While in the literature in many cases k-means is used to set these initialisations by finding cluster centres, Y.-T. Chen, Sun, and Lin (2020) comment that this approach has the disadvantage of k-means being affected by extreme values itself. Thus, Y.-T. Chen et al. (2020) present a novel EM algorithm for multivariate Gaussian mixture models which utilises concepts from density-based approaches discussed in Section 2.3.1. They demonstrate the usability of their methods through an application of the method to mobile data user behaviour of telecommunication companies.

Graph theory-based approaches

Graph theory-based clustering approaches define each datapoint as a node in a weighted graph and define the edges between them as their dissimilarity to each other. Different approaches now aim to cut these trees in such a way that they can define sub-trees of them as clusters. Schaeffer (2007) note that a common objective in graph clustering is for vertices within the clusters to be connected by internal edges. Internal edges are connections which only connect vertices within a cluster and not to different clusters. A high number of internal edges means that a cluster is densely connected in itself. There are two directions from which this problem can be approached: either the algorithm looks for connections which are weak links as they connect points too far away and chooses them to cut, for example by using a minimum spanning tree approach (Xu & Wunsch, 2008, p.81), or they approach the problem from the opposite perspective and define clusters as highly connected areas within a tree. This can also be thought of as either identifying empty space between clusters of points or identifying areas of high point density where clusters are located.

Zahn (1971) introduces a method within the former of the two categories. They begin by defining a spanning tree as a connected graph which contains all datapoints. They further define a minimum spanning tree as such a tree for which the sum of the weights of all the edges is minimal. In order to decide where a tree should be cut to separate clusters from each other, Zahn (1971) look for what they call inconsistent edges, that is edges which have a significantly larger weight than the average of the neighbouring edges. Cutting this edge will separate areas of closer points (i.e. clusters) from other areas. The result are individual graphs for each cluster. This approach is shown graphically in Figure 2.4. By detecting the inconsistent edge connecting clusters C_1 and C_2 and cutting it, two clusters are formed.

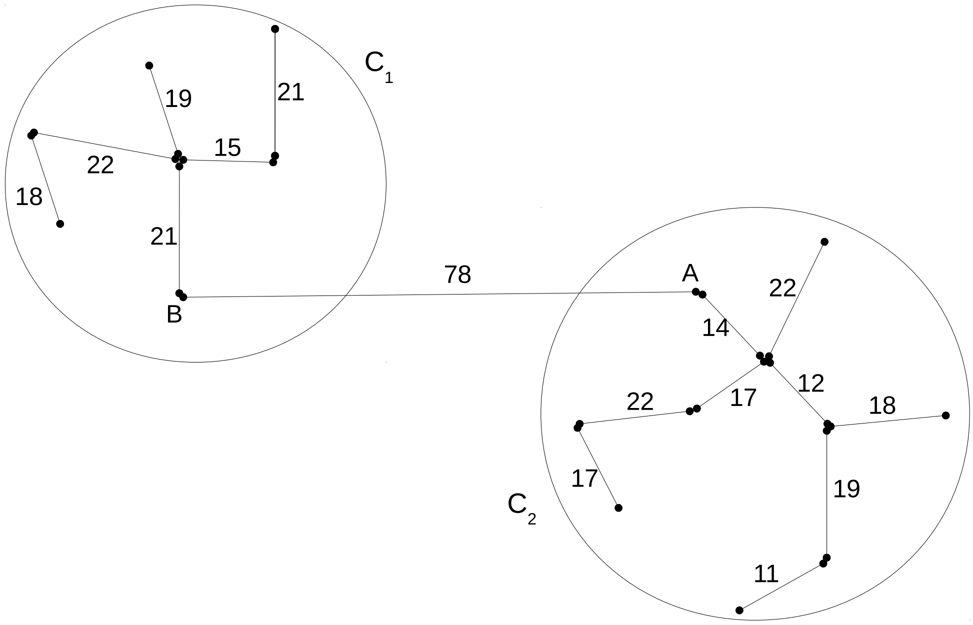

Figure 2.4: A minimum spanning tree and an inconsistent edge connecting two clusters C_1 and C_2. Own image based on Zahn (1971).

In contrast to that, Hartuv and Shamir (2000) introduce the Highly Connected Subgraphs algorithm which identifies areas within a tree which are highly connected and thus indicating clusters. The authors define high connectivity as subgraphs where the number of edges is more than half that of the number of vertices. The algorithm first uses a minimum cut function to divide the graph into subgraphs and then checks whether these subgraphs are highly connected. If that is the case they are being returned as clusters. While the algorithm can identify outliers (called "singletons" due to their nature as single vertices), it also allows for the later inclusion of these outliers in clusters. Another approach within the category of algorithms identifying highly connected areas as clusters are those which employ clique partitioning. In graph theory, a clique is such a set of vertices where there exists an edge between each pair of them (Bishop, 2006, p.385). Clique partitioning problems are not considered clustering algorithms, however their logic is very similar to that of a cluster problem as they also aim to identify a partition through an optimisation problem, similar to what has been described in Section 2.2.2. Specifically, they are trying to find such a partition that it maximises the sum of the edges within a clique. Indeed, Edachery, Sen, and Brandenburg (1999) introduce a popular distance-k clique clustering algorithm which makes use of this concept but extends the graph-theoretic concept of cliques, where connections must be direct, to allow for connecting paths of length k between members of a clique. Verma and Butenko (2013) also note that cliques are very restrictive in their definition. Due to a number of external factors some edges between points might be missing and a cluster would not be identified. They propose an approach which utilises k-communities, where all endpoints of edges within a subgraph must have a minimum number of k common neighbours in that subgraph. More recent developments in the area of clique relaxation include for example the introduction of a so-called small-world subgraph (Kim, Veremyev, Boginski, & Prokopyev, 2020).

Density-based approaches

Density-based clustering algorithms, such as DBSCAN (Ester, Kriegel, Sander, & Xu, 1996), use the distance measures to assess the density of points in a region to determine whether a cluster exists in that location. Identifying regions of high relative local density for cluster analysis is based on the logic that dense regions signify clusters which are divided by areas of lower density.

DBSCAN looks for these high-density neighbourhoods of points which are close to each other and segregated by areas of lower density by using different definitions for point connections. The relationships between points are defined as directly density connected, density connected and density reachable. Through the usage of these definitions DBSCAN is able to detect clusters of arbitrary shapes through a stringing along of cluster members, in a similar manner as hierarchical clustering with single linkage is performed however with a more limited length of such connection.

The algorithm first visits a random data point in data space D and searches for other points nearby. The ε-neighbourhood of a point p is defined as $N_\varepsilon(p)$, where

$$N_\varepsilon(p) = \{q \in D \mid \text{dist}(p, q) \leq \varepsilon\}. \tag{2.20}$$

Two parameters have to be pre-defined by the researcher, which are $N_\varepsilon(p)$ and η. They describe the spatial radius within which other points belonging to the same clusters are looked for and a threshold to describe a sufficiently dense area with. If $N_\varepsilon(p) \geq \eta$ the respective area is considered dense and a cluster is formed. Clusters consist of two types of points, p in Equation 2.20 is defined as a core point and subsequent points which are visited and assigned to the same cluster without fulfilling Equation 2.20 are considered border points. These assignments are relevant for the aforementioned definitions of point relationships.

A point q is *directly density-reachable* (DDR) from a point p if two conditions hold true, namely

1. $q \in N_\varepsilon(p)$ and
2. $|N_\varepsilon(p)| \geq \eta$,

with the second condition being called the *core point condition.* This means that q must be in the ε-neighbourhood of p, and p must be a core point. A point o is *density-reachable* (DR) from a point p if there exists a chain of points which are directly density-reachable from each other. This condition makes DBSCAN flexible for clusters of arbitrary shape due to a chaining effect which is, however, constricted to points which fulfil the two conditions for direct density-reachability. And lastly, a point t is *density-connected* (DC) to a point p if there is a point o which is density-reachable from both t and p.

Ester et al. (1996) then define a cluster C as a non-empty subset of a database D that fulfils the two conditions

1. $\forall p, q$: if $p \in C$ and q is density-reachable from p then $q \in C$, and
2. $\forall p, q \in C$: q is density-connected to p.

Another method using the concept of neighbours is CLARANS (Ng & Han, 2002), which is able to cluster not only points but also polygonal objects. There are two main advantages of density-based methods. The first is their ability to find clusters of arbitrary shape. The second is the fact that many density-based methods do not require the number of clusters as an input as does, for example, k-means. There are many adaptations to DBSCAN. These include the constraint-based C-DBSCAN (Ruiz, Spiliopoulou, & Menasalvas, 2007), which uses pre-knowledge about group memberships, sampling-based IDBSCAN (Borah & Bhattacharyya, 2004), which is especially suitable for very large spatial databases, and the spatio-temporal ST-DBSCAN (Birant & Kut, 2007). Another extension on the original DBSCAN algorithm is "Ordering Points to Identify the Clustering Structure" (OPTICS) (Ankerst, Breunig, Kriegel, & Sander, 1999), which produces not a clustering but an ordering of the database that represents the underlying clustering structure. Besides being a good tool for visualisation purposes, OPTICS's main advantage is its ability to handle varying densities by analysing the neighbours for each point. Similarly, "Locally Scaled Density Based Clustering" (LSDBC) (Biçici & Yuret, 2007) also uses neighbouring points, in this case to determine the local density maxima in regions which act as cluster centres. The algorithm, which is thought to be more robust towards parameter changes and background noise, connects density regions until the overall density falls below a pre-defined threshold. Rodriguez and Laio (2014) propose "Clustering by Fast Search and Find of Density Peaks" (CFSFDP), which identifies cluster centres by defining them as points with a high local density and a high distance to other such centres. Subsequent work includes approaches for finding the optimal values for these parameters (Mehmood, Zhang, Bie, Dawood, & Ahmad, 2016) and fuzzy implementations of the algorithm (Bie, Mehmood, Ruan, Sun, & Dawood, 2016).

The features or attributes which determine the membership to a cluster can vary depending on the context. They can present a combination of multiple features, including ones that are measured on different scales, and can carry certain assumptions made. Two special cases of this are being explored in the following two subsections.

2.3 Spatial statistics and clustering

Spatial statistics describes the use of statistical methods for the analysis of data in a typically two or three dimensional space. It is used for example in geospatial analysis, but also in other areas such as astrophysics where three dimensional patterns of stars and galaxies are being modelled. In their book on spatio-temporal data, Cressie (1992) explain one special property of spatial data based on an example of a spatial process. Assume $\mathbf{D}_s = \{s_0, s_0 + \Delta, ..., s_0 + 24\Delta\}$ which describes 25 observations $\{s_0, ..., s_{24}\}$ taken in space $\mathbf{D}$ where Δ describes the regular spacing with which the observations are taken along a transect. We neglect the temporal dimension and assume that observations are taken at a fixed time point t_0. The spatial process can therefore be described as:

$$Y = (Y(s_0), ..., Y(s_0 + 24\Delta))' \tag{2.21}$$

In spatial processes point independence is often replaced by including a spatial dependence. A simple way for doing this as described in Cressie (1992, p.19) is including a nearest neighbour dependence using Gaussian dependence distributions. Equation 2.22 utilises a Gaussian distribution with the mean set as the midpoint between points $Y(s_{i-1})$ and $Y(s_{i+1})$ and standard deviation $\sigma^2/1 + \phi^2$.

$$Y(s_i)|\{Y(s_j) : j \neq i\} \sim Gau((\phi/(1 + \phi^2))\{Y(s_{i-1}) + Y(s_{i+1}\}, \sigma^2/(1 + \phi^2)) \tag{2.22}$$

ϕ is then called the spatial dependence parameter which based on Equation 2.22 fulfills $|\phi| \leq 1$. This spatial dependence can be described as the correlation between spatial point i and its direct neighbouring point $i-1$, as shown in Equation 2.23.

$$corr(Y(s_i), Y(s_i - 1)) = \phi \tag{2.23}$$

with $i = 1, ..., 24$.

In this example, the observed value at each spatial location is dependent on the observed value of its directly neighbouring location. More generally speaking in spatial statistics, the dependency structure between points can also be described as not only affecting the direct neighbours but all points to diminishing degrees. This means that we assume that locations which are close influence each other more strongly than those further away from each other. The structure can be accounted for in different ways, for example through Markov random fields and mean-field models (Celeux, Forbes, & Peyrard, 2003; Zhang, 1992),as well as copula-based models as shown in combination with regression (Brunner, Furrer, & Favre, 2019; Musgrove, Hughes, & Eberly, 2016). Others have treated this interdependency by implementing a network setting (Hu, Li, Guo, van Gelder, & Shi, 2019) or used Self Organising Maps (SOMs) (P. Agarwal & Skupin, 2008; Andrienko et al., 2010; Kohonen, 1982, 1990).

2.3.1 Approaches to spatial clustering

Spatial clustering describes using cluster analysis on spatial data. For example, Coll, Moutari, and Marshall (2014) use hierarchical clustering to investigate patterns of vehicle collisions within an identified hot spot area in Northern Ireland. Copulas can also be used in spatial clustering to account for dependency structures through copula-based clustering (Di Lascio, Durante, & Pappada, 2017; Disegna, D'Urso, & Durante, 2017; Marbac, Biernacki, & Vandewalle, 2017). An important theme in spatial clustering is point density, as regions of high density are interpreted as clusters of points. Density in this context is usually defined as a region where the distance or dissimilarity between individual points is relatively lower compared to the surrounding area. Spatial regions of density have for example been used in criminology for routing of police vehicles to target specifically areas of high crime density (Moews, Argueta, & Gieschen, 2021). Accordingly, one group of approaches commonly used for clustering spatial data are density-based methods such as the DBSCAN algorithm (Ester et al., 1996) which was previously introduced in Section 2.2.2. DBSCAN as an algorithm is intuitively particularly useful for spatial data due to its use of neighbourhoods. It forms clusters by detecting other points in the two-dimensional vicinity of a reference point.

While density-based clustering methods such as DBSCAN are applied to a broad range of research areas, including epidemiology Gomide et al. (2011) and the analysis of road traffic Anbaroglu, Heydecker, and Cheng (2014), the algorithm and its adaptations like ST-DBSCAN face problems when trying to detect clusters in data spaces with varying point densities. This is due to the use of global parameters when looking for close neighbouring points that indicate existing clusters. If a data space has regions of high point density and low point density in different areas, clusters can not be found in both these areas at the same time. The algorithm is thus not able to find both clusters which are dense and clusters which are less dense, or in other words, clusters with a low average dissimilarity and clusters with a high average dissimilarity. In a spatial context this problem arises frequently if data is not collected or available in an even spatial distribution pattern, for example due to differences in more populated areas versus less ones. Density can mean different things in different contexts: a relative dense region which the algorithm should detect as a cluster in a city can look very different than a relative dense region in more rural areas. An urban cluster would likely have a higher density, defined as a lower average distance between points within the cluster, than a rural cluster. This issue has been addressed in the literature, including by Kriegel, Kröger, Sander, and Zimek (2011). They explain how density-based clustering methods often label points in less dense areas as noise or outliers when there are regions of high density separated by areas of lower density. As potential solutions to this problem, they propose hierarchical approaches of ordering points such as OPTICS (Ankerst et al., 1999), as well as nearest neighbour approaches (Ertöz, Steinbach, & Kumar, 2003; Pei, Jasra, Hand, Zhu, & Zhou, 2009). Birant and Kut (2007) discuss varying densities in their paper on ST-DBSCAN and outline the problem of identifying actual noise in a scenario of high-density areas surrounded by regions of lower density. They propose assigning each cluster a density factor, calculated by looking at the maximum and minimum distances within a respective cluster. This single factor might, however, not reflect the various different density regions which are present across each cluster, especially if the area described by one factor is relatively large. Similarly, Zelnik-Manor and Perona (2004) discuss the idea of local scaling in the context of their self-tuning spectral clustering algorithm. Based on this, Biçici and Yuret (2007) combine this local scaling with an adapted density-based clustering algorithm. Their density estimation, however, is based on a k-nearest neighbour approach requires the researcher to decide on the number of considered neighbours k.

2.4 Time series analysis and clustering

A time series is a set of sequential data which has been collected either at fixed intervals or continuously. A number of analytical methods have been developed especially for time series. They can be used for analysing factors and properties of the series such as its volatility, as well as the prediction and forecasting of time series data. This thesis is focused on clustering and the main issue to solve therefore lies in the comparison of two time series through using a suitable dissimilarity measure. However, a brief review of the methods used in forecasting will be provided first.

2.4.1 Time series models

Similar to the previously described spatial data properties in Section 2.3, time series data is often thought to be autoregressive or dependent on its own previous observations. One of the simplest time series models is called a random walk. It assumes that the current value is the previous value plus a random noise term, and it thus can be described as

$$Y_t = Y_{t-1} + W_t \tag{2.24}$$

with $t = 0, 1, ...$, for a time series $\{Y_t\}$ and a random noise process W_t which is independently normal distributed with mean $\mu_w = 0$ and variance σ_W^2.

If we assume this dependence to reach further than one step back in time, we can define an autoregressive (AR) process of order p with

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} + W_t \tag{2.25}$$

where $t = p, p+1, ...,$ W_t a random noise process with mean $\mu_w = 0$ and variance σ^2_W, and $\{\alpha_i : i = 1, ..., p$ is an unknown parameter set (Cressie, 1992, p.87). These parameters of the model, $\alpha_1, ..., \alpha_p$ and σ^2_W, can be estimated for example using ordinary-least-squares or maximum-likelihood approaches.

We can also assume the time series to be a realisation of a moving-average (MA) process of order q with

$$Y_t = W_t + \beta_1 W_{t-1} + ... + \beta_q W_{t-q} \tag{2.26}$$

where $\beta_1, ..., \beta_q$ are unknown parameters to be estimated and W_t is a random noise process with mean $\mu_w = 0$ and variance σ^2_W. A key property of this process is the fact that the terms of $\{W_t\}$ are random and unobservable. That means that we can assume for the time series $\{Y_t\}$ that $E(Y_t) = 0$ and $var(Y_t) = \sigma^2_W \sum_{i=1}^{q} \beta_i^2$ (Cressie, 1992, p.91). Again, maximum-likelihood approaches can be used to estimate the parameters, however Cressie (1992) note that due to nonlinearity of the parameters in the autocovariances of the process explicit solutions are not available.

If we combine the autoregressive (AR) and the moving average (MA) aspects discussed above and further add allowance for non-stationarity in the mean through differentiation, we arrive at the well-known ARIMA model. Introduced by Box (1970), the model continues to find application in a range of fields, most recently in epidemiological analysis of the COVID-19 pandemic (Benvenuto, Giovanetti, Vassallo, Angeletti, & Ciccozzi, 2020; Kufel et al., 2020).

2.4.2 Time series clustering

Clustering can be employed in time series analysis (De Angelis & Dias, 2014; Serra & Arcos, 2014). For clustering such data, emphasis in the literature has been put on the selection of the most suitable dissimilarity measure. In general, there are four ways of measuring the dissimilarity between time series, as reviewed by Serra and Arcos (2014).

Lock-step methods such as the Euclidean distance measure similarity by comparing values at the same time steps in each series, their main advantage being computational simplicity (Xing, Pei, & Philip, 2012). For two time series A and B, with values a_t at $t = 1, ..., T$ and b_t with $t = 1, ..., T$, the Euclidean distance is calculated as

$$d(A, B) = \sqrt{\sum_{T}^{t=1} (a_t - b_t)^2} \tag{2.27}$$

which compares the two time series by calculating the dissimilarity at each matching time step t.

J. Lin, Khade, and Li (2012) comment that lock-step methods have the disadvantage of being computationally expensive. They argue that methods which consider each time step individually, as shown in Equation 2.27, instead of using derived features, work well only for short time series. Two other disadvantages according to J. Lin et al. (2012) are the requirement that both time series are of the same length and the fact that this distance measure is sensitive towards even small shifts. If the same pattern is repeated in two time series but occurring even one time step later this would affect the dissimilarity measure.

Feature-based methods employ the same approach, but work on a transformation of the time series. This can, under some circumstances, improve the accuracy of the dissimilarity compared to the raw Euclidean distance (Montani, Portinale, Leonardi, Bellazzi, & Bellazzi, 2006; Serra & Arcos, 2014). This approach has been used by Inniss (2006) for seasonal clustering of time series applied to weather and aviation data. Räsänen and Kolehmainen (2009) comment on the suitability of feature-based time series clustering for grouping customers by their electricity usage that while the method performs well, the correct selection of the features is very important. X. Wang, Smith, and Hyndman (2006) follow a similar approach to Räsänen and Kolehmainen (2009) by using statistical measures as extracted features, which they consider an especially suitable approach for long time series where other methods might not perform well due to computational expense.

Model-based approaches assume an underlying model generating the time series and measure dissimilarity based on the model parameters. For example, mixture models are used by Povinelli, Johnson, Lindgren, and Ye (2004), who comment that the method is able to perform well on a range of application areas with minimal input tuning, and Dias, Vermunt, and Ramos (2015) use hidden Markov models to account for time-constant unobserved heterogeneity and hidden regimes within time series.

Elastic methods compare the time series by their overall shape over time instead of their absolute values measured at each time step. For example, Dynamic Time Warping (DTW) is one such elastic method and it aims to optimally align the two time series by minimising a warping cost. Warping is described as the alignment of values at different time steps, allowing for example for parts of the time series to be stretched to longer intervals to allow for alignment. The algorithm takes a warping window parameter which allows for tuning of how much warping is allowed, with a large window aligning far away points in time and a window of zero corresponding to the Euclidean distance case shown in Equation 2.27 (J. Lin et al., 2012). DTW has been used in various contexts, including for incomplete medical time series (Tormene, Giorgino, Quaglini, & Stefanelli, 2009) and to measure the dissimilarity between time series in combination with fuzzy clustering (Izakian, Pedrycz, & Jamal, 2015). DTW is also frequently used for functional data (K. Wang, Gasser, et al., 1997), for example in Arribas-Gil and Müller (2014) where, for computational efficiency, it is first used for pairwise and then global alignment of curves. This computational cost has been discussed by J. Lin et al. (2012) as one of the main disadvantages of this method. This method has been used in Gieschen, Ansell, Calabrese, and Martin-Barragan (2021) and will be used in the associated Chapter 3, as well as Chapter 5. We refer the interested reader to those chapters for a more in-depth discussion of its application and statistical background. Other elastic methods include “Edit Distance on Real Sequence” (EDR), which calculates dissimilarity as the number of edit operations necessary for transforming one time series into the other (L. Chen, Özsu, & Oria, 2005). *k*-Shape and *k*-MS are iterative time series clustering algorithms which use shape-based distance (SBD) as their dissimilarity measure (Paparrizos & Gravano, 2017).

2.5 Spatio-temporal analysis

Spatio-temporal data describes data which has both a spatial and a temporal dimension. This type of data can therefore become very complex, as spatial considerations such as spatial autocorrelation can be included at the same time as temporal autocorrelation or different approaches to dissimilarity calculation. In their review of approaches to spatio-temporal clustering, Kisilevich, Mansmann, Nanni, and Rinzivillo (2009) outline how methods within this group can be classified depending on the nature of the spatial and temporal components (Figure 2.5). The Figure shows on the x-axis the increasing complexity on the temporal dimension from single instances to continuously updated time series. On the y-axis, the Figure describes increasing complexity in the spatial dimension from fixed spatial locations to moving ones. The different types of spatio-tempotal data are described below.

Spatio-temporal (ST) events describe data which has both a temporal and a spatial dimension, however both dimensions are static and are captured as single moments in time and space. Such events are for example natural disasters which occur at a fixed spatial location and a fixed point in time. They can be clustered according to their closeness in both space and time, with methods such as ST-GRID and ST-DBSCAN (M. Wang, Wang, & Li, 2006), the latter of which will be discussed in more detail later in this section. **Geo-referenced variables** describe situations in which the spatial locations are static, but there are multiple observed values of some variables over time. In some cases, the number of values can be fairly limited. Kisilevich et al. (2009) describe how in these situations a special area of interest can be clusters changing over time with each step-wise update of the temporal data. If the complete history describing

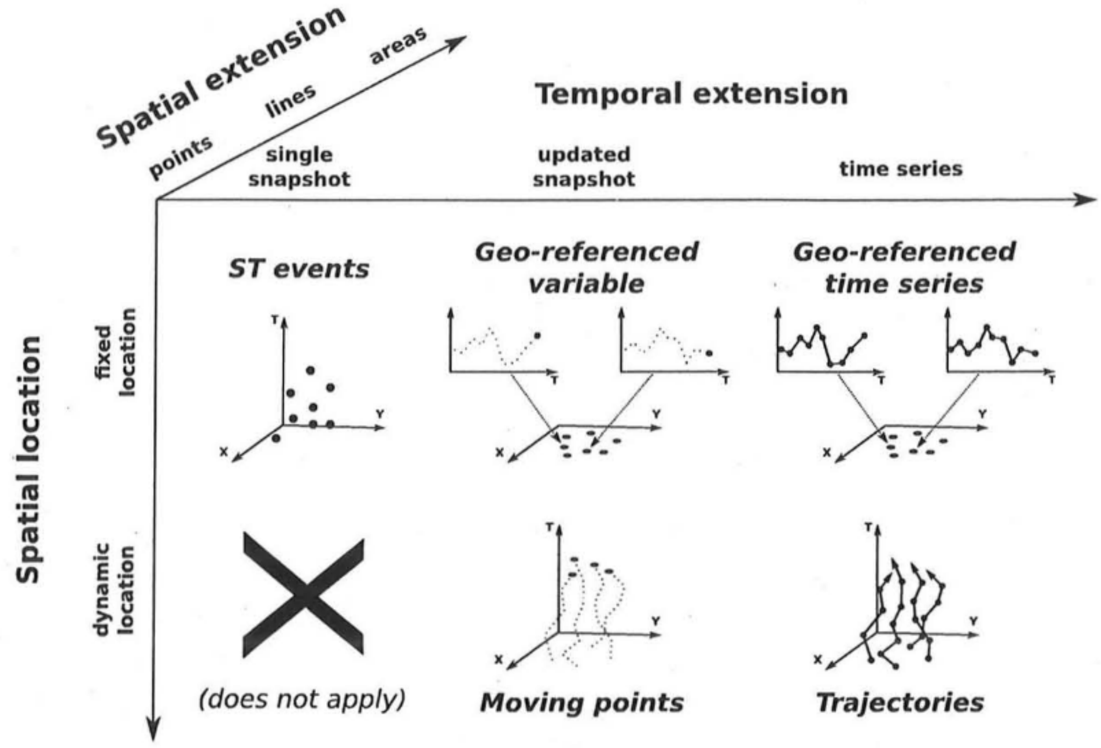

Figure 2.5: Types of spatio-temporal data. The x-axis describe the complexity of the temporal dimension, the y-axis the complexity of the spatial dimension, and the z-axis hints towards further complexity that can be added through the different nature of the observed objects themselves. Figure from Kisilevich et al. (2009).

the values of this variable are available this can be described as a **geo-referenced time series**. Methods within this group of data often employ elements from both spatial and time series clustering as outlined earlier in Sections 2.3 and 2.4. As an example for how this is approached, Disegna et al. (2017) introduce the COpula-based FUzzy clustering algorithm for Spatial Time series (COFUST) which takes into account both spatial dependency structures and time series dissimilarity through copulas.

Moving points and **trajectories** methods deal with highly complex movement through space and time. Examples for approaches include the Dynamic Time Warping Fuzzy-Medoids for Spatial–Temporal Trajectories (DTW-FCMd-STT) clustering algorithm (D'Urso, De Giovanni, Disegna, & Massari, 2019), approaches which utilise visual methods (Andrienko et al., 2009), and those which make use of density-based systems (Nanni & Pedreschi, 2006). Kisilevich et al. (2009) have a more in-depth discussion of the different methods available for trajectory clustering, to which we refer the interested reader.

We will focus in this thesis on the first row of Figure 2.5, with Chapter 3 working with geo-referenced time series, Chapter 4 with ST events and Chapter 5 with what can most accurately be described as geo-referenced variables.

2.5.1 Spatio-temporal clustering

One approach to spatio-temporal clustering which we will look at in more detail now is an adaptation to the DBSCAN algorithm (Section 2.2.2) called ST-DBSCAN (Birant & Kut, 2007). The main difference to DBSCAN lies in the use of not one but two neighbourhood parameters ϵ_1 and ϵ_2 and, following that, two neighbourhoods N^1 and N^2. Where ϵ_1 continues defining the spatial neighbourhood of the data, ϵ_2 does the same for the temporal neighbourhood. In order to be considered a member of a cluster, a data point q has to be in both the spatial and the temporal neighbourhood of point p. Accordingly, $dist_1$ will refer to the spatial distance and $dist_2$ to the temporal dissimilarity between points p and q. This extends Equation 2.20 as illustrated in Equation 2.28.

$$
\begin{aligned}
N^1_{\varepsilon_1}(p) = \{q \in D \mid \mathrm{dist}_1(p,q) \leq \varepsilon_1\} \text{ and} \\
N^2_{\varepsilon_2}(p) = \{q \in D \mid \mathrm{dist}_2(p,q) \leq \varepsilon_2\}.
\end{aligned}
\tag{2.28}
$$

With the two neighbourhoods a point q is directly density-reachable from a point p if

1. $q \in N^1_{\varepsilon_1}(p) \cap N^2_{\varepsilon_2}(p)$ and
2. $|N^1_{\varepsilon_1}(p)| \geq \eta$ and
3. $|N^2_{\varepsilon_2}(p)| \geq \eta$.

What that means practically is that two data points can only be in the same cluster if they are similar both in terms of spatial closeness and in terms of temporal similarity. The method continues suffering from a major drawback that the original DBSCAN has shown in terms of choosing a single parameter ϵ_1 for a global data space, which was the main motivation to us introducing the adaptation of the algorithm shown in Chapter 3.

2.6 Conclusion and research gaps

Clustering allows the researcher to group together unlabelled data points based on a measure of similarity or closeness. Different algorithms have been proposed in the literature depending on the data type and objective. As we have seen, spatial clustering and time series clustering present unique challenges which gave way to specialised methods for them. There are, however, also similarities between the two, namely in the autocorrelation or contagion effects and in the decision of what constitutes closeness or similarity. The simultaneous consideration of both spatial and temporal data for cluster analysis resulted in spatio-temporal clustering. Through a survey of the literature, we have identified a number of existing gaps that we intend to address in this thesis:

While a popular algorithm for spatial data, DBSCAN (Section 2.3.1) suffers from problems when presented with spatially unevenly distributed data. This is a common problem which can occur in real-life situations, as data in the real world is in many cases not evenly spread out. In a social science context this might be reflected by situations such as urban and rural settlements and how a population is spread out across a country. Furthermore, we have seen that temporal data requires its own considerations when being clustered in Section 2.4.2. The ST-DBSCAN algorithm extends the use of DBSCAN into the spatio-temporal realm, however, methods which consider shifted and warped time series such as DTW have not been used in connection with this approach. As shifted time series are a common complication of real life data, we consider this an additional gap to address in the pursue of the overarching objective of this thesis of making clustering methods applicable to real life scenarios. We will address these two gaps in Chapter 3 of this thesis.

Section 2.1.1 introduced the reader to the way that logistic regression can be used to predict a binary outcome based on a number of explanatory variables. When connecting this concept to spatial statistics, spatial logistic regression can be used to include autocorrelation to explain how effects are spread spatially with diminishing effects with an increasing distance. This model, however, assumes that all data points affect each other based on their distance to each other. In real-world scenarios, one can imagine a number of situations in which these effects only affect a limited number of close data points depending on some other factor which creates a connection between them or functions as a common factor impacting them. We will present a possible solution to this research gap in Chapter 4.

Lastly, we have seen the impact the data type has on the methodology as well as the importance of the concept of closeness to clustering methods in Section 2.2.1. The combination of data from multiple sources has become a crucial step for many research projects. This can be an even bigger challenge if the data comes in different formats, scales, is of different length in the case of timer series, or experiences missing values. As a major challenge in the application of machine learning techniques to real-world data, we will address some of those challenges and present possible solutions in Chapter 5. In this Chapter, we will also introduce a novel concept of reachability which addresses the question of what constitutes closeness in geospatial situations when geo distance is not sufficient.

Chapter 3

Modelling antimicrobial prescriptions in Scotland: A spatio-temporal clustering approach

This chapter utilises clustering in an effort to derive group structures of Scottish GPs based on their geographical location as well as their prescription behaviour. In doing so, we will demonstrate that spatio-temporal clustering can be used to take into account both geographic and time series variables and produces valuable insights for policy makers. The data provides additional challenges due to Scotland's geography and uneven population distribution, which led to methodological advancements of existing methods to make them suitable for such data. The existing ST-DBSCAN algorithm will be adapted to take into account the spatial density of GP locations for calculating their distance, making regions of higher and lower spatial density comparable. For this, a Kernel density estimation (KDE) based approach is introduced which derives weights from local point density and uses them to weight spatial distances. The usefulness of the introduced KDE based approach is demonstrated on both simulated and empirical data.

Parts of this chapter appear in a paper of which a peer-reviewed version has been accepted for publication in Risk Analysis[1]. This research has been supported by The Data Lab [project registration number Reg-17669]. It was conducted in collaboration with Wallscope[2] and supported by NHS Scotland.

3.1 Introduction

Harmful outcomes are a potential risk due to unwarranted variation in drug as prescriptions highlighted in 2018 by Catherine Calderwood, the former Chief Medical Officer for Scotland, in her report titled ‘Practising Realistic Medicine’ (Healthcare Quality and Improvement Directorate, 2018). With antibiotics there is the specific risk of harm in their overuse due to a possible loss of effectiveness, as is frequently highlighted due to their use in farm animals (Collineau et al., 2018; George, Stewart, Evans, & Gibson, 2020). Thus, a number of schemes and campaigns in the UK are targeted specifically at the overuse of antibiotics such as the “Keep Antibiotics Working” campaign (Public Health England, 2019). Campaign groups such as the Scottish Antimicrobial Prescribing Group (SAPG) emphasise the dangers of respiratory and urinary tract infections and provide guidance about antimicrobial drugs to General Practitioners (GPs) (Scottish Antimicrobial Prescribing Group, 2020a). Along with National Health Service (NHS) Education for Scotland (NES) they targeted Primary Care (GP practices) in 2013, with the Scottish Reduction of Antibiotic Prescribing programme (ScRAP) and, following that, ScRAP 2 (NHS Education for Scotland, 2020). Scottish Antimicrobial Prescribing Group (2018) also acknowledged that prescription data would be useful in peer comparison amongst GPs to assist in reduction.

The motivation of the research is to explore prescribing behaviour of GP practices in Scotland which can be used for the risk management of over prescription of antimicrobial drugs, but can

[1] Gieschen, A., Ansell, J., Calabrese, R. and Martin-Barragan, B. (2021), Modeling Antimicrobial Prescriptions in Scotland: A Spatiotemporal Clustering Approach. *Risk Analysis*. Advance online publication. https://doi.org/10.1111/risa.13795

[2] https://www.wallscope.co.uk/

also be applied to under-prescription of drugs where the drugs are warranted. Insights gained from this analysis can then inform strategies to deal with the risks arising from prescription behaviour, and may indicate areas where best practice is seen, presenting evidence for distinct training and communication needs. Highlighting extreme behaviour in the form of outliers or within clusters can offer valuable insights into the behaviour of specific GPs and specific relatively small-scale areas. In a broader sense, our findings can be used in connection with a disease recognition and, ideally, prevention system. They can also be used to inform GPs in a peer-comparison about their own and others' prescription behaviour. The goal of our research is to inform more effective antimicrobial prescriptions by identifying unwarranted variation and to minimise the risk of harm. Another aim is to inform an efficient reallocation of resources, considering that prescription costs are covered by the NHS. The specific focus on Co-Amoxiclav or Amoxicillin is given as an example for a common antibiotic targeted by awareness and information campaigns such as those mentioned above.

Prescribing practice will be effected by spatial and temporal factors. Spatial aspects of prescribing behaviour arise from socio-economic factors (Kjærulff, Ersbøll, Gislason, & Schipperijn, 2016; Mölter et al., 2018) such as areas of deprivation. Temporal factors come from seasonal patterns of disease, such as influenza occurring more frequently in winter months (Durkin et al., 2018). Several authors have considered spatial and spatio-temporal models to analyse diseases and drug prescriptions (C. Anderson, Lee, & Dean, 2016; Blangiardo, Finazzi, & Cameletti, 2016), as well as spatial models for the risk management of epidemics (Zagmutt, Schoenbaum, & Hill, 2016). Katz et al. (2010) analysed individual patients' prescriptions whereas the focus of this research is GP practice aggregated data, since we want to understand their prescribing behaviour. A feature of Scotland's distribution of GPs lies in its split between urban and rural geographies. Such differences have been explored before, for example Cairns, Marshall, and Kee (2011) employed simulation to analyse first-responder schemes for cardiac arrests in Northern Ireland.

In this chapter we present a method of identifying unwarranted variation in antimicrobial prescriptions through forming peer groups of GPs, while taking into consideration differences in their spatial prevalence and prescription seasonality. Specifically, we will employ spatio-temporal density clustering to identify the cluster amongst GP practices using their prescription behaviour to define their peer groups.

As explored in the literature review in Section 2, density-based clustering algorithms, such as DBSCAN, use the distance measures to assess the density of points in a region to determine whether a cluster exists in that location. These algorithms, however, suffer from a drawback. Varying densities in spatial dimensions can lead to the algorithm not being able to identify clusters in both very dense and not dense areas. A region of high density can only be identified as such in relation to the surrounding density. Therefore, relative local density is an important concept if a density-based clustering algorithm is supposed to detect clusters across a spatial area. This becomes relevant when comparing urban and rural regions due to different spatial prevalence across areas, as is the case, for example, in Scotland with a more densely populated South compared to the North. A cluster in an urban region will be more dense compared to a cluster in a rural region, due to both possessing a relatively higher density compared to their surrounding area. In order to solve this, related research usually focusses on using a nearest neighbour approach to determine density, such as DECODE (Pei et al., 2009), or the use of shared nearest neighbours (Ertöz et al., 2003). Ertöz et al. (2003) propose to define the similarity between two points based on the number of shared nearest neighbours. For this solution, however, one needs to decide the number of nearest neighbours, which can strongly impact the clustering results. Instead, we propose a methodological innovation that uses a KDE based approach for determining the density of an area for each point and uses this factor for weighting the distances between points.

Clustering of time series often uses transformations of the time series by representing them in a stochastic way for example with Markov Models (model-based approaches) or, alternatively, features (feature-based approaches) (De Angelis & Dias, 2014). Alternatively shape-based approaches compare the raw, untransformed, time series by trying to match them by stretching or shifting them (Aghabozorgi, Shirkhorshidi, & Wah, 2015). In all three cases, one needs to determine the measure of similarity between time series. This can be difficult, because effects may not be synchronised, but delayed or shifted. Due to the nature of our data we can expect the occurrence of shifts in the time series, for example due to the increase in the number of cases of a disease like the flu, which might occur in different regions at separate times. The challenge of shifted time series has been discussed in the literature and lead to the introduction of methods such as dynamic time warping (DTW) (Serra & Arcos, 2014). DTW calculates the dissimilarity of two time series by minimising the 'cost' needed to match them. While it is a popular method in time series clustering, to the best of our knowledge it has not been used in connection with spatial data in the DBSCAN algorithm.

Our proposed method is able to handle both of these identified challenges, namely clusters of varying densities and shifted time series. In contrast to the nearest neighbour solutions in the literature (Ertöz et al., 2003; Kriegel et al., 2011), our approach is continuous, covering points with a smooth function, without considering a specified number of neighbours. The continuous weights have the advantage that they avoid the need for over-smoothing any small variations through binning. Furthermore, this approach does not need to define the number of nearest neighbours to be considered. Instead, our approach makes use of KDE to give us a smooth density estimate in every location of our map (Silverman, 1986; Terrell & Scott, 1992) and subsequently weights the spatial distance matrix with values derived from these estimates. This makes our approach specifically useful for real-life applications in which spatial clusters exhibit varying densities and for organisations which are looking for ways of geographically representing results. While we introduce our approach using ST-DBSCAN, a spatio-temporal version of the density-based clustering algorithm DBSCAN (Birant & Kut, 2007), it is very flexible. It can be implemented for a variety of density-based clustering algorithms which consider spatial information in the form of a distance matrix. Furthermore, regarding the temporal aspect, this approach is combined with a flexible way of measuring temporal dissimilarity by employing DTW. This results in a spatio-temporal clustering algorithm suitable for the analysis of time series which are spatially unequally distributed.

We test our approach against the original ST-DBSCAN algorithm on both simulated and real-life spatio-temporal data. In the simulation, our results show an improved performance regarding the number of correctly clustered elements, with 1994 of 2000 correctly clustered data points with the proposed method versus 1614 of 2000 with the original method. Both methods are then applied to spatio-temporal antibiotic prescription data of Scottish GPs. It can be seen that our approach is able to cluster both the dense and sparse areas of Scotland simultaneously, while the original ST-DBSCAN algorithm either over-smoothed them or identified a very high number of small clusters. The findings can be used to identify peer-groups of GPs within which they can compare their behaviour to others, and highlight regions of high or low prescription volume and therefore the success of risk management efforts or need for further information and training.

This chapter is organised as follows. Section 3.2 describes the proposed methodology in detail. Our approach is tested against the original ST-DBSCAN to show its improvements in a simulation in Section 3.3. In Section 3.4 our method is applied in a case study of GP prescription behaviour in Scotland to demonstrate its performance for spatio-temporal clusters with varying densities and shifted time series. The chapter concludes and offers suggestions for further research in Section 3.5.

3.2 Methodology

The objective of this chapter is to enable the spatio-temporal clustering of data points with varying spatial density and each associated with a univariate time series. Our approach can be used in connection with a variety of clustering algorithms but will be demonstrated with ST-DBSCAN (Birant & Kut, 2007), chosen due to its computational efficiency and its flexibility with regard to the addition of more variables alongside the temporal and spatial information. In Section 3.2.1 we briefly revise the DBSCAN and ST-DBSCAN algorithms and discuss their limitations. In Section 3.2.2 we propose to manipulate the input spatial distance matrix using weights based on the location density derived through KDE, in order to tackle ST-DBSCAN's inability to handle unevenly distributed spatial points. This enables us to apply the algorithm to the analysis of spatio-temporal antimicrobial prescription data with the objective of identifying opportunities to manage risks associated with that.

3.2.1 Clustering using DBSCAN and ST-DBSCAN

In Sections 2.2 and 2.5 of the literature review, we have introduced the DBSCAN and ST-DBSCAN algorithms for the clustering of spatial and spatio-temporal data respectively. DBSCAN operates by visiting a random data point and looking for neighbouring points within a pre-defined maximum spatial radius ε_1. ST-DBSCAN operates in a similar manner as DBSCAN, but in addition to the spatial parameter for determining the ε-neighbourhood, ε_1, the algorithm takes an additional parameter ε_2, which determines the temporal neighbourhood. In order to be considered a member of a cluster, a data point q has to be in both the spatial and the temporal neighbourhood of point p.

Computationally, DBSCAN and ST-DBSCAN both look for neighbouring points for each data point separately, which means that the distance matrices are calculated within the main algorithm body. If one can assume that the spatial distances are stable over time, and that the time series are not updated very frequently, it makes sense to take this matrix calculation outside of the neighbourhood calculation. The two matrices are calculated and stored before running DBSCAN or ST-DBSCAN, after which the algorithm then just has to access the stored information relevant for the respective visited data point.

Both DBSCAN and ST-DBSCAN suffer from the drawback of choosing ε_1 for $N_{\varepsilon_1}(p)$. Note that ε_1 is global, in the sense that it is constant across the whole region of the study. If the spatial points are distributed with varying densities, as it is, for example, the case in our application in Scotland, then one ε_1 is not be able to capture structures in both dense and sparse areas. If ε_1 is chosen too small, it will not be able to identify areas of relatively high density in less dense areas. If it is chosen too large, it will not be able to distinguish between structures in high-density areas. This drawback is the primary motivation for the development of our method. We propose that this issue can be solved by using a distance that is modified according to the density of an area, enabling the use of a global ε_1 for the whole space.

3.2.2 ST-DBSCAN with KDE-based local scaling

In order to make more and less dense areas comparable and, therefore, a global parameter suitable, we employ KDE as a non-parametric way of estimating an area's density. KDE is a density estimation technique which approximates the probability density function from samples. An advantage of KDE is its continuous nature, as it approximates the density function at all locations of the data space, not just at the locations with existing data points. This makes our approach also suitable for situations in which new data points are added in between existing ones. KDE is often used to identify hotspots which are areas of high point density, for example for road accidents (T. Anderson, 2009; Eslinger & Morgan, 2017), infection cases (Smith, Lessells, Grant, Herbst, & Tanser, 2018), the occurrence of fires (González-Olabarria, Mola-Yudego, & Coll, 2015), or crime (Gerber, 2014).

Generally, the estimation takes the following form (Matioli, Santos, Kleina, & Leite, 2018; Terrell & Scott, 1992):

$$\widehat{f}(y) = \frac{1}{nh^d} \sum_{i=1}^{n} K\Big(\frac{x_i - y}{h}\Big), \tag{3.1}$$

for a random sample $\{x_1, x_2, ..., x_n\}$, with $x_i \in \mathbb{R}^n$, h is a smoothing parameter, and K is the kernel function which satisfies the conditions

$$\int_{-\infty}^{\infty} K(x)dx = 1 \tag{3.2}$$

$$K(x) \geq 0. \tag{3.3}$$

KDE is used to estimate the density function from observed data. As a non-parametric approach, it assumes that the data originates from a distribution with the probability density f, and then directly uses the data to estimate it. For this, a suitable kernel function is selected, which models the influences of the data points in the data space, with the Gaussian kernel as a commonly used option (Hinneburg & Gabriel, 2007). The sum of all kernels at a location then provides an estimate of the density function at this location.

For the Kernel function a bandwidth parameter h has to be selected which affects the smoothness of the surface laid over the points. In other words, it is the influence a data point has on more distant regions (Hinneburg & Gabriel, 2007; Terrell & Scott, 1992). h is often chosen by cross validation or "Scott's Rule" (Scott, 1992), which we also employ here. It has previously been shown to perform well compared to other approaches (Scott, 2009), and is frequently used as a way of determining the optimal bin width, for example by Bernacchia and Pigolotti (2011).

Let now $\{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}\}$ be a set of observations in a two-dimensional space, with each element being a pair of coordinates. Using these observations, we estimate the density function with KDE as described above. Our aim is to re-scale distances in such a way that the result is a distance matrix in which entries for more and less dense areas are comparable. A logistic function is chosen for the calculation of the weights from the density estimates. The reasons for this is that the slope of the function can be adjusted, which makes it possible to change the impact the density estimate has on the weights. Other reasons for choosing a logistic function are its continuous nature and the fact that it can be symmetrically centred around one. Specifically, we choose a logistic function with the curve's midpoint to be the density estimates' mean and the curve's maximum to be two. This centres the curve around one on the vertical axis, which gives us for the density d of point i

$$f(d_i) = \frac{2}{1 + e^{-k \cdot (d_i - \bar{d})}}. \tag{3.4}$$

This means that if the density estimate for an area is equal to the overall average of all areas, these distances are multiplied with a weight of one, resulting in no modification. The two extreme ends represent the most dense and least dense areas, where the weights are two and zero, respectively. We choose the value two as a maximum in the highest-density area for symmetry reasons, to account for zero being the minimum and one being the mean weight. The weight of exactly zero is only used to scale the distance between spatially identical points, in which case the distance is already zero. The steepness k of the curve can be adjusted depending on the application. When choosing k it is advised that multiple values of the parameter are tested and the results are compared. A higher value of k will result in the combination of neighbouring regions and a smoother result. The application context and possible policy implications might demand a higher degree of smoothing with larger resulting clusters or a more fragmented and detailed view of smaller groups. The parameter sensitivity to changes will depend on the scale of the distances.

A distance matrix S with elements s_{ij} for the distance between points i and j is first calculated for example using a great-circle distance, which is the shortest distance between two points on a sphere such as the Earth (Williams, 2011). It is then modified with

$$w_{ij} = s_{ij} \cdot \Big(\frac{f(d_i) + f(d_j)}{2} \Big), \tag{3.5}$$

for a point pair i and j, and their respective distance s_{ij} in the spatial distance matrix S. Equation (3.5) calculates the weights for both i and j using Equation (3.4), takes the mean of these two weights, and multiplies this combined weight with the respective distance s_{ij}.

After computing the distance matrix, the ST-DBSCAN algorithm as introduced in Section 2.5 is applied to the data.

3.3 Simulation

In order to test the properties of the proposed method, we first apply it to a simulated dataset. The main purpose of this simulation study is to determine whether our approach is able to handle data with varying point densities between different areas, with an improvement in the results when compared to the original algorithm. An implementation with real world data can be found in Section 3.4.

3.3.1 Set up of the simulations

We generate 2,000 data points, each of them associated with both a two-dimensional location (spatial dimension) and a time series of length 100 (temporal dimension). The data points are generated in such a way that they would naturally fall within four clusters with 500 data points in each. These clusters can be seen in Figure 3.1. Each cluster consists of 500 sampled points from two overlapping two-dimensional Gaussians. 250 of which are drawn from a Gaussian with a small standard deviation of $\sigma = 1$ and 250 drawn from a Gaussian with a larger deviation of $\sigma = 6$, and both of which with the same mean to ensure overlap.

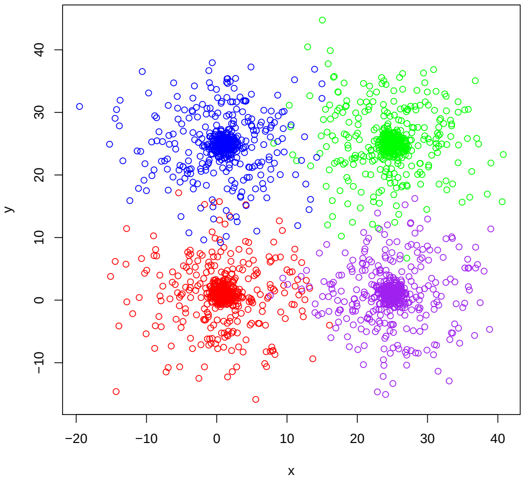

Figure 3.1: Spatial representation of simulated data. 2,000 data points in total are generated for four equal sized clusters. Each cluster consists of two overlapping Gaussians to result in a very dense center and a less dense surrounding area. Cluster 1 (red), cluster 2 (green), cluster 3 (blue), and cluster 4 (purple).

The temporal dimension consists of 2,000 randomly generated time series, each of length 100. For each cluster 500 time series which are similar to each other are generated. Figure 3.2 shows a sample of one of such time series for each cluster. They are generated using random Gaussians at varying time steps. They are added to $t = 10$ for Cluster 1, $t = 30$ for Cluster 2, $t = 50$ for Cluster 3 and $t = 70$ for Cluster 4. This results in different peaks in the time series for each of the clusters. The remainder of the time series is set to 0.

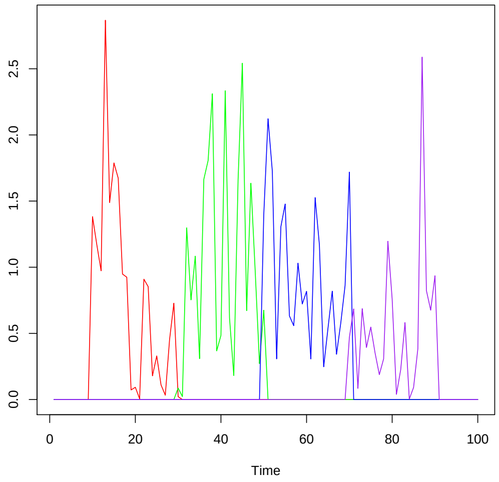

Figure 3.2: Examples of randomly generated time series. Each cluster consists of 500 points and each one is associated with a time series of length 100. The figure shows four examples of these time series, one for each of the four clusters: Cluster 1 (red), cluster 2 (green), cluster 3 (blue), and cluster 4 (purple).

The simulated dataset is clustered using the original ST-DBSCAN as well as the proposed method of ST-DBSCAN with distance weighting. The chosen methods allow us to compare our approach with the original ST-DBSCAN algorithm (Birant & Kut, 2007) to demonstrate its improved performance under certain conditions. The results, as displayed in Figure 3.3, are compared based on their ability to identify the more dense clusters while still accounting for any clusters in the less dense areas surrounding them. For assessment purposes, the results are compared regarding their number of outliers and the number of correctly classified points. Furthermore, a comparison with k-means as a common baseline model is conducted.

3.3.2 Results of the simulations

We aim to choose parameters for both methods using common approaches, comparing them under realistic conditions. This results in ST-DBSCAN with $\varepsilon_1 = 19.9$ and $\varepsilon_2 = 1.7$, chosen according to the nearest neighbour plot as proposed by Ester et al. (1996), and $\eta = 240$. For the proposed adapted ST-DBSCAN, we choose the same parameters to compare any changes to the results occurring under the same circumstances. The parameter k of the logistic function is set to 55. This parameter can be tuned depending on the application context and any existing expert knowledge as it describes the amount of impact nearby points have on a chosen location. For k-means we select the known correct number of clusters, $k = 4$, and use 100 as the number of random starts to allow the algorithm to pick the best starting location for its center points.

The initial assessment is a general visual comparison of the results as shown in Figure 3.3. As the data is generated to represent four clusters, we are able to assess whether the two methods correctly identify these four groups.

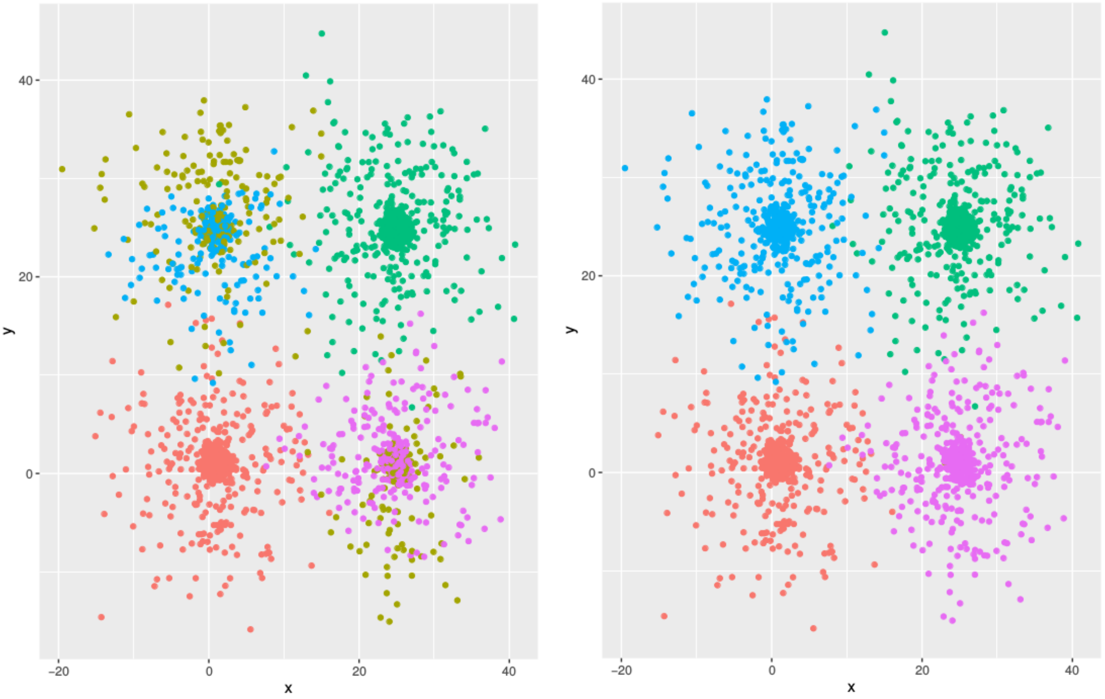

Figure 3.3: Clustering results for the tested algorithms using the generated simulation data. The original ST-DBSCAN on the left and the adapted method on the right. The axes refer to the random variables x and y from which the Gaussians were sampled as shown in Figure 3.1. The colours refer to the assigned cluster. Cluster 1 (red), cluster 2 (green), cluster 3 (blue), and cluster 4 (purple). Points in mustard-yellow are outliers.

Figure 3.3 shows that the original ST-DBSCAN is able to cluster two of the clusters (one and two, shown in red and green) well. However, points which would belong to the other two clusters (three and four, shown in blue and purple) are partly classified as outliers (mustard-yellow). Compared to that, the proposed ST-DBSCAN using the same parameters has a much smaller number of outliers and is able to cluster all four groups. Figure 3.4, showing the results of the k-means algorithm on the simulated data, shows a good clustering result which is able to capture all four of the clusters. Visually difficult to detect is a small number of incorrectly clustered elements. This becomes apparent when looking at Table 3.1 which counts these elements assigned to the wrong, neighbouring cluster.

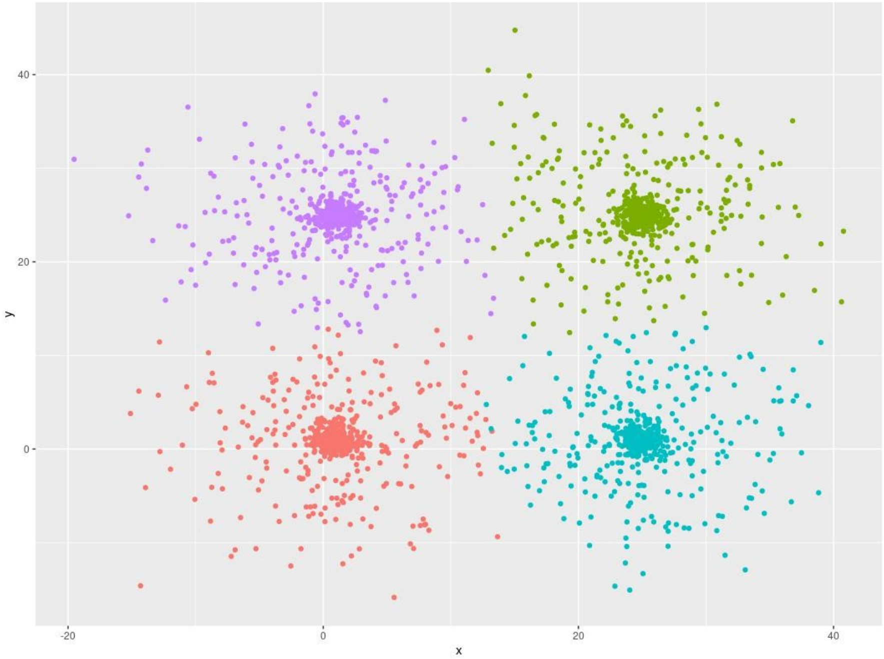

Figure 3.4: Clustering results for the k-means algorithm using the generated simulation data. The axes refer to the random variables x and y from which the Gaussians were sampled as shown in Figure 3.1. The colours refer to the assigned cluster.

	K-means	Original ST-DBSCAN	Proposed method
Number of outliers	0	385	6
Correctly clustered elements	1995	1614	1994
Incorrectly clustered elements	5	1	0

Table 3.1: Number of outliers and number of correctly and incorrectly clustered data points for each algorithm when tested on the generated data set. The proposed method has a higher number of correctly clustered elements than the original ST-DBSCAN. This is mostly driven by a higher number of outliers in the results of the original ST-DBSCAN. K-means as a baseline method performs slightly worse than either of the methods in terms of correctly classified elements.

When reading Table 3.1, it is noticeable that both methods do well in assigning the correct cluster; there is only one incorrectly clustered element. The main problem identified here is assigning a cluster to those points which are classed as outliers by the original method. The comparison with the k-means algorithm shows a very small number of incorrectly clustered elements and, as was to be expected from this algorithm, no outliers or not clustered data points. In the presented simulation, this number of incorrect elements has no major impact on the overall good performance of the algorithm. However, it shows a potential weakness of k-means when presented with cases such as this where clusters somewhat overlap in a low density area between them. Our proposed method thus outperforms the baseline k-means model by capturing those data points correctly, however it shows instead a number of outliers similar to the number of incorrect elements.

3.4 Empirical application

Our proposed method is used to analyse prescription data on GPs from the Scottish NHS from October 2015 to September 2017. This data plays a fundamental role in minimising the risk of antimicrobial over-prescription. There are four main objectives for this application: To improve the general understanding of GP behaviour over time and in different locations, produce insights which can be used for training purposes by organisations such as SAPG and for peer-comparison by the GPs themselves, enable organisations to generate representative samples of GPs by sampling from clusters to replicate existing underlying structures, and to develop a software tool making our findings accessible for users from different areas to improve their data understanding in an interactive and real-time manner.

3.4.1 Data source and description

The data sources are NHS Scotland publications and the Open Data Platform of their Information Services Division (ISD)[3]. Using open data has multiple advantages, such as being readily accessible via the ISD's website which also makes our research easy to replicate.

NHS Scotland covers 14 territorial healthboards of Scotland as shown in Figure 3.5 and 4,994 GPs as of September 2018 (NHS ISD, 2018). The number of GPs varies across the different healthboards, as seen in Table 3.2.

Healthboard	Number of GPs	Healthboard	Number of GPs
Ayrshire & Arran	338	Borders	126
Dumfries & Galloway	120	Fife	280
Forth Valley	270	Grampian	533
Greater Glasgow and Clyde	1,076	Highland	398
Lanarkshire	441	Lothian	941
Orkney	36	Shetland	30
Tayside	395	Western Isles	32

Table 3.2: Number of GPs for each territorial healthboard in Scotland. Each of the fourteen healthboards is listed with their respective GP ‘headcount’. The number of GPs varies greatly between the different healthboards, from as low as 30 to as high as 1076.

[3] https://www.opendata.nhs.scot/

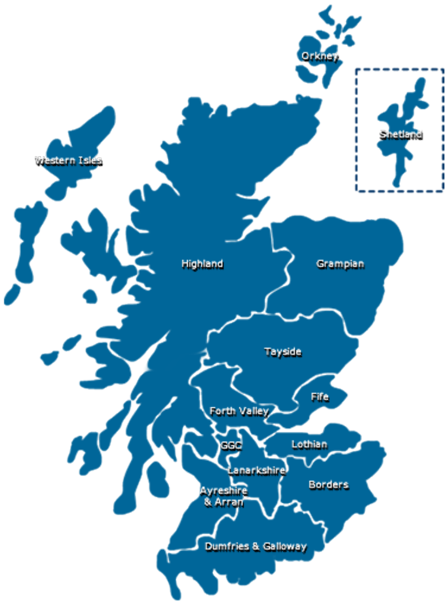

Figure 3.5: Healthboard areas of NHS Scotland. The figure shows a map of Scotland and the fourteen territorial healthboards into which it is divided. Source: NHS Scotland (2013)

The prescription data from NHS ISD is merged with additional open data containing the GPs' locations given by their postcode and corresponding latitude and longitude. We filter the data table for one specific drug, Amoxicillin, which is an antibiotic that is used to treat bacterial infections. The time series for each GP is then calculated by aggregating the prescription instances for different types of the drug on a monthly basis. Based on a two-year period from October 2015 to September 2017 with monthly data, this gives us a time series of length 24 for each GP. Additional variables are a unique identifier (practice code), post code, latitude, and longitude, which account for four more columns. There are 1081 rows in total, which includes all GPs who prescribed any type of Amoxicillin over the period of two years. Due to missing values in their coordinates, a number of rows had to be removed, resulting in a 928×28 matrix.

v1	v2	v3		v26	v27	v28
Practice code	Postcode	201510	...	201709	Latitude	Longitude

Table 3.3: Variables in the aggregated data matrix after filtering. Each GP practice which prescribed Amoxicillin between October 2015 and September 2017 is represented by a table row using a unique practice code as an identifier. The table lists their postcode and the corresponding latitude and longitude values, as well as their individual time series of summed up Amoxicillin prescriptions per month.

When plotting the locations of the 928 analysed GPs using their latitude and longitude, as shown in Figure 3.6, it becomes clear the distribution of their locations is not even across the country. This is likely to be the case due to the uneven population distribution in Scotland. The "belt" of more densely populated areas towards the South is clearly visible, as are the much less populated northern regions.

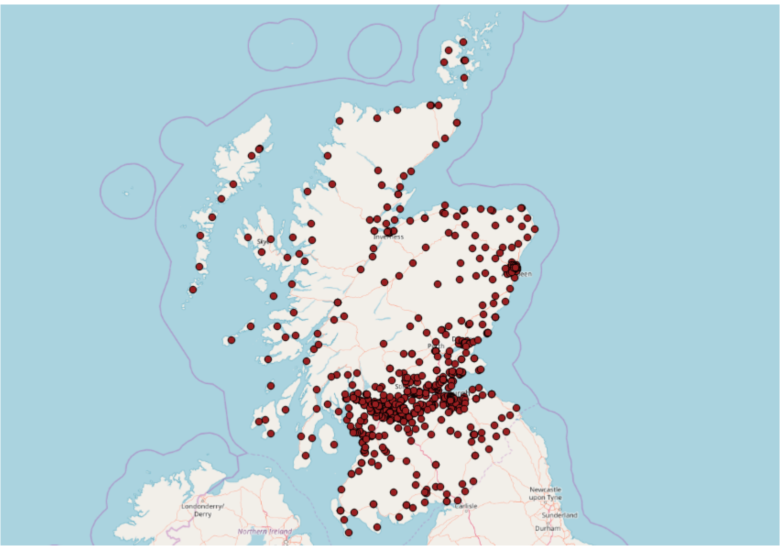

Figure 3.6: Map of Scotland with relevant GP locations. The map shows the location of each GP in Scotland who prescribed Amoxicillin between October 2015 and September 2017. The more densely populated South of the country as well as the more sparsely populated North and North-West are clearly visible.

Figure 3.7 shows the mean sum of prescriptions over all GPs per month. As can be seen, more Amoxicillin is prescribed during the late autumn and winter months (October to January) compared to the summer months (May to August).

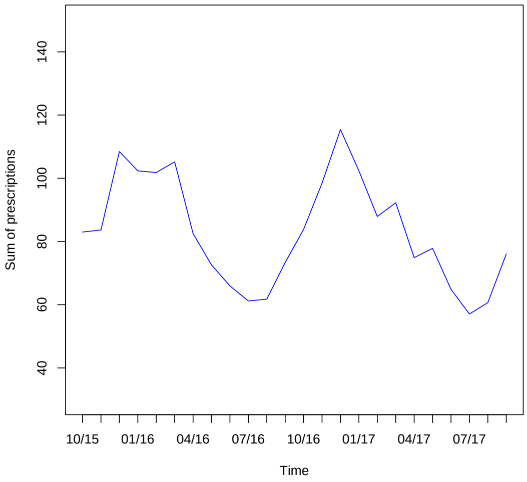

Figure 3.7: Amoxicillin prescription counts in Scotland from October 2015 to September 2017. The figure shows the average sum of prescriptions per month for all analysed GPs. The line shows seasonality, with higher prescription volumes during the winter months compared to the summer.

3.4.2 Clustering process and results

We use two algorithms to cluster our data. In a first step, we use the original ST-DBSCAN algorithm. The spatial distance matrix is calculated using the 'Meeus' great-circle distance from the R package 'geosphere' (Hijmans, 2017). The temporal dissimilarity between the time series is calculated using the DTW function 'DTWDistance' from the R package 'TSDist' (Mori, Mendiburu, & Lozano, 2016). These two dissimilarity matrices are then used by the algorithm to define ε-neighbourhoods as described in Section 3.2.1.

The time series are first normalised using the min-max approach. They are subsequently clustered by calculating the temporal dissimilarity matrices using DTW. Due to the expected seasonal variations in the prescription levels of Amoxicillin, DTW was chosen as it is able to detect similarity between time series where peaks in prescription volume occur shifted. This seasonality is shown in Figure 3.7. The ST-DBSCAN algorithm calculates the dissimilarity matrix outside of the main algorithm body and refers back to it during the N_{ε_2} search. $\varepsilon_2 = 15$ is chosen using the nearest neighbour distance plot as proposed by Ester et al. (1996).

The uneven spatial data distribution does, however, lead to a problem. The global spatial parameter ε_1 is not able to capture small-scale clusters in the South without oversmoothing them. If chosen too small, the algorithm will form many (usually around eleven) very small and dense clusters, with many points within less dense areas marked as outliers. Alternatively, the algorithm will select all points in the denser areas to belong to the same cluster when the chosen parameter is too large, as seen in Figure 3.8.

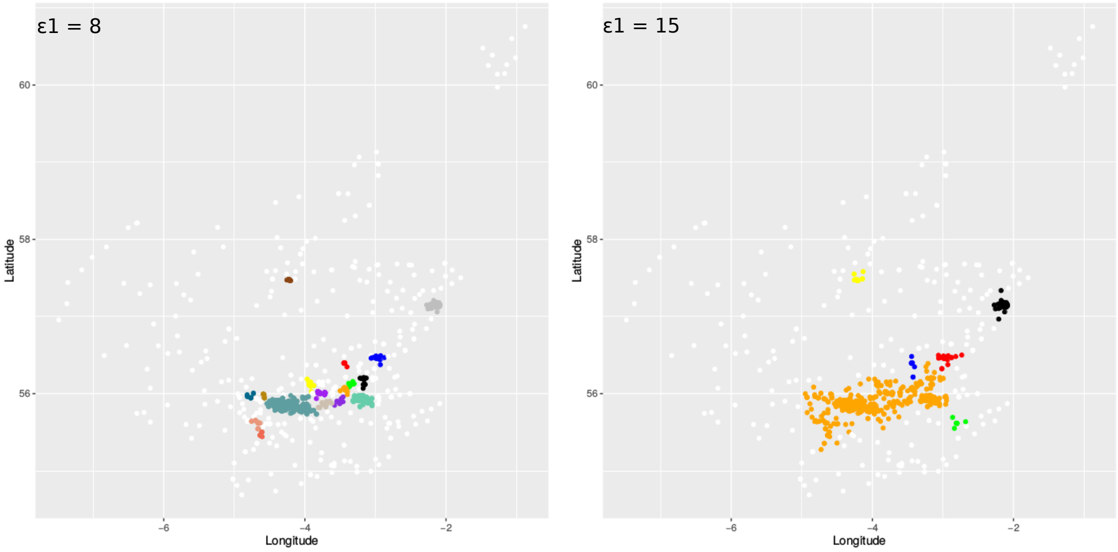

Figure 3.8: Clustering results for the original ST-DBSCAN algorithm. The left figure shows the results with $\varepsilon_1 = 8$, which means an 8km radius. Each colour and shape represents one cluster, the white data points represent outliers. With the smaller radius, the algorithm is able to cluster mostly dense areas into several small clusters, especially in the "belt" area which consists of Edinburgh and Glasgow. The right figure shows results with $\varepsilon_1 = 15$, a 15km radius. With the larger radius, the algorithm is not able to cluster dense areas. Instead, it combines them into one large cluster marked in orange.

The overall goal of our proposed approach is for the algorithm to not overemphasise clusters in dense regions such as cities. This thought leads to the idea of scaling the distances based on their surrounding point density. For this, we use the process described in Section 3.2.2. We calculate the density estimates for all regions based on the amount of data points present in each area and use them to derive weights with which we multiply our spatial distance matrix. The updated distance matrix is then used as an input for the original ST-DBSCAN algorithm.

The results for our implementation show a more realistic clustering result as seen in Figure 3.9. For both the original ST-DSBCAN algorithm and our proposed method, we choose ε_1 =

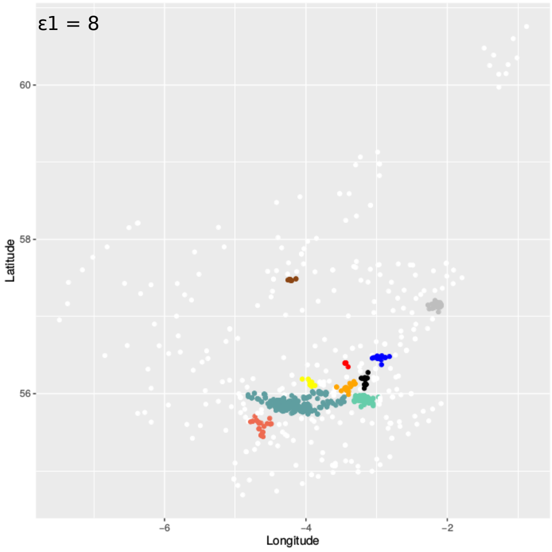

Figure 3.9: Adapted approach with $\varepsilon_1 = 8$, the same ε_1 as the left panel of Figure 3.8. Each cluster is represented by a different colour and shape. Our approach is able to cluster the dense area and combine some of the smaller clusters in the latter into two larger clusters, marked in dark turquoise. In the case of Glasgow, the high density of the city meant that in Figure 3.8 there are multiple smaller clusters surrounding it that were considered separate clusters. By adapting the algorithm in such a way that urban and rural areas could be compared more equally, our approach combines GPs in Glasgow with several of GPs in the surrounding area.

8, $\varepsilon_2 = 15$, and $\eta = 7$ according to nearest neighbour distance plots as proposed by Ester et al. (1996). With the same parameters, our adapted ST-DBSCAN method captures 10 clusters compared to the original 17. A summary of the resulting clusters of both methods is shown in Table 3.4. When analysing the movements of GPs from the original to the adapted solution, we observe the behaviour shown in Table 3.5.

Our proposed approach is able to highlight for example extreme behaviour such as particularly high or low prescription volumes of single or groups of GPs, or their differences in their behaviour over time. This makes the results especially interesting for policy makers and the NHS as a whole, as they could indicate regions of focus for the risk management of over- and under-prescription. The highlighted regions can also indicate a need for further communication and training efforts by organisations such as SAPG to further minimise these risks. Figure 3.9 shows, however, that many points are still classified as outliers. This can be explained due to the chosen variables, meaning prescription volume over time and location, not being informative enough to fully cluster GPs. Additional variables such as socio-economic factors can potentially add more useful

	Original	Adapted	
Cluster	Count	Count	Colour in plot
-1	341	313	white
1	9	9	red
2	25	27	blue
3	8	-	green
4	11	21	orange
5	15	16	black
6	13	12	yellow
7	19	-	purple
8	31	31	grey
9	300	373	cadetblue
10	9	11	chocolate4
11	13	-	antiquewhite3
12	83	84	aquamarine3
13	11	-	blueviolet
14	10	31	coral2
15	11	-	darksalmon
16	9	-	darkgoldrod
17	10	-	deepskyblue4

Table 3.4: Comparison of the resulting clusters when using the original and the adapted ST-DBSCAN. As the number of clusters is smaller with the adapted method not all clusters of the original have a corresponding one. -1 describes the number of outliers. The corresponding colour used in the plots has been added for geographic locating of the clusters, the names correspond to the R colour names.

	2	4	5	6	9	10	12	14
-1	2	2	1	1	9	2	1	10
3	0	8	0	0	0	0	0	0
6	0	0	0	0	2	0	0	0
7	0	0	0	0	19	0	0	0
11	0	0	0	0	13	0	0	0
13	0	0	0	0	11	0	0	0
15	0	0	0	0	0	0	0	11
16	0	0	0	0	9	0	0	0
17	0	0	0	0	10	0	0	0

Table 3.5: Table showing GP movement between clusters when comparing original and adapted method. The original clusters are shown on the y-axis and the adapted clusters on the x-axis. It can be seen that some of the outliers (28 of them, shown in row 1) of the original method were assigned to different clusters, explaining the lower number of outliers in the adapted method result. Furthermore, the most number of moving data points from other clusters were assigned cluster 9 in the adapted result (73 of them, shown in column 6). This corresponds to the now larger city area cluster, shown in turquoise in Figure 3.9.

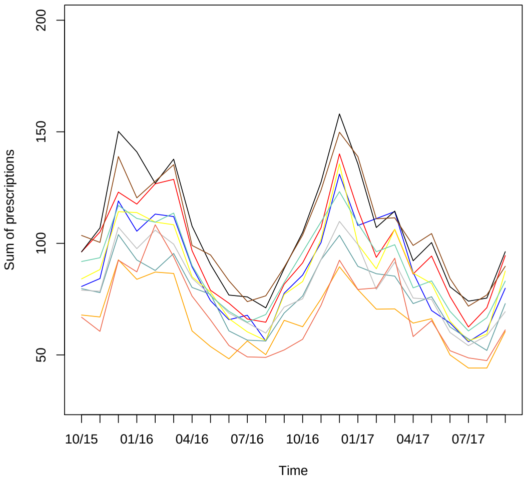

Figure 3.10: Average Amoxicillin prescription volume over time in each of the nine identified clusters. The clusters show a similar seasonality pattern which mostly differs in volume and to smaller degrees, for example, in the point in time at which a peak occurs. The colours correspond to the colours of the clusters in Figure 3.9, the outliers are not shown.

information to this problem. Figure 3.10, which shows the average prescription amounts in each cluster and their seasonal changes over the analysed time period, illustrates this issue. Each cluster has a very similar time series, which mostly differs in total prescription volume and not, as one would expect, in the point in time where peaks occur. This exemplifies the need for further research, for example by analysing a different drug or adding additional factors.

3.5 Conclusion

Unwanted variation in drug prescribing poses serious risk of harmful consequences, for example for antibiotics, which motivated campaigns groups to address this issue through monitoring and communication efforts (Scottish Antimicrobial Prescribing Group, 2020a). Clustering is one possible way to monitor prescribing behaviour. In this chapter we not only present a way of using an adapted ST-DBSCAN algorithm (Birant & Kut, 2007) for this purpose, but also present a novel way of adapting the algorithm for our application to spatial data with varying densities.

The main methodological contribution of our work lies in the use of distance scaling based on local densities using KDE. This enables the algorithm in our approach to handle areas of high density, in our context indicating a large population, and sparse areas. Our approach can identify clusters of relative high density compared to their surrounding area. Other approaches that tackle this issue do not produce a continuous density map, which is not only useful for measuring the density for any location or new point, but also has the advantage of good visualisation possibilities. In contrast to nearest neighbour approaches, our proposed method does not require a specification of the number of nearest neighbours. We thus see our algorithm as an improvement, especially in terms of interpretability and visualisation. To the best of our knowledge, this is the first analysis that uses KDE for addressing this problem in the context of clustering. With respect to the temporal component, due to the nature of prescriptions our data experiences seasonality effects. For that reason, we employ DTW which as an elastic distance measure performs better in such a situation. To the extent of our knowledge, we are the first to use DTW in connection with ST-DBSCAN.

With regard to the empirical contribution of this chapter, we demonstrate how clustering can be used to identify peer-groups of GPs as well as highlight regions of extreme behaviour which carry an increase in risk. We adapt the methodology to the Scottish case by introducing a KDE-based process to ST-DBSCAN. The identified clusters can be used by GPs to self-assess their prescription behaviour by comparing it to their peers. This has been acknowledged as an important issue by both the NHS and, in the special case of antibiotics, the British government (Department of Health and Social Care, 2016; Healthcare Quality and Improvement Directorate, 2018), showcasing the increasing awareness of policy makers. Specifically, our approach and findings can be used for training and peer comparison purposes by organisations such as SAPG and NES. SAPG has a number of tools available for GPs to monitor their own prescription behaviour, which lends itself to offering them a way of comparing their behaviour to their similar peers within their cluster (Scottish Antimicrobial Prescribing Group, 2020b). The spatial nature of the clusters can also help to identify regions with a need for additional training and communication regarding antibiotic prescriptions. Our application to prescription data has implications for the analysis of prescription behaviour in different areas and over time. This could, for example, be a crucial step in understanding the effects of COVID-19 on the management of health conditions and the access to healthcare providers during lockdown.

We see potential for future research in exploring how the smoothing factor in KDE affects the results and can be adjusted depending on the application case. When interpreted in the context of our approach, the smoothing factor accounts for the amount of impact a group of data points has on their surroundings, or how far their "density-impact" reaches spatially. From an application point of view, this should be considered by researchers on a case-by-case basis. Another additional research line is the inclusion of additional explanatory variables. In this chapter we have focused on only one drug and its spatial and temporal prescription patterns. Additional explanatory variables may shed more light on this behaviour, for example socioeconomic factors. Research similar to this proposal is, for example, introduced by Kjærulff et al. (2016) and the consideration of multi-impact variables in spatial risk analysis has been called for by Ferretti and Montibeller (2019). This will help to further expand the understanding of prescription behaviour for different medications and in different regions, offering valuable insights for organisation such as the NHS and policy makers alike.

In terms of the potential for further research based on the introduced methodology, we believe our adaptation to the ST-DBSCAN algorithm makes this method worth considering for various applications. When analysing group dynamics in data sets with both a spatial and a temporal component, taking both of these into consideration simultaneously can be a challenge. ST-DBSCAN allows for the grouping together of data points which can be described in terms of their geographical closeness to each other and their similarity expressed in a one dimensional time series. This is a situation which can be encountered in multiple scenarios in the social sciences, as well as beyond that area in the more general geo-sciences and spatial statistics. Clustering is a useful method for such situations not only due to its ability to classify unlabelled data and discover group structures, but also its way of doing so based simultaneously on variables of different format such as geographic information and time series. Our introduced adaptation to the algorithm allows for both varying spatial densities and shifted or warped time series, making our method dynamic and adaptable to more situations compared to the original DBSCAN and ST-

DBSCAN algorithms. Furthermore, beyond the introduction of this methodological adpatation to ST-DBSCAN, our method of stretching and compressing distances based on local point density can also be used in connection with other methods which use distances matrices. As our proposed method is basically a two-step process in which first the elements in the distance matrix are multiplied with a density factor, followed by the use of the matrix in ST-DBSCAN, this first step can be separated from the process. Instead of using ST-DBSCAN, other clustering approaches such as k-medoids could be used depending on the application context. This makes our presented finding generalisable beyond both the application context of public health and the methodological context of ST-DBSCAN.

Chapter 4

The role of interdependency between UK SMEs in the challenge of access to finance

This chapter introduces an approach combining clustering with spatial logistic regression. The purpose is to explain SMEs ability to access external finance. Grouping SMEs based on their similarity and geographical location allows us to take into account regional contagion effects between them in the logistic regression model. We demonstrate the existence of contagion effects between companies depending on their geographic closeness, that is, companies which are physically close show a similar likelihood of being able to access or not access external finance in the form of bank loans and overdrafts. For this purpose, the spatial regression equation includes a similarity matrix **W** which consists of distances between companies. If we were to directly use SME distances in the **W** matrix, this matrix would become dense which would make the estimation of the logistic model challenging. Instead, we here propose using clustering to determine which entries in the **W** matrix should be non-zero. Only those pairs of companies which are spatially close and similar in terms of turnover and sector have a non-zero entry in the **W** matrix. This ensures that **W** is a sparse matrix, allowing for fast computation times on the logistic regression.

It also crucially has the added benefit of simultaneously taking into account spatial factors and other company characteristics when analysing contagion effects. Our findings show that spatial contagion effects between spatially close and similar SMEs exist, which help to explain access to finance and thus present important insights for policy makers looking to improve the financial situation of small companies which form an important part of many regional economies.

4.1 Introduction

Small and medium enterprises (SMEs) are considered a crucial part of the UK economy. “[They] make up 99 percent of the UK’s business population by number, and it is vitally important to [the UK’s] economy to enable these businesses to prosper and grow.” (British Business Bank, 2018). The government-owned British Business Bank outlines its objective of increasing the financing available to small and medium businesses in the UK. They work together with banks and leasing companies, connecting them to small businesses looking for capital to start or grow. Due to this cooperation, these companies are able to lend more to small businesses, including to those considered more risky. A similar initiative exists in Scotland, where the Scottish investment bank Scottish Enterprise fulfils this role. From a council perspective, sufficient funding for local businesses creates jobs for the community. From the perspective of the country as a whole, small and medium businesses drive exports and establish the region’s importance in a competitive global market (Scottish Enterprise, 2019). The importance of providing finance for SMEs is well understood (Beck, Demirgüç-Kunt, Laeven, & Maksimovic, 2006). However, research into how successful SMEs are in accessing finance is crucial in order to effectively support the work of organisations such as the British Business Bank and Scottish Enterprise.

This chapter explores access to credit for SMEs by making two important contributions to the literature on access to finance. First, our empirical findings using clustering shed light on hidden structures among SMEs which we connect to their their ability to access bank loans and overdrafts. We account for spatial location as well as annual turnover and primary operating sector, and construct groups of SMEs which are similar to each other. These groups can be used to understand differences in their access to finance, and therefore support local policy making tailored to their specific needs. This has been highlighted previously as an important aspect to consider in the SME financing landscape of the UK (Mwaura & Carter, 2017). Second, our findings add a crucial new consideration to previous models of access to finance. Networks and connectedness are frequently considered in the analysis of entrepreneurs (Manolova, Manev, & Gyoshev, 2014). It has also been established that networks and connections play a large role for SMEs as their resources are more limited compared to large companies. The consideration of interconnectedness and networks for SMEs, however, is usually limited to topics such as knowledge exchange (F.-J. Lin & Lin, 2016), productivity (Tiwasing, Gorton, Phillipson, Maioli, & Newbery, 2019) and defaults (Calabrese, Andreeva, & Ansell, 2019; Fernandes & Artes, 2016). The consideration of networks and connectedness in the context of access to finance is limited. We argue here that the interconnectedness of SMEs should be taken into account when analysing this. Therefore we present an approach of using the clustering results mentioned above for constructing a spatial dependency matrix. With this we are able to account for connectedness between SMEs which are similar and spatially close to each other and therefore affect each other. We find that spatial effects within these groups do occur and affect access to finance to a significant degree. Our findings therefore support and follow previous calls for more consideration of connectedness in the analysis of SMEs (Vos, Yeh, Carter, & Tagg, 2007) and are connected to previous research which argues that finance could be distributed through spatially bound networks of SMEs (Lee & Luca, 2019). In addition, our results also support previous findings that the age and size of an SME have a positive effect on its access to bank finance in the form of loans and overdraft acceptance (Artola & Genre, 2011; Ferrando & Griesshaber, 2011; Owen, Botelho, & Anwar, 2016; Presbitero, Udell, & Zazzaro, 2014).

This chapter therefore contributes not only to the literature on access to finance, but through its findings can also help to inform policy makers as well as organisations such as the British Business Bank or Scottish Enterprise, to tailor local support for SMEs. It provides insight into the UK SME bank financing landscape, both in a geographic and a metaphorical sense, as well as factors impacting the SME's application success and their dependencies among each other.

This chapter is structured as follows. The next section, Section 4.2, will introduce the relevant literature on access to finance as well as contagion effects. The data is described in Section 4.3. Section 4.4 will outline our proposed methodology based on spatial regression models and clustering. Section 4.5 outlines the results. Section 4.6 will conclude with a discussion of the results and their implications.

4.2 Background

When companies are credit constrained, research suggests that this leads to cuts in technology spending, employment, and capital spending (Campello, Graham, & Harvey, 2010). It is therefore important to explore why companies are credit constrained. The access to finance for SMEs specifically is an issue of interest to policy makers as well as research (Wehinger, 2012). Financing patterns differ between large sized businesses and those of smaller or medium size. Beck, Demirgüç-Kunt, and Maksimovic (2008) find that leasing and supplier finance is not able to close the financing gap for small firms, highlighting the need to improve small firms' access to external finance through institutional reforms. Some research asks for these institutional interventions to be specifically focused on a local market level, for example through Small Business Administration guaranteed loans (Armstrong, Craig, Jackson III, & Thomson, 2014; Craig, Jackson III, & Thomson, 2007).

When analysing access to finance one aspect to consider is how to differentiate between supply and demand effects. One approach is to use credit registry data on firms with multiple lenders to control for demand effects (Jiménez, Ongena, Peydró, & Saurina, 2012), apply a disequilibrium (Carbó-Valverde, Rodrıguez-Fernández, & Udell, 2016) or use survey data which contains information on both applications and whether these have been successful (Popov & Udell, 2012; Presbitero et al., 2014). Another decision to make is if and how to differentiate between whether an SME is *discouraged* and refrains from trying to apply for finance from banks in the first place due to for example high lending costs or not believing they would receive funding. Other options are whether their application for finance to a bank has been *successful or not*, and whether they have been approved for the *full amount* asked for or not. Ferrando and Griesshaber (2011) model a firm's response indicating that access to finance is the most pressing problem.

However, Artola and Genre (2011) show that there can be a difference between firms perceptions about financial constraint and those who actually experience it. Presbitero et al. (2014) find that with the beginning of the financial crisis loan applications by SMEs in Italy decreased. This could either be driven by a decrease in demand or the firms could feel discouraged and refrain from applying. The findings of Owen et al. (2016) also show that almost one in ten SMEs were discouraged from applying for a loan even though they had financing needs. All of these aspects can best be answered using survey data, as it focuses on the demand side while taking multiple additional factors related to the company into consideration.

4.2.1 Access to bank finance

Brown, Ongena, Popov, and Yeşin (2011) outline that whether a firm applies for a loan will depend on four factors: the firm's need for financing, its access to alternative funding sources, the costs of applying for a loan, and the probability that the application will be approved. Their findings show that half of the firms in Eastern Europe and almost half the firms in Western Europe do not apply for bank credit. The reasons for this differ between Eastern and Western countries, with SMEs in the Eastern European countries quoting discouragement due to lending conditions and Western European SMEs not requiring loans.

Cole (2008) distinguishes between four types of companies: those who did not apply for credit because they do not need it (non-borrowers), those who did not apply for credit because they were discouraged (discouraged borrowers), those who did apply but were rejected (denied borrowers), and those who did apply and were approved (approved borrowers). They argue that there are major differences between non-borrowers and discouraged borrowers. The first group behaved more similarly to approved borrowers.

Casey and O'Toole (2014) use SAFE survey data in their analysis which asks firms whether they applied for bank loans and, if so, whether they were rejected, given between 75-99 percent of their application, given between 1-74 percent of their application, or rejected the offer due to high costs. Casey and O'Toole (2014) then define two types of credit constraint. They use the term credit constrained (denied finance) for firms that are credit-rationed due to denied finance i.e. refused loans, overdrafts or credit lines for which they applied, or for which they received less than 75 percent of what they sought. They use the term cost too high (self rationing) for firms that applied for loans, overdrafts or credit lines and were made an offer, which they rejected due to high costs. They furthermore distinguish between whether firms were denied finance for investment or denied finance for working capital.

Wehinger (2012) show that there is still a low confidence in the banking system and that companies faced bureaucratic hurdles when applying for loans. The authors also outline how policies in the UK have the aim of helping SMEs diversify their funding sources to reduce their dependency on banks.

4.2.2 Factors impacting access to finance

The literature shows a diverse range of variables to use when analysing access to (bank) finance. These include company characteristics, such as their financial situation, size, age, organisational structure and sector. Furthermore, some studies use characteristics of the owner. The spatial location of the company in itself or in connection with the location of the banking sector or environmental factors such as the local economic situation has also been highlighted as important.

The financial situation of a company has been measured in various ways, including capital expenditure and profits (Cosh, Cumming, & Hughes, 2009), growth objectives (Cosh et al., 2009), ratio of liabilities to assets (Cole, 2008), bankruptcy (Cole, 2008), number of bank loans (Carbó-Valverde, Rodrıguez-Fernández, & Udell, 2008), diversification of funding sources (Carbó-Valverde et al., 2008), and liquidity (Presbitero et al., 2014). For example, Carbó-Valverde et al. (2008) find that constrained firms have a slightly higher proportion of bank loans and a lower cash flow generation. Their findings furthermore show that unconstrained firms experience a larger diversification of funding sources. Presbitero et al. (2014) demonstrate that financing constraints decrease with the level of demand for products, but financial health as measured by liquidity is not associated with constraints.

Company size measured by either number of employees (Owen et al., 2016) or annual sales (Cole, 2008) is also frequently used in the analysis of access to finance. Owen et al. (2016) find that in the UK larger SMEs were more successful in receiving finance. Smaller SMEs, especially those with less than 10 employees, were discouraged from applying in the first place. The finding that smaller companies struggle more has also been demonstrated by other studies (Artola & Genre, 2011; Presbitero et al., 2014). Psillaki and Eleftheriou (2015) find that during the financial crisis in France bank finance was more difficult for small businesses to obtain, while medium sized ones increased their bank borrowing. Different explanations have been offered for this behaviour. Including the smaller amount of significant assets in smaller firms (Cosh et al., 2009), smaller firms using more informal lending sources (Chavis, Klapper, & Love, 2011; Gama & Van Auken, 2015), and smaller firms making less applications for bank finance (Brown et al., 2011). In contrast, Ferrando and Griesshaber (2011) find that a firm's size does not have an impact on their likelihood of reporting access to finance as a problem in Euro countries, but they acknowledge that this might differ for other countries. They also discuss whether their model including ownership structure as an important variable might influence this result.

The age of a company is another factor considered relevant by numerous studies. This is for example the case when showing that younger firms are less likely to apply for bank finance in the first place due to them preferring alternatives (Chavis et al., 2011). In contrast, other studies have found that younger SMEs were more likely to apply but less likely to be successful (Owen et al., 2016). Furthermore, it has been shown that young SMEs experienced more credit restrictions during financial turmoil (Artola & Genre, 2011). Artola and Genre (2011) do not find a significant difference between companies that are less than five or between five and ten years old. Firms under ten years of age, however, had a significantly higher chance of experiencing problems accessing finance compared to those which were 20 years or older. Ferrando and Griesshaber (2011) confirm that young firms have a higher probability of reporting access to finance as their most pressing problem.

The company's organisational form is frequently used to explain access to finance (Cole, 2008; Ferrando & Griesshaber, 2011; Owen et al., 2016). The operating sector of a company is also considered relevant due to investment opportunities, capital requirements, doubtful loan ratio and the companies' use of informal lending (Brown et al., 2011; Chavis et al., 2011; Jiménez et al., 2012; Psillaki & Eleftheriou, 2015). Carbó-Valverde et al. (2008) show that there is a high degree of heterogeneity among firms from different sectors, with less constrained firms in transport services or construction, and a larger amount of constrained firms for example in textile manufacturing. The findings of Lee and Brown (2017) show that innovative SMEs are more likely to be discouraged from applying for bank finance, which is especially true for those in peripheral areas.

Other factors include owner characteristics such as the owner's gender (Brown et al., 2011; Owen et al., 2016), ethnicity and race (Cole, 2008; Owen et al., 2016), age, business experience in years, educational attainment, and personal bankruptcy (Cole, 2008). Furthermore, some studies include characteristics relating to the bank such as their liquidity and risk aversion (Jiménez et al., 2012), market power (R. M. Ryan, O'Toole, & McCann, 2014) and the presence of foreign vs national banks (Gilje, 2017; Presbitero et al., 2014). The relationship between company and bank has also been subject to various studies regarding its duration and the number of banks that the company is working with (Agostino & Trivieri, 2018; Chakraborty & Hu, 2006; Giannetti, Burkart, & Ellingsen, 2011; Jiménez et al., 2012; Petersen & Rajan, 1997), with some discussions on the impact of new lending technologies on the importance of traditional relationship lending (Lee & Brown, 2017).

With respect to spatial locations, local lending conditions impact both the applications made and the discouragement of companies (Brown et al., 2011; Lee & Brown, 2017). Lee and Brown (2017) also shows that there is a higher demand for bank finance for innovative firms in peripheral areas.

The importance of spatial factors is not only important due to differences in the local lending markets, but also due to economic conditions (Artola & Genre, 2011; Casey & O'Toole, 2014; Jiménez et al., 2012; Popov & Udell, 2012). Other factors relating to the spatial location of companies take into consideration the physical closeness to the banking sector in the form of bank branches and headquarters (S. Agarwal & Hauswald, 2010; Gilje, 2017; Jiménez et al., 2012; Lee & Brown, 2017; Presbitero et al., 2014). This demonstrates that there are interactions in relation to access to finance that take place due to physical closeness of SMEs and banks. We propose therefore that the closeness of similar SMEs to each other can also affect this.

4.2.3 Contagion effects

Vos et al. (2007) emphasise the importance of considering connectedness as a basis for any SME finance analysis, and propose that SME researchers should consider this *connected* presumption as a potential underlying research paradigm. It seems logical that this holds true in the context of access to finance. We believe this is the case because it might explain why groups of SMEs which are in some way interdependent might experience structural issues when accessing external finance sources and thus support organisations and policy makers addressing these problems. contagion effects could also indicate that an SME's network leads to the acquisition of knowledge about financing opportunities, both institutional and social. Previously, the consideration of dependencies has been adapted in other contexts of financing, such as defaults, or similar situations of economic distress of firms (Calabrese et al., 2019; Fernandes & Artes, 2016; Giesecke & Weber, 2006). Giesecke and Weber (2006) argue that there are two reasons for this: 1. there are common factors impacting all firms simultaneously, such as macro-economic factors like energy prices, and 2. there exist business ties between firms, channelling economic distress from one firm to another. One should note here that the existence of networks as an explanation for contagion effects is one of multiple possible theories. While we hypothesise in this chapter that SME networks are responsible, we can only study existing correlation effects without making definite statements about the causal relationship.

In order to model dependence structures, a range of methods are available. Factor models assume that one or more common factors, usually representing fundamental economic effects, influence, for example, all portfolio positions. The correlation between factors is the only source of dependence, so that the default probabilities themselves are, conditionally on the value of the factors, independent. However, empirical evidence shows that common factors are not sufficient to explain the clustering of default events (Barro & Basso, 2010). Calabrese et al. (2019) also comment that factor models have been proposed for corporates which are classified by rating agencies, while SMEs are noted to be a single category of business.

Giesecke and Weber (2006) study distress using a business partner network between firms as a d-dimensional lattice where the nodes represent the firms and the edges represent the relationships between them. They then model how the economic distress of a firm, in the form of low available liquidity, impacts its business partners. The more distressed business partners a firm has, the higher the probability that they themselves also suffer a liquidity shortage and become distressed. In their model, however, firms are homogeneous in their individual characteristics, with an equal number of business partners and equal size. Furthermore, these types of networks depend on the amount of knowledge one has about the business relations of a firm. Acquiring this kind of information is especially difficult for SMEs.

From a spatial and geographical perspective, it has been established that SMEs in large cities perceive less constraints to accessing financing (Lee & Luca, 2019). The reason for that is that finance seems to be distributed through spatially bound networks. Major cities become focal points of these networks which leads to an improved information exchange. Firms outside of these networks could have problems accessing finance. Fernandes and Artes (2016) showed that incorporating interdependency, as obtained through ordinary kriging, when predicting default probability using factors including the local economic situation improved their model. Barro and Basso (2010) introduce a model consisting of counterparty risk in a network of interdependent firms. Counterparty risk has been defined as the "risk that the default of a firm's counterparty might affect its own default probability" (Barro & Basso, 2010). This network of business relations is simulated to include the economic sector and geographical location of firms. For the spatial location they use an entropy spatial interaction model. Entropy models are included in the wider class of spatial interaction models. The advantage of entropy models is that they can be obtained by maximising the entropy of the system as studied by Wilson (1970). The network itself in Barro and Basso (2010) is constructed as a weighted network. Their Monte-Carlo simulated networks consider the spatial dimension using the entropy spatial interaction model and the economic sector of the firms. For the spatial location of companies in the entropy spatial interaction model they consider both distances and economic weights in the respective area.

The consideration of spatial dependencies is of special interest due to the impacts of local policy making. Mwaura and Carter (2017) show regional differences in the UK SME financing landscape. With their analysis, they highlight the importance of local over national level policies for improving the support for these distinct and different groups of companies.

4.3 Data description

The SME Finance Monitor (BDRC Continental, 2018) is a large, independently collected dataset on access to finance of UK SMEs. The quarterly collected surveys cover questions on applications for various finance sources and success thereof, as well as the financial situation, plans and optimism of the companies. The advantage of using this dataset lies not only in its company and geographic diversity, but also in its questionnaire nature. Previous research has shown that perceived financial constraints differ from real ones (Artola & Genre, 2011), making surveys directly asking companies about their applications as well the corresponding success or rejection so valuable.

The chosen dataset covers Q1 2015 - Q2 2017. For the purpose of this analysis, no temporal effects are considered, instead this time period is considered as one by ignoring the exact application date and considering the latest application any time in this period. The complete dataset consists of 46,019 companies (rows) and 939 variables (columns). The variables range from questions on the business size in terms of numbers of employees, to the forms of finance currently in use, the size of the loan or overdraft applied for, and the financial performance of the company. Considering the large amount of available variables, a selection for the model has to be made first. The variable selection is based on literature findings covered in section 4.2.2. For the spatial component, a two-level postcode is used as this was the available level.

For pre-processing, the dataset is first filtered for records with missing values in their geographic coordinates, as well as missing values or refused answers in the application outcome. Due to a differing postcode system in Northern Ireland, we restrict the scope of the study to Great Britain. This results in a dataset of 3,227 companies, with a very large number of missing values of attributes. The locations of all 3,227 SMEs are shown in Figure 4.1. A small amount of random noise is added to latitude and longitude of each to avoid overlapping in the graphic.

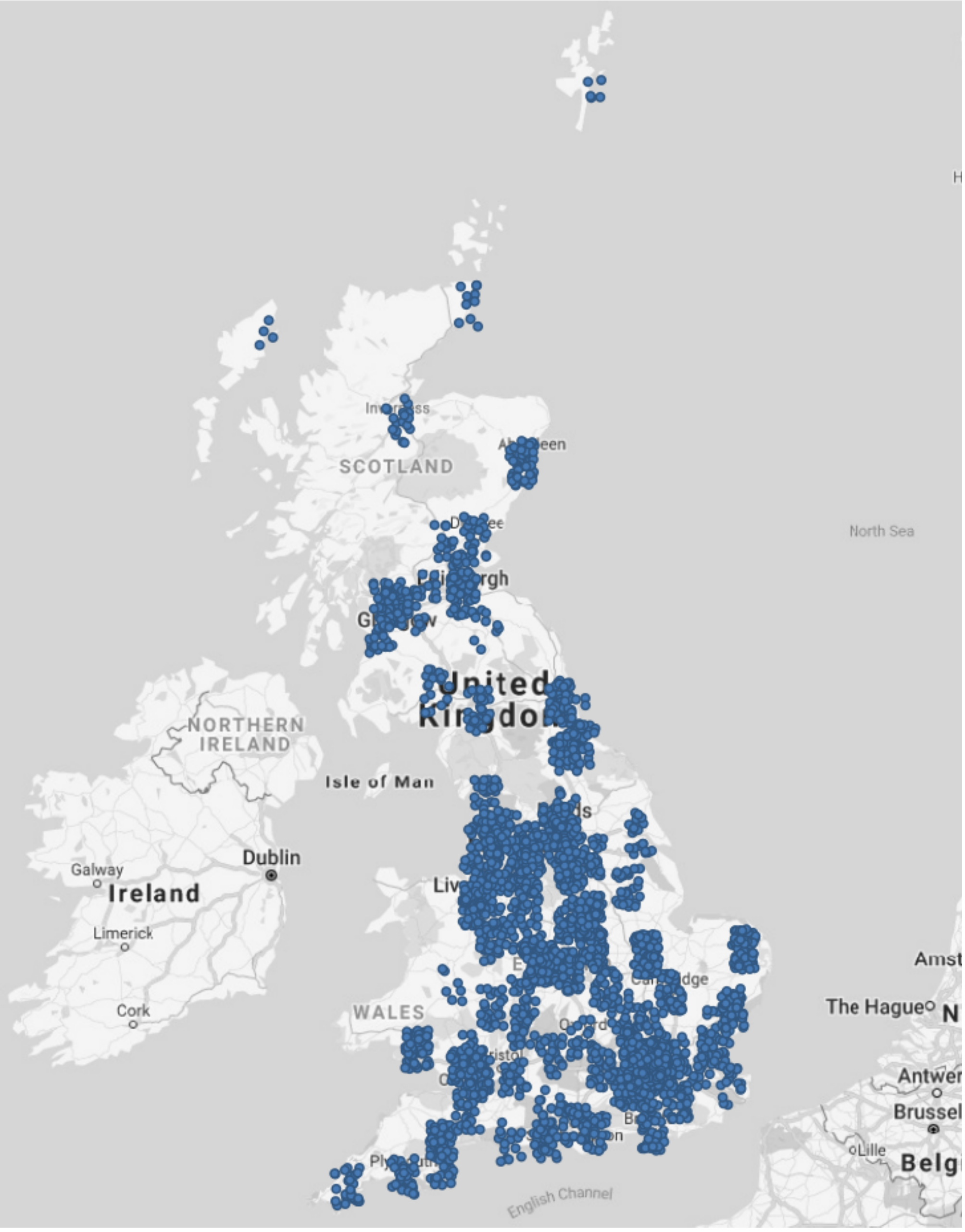

Figure 4.1: The locations of all 3,227 analysed SMEs across the UK by their two digit postcode. A small amount of noise was added to the respective latitude and longitudes to avoid overlapping. The figure shows a relatively even spread across England, with smaller numbers of SMEs in the North of England and the North of Scotland.

4.3.1 Variable description

Our primary variable of interest is the access to bank finance in the form of loan and overdrafts. There are four possible outcomes to the question on loan acceptance (“loan final outcome”) (a_l) and overdraft acceptance (“overdraft final outcome”) (a_o), coded as numbers 1 to 4. We define 'acceptance' for two of these cases.

- Acceptance
  - Bank offered what SME wanted and SME took it : $a_l = 1$ or $a_o = 1$
  - SME received overdraft/loan after issues : $a_l = 2$ or $a_o = 2$
- No acceptance
  - SME took other funding : $a_l = 3$ or $a_o = 3$
  - SME received no overdraft/loan after issues : $a_l = 4$ or $a_o = 4$

For the purpose of this analysis, the loan and overdraft acceptance variables are transformed into a binary dependent variable $\boldsymbol{Y}_i$ for a sample of size $n = 1, ..., i$, where

$$\boldsymbol{Y}_i = \begin{cases} 1 & \text{if } a_l \in \{1,2\} | a_o \in \{1,2\} \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

.

This results in a binary vector which is 1 if the SME was able to secure either a loan or overdraft, or both, with or without issues while doing so. One has to note that a company is assigned a value of 1 (able to access) even in cases in which one of the applications was not successful. If, for example, a company applied for both a loan and an overdraft but was able to secure only one, this would still be counted as "accepted" in this analysis. Table 4.1 shows the numbers of SMEs for the two respective situations. It can be seen that the large majority of SMEs (91 percent) were able to access bank finance as defined by us, with only 9 percent being unable to do so.

One should note that we define the case in which the SME decides to take funding from a different source as 'no acceptance'. This is due to the assumption that if an application was

Y	Number of companies	Percentage
$\boldsymbol{Y}_i = 1$	2,937	91 percent
$\boldsymbol{Y}_i = 0$	290	9 percent

Table 4.1: Loan/overdraft acceptance Y. $\boldsymbol{Y}_i = 1$ indicates a successful loan or overdraft application. $\boldsymbol{Y}_i = 0$ indicates an unsuccessful or otherwise unaccepted application. For company i of sample $n = 1, ..., i$, with 3,227 SMEs in total.

made, bank finance is considered the preferred choice for the SME. Taking funding from an alternative source implies to us that the alternative was chosen because bank finance was either not available or the conditions with which it was offered made it impossible for the SME to accept it. This does not, however, take into account that there are multiple possible reasons as to why the company might choose other funding, including more favourable conditions which do not automatically imply no access to bank finance. We therefore define this category as "no acceptance" instead of "refusal".

Figure 4.2 displays the geographic distribution of both accepted (white) and not accepted (red) SMEs. A split by acceptance can be seen in Figures 4.3 and 4.4, which show only accepted and not accepted SMEs, to visualise geographic differences in their distribution. Figure 4.4 presents a relatively even spread of not accepted SMEs across the UK with some apparent clustering around London, as well as the North of England around Manchester. This could, however, be driven by the larger number of SMEs present in the area. The question whether the location of the SMEs drives their (non) acceptance remains therefore open after this visual examination, motivating further analysis into the drivers of the access to bank finance.

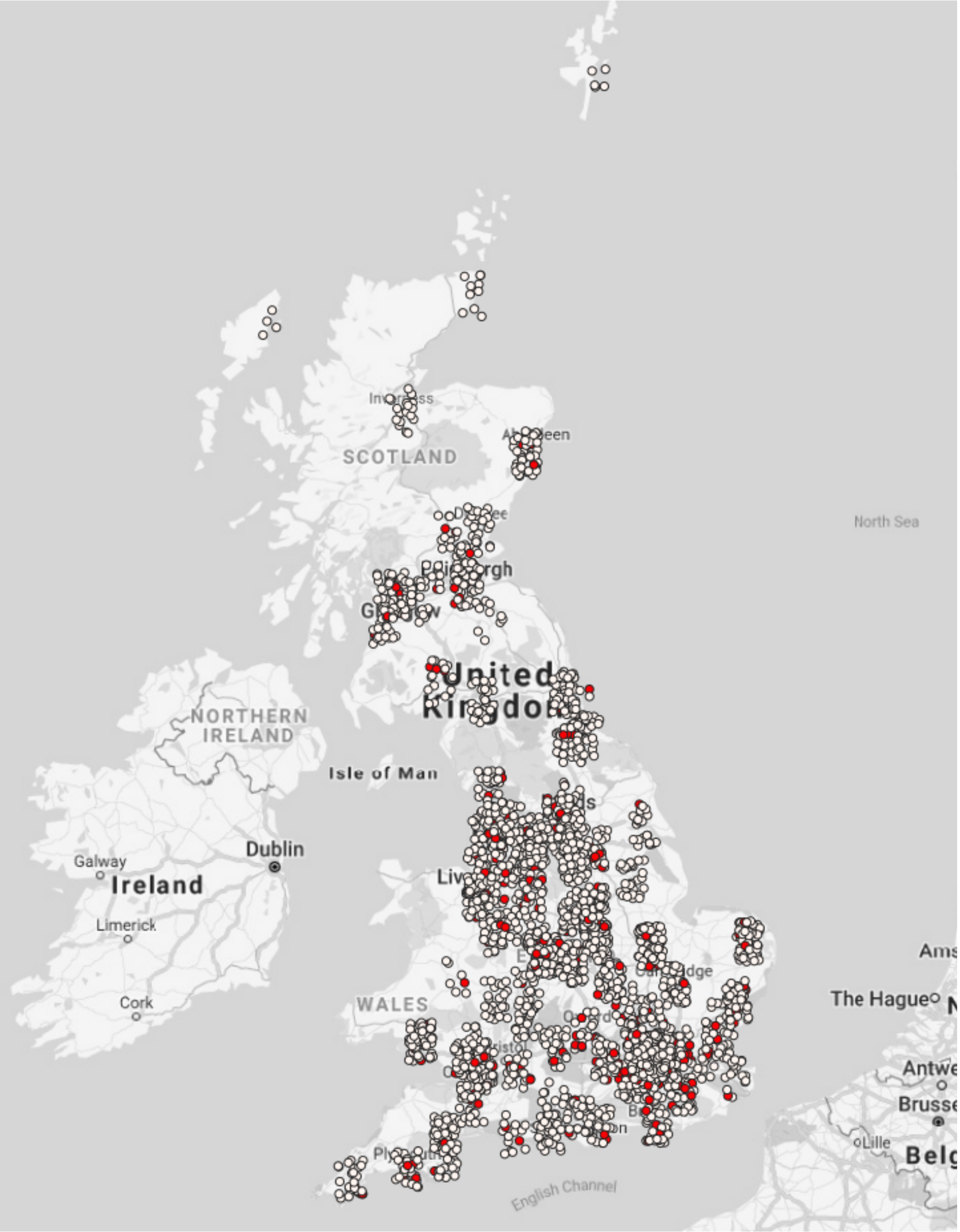

Figure 4.2: Acceptance status of all 3,227 analysed SMEs. SMEs which were accepted for loan and/or overdraft are shown in white (2,937), not accepted SMEs are shown in red (290).

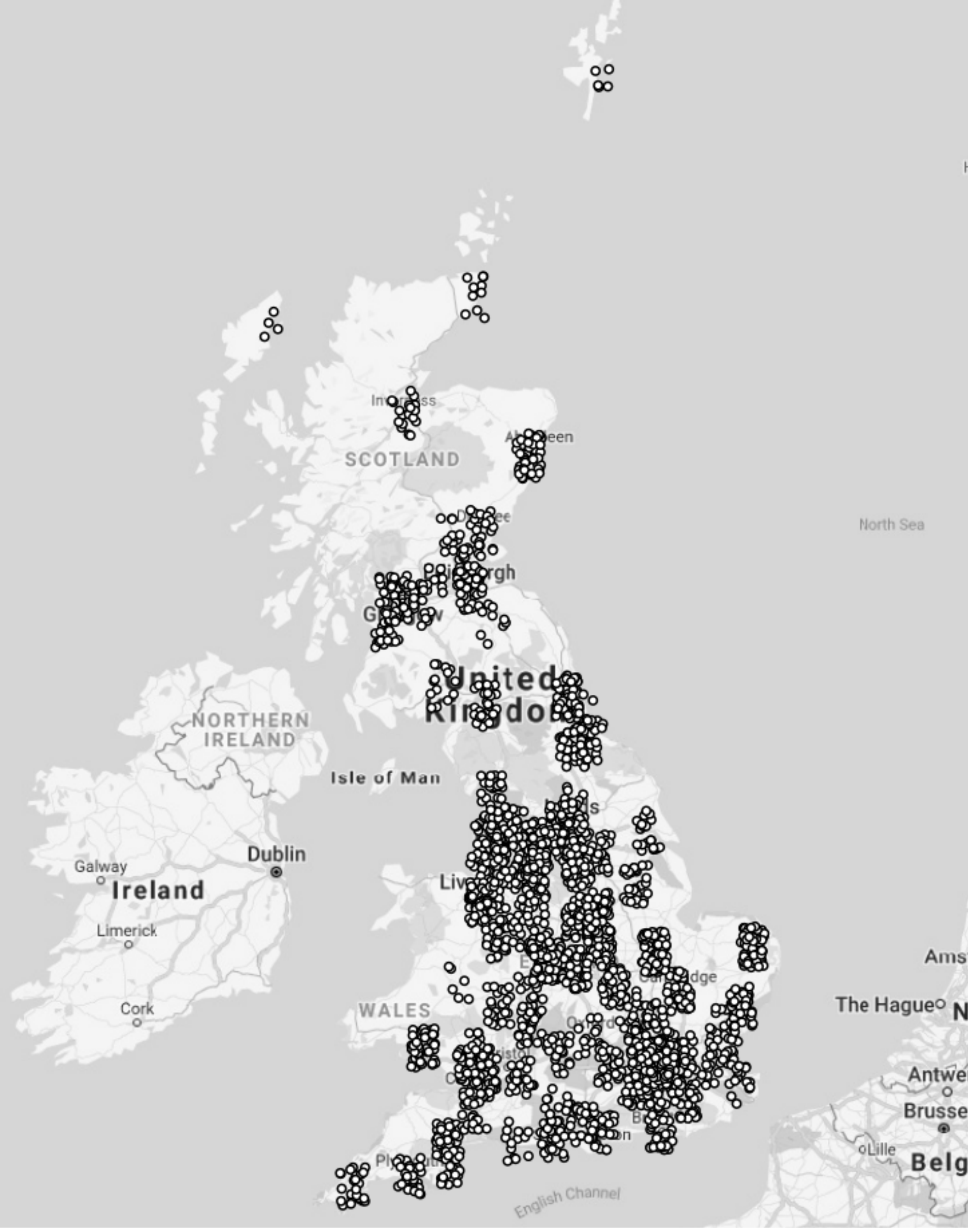

Figure 4.3: All 2,937 SMEs which had their loan and/or overdraft application accepted with or without difficulties.

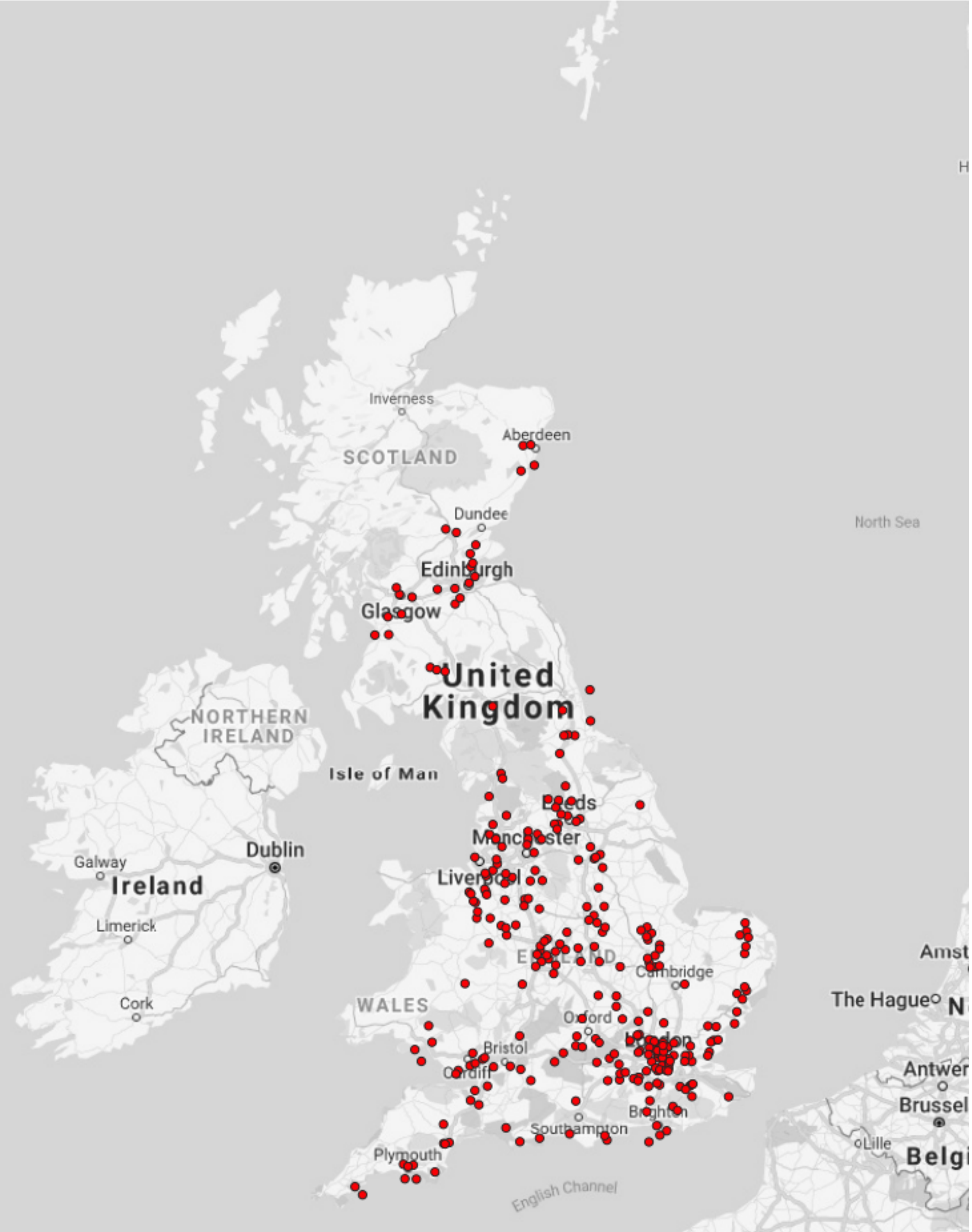

Figure 4.4: All 290 SMEs which had their loan and/or overdraft application not accepted or which did not choose to take the offer for a different reason.

In order to explain acceptance and non-acceptance, the following variables are selected from the dataset based on the literature (Section 4.2.2).

- Age of company
- Company size, by number of employees
- Annual turnover
- Primary operating sector
- Legal status

Table 4.2 shows the respective variable descriptions and distributions. The majority of SMEs in our sample were established more than ten years ago (74.41 percent) with only a small number being recently established (2.23 percent less than 12 months ago). Nevertheless, the majority of companies analysed (84 percent) had 50 employees or less, making them relatively small in size. The annual turnover of most companies (22.81 percent) was between £1m and £1.9m, with 62.04 percent of SMEs reporting an annual turnover of between £100k and £1.9m. All sectors were presented in relatively even numbers, with the largest group of SMEs operating in Construction (16.98 percent) and the smallest in Health and Social Work (6.32 percent). Lastly, the majority of SMEs (64.33 percent) were registered as a Limited Liability Company.

Variable	Label	Frequency	Percentage	Acceptance
Region	Scotland	340	10.54	92.35
	North/North East	182	5.64	**95.60**
	Yorkshire/Humberside	245	7.59	92.24
	North West	311	9.64	90.35
	West Midlands	294	9.11	91.50
	East Midlands	261	8.09	94.64
	East Anglia	259	8.03	90.35
	Wales	204	6.32	91.67
	South West	371	11.50	92.45
	London	340	10.54	87.65
	South East	**420**	**13.02**	86.67
Age: Company was first established...	Less than 12 months ago	72	2.23	66.67
	Over 1 but under 2 years ago	125	3.87	68.80
	2 - 5 years ago	292	9.05	77.40
	6 - 9 years ago	337	10.44	89.32
	10 - 15 years ago	554	17.17	93.32
	More than 15 years ago	**1847**	**57.24**	**95.24**
Size: Number of employees	1 employee	326	10.10	79.75
	2-10 employees	1106	34.27	88.07
	11-50 employees	**1279**	**39.63**	93.82
	51-100 employees	355	11.00	97.46
	101-200 employees	135	4.18	**97.78**
	201-250 employees	26	0.81	96.15
Turnover: Last annual turnover	Less than £25,000	158	4.90	67.72
	£25,000 - £49,999	191	5.92	80.10
	£50,000 - £74,999	166	5.14	89.76
	£75,000 - £99,999	152	4.71	84.21
	£100,000 - £249,999	417	12.92	89.69
	£250,000 - £499,999	406	12.58	91.87
	£500,000 - £999,999	443	13.73	90.29
	£1m - £1.9m	**736**	**22.81**	96.33
	£2m-4.9m	290	8.99	97.93
	£5m - £9.9m	135	4.18	96.30
	£10m - £14.9m	65	2.01	**98.46**
	£15m-24.9m	68	2.11	97.06
Sector: Primary operating sector of company	Agriculture, Hunting and Forestry, Fishing	349	10.81	**95.70**
	Manufacturing	317	9.82	91.48
	Construction	**548**	**16.98**	88.14
	Wholesale / Retail	548	10.66	93.02
	Hotels and Restaurants	266	8.24	86.09
	Transport, Storage and Communication	314	9.73	87.58
	Real Estate, Renting and Business Activities	531	16.45	92.66
	Health and Social Work	204	6.32	92.16
	Other Community, Social and Personal Service Activities	354	10.97	92.09
Legal status	Sole Proprietorship (single owner)	516	15.99	83.53
	Partnership	432	13.39	**96.06**
	Limited Liability Partnership	203	6.29	93.10
	Limited Liability Company (private limited company, public)	**2076**	**64.33**	91.62

4.4 Methodology

The literature has highlighted the importance of considering both company related characteristics, as well as taking into account spatial interdependencies and inter-connectedness of SMEs. We therefore propose the use of a spatial regression model which enables us to take into account explanatory variables in the form of company characteristics, as well as interactions due to the spatial location of SMEs. In addition, we introduce the use of clustering in order to include connectedness of SMEs through their similarity. All implementations are done in R (R Core Team, 2020) using RStudio Version 1.0.44.

4.4.1 Spatial regression model

We have previously introduced regression models for classification in Section 2.1 and will now expand upon that idea through concepts of spatial statistics. Spatial regression models are useful for the analysis of spatial data as they take into account spatial dependence (LeSage, 2009). These models can thus be used to analyse contagion effects as in our problem at hand. Similarly, in the analysis of SMEs, spatial regression models have previously been used to model defaults in terms of the company's location as well as company characteristics (Calabrese et al., 2019). Our objective is to estimate access to finance in the form of loan and overdraft acceptance in terms of an SME's spatial dependencies as well as additional explanatory variables. This model will provide insight into the factors that impact whether an SME is able to access bank finance, as well as the impact that their spatial location and therefore their neighbouring SMEs have on this.

When implementing a spatial regression model to estimate a binary outcome as in our case, one can choose between a logit or a probit model. These models are very similar and mainly differ in their link function which is used to map the continuous value to a binary outcome. While the logit model uses the logit transform, the probit model utilises the inverse normal cumulative distribution function. Differences in the outcome are mostly observed where extreme behaviour exists in the data, and if the interest in the outcome is mainly on classification or prediction instead of analysing the odds, a probit may even be preferred over a logit model

(Hardin, Hardin, Hilbe, & Hilbe, 2007). As we expect no major difference between the outcome of our model for the given data, we here choose a probit model due to ease of implementation through the availability of estimation methods described below.

Specifically, we use a spatial autoregressive regression (SAR) probit model as shown in Equation 4.2 (LeSage, 2009, p.32). $\boldsymbol{Y^*}$ is a vector of latent continuous variables $(\boldsymbol{Y}_1^*, \boldsymbol{Y}_2^*, ..., \boldsymbol{Y}_n^*)^t$ with sample size n. $\boldsymbol{Y}_i$ represents the binary application outcome of an individual i with $\boldsymbol{Y}_i = 1$ if the latent continuous variable $\boldsymbol{Y^*} > 0$ and $Y_i = 0$ otherwise. In our proposed methodology, the explanatory variable matrix $\boldsymbol{X}$ consists of company characteristics as described in Section 4.2.2. The corresponding parameter vector $\boldsymbol{\beta}$, which is to be estimated, will therefore demonstrate the importance of each of these factors in determining whether an SME is able to access bank finance. As the elements of $\mathbf{W}$, w_{ij}, represent the relationship between SMEs i and j, they are usually the distance between them. In our approach, we will consider alternative representations of such relationships through clustering which captures geographical closeness as well as other factors. Therefore, the spatial autocorrelation term ρ will represent spatial contagion effects between SMEs regarding their success in accessing bank finance. γ represents the noise which is assumed to be independently distributed and follows a normal distribution with mean 0 and variance σ_γ^2.

$$\begin{aligned} \boldsymbol{Y^*} &= \rho \mathbf{W} \boldsymbol{Y^*} + \boldsymbol{X\beta} + \gamma \\ \gamma &\sim \mathcal{N}(0, \sigma_\gamma^2 \boldsymbol{I}_n) \end{aligned} \tag{4.2}$$

The data generating process for $\boldsymbol{Y^*}$ is described as in Equation 4.3 with the identity matrix I_n of size n by n.

$$\begin{aligned} \boldsymbol{Y^*} &= (\boldsymbol{I}_n - \rho \mathbf{W})^{-1} \boldsymbol{X\beta} + (\boldsymbol{I}_n - \rho \mathbf{W})^{-1} \boldsymbol{\zeta} \\ \boldsymbol{\zeta} &\sim \mathcal{N}(0, \boldsymbol{I}_n) \end{aligned} \tag{4.3}$$

The process of estimating spatial probit models is subject of numerous studies as it can present a challenging task. Common approaches include Maximum Likelihood (ML), Generalised

Method Of Moments (GMM) (both available for example in McSpatial[1]) and Bayesian Markov Chain Monte Carlo estimation (MCMC) (available for example in spatialprobit[2]). In our approach we use the linearised GMM estimation provided in McSpatial (McMillen & McMillen, 2013) due to its ability to handle large datasets with relative high speed. This approach was previously presented in Klier and McMillen (2008) as a linearised version of Pinkse and Slade's spatial GMM estimator (Pinkse & Slade, 1998), specifically for handling large datasets.

A particular challenge due to the current increase in size of datasets lies in the computation of the inverse of $(\boldsymbol{I}_n - \rho\mathbf{W})^{-1}$ in the estimation of the probit model when $\mathbf{W}$ is dense. For this reason, Wilhelm and de Matos (2013) outline the increasing importance of a sparse representation of the $\mathbf{W}$ matrix. We here propose the use of cluster analysis to achieve a very sparse version of the matrix, while capturing not only spatial dependencies but also similarity based on other factors. We detail our proposal in the next section.

4.4.2 Clustering-based W matrix

We have previously reviewed how clustering can be used for various types of data including spatial data in the form of geographic location as measured by latitude and longitude (Coll et al., 2014; Ester et al., 1996). Previously, clustering was also used for the analysis of financing patterns of SMEs (A. Moritz, Block, & Heinz, 2016).

We propose the use of clustering specifically to account for the connectedness of similar SMEs in the construction of the $\mathbf{W}$ matrix for the spatial regression model. As described in Section 4.4.1, the estimation of the spatial regression model benefits from the sparsity of the $\mathbf{W}$ matrix. We propose the use of clustering to construct a sparse $\mathbf{W}$ matrix based on the assumption that contagion effects between SMEs only take place if the SMEs are similar to each other and geographically close, and in such a case, the effect will be proportional to the distance. In a first step, we run a clustering algorithm on a selected number of SME characteristics and geographical location, while using the remaining SME characteristics as explanatory factors in the probit model. We then define SME similarity in terms of combining geographical distance and cluster membership. The elements of $\mathbf{W}$ between two SMEs i and j, denoted w_{ij}, are there-

[1] R package McSpatial (McMillen & McMillen, 2013)

[2] R package spatialprobit (Wilhelm & de Matos, 2013)

fore the geographic distance between i and j if and only if the SMEs are in the same cluster, and zero otherwise. This means that the access to finance of one company will only affect or be otherwise connected to the access of another if they are similar companies regarding some chosen characteristics. Our approach has the advantage of ensuring the sparsity of the matrix by only accounting for connections between SMEs in the same cluster. Another advantage that this method provides lies in its added analysis of hidden group structures among the SMEs. This can give us further insights into how SMEs in the UK behave, and can therefore inform targeted interventions in the form of local policy making.

For our purpose we focus on non-overlapping partitioning algorithms, specifically k-means. This algorithm has the advantage that it can handle multiple dimensions and large datasets with relative ease, as well as producing an outcome with all points being assigned to a cluster. The clustering variables are the SME's location (latitude and longitude) and additionally its turnover and sector. We choose the geographic location because we assume that the contagion effects present are spatially restricted in some way. This means that companies affect only other companies which are not only similar to them, but also relatively close in space. Additionally, we choose turnover and sector to add a notion of similarity. Specifically turnover and sector are chosen from the available variables, as we interpret turnover to be a good measure of viability which in turn we expect to have a strong impact on the decision making in the lending process. Sector is chosen because external factors impacting for example a specific industry could best be captured with it (Carbó-Valverde et al., 2008). In summary, similarity is defined as companies which are close to each other, are in the same sector and have a similar annual turnover.

Due to the mixed type data with numeric (latitude and longitude), ordinal (turnover) and categorical variables (sector), the distance matrix of size 3,227 x 3,227 is constructed using Gower distances[3]. This allows us to construct each distance between two SMEs in terms of their location, turnover and sector, while taking into consideration that these variables are all of differing types. The number of clusters is chosen using the Elbow criterion[4]. It provides us with

[3] `daisy` function in R package `cluster` (Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2018)

[4] R package `NbClust` (Charrad et al., 2014)

the optimal number of clusters of seven. Using this chosen number of clusters, k-means[5] is then run on the distance matrix to construct these seven clusters. In order to identify each SME the cluster result is connected back to the original dataset using a unique ID.

4.5 Analysis

The results for our analysis are two-fold. The clustering results shed light on hidden structures and groups among the analysed SMEs. The spatial regression model explains access to finance as a function of spatial effects and additional explanatory characteristics of the SMEs.

4.5.1 Clustering results

Based on the Elbow criterion, seven clusters are formed as shown in Table 4.3. The spatial locations of each cluster can also be seen in Figure 4.5.

Cluster	1	2	3	4	5	6	7
SMEs	491	421	515	383	358	688	371

Table 4.3: Number of SMEs in each cluster. It can be seen that the clusters are similar in size.

One of the primary advantages of using clustering in this context is the ability to analyse and describe the formed groups. This gives us an added insight into the hidden structures of SMEs in the UK. Short descriptions of each cluster follow in Table 4.4. In Appendix A a detailed table outlines how each cluster is composed, while in Appendix B boxplots of each cluster are shown for the three numeric variables. While defining these clusters one has to take into consideration that the underlying data is only a small sample of the SME sector in each area. This means that the analysed SMEs are not necessarily always representative for the whole SME sector in that specific cluster. Hence the interpretation also may not reflect an intuitive understanding.

[5] R package `stats` (R Core Team, 2020)

Cluster	1	2	3	4
Acceptance ratio	91.04	81.95	95.15	91.91
Cluster	5	6	7	
Acceptance ratio	92.74	95.2	85.18	

Table 4.4: Percent of accepted SMEs in each cluster. The overall average for all SMEs is 91 percent acceptance. It can be seen that Cluster 2 has the lowest value of acceptance (around 82 percent), and Cluster 6 the highest (95.2 percent).

- **Cluster 1** is driven by location and sector, being located in the South-East of England and East Anglia and having mostly construction and manufacturing companies. It consists of mostly older SMEs (established more than ten years ago) which have between two and 50 employees. Most of them have a turnover between £500k and £1.9m, and operate as a Limited Liability Company. With 91.04 percent acceptance rate it lies right at the overall average of 91 percent.

- **Cluster 2** is constructed mostly through turnover and sector, with a relatively low turnover and many SMEs operating in real estate or community services. Geographically, it is a more widely spread cluster over all of England, but specifically around London and the South East. A majority of the SMEs are older than 15 years, but there is also a large group of younger SMEs which were established two to five years ago. They are mostly smaller SMEs with one to ten employees, and a lower annual turnover of less than £250k. Many of these companies operate in the Real Estate, Renting and Business Activities sector, with another large group in the Community, Social and Personal Service Activities. They do so either as a Sole Proprietorship with a single owner, or as a Limited Liability Company. With 81.95 percent acceptance rate they have the lowest value from all clusters, and are below the average of 91 percent.

- **Cluster 3** is mostly driven by a high turnover. These SMEs are located in the North West of England and around Yorkshire. It consists of mainly older SMEs which were established ten or more years ago, with most being more than 15 years old. They are mostly of small to medium size with two to 50 employees. All of them report an annual turnover of more than £250k, with the majority having a turnover of between £1m and £1.9m. The major sectors are Construction, Wholesale/Retail, as well as Manufacturing. The most prominent legal form is the Limited Liability Company. The acceptance rate of Cluster 3 is 95.15 percent and therefore higher than the overall average of 91 percent.

- **Cluster 4** is constructed using the geographic location, with most SMEs located in the South West of England and Wales. Most of these SMEs are older than 15 years, but a fair number were established six to 15 years ago. They have two to 50 employees and a wide range of turnover bands, with most reporting an annual turnover of between £250k and £1.9m. All sectors are present, the largest ones being Hotels and Restaurants as well as Real Estate, Renting and Business Activities. A wider range of legal forms is also present, with most SMEs being registered as a Partnership or a Limited Liability Company. With 91.91 percent acceptance, Cluster 4 has an only slightly higher percentage of acceptance than the average (91 percent).

- **Cluster 5** is mostly driven through its location in Scotland, with some spread towards the Scottish Borders region. The majority of SMEs are older than 15 years, with some being older than six years. Most of them also have between two and 50 employees, and a wide range of reported annual turnovers, with most being between £100k and £1.9m. The two most prominent sectors are Agriculture, Hunting and Forestry, Fishing, as well as Construction. While the majority does so as a Limited Liability Company, a relatively large share is registered as a Sole Proprietorship with a single owner or a Partnership. The acceptance rate of 92.74 percent is higher than the average of 91 percent.

- **Cluster 6** is a widely spread cluster, which was mostly constructed through a high turnover and operations in the real estate and community services. It is located all over England, with the majority in London and the South East. A majority of SMEs in this cluster are older than 15 years, with some being older than ten. The companies range in size mostly between two and 100 employees. In this cluster, all SMEs report an annual turnover of over £250k, with most having an annual turnover of between £1m and £1.9m. Most operate in the Real Estate, Renting and Business Activities or Community, Social and Personal

Service Activities sector. Almost all do so as a Limited Liability Company. With 95.2 percent acceptance, this cluster has the highest acceptance rate and above the average of 91 percent.

- **Cluster 7** is defined by the lower turnover and companies operating in agriculture and construction. Its locations are widely spread with some majority in the West Midlands and North West. Most SMEs are more than 15 years old, but a relatively large number is only between two and five years old. These SMEs are mostly small, with between one and ten reported employees. They all report an annual turnover of below £500k, with the majority in the £100k to £250k bracket. The two dominant sectors of analysed SMEs in this cluster are Agriculture, Hunting and Forestry, Fishing, and Construction. This is perhaps surprising due to the importance of manufacturing in this region of the country, however this could be explained with the large geographic spread of this cluster. The majority of the SMEs operates as a Sole Proprietorship with a single owner, with the other larger group operating as a Limited Liability Company. The acceptance rate of Cluster 7 is below the average of 91 percent, with 85.18 percent of SMEs being accepted for loan and/or overdraft.

4.5.2 Empirical results

While the clustering results themselves already provide us with interesting insights into the structures of access to bank finance across the UK, they also enable us to construct a sparse **W** matrix for the spatial regression model such as the one described in Section 4.4.1. In this Section, we explore the results of analysing the data according to the model defined by Equations 4.2 and 4.3, when **W** is a sparse matrix constructed using the procedure described in Section 4.4.2.

From the available variables, turnover and sector are used in the clustering process for the **W** construction as done to construct the clusters in the previous Section. As outlined previously, the elements of this matrix, w_{ij} are the distance between i and j for any point pair ij which are in the same cluster. All other elements of **W** are zero. As turnover and sector are included in

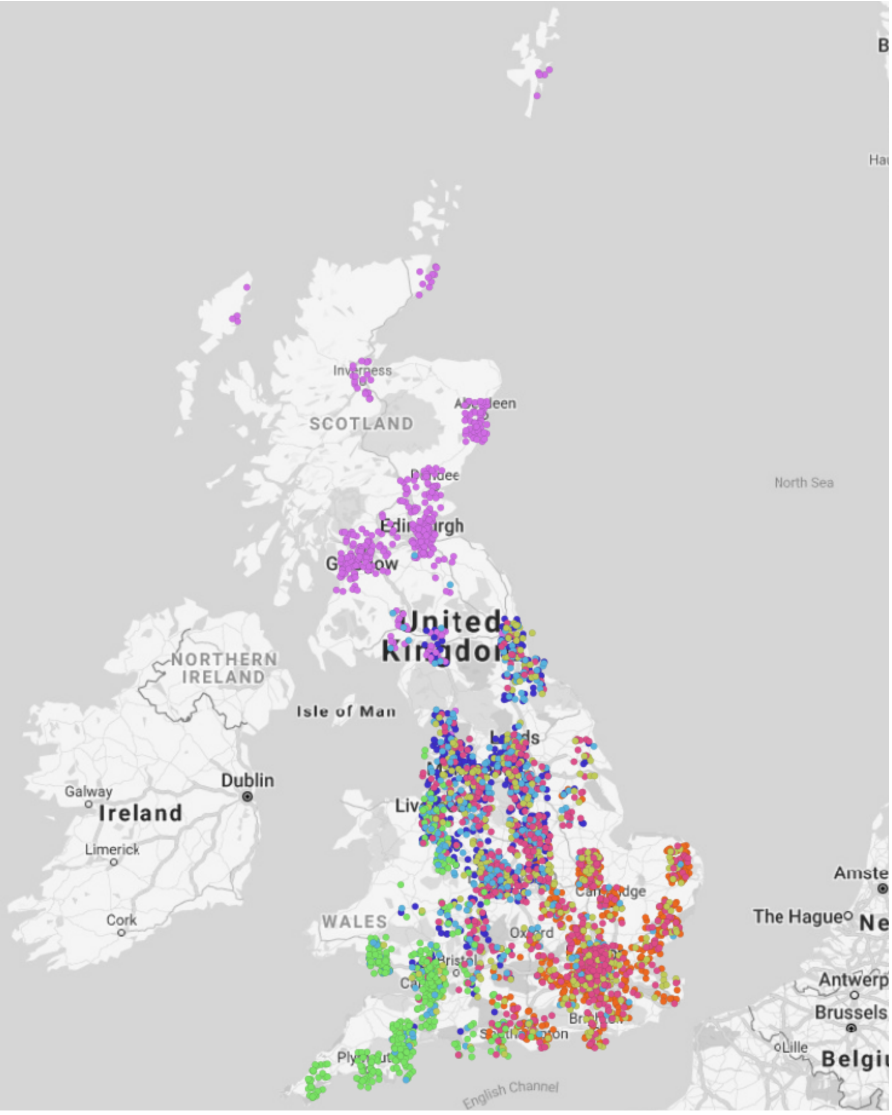

Figure 4.5: Clustering result. Cluster 1 (orange). Cluster 2 (pastel green). Cluster 3 (dark blue). Cluster 4 (strong green). Cluster 5 (purple). Cluster 6 (magenta). Cluster 7 (petrol).

the model within the **W** matrix, the remaining variables, age, legal form and size, are put into the model as explanatory variables in $\boldsymbol{X}$. Legal form is included in the model in the form of three dummy variables. The final model has the form shown in Equation 4.4.

$$\boldsymbol{Y}^* = \rho \mathbf{W} \boldsymbol{Y}^* + \beta_1 age + \beta_2 size + \beta_3 legal + \gamma \tag{4.4}$$

We estimate the spatial autocorrelation ρ and the parameters β_1, β_2 and β_3. Table 4.5 shows the estimates of the implemented spatial probit model acquired using linearised GMM, as described in Section 4.4.1.

	Estimate	std. error	z-value	p-value
Int.	0.79	0.22	3.57	0.00036
age	0.30	0.02	12.59	0.00000
size	0.18	0.03	6.27	0.00000
legal 2	0.17	0.03	5.73	0.00000
legal 3	0.03	0.03	1.01	0.31467
legal 4	0.02	0.03	0.58	0.56460
W	0.47	0.15	3.19	0.00143

Table 4.5: Estimates of the spatial probit model through linearised GMM estimation.

The model fit is assessed using the area under the curve (AUC) measure (Basel Committee on Banking Supervision, 2005). We receive a measure of 0.7357, indicating a fairly good fit when compared to random (0.5).

There are two major findings from the results of this model. We can see a positive relationship between age ($\beta_1 = 0.30$) and size ($\beta_2 = 0.18$) of an SME and its ability to access bank finance. Both these effects are significant ($p = 0.00000$). This implies that older SMEs have a higher chance of being accepted for a loan or overdraft from a bank. The same effect is true for the size of the SME, though this effect is slightly weaker than the one for age. We see mixed results for the legal form of the SME on its access to finance. Legal status 2 (partnership) in comparison with the baseline legal status 1 (sole owner) shows a significant result in its positive impact on access to finance.

With respect to the main focus of this chapter, our findings confirm there are spatial effects taking place which affect the SMEs' access to finance. These effects are positive and significant (0.47, $p = 0.00143$). This indicates that if an SME is able to access bank finance in the form of loans or overdrafts, other SMEs which are both similar and close to them also have a higher chance of doing so. We defined this similarity in terms of their annual turnover and sector, as well as physical closeness. Our findings therefore confirm that taking into account connectedness between SMEs, but especially the connectedness between *similar* SMEs, is of high importance when discussing their access to finance. It is important to note here that we make statements about a correlation effect taking place and can not make absolute statements about the reason behind them or the causality of them. One also has to take into account that by using turnover and sector as clustering variables, the pure spatial effects are difficult to extract. We therefore compare this model with two further models with $\mathbf{W}$ being constructed without additional variables.

4.5.3 Model comparison

In order to dissect the spatial effects from the effects of turnover and sector, we compare our model with three other approaches. The first comparison model is a simple logistic regression with no spatial element or W matrix. This model is considered a baseline approach if no spatial effects are considered for the explanation of loan and overdraft acceptance. In the second model, the $\mathbf{W}$ is a binary contingency matrix which indicates neighbouring postcodes with a 1 in the respective field and zero otherwise. Finally, the third model is similar to our proposed model, but the clustering of SMEs is done using their spatial location only.

The estimates of comparison model 1 in Table 4.6 shows that turnover and age continue to be significant when compared to the proposed model, with $p = 0.0000$, and they have a positive correlation with a successful loan or overdraft application. Some of the dummy variables for the legal status and operating sector of the SME also show a significant correlation at a chosen threshold of $p < 0.05$.

The comparison with model 2 (binary $\mathbf{W}$) in Table 4.7 shows that compared to our proposed model in Section 4.5.2 some factors still have a significant effect such as age, size and some of

	Estimate	std. error	z-value	p-value
Int.	0.91	0.00	189.87	0.0000
turnover	0.04	0.01	5.89	0.0000
age	0.06	0.01	10.79	0.0000
legal factor 2	0.02	0.01	3.01	0.00260
legal factor 3	0.00	0.01	0.54	0.59098
legal factor 4	0.00	0.01	0.01	0.99502
size	0.00	0.01	0.52	0.60169
sector factor 2	-0.01	0.01	-1.24	0.21539
sector factor 3	-0.02	0.01	-2.51	0.01197
sector factor 4	0.00	0.01	-0.71	0.47972
sector factor 5	-0.01	0.01	-2.34	0.01942
sector factor 6	-0.02	0.01	-2.64	0.00846
sector factor 7	0.00	0.01	-0.66	0.50986
sector factor 8	0.00	0.01	0.25	0.80643
sector factor 9	0.00	0.01	-0.56	0.57374

Table 4.6: Estimates of comparison model 1 as a simple logistic regression with no spatial elements and all variables in the explanatory matrix X.

the legal statuses and sectors. This is also similar to the comparison baseline model 1. However, compared to the proposed model the spatial effects as measured by ρ, while still present (0.00145), is not significant anymore. This implies that the significance of the **W** in our model could either be coming through the inclusion of turnover in its construction, considering that turnover is shown in comparison model 1 as having a significant effect. The other explanation is that a binary **W** is not sufficiently detailed to capture the connections between SMEs. Being in neighbouring postcode areas may not be enough to explain connectedness.

Comparison model 3 shows the same results for the variables age and size, in that both are significant in their positive effect on access to finance in the form of loan and overdrafts. The spatial effect in the form of **W** is still not significant.

These two spatial models demonstrate that being merely spatially close is not enough to explain interdependencies between SMEs regarding their access to finance. Spatially close SMEs which are different in turnover and sector are not necessarily connected in regards to their access or their connection does not impact it. In order to experience significant contagion effects SMEs do not only have to be physically close, but in addition they also have to be similar to each other.

	Estimate	std. error	z-value	p-value
Int.	1.52	0.07	20.54	0.00000
turnover	0.27	0.04	7.42	0.00000
age	0.26	0.02	10.67	0.00000
legal 2	0.14	0.03	4.49	0.00001
legal 3	-0.01	0.03	-0.33	0.74237
legal 4	-0.05	0.04	-1.39	0.16734
size	0.08	0.04	2.25	0.02448
sector 2	-0.08	0.03	-2.53	0.01117
sector 3	-0.14	0.04	-3.96	0.00007
sector 4	-0.04	0.03	-1.07	0.28612
sector 5	-0.12	0.03	-3.90	0.00010
sector 6	-0.12	0.03	-3.66	0.00025
sector 7	-0.04	0.04	-1.09	0.27649
sector 8	-0.01	0.03	-0.28	0.77733
sector 9	-0.05	0.03	-1.56	0.11795
W	0.00001	0.00	0.01	0.99318

Table 4.7: Estimates of comparison model 2 with a binary **W**.

4.6 Discussion and implications

This chapter makes two important contributions to the discussion of access to finance for small and medium-sized enterprises. One, when explaining access to bank finance, this chapter finds evidence of spatial effects impacting it when taking into account connectedness of SMEs through similarity. Our findings confirm that contagion effects exist, but between SMEs which are both similar in terms of turnover, sector and geographically close. The second contribution is that, through exploring structures in the UK SME landscape with clustering we find evidence for groups of SMEs which differ regarding their access to finance. We describe these groups in terms of their geographic location, operating sector and turnover, and show differences in their acceptance rates.

Our findings confirm previous research in the area of access to finance for SMEs in showing that the age and size of a company have a positive relationship with access to finance (Artola & Genre, 2011; Ferrando & Griesshaber, 2011; Owen et al., 2016; Presbitero et al., 2014). This shows that older and larger SMEs have easier access to bank loans and overdrafts in the UK. This chapter, however, makes the important contribution of showing that additionally there are also spatial effects. We detect these spatial effects through the implementation of a spatial regression

	Estimate	std. error	z-value	p-value
Int.	0.67	1.21	0.55	0.57997
turnover	0.27	0.04	7.54	0.00000
age	0.26	0.02	10.62	0.00000
legal 2	0.15	0.03	4.74	0.00000
legal 3	-0.01	0.03	-0.39	0.69385
legal 4	-0.05	0.04	-1.43	0.15359
size	0.07	0.03	2.11	0.03457
sector 2	-0.07	0.03	-2.34	0.01940
sector 3	-0.13	0.03	-3.87	0.00011
sector 4	-0.03	0.03	-0.84	0.40029
sector 5	-0.11	0.03	-3.75	0.00017
sector 6	-0.11	0.03	-3.60	0.00032
sector 7	-0.03	0.04	-0.79	0.43069
sector 8	-0.01	0.03	-0.22	0.82907
sector 9	-0.04	0.03	-1.40	0.16082
W	0.56	0.79	0.70	0.48236

Table 4.8: Estimates of comparison model 3 with a spatially clustered **W**.

model and show that positive contagion effects between SMEs exist. Previous research has called for the consideration of connectedness and networks between SMEs (Vos et al., 2007), and this notion was used in other SME related contexts (Calabrese et al., 2019; Fernandes & Artes, 2016; F.-J. Lin & Lin, 2016; Tiwasing et al., 2019). However, until now it has found little application in access to finance. We close this gap by considering connectedness of SMEs using clustering and its notion of similarity in considering dependency structures. This allows us to account for spatial effects taking place between companies which are similar to each other and therefore *connected*. Previously, Lee and Luca (2019) claim that their finding of easier access to finance in urban regions could potentially be evidence for access being dependent on spatially bound networks. Our findings show that spatially dependent effects do exist between SMEs. We therefore also support and strengthen the call for accounting for spatial dimensions when analysing the finance landscape for SMEs. This is not only of high importance when trying to understand the behaviour of businesses, but can also inform policy making on a more detailed and local level.

One thing to consider when discussing our results is that the variables turnover and sector were used for the clustering and therefore not included in the explanatory variables of the regression model. Based on this, the spatial contagion effects that we see in our model may be

coming from two possible sources. One possibility is that any effects that we see in the spatial dependencies are purely coming from the effect turnover and sector have on access to finance, with no variable-independent spatial effects. The other possibility is that turnover and sector are merely strengthening the effect of spatial contagion. One should keep in mind that spatial effects themselves could be understood as originating from spatial structures of other variables such as turnover. Spatial dependencies in turnover or sector can be seen as detached from their spatial dimensions as pure variable effects or as spatial effects driven through underlying dynamics. We here make the claim that the spatial distribution and dependencies of financial variables such as turnover area important to consider and include in modelling efforts.
The comparison of our approach with three models, a simple logistic regression and two spatial regressions one with binary connections indicating neighbouring postcodes and one with connections through spatial closeness without other similarity factors, showed that these changes to the construction of the spatial dependency lead to no presence of significant spatial effects. The significant positive effects of age and size, however, stayed the same. This implies that the inclusion of turnover and sector, and probably mostly turnover when considering that its inclusion in the explanatory variables in comparison models 1, 2 and 3 showed a strong significant effect, impacted the spatial contagion effects significantly.
Further research into the interpretation of the role of spatial effects and their connection to potentially spatially dependent variables such as turnover is needed to deepen our understanding on this further.

In the interpretation of our results it is important to keep in mind that while correlation has been detected in the form of contagion effects, we can not make definite statements about the source of those effects. There are a number of possible reasons for similar and close companies to experience similar success in loan and overdraft applications. One explanation is the existence of networks between SMEs through which these positive effects are distributed, which we here imply through our use of the word 'contagion'. However, an alternative explanation would be for banks to apply a similar decision making process to similar companies. While having the same outcome, this effect would be stemming from the banks instead of spreading between the SMEs independently. Another limitation of the present research is that the model variables so far are limited to location, age, size and legal status of the SME. These factors have been highlighted by the literature as being important, which our findings further confirm, but additional insights can be gained by including variables such as growth indicators and characteristics of the owner. Besides adding additional variables, future research could also consider temporal effects and changes. The available dataset spans eleven years and is updated quarterly, offering a rich and dynamic picture of the UK financing landscape. It could also be argued that the existing spatial contagion effects only take place in one directon time-wise, that is into the future. Furthermore, the addition of the spatial dimension to this analysis of access to finance also opens up further opportunities in the consideration of economic and environmental factors surrounding and affecting the SMEs. Mwaura and Carter (2017) have previously shown that the use and demand of bank finance in the UK is very heterogeneous and that this diversity has to be taken into account when analysing SMEs' access to finance. Our research can inform policy makers in formulating these local level policy changes. Through this SMEs can be supported with targeted and tailored help, allowing them to strengthen local communities and continuing to be the stable backbone of the wider UK economy.

In addition to these contributions to the literature on SME financing as well as the use of clustering to discover groups of such SMEs for policy making, our methodological findings carry further weight beyond this application context. We introduce a two-step process to spatial logistic regression which utilises cluster analysis as a first step, in order to restrict the detection of spatial contagion effects to happen within groups of similar and geographically close companies. This method of using clustering as a first step to cluster the W matrix has, to the best of our knowledge, not been used before in the literature. However, we believe it could be a valuable method to use for researchers who are looking to add a level of granularity to their regression analysis. The use of clustering has a number of advantages, mainly in its ability to detect groups of data points without the need of labelling and its ability to include a number of variables of different format and source. For example, we have previously demonstrated that clustering can also be used for time series clustering, which could add another layer of information to this analysis. For instance, it would allow the researcher to separate spatial autocorrelation effects to affect only data points exhibiting a similar behaviour at the same point in time. The second modelling step does not need to be a spatial logistic regression model either. It would be possible by using clustering as a first step in a modelling process to allow for models to detect within group dynamics and separate them from between group ones. Clustering can also detect potential networks among group members, allowing for the analysis of within network effects that only affect members of a group. This shows the value that clustering can bring to the analysis of various scenarios beyond the analysis of SME financing.

Appendix A

Table 4.9: Cluster description of resulting eight groups

Variable	Label	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Region	Scotland	-	-	2 (0.39)	-	**331 (92.46)**	1 (0.15)	6 (1.62)
	North/North East	-	29 (6.89)	68 (13.2)	2 (0.52)	21 (5.87)	25 (3.63)	37 (9.97)
	Yorkshire/Humberside	-	37 (8.79)	104 (20.19)	-	1 (0.28)	61 (8.87)	42 (11.32)
	North West	-	42 (9.98)	**130 (25.24)**	11 (2.87)	3 (0.84)	66 (9.59)	59 (15.9)
	West Midlands	-	42 (9.98)	87 (16.89)	25 (6.53)	-	66 (9.59)	**74 (19.95)**
	East Midlands	33 (6.72)	40 (9.5)	71 (13.79)	-	1 (0.28)	79 (11.48)	37 (9.97)
	East Anglia	137 (27.9)	42 (9.98)	-	-	-	71 (10.32)	9 (2.43)
	Wales	1 (0.2)	10 (2.38)	32 (6.21)	110 (28.72)	-	12 (1.74)	39 (10.51)
	South West	11 (2.24)	30 (7.13)	20 (3.88)	**232 (60.57)**	1 (0.28)	43 (6.25)	34 (9.16)
	London	110 (22.4)	**78 (18.53)**	-	-	-	**140 (20.35)**	12 (3.23)
	South East	**199 (40.53)**	71 (16.86)	1 (0.19)	3 (0.78)	-	124 (18.02)	22 (5.93)
Mode	Most frequent region	South East	London	North West	South West	Scotland	London	West Midlands
Age: Company was first established...	Less than 12 months ago	8 (1.63)	30 (7.13)	3 (0.58)	7 (1.83)	7 (1.96)	4 (0.58)	13 (3.5)
	Over 1 but under 2 years ago	17 (3.46)	42 (9.98)	2 (0.39)	13 (3.39)	9 (2.51)	13 (1.89)	29 (7.82)
	2 - 5 years ago	34 (6.92)	79 (18.76)	34 (6.6)	36 (9.4)	20 (5.59)	42 (6.1)	47 (12.67)
	6 - 9 years ago	46 (9.37)	44 (10.45)	48 (9.32)	49 (12.79)	36 (10.06)	77 (11.19)	37 (9.97)
	10 - 15 years ago	99 (20.16)	68 (16.15)	73 (14.17)	64 (16.71)	63 (17.6)	129 (18.75)	58 (15.63)
	More than 15 years ago	**287 (58.45)**	**158 (37.53)**	**355 (68.93)**	**214 (55.87)**	**223 (62.29)**	**423 (61.48)**	**187 (50.4)**
Median	Median of company age	6 (¿15 years)	5 (10-15 years)	6 (¿15 years)	6 (¿15 years)	6 (¿15 years)	6 (¿15 years)	6 (¿15 years)
Size: Number of employees	1 employee	16 (3.26)	135 (32.07)	5 (0.97)	23 (6.01)	32 (8.94)	3 (0.44)	112 (30.19)
	2-10 employees	179 (36.46)	**228 (54.16)**	104 (20.19)	146 (38.12)	120 (33.52)	118 (17.15)	**211 (56.87)**
	11-50 employees	**212 (43.18)**	51 (12.11)	**275 (53.4)**	**159 (41.51)**	**151 (42.18)**	**388 (56.4)**	43 (11.59)
	51-100 employees	67 (13.65)	3 (0.71)	88 (17.09)	33 (8.62)	41 (11.45)	120 (17.44)	3 (0.81)
	101-200 employees	14 (2.85)	3 (0.71)	34 (6.6)	20 (5.22)	11 (3.07)	51 (7.41)	2 (0.54)
	201-250 employees	3 (0.61)	1 (0.24)	9 (1.75)	2 (0.52)	3 (0.84)	8 (1.16)	-
Median	Median of number of employees	3 (11-50)	2 (2-10)	3 (11-50)	3 (11-50)	3 (11-50)	3 (11-50)	2 (2-10)
Turnover: Last annual turnover	Less than £25,000	-	79 (18.76)	-	6 (1.57)	13 (3.63)	-	60 (16.17)
	£25,000 - £49,999	4 (0.81)	94 (22.33)	-	9 (2.35)	21 (5.87)	-	63 (16.98)
	£50,000 - £74,999	1 (0.2)	64 (15.2)	-	11 (2.87)	20 (5.59)	-	70 (18.87)
	£75,000 - £99,999	11 (2.24)	56 (13.3)	-	22 (5.74)	14 (3.91)	-	49 (13.32)
	£100,000 - £249,999	60 (12.22)	**128 (30.4)**	-	66 (17.23)	55 (15.36)	-	**108 (29.11)**
	£250,000 - £499,999	84 (17.11)	-	53 (10.29)	**85 (22.19)**	56 (15.64)	107 (15.55)	21 (5.66)
	£500,000 - £999,999	96 (19.55)	-	106 (20.58)	69 (18.02)	52 (14.53)	120 (17.44)	-
	£1m - £1.9m	**137 (27.9)**	-	**188 (36.5)**	73 (19.06)	**70 (19.55)**	**268 (38.95)**	-
	£2m-4.9m	54 (11.0)	-	75 (14.56)	31 (8.09)	29 (8.1)	101 (14.68)	-
	£5m - £9.9m	24 (4.89)	-	41 (7.96)	10 (2.61)	18 (5.03)	42 (6.1)	-
	£10m - £14.9m	13 (2.65)	-	22 (4.27)	1 (0.26)	4 (1.12)	25 (3.63)	-
	£15m-24.9m	7 (1.43)	-	30 (5.83)	-	6 (1.68)	25 (3.63)	-
Median	Median of turnover	7 (£500k-999k)	3 (£50k-74.9k)	8 (£1m-1.9m)	6 (£250k-499k)	6.5 (£250k-499k)	8 (£1m-1.9m)	3 (£50k-74.9k)
Sector: Primary operating sector of company	Agriculture, Hunting and Forestry, Fishing	79 (16.09)	-	63 (12.23)	23 (6.01)	**62 (17.32)**	-	**122 (32.88)**
	Manufacturing	90 (18.33)	-	106 (20.58)	15 (3.92)	27 (7.54)	-	79 (21.29)
	Construction	**165 (33.6)**	-	**152 (29.51)**	58 (15.14)	56 (15.64)	-	117 (31.54)
	Wholesale / Retail	88 (17.92)	10 (2.38)	116 (22.52)	52 (13.58)	41 (11.45)	-	37 (9.97)
	Hotels and Restaurants	53 (10.79)	34 (8.08)	47 (9.13)	64 (16.71)	45 (12.57)	7 (1.02)	16 (4.31)
	Transport, Storage and Communication	16 (3.26)	77 (18.29)	29 (5.63)	45 (11.75)	37 (10.34)	110 (15.99)	-
	Real Estate, Renting and Business Activities	-	**127 (30.17)**	2 (0.39)	**65 (16.97)**	48 (13.41)	**289 (42.01)**	-
	Health and Social Work	-	69 (16.39)	-	29 (7.57)	13 (3.63)	93 (13.52)	-
	Other Community, Social and Personal Service Activities	-	104 (24.7)	-	32 (8.36)	29 (8.1)	189 (27.47)	-
Mode	Most frequent sector	3 (Constr.)	7 (Real Est.)	3 (Constr.)	7 (Real Est.)	1 (Agr.)	7 (Real Est.)	1 (Agr.)
Legal status	Sole Proprietorship (single owner)	40 (8.15)	172 (40.86)	22 (4.27)	47 (12.27)	66 (18.44)	24 (3.49)	145 (39.09)
	Partnership	57 (11.61)	45 (10.69)	45 (8.74)	61 (15.93)	71 (19.83)	61 (8.87)	92 (24.8)
	Limited Liability Partnership	38 (7.74)	19 (4.51)	36 (6.99)	20 (5.22)	29 (8.1)	51 (7.41)	10 (2.7)
	Limited Liability Company (private LC, public)	356 (72.51)	185 (43.94)	412 (80.0)	255 (66.58)	192 (53.63)	552 (80.23)	124 (33.42)
Mode	Most frequent legal status	LLC	LLC	LLC	LLC	LLC	LLC	SP
Acceptance	Percent of SMEs accepted	91.04	81.95	95.15	91.91	92.74	95.2	85.18
Total	Number of SMEs total	491	421	515	383	358	688	371

Appendix B

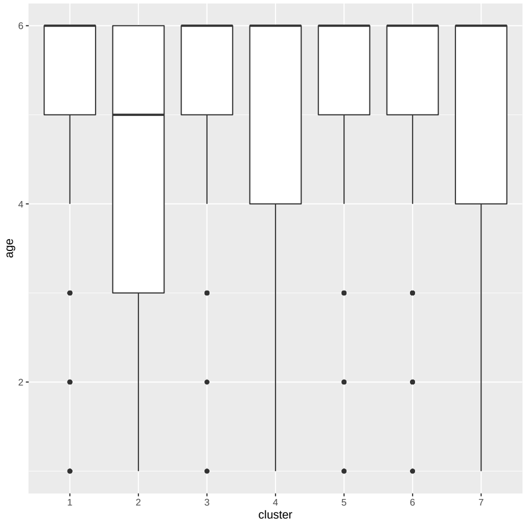

Figure 4.6: Age distribution per cluster.

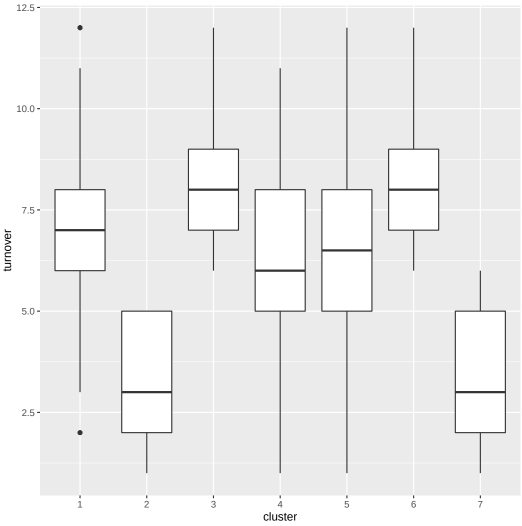

Figure 4.7: Turnover distribution per cluster.

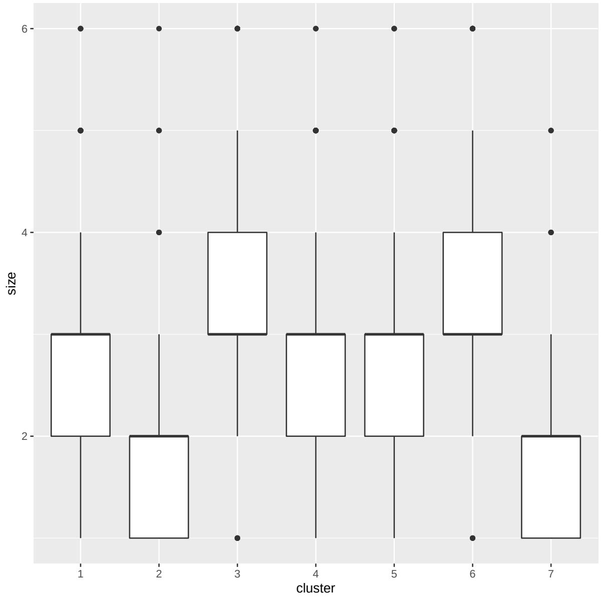

Figure 4.8: Size distribution per cluster.

Chapter 5

Understanding the impact: Segmenting countries by their intention to visit tourist destinations post COVID-19

This chapter will demonstrate the use of clustering in tourism by segmenting countries based on how likely they are to visit Edinburgh and Scotland in the future. We will explore why this is an important topic in the recovery of the Scottish economy after the COVID-19 pandemic, and identify three factors which can be used to analyse tourist behaviour: intention, affordability and reachability. With the latter, this Chapter will contribute to the previously explored idea of closeness in Chapter 3, by defining closeness or reachability based on flight connections instead of geographic distance. This will be achieved by introducing a reachability index which calculates the distance of two locations based on the number of daily flights and their directness.

This chapter is based on a research project funded by the Data-Driven Innovation initiative. It was supported by the Edinburgh Tourism Action Group (ETAG) and its partners. The resulting tool has been published under the domain `https://tourism-insights-scotland.co.uk`.

5.1 Introduction

On 30 January 2020, the Director-General of the World Health Organization declared the outbreak of COVID-19 a public health emergency of international concern (World Health Organization, 2020). Since then, the global COVID-19 pandemic in 2019-2021 had and will have profound impacts on countries' economies, societies, and daily lives. Case numbers reached 115m confirmed cases worldwide in March 2021, 4M of which are in the United Kingdom (Johns Hopkins University, 2020). Countries are being affected at different rates with some regions showing recovery and others reporting accelerating case numbers. On 23 March 2020, the UK Foreign Secretary advised all British travellers to return to the UK immediately (Foreign & Commonwealth Office, 2020). Other countries such as Germany also advised their citizens to refrain from any international travel (The Federal Government of Germany, 2020), resulting in international tourism in Europe coming to a de-facto standstill. The European Union decided to lift their worldwide travel warning on the 15 June 2020 and replaced it with country specific warnings, indicating that tourism would be slowly starting again in the months to come. However, in early 2021 international restrictions were re-introduced due to rising case numbers and new variants of the virus found in different parts of the world.

From a business perspective, these months of travel warnings and restrictions mean uncertainty. Businesses in the tourist industry and those heavily depending on tourists as customers are especially hard hit. Even after restrictions and warnings are being lifted eventually, it is uncertain how tourism will recover. Countries are being affected by the virus at different rates, and restrictions are being put into place on national level. Further lockdowns across Europe and the UK in autumn and winter of 2020 show that the expected recovery will not follow a linear path with frequent setbacks and changes being expected.

Tourism is an important part of the Scottish economy. The Scottish government's Tourism and Events Division reports that tourism contributes £12 billion of economic activity for the wider Scottish supply chain and around £6 billion or 5% to Scottish GDP. In doing so the sector also supports 217,000 jobs in 2015 (Tourism and Events Division, 2020). Edinburgh as the capital of Scotland with frequent travel connections to other countries works as a hub for tourists to visit the country and relies heavily on tourism itself.

Understanding tourists' intentions to visit can inform businesses on whom their communication efforts should focus and what to expect in terms of business effects. This is of crucial interest especially in the context of the ongoing pandemic. Common patterns of tourist behaviour are being disrupted and change on an ongoing basis. The analysis of tourism demand and behaviour within and outside crises has been the subject of a diverse set of analyses in the literature. Models often employ time series methods to allow for the consideration of trends and seasonality. Yet they often focus on forecasting the aggregated number of tourists visiting a destination instead of taking a broader view. From a company's perspective it can be helpful to understand whether their usual pattern of visitors can be expected and which countries should be targeted with marketing efforts.

The objective of this chapter is to group countries according to the likelihood of their inhabitants visiting Edinburgh and Scotland. For this, clustering is used in an effort to group together countries which show a similar expected visiting behaviour. Segmentation in marketing is a useful tool for allowing companies to target groups of customers or, in this case, groups of countries with targeted marketing strategies instead of addressing the whole audience with a single message. In our case, these groups of countries can be interpreted in terms of how likely or unlikely tourists are to visit, which can be useful for strategy formulation by local companies and policy makers, for example in terms of which countries should be addressed first as they are the most likely to visit. Clustering was chosen as a method in this case because it enables us to group together countries in such a way that they show a similar behaviour to each other that is different from the other groups. These groups can be formed taking into account different variables and dimensions in this formulation of similarity and without the need for pre-labelling countries for belonging to any group. This allows us to approach this question with a relatively limited amount of expert knowledge in this area.

The approach to clustering we propose in this chapter uses multiple relevant behavioural features simultaneously. The three features chosen were Intention, Affordability and Reachability. These factors were chosen as they can be used to explain different aspects of customer behaviour. Intention has been characterised as the planning of a trip and the perceived risk of doing so. Affordability represents the financial ability of the tourists to travel, impacted for example by the economic situation. A straightforward approach would consider reachability as defined by the geographic distance. However, we believe that the possibility to travel according to available travel routes is a better measure of such concept. We hence propose in this chapter a new reachability index based on flight connections.

Three data sources are used for the analysis, with Google search data as a proxy for intention, the Consumer Confidence Index for affordability and airport connections from Skyscanner for reachability.

The main contribution of this chapter is to introduce a novel reachability index which takes into account both the number and directness of available flight connections between two locations. This is chosen instead of geographic distance because of the way that geographical closeness of a location does not necessarily accurately represent whether it can be reached by a tourist. If no flights are available, it would be very inconvenient for someone to travel even if they were geographically close. Therefore, in our presented case whether someone is able to travel from their country to Edinburgh depends more on available flights, which could for example be impacted by things like the pandemic, affecting how many flights an airline offers.

Another challenge arises due to the spikiness of Google search data. This data is used to take into account customer intentions through the analysis of time series data of their Google searches. As we find these times series to be very spiky, we compare three smoothing approaches: exponential smoothing, cubic smoothing splines and weekly averages.

We finally compare our model with seven other model configurations in order to discuss the impact that the three dimensions in different combinations, as well as the smoothing of the volatile time series, have on our findings.

This chapter will present an approach useful for practitioners in their forecasting and decision making process. We find that while some countries show both increased interest in Edinburgh and Scotland, not all of them are in the economic position to visit and feel that Edinburgh is a reachable destination, and others are in a different situation regarding one or multiple of these aspects. Businesses can use these groups of countries when designing their marketing strategy and focus on targeting those customers that are most likely to travel. By using time series data the model can be updated on an ongoing basis and reflect the frequent changes that are to be expected during the recovery process. As such, it can also be used by policy makers in an effort to understand how the Scottish economy will be impacted and what a possible route to recovery can look like.

The chapter is organised as follows: Section 5.2 will explore the existing literature on the impact crises in general and COVID-19 in particular have on tourism, as well as how tourism behaviour is analysed using search and flight connection data and economic indices. In Section 5.3 we will introduce our method of analysing the impact of the COVID-19 pandemic on tourist's intention and ability to visit Edinburgh and Scotland, with results presented in Subsection 5.5. Finally, we will discuss and conclude based on our findings and their implications.

5.2 Background

As much as the COVID-19 pandemic situation is rapidly evolving, so is the existing literature on its impact on economies around the world and on tourism. Pandemics and other international crises have previously impacted the tourism industry. Generally, we can distinguish between natural disasters which include pandemics as well as earthquakes or tsunamis, and human-made disasters including terrorist attacks and economic crises (Zenker & Kock, 2020). Some authors, such as Zenker and Kock (2020), argue that we can draw from the existing literature on past disasters and their impact on the tourism industry for analysing the impact that the COVID-19 pandemic will have. They do, however, make the point that while we can use this existing base as a starting point, some aspects of COVID-19 are unprecedented, mostly due to its complexity. Gössling, Scott, and Hall (2020) note that none of the other major events during 2000 and 2015, namely the 9/11 terrorist attacks, SARS and MERS outbreaks, or the economic crisis of 2008/09, had any significant or long-lasting effect on the global tourism. Yet the COVID-19 pandemic presents a unique mixture of a multitude of different forms of crises, including, of course, tourism demand but also socio-economic and economic aspects (Zenker & Kock, 2020).

Tourism has been significantly affected by the COVID-19 pandemic through travel restrictions, uncertainty around the safety of travelling, imposed quarantine, the cancellation of major events, as well as the closure of hotels and gastronomy (Gössling et al., 2020). It was estimated that up to 90% of the world's population was impacted by some form of travel restrictions (Gössling et al., 2020). The airline industry has been particularly hard hit by the COVID-19 pandemic, with Eurocontrol (2021) reporting a reduction of 90% in the number of daily flights and a full recovery to previous levels not expected in Europe before 2024.

However, while the pandemic has affected the overall mobility of tourists, countries are affected at different rates. Restrictions, which restrict people from travelling out of countries as well as into countries, are put in place on national and sometimes even sub-national level. As an example for how complex the situation is, professional service network Deloitte continued publishing COVID-19 mobility updates which outline changes to country specific restrictions (Deloitte, 2020). And in their discussion of travel restrictions from a geo-political perspective, Seyfi, Hall, and Shabani (2020) describe the situation for the case of Australia where individual states were making their own judgements and decisions regarding their border policy: "The New South Wales state premier, Gladys Berejiklian, argued that interstate travel would be an important part of economic recovery from the pandemic and suggested that other states would miss out if they kept their borders closed. At the time the states of Queensland, South Australia, and Tasmania, as well as the Northern Territory, had their borders closed, and were yet to announce when they would be reopened." As such, when analysing how tourism will recover rather than a global view a more nuanced level view will be needed.

Not only restrictions but also the perceived risk means that COVID-19 has a significant impact on the intentions of tourists to travel. Matiza (2020) note that the impact of the pandemic on the demand side in the form of consumers' perceptions of risk is particularly important for the tourism industry. In the context of the hospitality industry, Gursoy and Chi (2020) state that factors such as test and trace systems, local case numbers, and available vaccinations all impact if and when people feel comfortable travelling again. These are all factors which are not only decided by individual businesses and their implemented safety features, but also refer to country-level regulations and situations. Kourgiantakis, Apostolakis, and Dimou (2020) find that while most people surveyed still intent to travel, the pandemic impacted both their certainty about if and when they would pursue their plans and whether they would travel more locally. Similarly, Graham, Kremarik, and Kruse (2020) found that 60% of their sample of elderly UK residents were planning to travel within 12 months when asked in 2020, however 30% planned more local trips and 60% planned overall less trips than usual. Bulchand-Gidumal and Melián-González (2021) find that 28% of their respondents will delay buying a flight ticket for their travels, indicating how other sources of information such as search data could offer more insights into travel intention than actual booking information. The use of air travel data in tourism recovery efforts has been shown by Gallego and Font (2020). Specifically, they assessed future travel demand through the analysis of Skyscanner search and pick data. Gallego and Font (2020) find that demand as measured by the number of flight searches on Skyscanner has dropped by 30% in Europe and the Americas, and by about 50% in Asia. They do acknowledge that the applicability of their research to policy makers depends on their willingness and ability to employ Big Data sources and subsequent analysis techniques.

In our digital age, travel plans often start on the internet. Tourists use search engines to gather information and plan an upcoming trip. As such, search data can be an indicator for future behaviour. Google search data, which is accessible to the public in the form of Google Trends[1], has been used to predict economic behaviour in different ways. Choi and Varian (2012) demonstrate that Google Trends data can be used to make short-term predictions of automobile sales, unemployment benefits claims, consumer confidence index, and visitor arrivals to Hong Kong. They show that Autoregressive (AR) models including Google Trends data outperform those not containing it. This is explained with the correlation of Google searches and various economic indicators. In a tourism context specifically, Bangwayo-Skeete and Skeete (2015) incorporate Google Trends data into their Autoregressive Mixed-Data Sampling (AR-MIDAS) model predicting the number of tourists arriving from US, Canada and UK to five different countries in the Caribbean. Their work supports previous findings and shows that incorporating Google search data resulted in an improvement of the model's prediction as measured by Root Mean Square Error (RMSE) and mean absolute percentage error (MAPE). In a similar manner, Google Trends has been used to forecast tourism demand and arrivals for a variety of countries including Germany, the Czech Republic, Switzerland, Austria and Belgium (Bokelmann & Lessmann, 2019; Havranek & Zeynalov, 2021; Siliverstovs & Wochner, 2018; Önder, 2017) and aspects related to tourism such as hotel room demand (Pan, Wu, & Song, 2012). Y. Yang, Altschuler, Liang, and Li (2020) use Google search data in the development of their COVID19tourism index as an indicator for tourism interest in travelling. When analysing data in a global context it is important to understand differences in the usage of different platforms. The pre-dominant search engine in China is not Google but Baidu. Search data taken from that platform has thus been used for forecasting tourism demand from Chinese visitors (Huang, Zhang, & Ding, 2017) and has been shown better performance when compared to Google (X. Yang, Pan, Evans, & Lv, 2015).

We have seen that the COVID pandemic has profound effects on the tourism industry. Understanding and planning for tourist behaviour is therefore important for businesses to recover. Polyzos, Samitas, and Spyridou (2020) estimate the impact of the COVID-19 pandemic on tourism from China to the USA and Australia using a type of Neural Network called Long Short Term Memory (LSTM) model. Analysing monthly arrival data of Chinese tourists to the USA and to Australia spanning 2003-2019, their time series data includes the SARS outbreak and ends

[1] https://trends.google.com/

just before the COVID-19 pandemic. Based on their findings they estimate that it could take up to one and a half years for tourism to return to pre-pandemic levels. This shows that approaches addressing tourism recovery will need to be implementable long-term and be adaptable to changes that occur during this time.

5.2.1 Factors impacting tourism behaviour

Tourist behaviour is complex and multi-faceted, with a number of factors impacting it such as the economic situation of the country the tourist is travelling from or the reachability of the travel location.

The expected economic situation of a person impacts their plans and behaviour in the future. Therefore, a measure of optimism about the future can be used to understand future consumption behaviour including in tourism. An important indicator for the expected future financial situation of consumers is the Consumer Confidence Index which is published internationally by the OECD (OECD, 2020). A value over 100 indicates optimism in the future and thus consumers are more likely to spend, while a value under 100 indicates pessimism and thus more inclination to save. The index is often used for forecasting stock market prices, for example in Demir and Ersan (2018) when analysing stock prices of tourism companies listed on the Turkish stock market and M.-H. Chen (2015) for stock prices of hospitality companies in Taiwan. It is also a suitable index for forecasting consumer spending (Juhro & Iyke, 2020) and, within that, tourism behaviour (Xie, Qian, & Wang, 2021). The COVID-19 pandemic presents a special case due to its impact on global consumer confidence and the slow recovery. In the early phase of 2020 the overall global index showed a drop from its previous level of around 100 to 97.6 in May 2020, indicating global pessimism, which is now slowly recovering to 98.6 in March 2021 (OECD, 2020). Country specific indices allow for a comparison of the recovery in different regions of the world.

In an increasingly connected world the concept of spatial distance is of decreasing importance. Affordable and direct long haul flights connect destinations much quicker and more comfortable than in the past. Instead of considering spatial distance, we believe the concept of reachability will play a more important role. The concept of perceived distance has been discussed in the context of tourist mobility within destinations when visitors build networks of locations they are visiting (Asero, Gozzo, & Tomaselli, 2016). Hwang, Gretzel, and Fesenmaier (2006) note that transportation links are an important part of defining perceived distance in multi-city travel. In their paper on a stronger focus on sustainable tourism as a possible outcome from the COVID-19 pandemic, Romagosa (2020) discuss how they believe the pandemic will lead to an increase in so-called proximity tourism. This concept describes tourism and travelling that is conducted closer to the home of the tourist. It shows the relevance of a concept of closeness as well as reachability in tourism. The authors believe this will be driven both by a feeling of increased security when the traveller is staying closer to their place of residence, but also the availability of flights to locations further away which might be impacted for a long period of time. The consideration of reachability has also been discussed by Liu, Vici, Ramos, Giannoni, and Blake (2021), who implement a model to forecast the recovery of tourism demand based on two indices which they call “accessibility risk” and a country’s “self-protecting measures”. The accessibility risk index takes one of three categories for each country, indicating whether they can be reached by land/water transport under six hours, short haul flights under three hours, or long haul flights over three hours. Based on this analysis the authors assign for example Australia a high accessibility risk value due to their reliance on long haul flights for tourists to reach them.

From the literature review it is clear that there seems to be a missing connection between different aspects of tourist behaviour, especially when analysing it during a crisis such as the COVID-19 pandemic. The economic situation and the location reachability have been shown to be important factors in the analysis of behaviour. Interest in visiting , which would more directly reflect travel plans, has not been used in the literature in combination with more economic and distance based factors. Furthermore, in order to be useful for companies to work on, the results of the analysis have to provide insights into structures of countries. Such structures can be reflected in the form of groups of countries. A way to achieve this is through cluster analysis as a tool for segmenting countries, which results in groups with similar behaviour based on chosen factors.

5.2.2 Clustering in tourism analysis

Clustering has been used in the context of tourism mostly to cluster tourists based on their behaviour, intentions or feelings (Hosany & Prayag, 2013; Ramires, Brandão, & Sousa, 2018; C. Ryan & Huyton, 2000). It has also been used on a within-country level to cluster destinations or regions based on the number of tourist attractions and the length of tourist stays (Lascu, Manrai, Manrai, & Gan, 2018) or regional indicators such as resident demographics and available hotel beds (Perles-Ribes, Ivars-Baidal, Ramón-Rodríguez, & Vera-Rebollo, 2020).

The importance of spatial considerations in these contexts have been highlighted by research such as "Spatial patterns of cultural tourism in Portugal" (2015). They note the importance of tour operators in explaining spatial tourist patterns, indicating that reachability of a location through existing infrastructure that is here provided through such tours might be an important factor.

Regarding the segmentation of whole countries from which tourists are visiting, Mazanec (2010) use clustering and network analysis for grouping and connecting countries through a semantic analysis of web pages. Due to their focus on a country's image as considered in the area of branding, they choose traits such as "adoration", "fun" or "serenity" to determine the groups. While insightful and important for a country to consider in their marketing efforts, these traits do not tell us anything about concrete plans of tourists to visit them, which is the focus of this chapter. Gabor, Conţiu, and Oltean (2012) use hierarchical and subsequent k-means clustering to group together European countries based on their tourism competitiveness. The focus of this study clearly lies on the activities of the countries instead of the tourists, however interestingly air transport infrastructure was found to be an important factor for explaining competitiveness, seemingly supporting our choice of airport connections as a variable.

5.3 Data sources and description

In a first instance, countries which are considered key markets of the Scottish tourism industry are identified by industry partners of the project based on ticket sales for local attractions. After excluding some of them due to data availability issues, 31 countries are identified for analysis (Table 5.1). Three main aspects are considered in analysing the impact of COVID-19 on the intentions of tourists to visit Edinburgh and Scotland: interest, affordability and reachability. We have identified these three dimensions as being representative for a number of different behaviours and factors that impact tourism behaviour. Interest demonstrates planned behaviour outcoming from the tourists themselves, affordability describes external factors and the financial situation impacting them, and reachability describes whether tourists are actually able to reach a destination if planning to do so. These factors thus have wider reaching impact which we consider as being able to take into account longer and shorter term effects simultaneously.

5.3.1 Data for factor interest

Interest is taken into account using Google search trend data for the key words “Edinburgh” and “Scotland” in the travel category.

One has to take into account that Google does not report the number of absolute searches but reports a relative value for the considered time period. Specifically, “[e]ach data point is divided by the total searches of the geography and time range it represents to compare relative popularity. [...] The resulting numbers are then scaled on a range of 0 to 100 based on a topic’s proportion to all searches on all topics.” (Google, 2020). For our purpose this means that the results have to be considered as trends for each individual country and can only be taken as a proxy for the change in behaviour and its strength during the analysed time period, but not in absolute numbers of searches and thus potentially interested tourists. Siliverstovs and Wochner (2018) also highlight that using a single time series carries the risk of misinterpreting the search volumes. By using two search terms instead of a single one we take this into account.

One problem that arises with Google Trends data is the fact that the time series can be very spiky. As they depict daily trends in search volume, there are strong fluctuations depending on factors such as day of the week or external events happening and affecting the search volume of users. Search behaviour can vary strongly from day to day due to for example weekdays and weekends experiencing different interest. Another reason could be the way that Google reports the volume as relative numbers in relation to other search terms. If for a short period another search term experienced an extremely high volume this would affect the searches for our terms relatively for that period. This "spikiness" of the data affects the performance of analytical methods used on them.

While Google is of immense importance in most of the countries within our analysis, it has a much smaller user base in China. The most used search engine there is Baidu[2] (Vaughan & Chen, 2015). In order to gain a sufficient insight into the Chinese market, we use Baidu search data for the same search terms. Vaughan and Chen (2015) discuss several things to take into account when comparing Google and Baidu results. There are three main areas through which problems can arise for our particular study. The first challenge is the translation of terms which we solve by involving a Chinese native speaker with fluent English language skills to translate any search terms. The second challenge lies in exact versus partial matching of search terms. The Chinese language differs majorly from the English language in its combination of multiple words, and Baidu Index requires a complete match of all search terms (Vaughan & Chen, 2015). We approach this issue by sticking to single word search terms only. The third possible problem arises through the way Baidu and Google report search results. While Baidu reports absolute searches for each term, Google only publishes relative numbers as discussed earlier. Through the use of normalisation to normalise all values to a range of [0,1], Baidu and Google data is being made comparable. However, this leads to a limited comparability of numbers between records, in our case representing countries. Domain knowledge can be employed to derive proxies which can be used to estimate total numbers based on these values. In our case, we use a factor derived from a sample of tourists provided by an industry project partner. The number of tourists to a sample location in Edinburgh and their country is given, from which we calculate a rank by

[2] https://www.baidu.com/

ordering the countries from most common to least. We multiply this rank factor with the time series, creating a weighted time series of interest and estimated absolute number of tourists. In order to provide us with a baseline of normal search behaviour outside of crises situations we are not only including 2020 search behaviour data but also 2019 data for the same time period. The exact time periods analysed are 17.01.2020 - 9.10.2020 and 17.01.2019 - 9.10.2019. This enables our approach to take into account normal time series levels as well as expected extraordinary levels during the COVID-19 pandemic.

5.3.2 Data for factor affordability

Affordability is incorporated in our analysis using Consumer Confidence Index (CCI) data. This data is available from the Organisation for Economic Co-operation and Development (OECD) and reports are published monthly, with some countries reporting their CCI less regularly or in quarterly bursts. The CCI gives an indication of the expected future financial situation of consumers, and thus whether consumers are more likely to spend or save (OECD, 2020). It varies around a value of 100, with a value above 100 indicating optimism about the future and a value below 100 indicating consumer pessimism. Not all countries considered for analysis have a CCI value available for the time period. This leads to the exclusion of 13 countries from 44 key markets which were identified through the industry partners. The final selection of countries can be found in Table 5.1. It has to be noted that not all countries report monthly values, Russia in particular reports values less frequently than other countries meaning that the only available CCI values in 2020 for them are in January and February 2020.

5.3.3 Data for factor reachability

Reachability could be included in the analysis as the geographic distance between visitor countries and Edinburgh. However, this would require a calculation of the distance between a country and a city. One possible method of doing this could be to use the centre point of the respective country. With countries such as Russia included in the analysis this could lead to an induced bias, as the majority of the Russian population is located in the Western most part of the country.

Australia	Japan
Austria	Mexico
Brazil	South Africa
Belgium	South Korea
Canada	Spain
China	Sweden
Czech Republic	Switzerland
Denmark	Poland
Finland	Portugal
France	Russia
Germany	the Netherlands
Greece	United Kingdom
Hungary	United States of America
Ireland	Turkey
Israel	New Zealand
Italy	

Table 5.1: List of 31 key visitor countries to the UK which are considered in this analysis.

Another way could be to use the geographic distance between the capital of the visitor country and Edinburgh.

One could argue, however, that the concept of reachability relies on more than just the geographic distance. Whether a destination is truly reachable for a tourist depends on the availability of means of travel. This is especially true in cases in which travel restrictions, such as those during the COVID-19 pandemic, impact whether travel is allowed to happen. We are therefore proposing to use flight connections as a measure of reachability. To be precise, we are analysing flight connection data taken from travel booking platform Skyscanner, which is one of the leading platforms in the market with 100 million monthly users (Skyscanner, n.d.). We are thus utilising flight data for the identified key markets in the time period 1 July 2020 to 29 October 2020. The reason that this time period is later than the one for the search data is that we expect a planning period to take place between searching for a destination and the actual travel to take place. The data includes the number of daily flights between a key country and Edinburgh Airport.

Flight connections can be defined as a bookable itinerary between an airport and Edinburgh even if a stopover is required, as long as they are booked together. The data includes an indicator whether on this day direct flights were available at all or not, but does not specify how many of the available flights are direct. A distance measure is proposed based on this information as described in Section 5.4. This distance describes how reachable Edinburgh is in terms of flight connections. One should note that in the case of domestic flights from the rest of the UK to Edinburgh, London Heathrow has been used as a common “gateway” airport carrying international travellers to other parts of the UK. This leads to a large portion of domestic flights not being considered. However, as the focus of this research lies largely on international tourism we consider this a suitable proxy for tourists arriving at Heathrow and continuing on into Scotland.

To illustrate how our approach compares to using the geographic distance between countries, we are also implementing the the approach calculating the geographic distance between Edinburgh and the capitals of the visitor countries and compare the results in Section 5.5.2.

5.4 Methodology

5.4.1 Interest calculation

Plotting example time series of Google and Baidu search data for different countries shows strong fluctuations as seen in Figure 5.1. This spikiness of the time series can strongly affect later analysis of the time series. Three different smoothing methods are thus considered to smooth out spikes: exponential smoothing, smoothing splines and weekly averages.

Exponential smoothing describes a class of methods commonly used for time series forecasting, the name deriving from the property of these methods that give an exponentially decreasing weight of importance to past observations in the forecasting of future ones (Hyndman, Koehler, Ord, & Snyder, 2008, p.5). A straightforwards state space exponential smoothing model with linear innovations as described in Hyndman et al. (2008) takes the following form:

$$\mathbf{y}_t = \mathbf{w}'\mathbf{v}_{t-1} + \boldsymbol{\epsilon}_t, \tag{5.1}$$

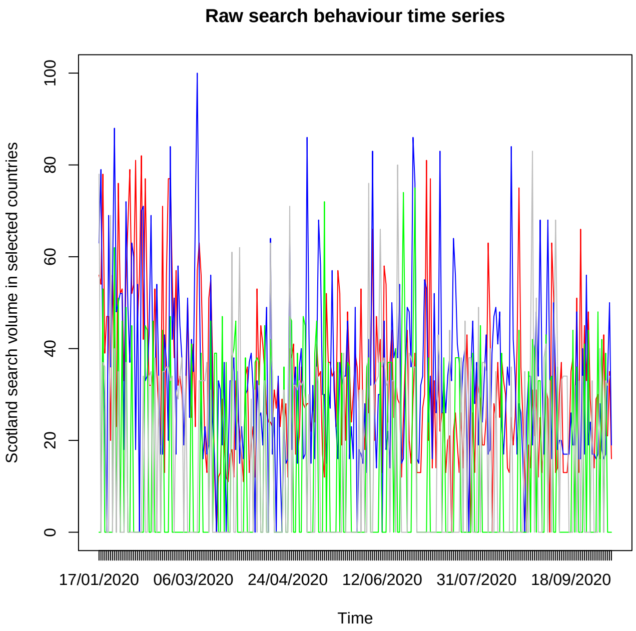

Figure 5.1: Google search behaviour for "Scotland" in the travel category. Shown for four selected countries: Australia (red), Belgium (blue), South Korea (green) and Turkey (grey).

$$\mathbf{v}_t = \mathbf{F}\mathbf{v}_{t-1} + \mathbf{g}\boldsymbol{\epsilon}_t, \tag{5.2}$$

where $\mathbf{y}_t$ is the time series which we are estimating. $\mathbf{v}_t$ is called the state vector and it contains other describing elements of the time series such as a trend or seasonality effect. $\mathbf{w}$ and $\mathbf{F}$ are the measurement vector and transition matrix respectively. $\mathbf{g}$ is called the persistence vector and contains the smoothing parameters. Lastly, $\boldsymbol{\epsilon}_t$ is a white noise (random error) series (Hyndman et al., 2008; Svetunkov, 2020).

It should be noted that these ES models using innovations as above assume that all error sources are perfectly correlated, shown by using identical errors $\boldsymbol{\epsilon}_t$ called *innovations* for the observation (Equation 5.1) and the transition part (Equation 5.2). This type of model is sometimes called a single source of error (SSOE) model.

The type of ES model varies further depending on the components considered in its construction. Typically there are four factors with which a time series can be described: error term (E), trend (T), seasonality (S), and cycle (C) (Hyndman et al., 2008, p.10). Here, we are considering further those models which consider the first three of those, commonly referred to as ETS model. Cycle is not considered in this analysis as this refers to dynamics which occur over a longer period of time of at least more than one year.

Depending on whether any of those components is present in the time series or not, and whether their effect is additive or multiplicative, we can distinguish between a large number of possible models. A purely additive model ETS(A,A,A) would take the form $y = T+S+E$ whereas a purely multiplicative model ETS(M,M,M) takes the general form of $y = T \times S \times E$ (Hyndman et al., 2008, p.10). Combinations are also possible, such as ETS(A,M,M) which describes an ETS model with additive errors and multiplicative trend and seasonality, or ETS(M,N,A) which represents a model with multiplicative errors, no trend and additive seasonality.

It is clear how deciding which model is best suited for a given data case can be challenging. As we do not yet know which kind of model best fits our data, we are looking for a way of automatically choosing one. For this purpose, we implement an exponential smoothing in SSOE state space model as described above using the `es` method from the R package `smooth` (Svetunkov, 2020).

As seen in the basic model in Equations 5.1 and 5.2, we will continue to have $\mathbf{v}_t$ as the state vector and error term $\boldsymbol{\epsilon}_t$. However, we now define a measurement function $w(\cdot)$, transition function $f(\cdot)$, persistence function $g(\cdot)$ and error function $r(\cdot)$, as the model can adapt to different forms of models and switch between additive and multiplicative forms. Svetunkov (2020) also introduce a vector of lags $\mathbf{l}$ which can affect different elements of the time series differently, for example seasonality can be lagged differently than trend. This view of the model further includes a vector of exogenous variables $\mathbf{x}_t$, which can optionally be included in the model, with their exogenous variable function $h(\cdot)$ and corresponding parameter vector $\mathbf{a}$. $\mathbf{F}_X$ and $\mathbf{g}_X$ are the transition and persistence matrices for exogenous variables. Lastly, the vector $\mathbf{o}_t$ is a Bernoulli distributed binary variable which is equal to 1 when y_t is observed, which allows the model to account for intermittent data. If that is not the case, $o_t = 1$ for all observations (Svetunkov, 2017).

$$\mathbf{y}_t = o_t(w(\mathbf{v}_{t-l}) + h(\mathbf{x}_t, \mathbf{a}_{t-1}) + r(\mathbf{v}_{t-l})\boldsymbol{\epsilon}_t) \tag{5.3}$$

$$\mathbf{v}_t = f(\mathbf{v}_{t-l}) + g(\mathbf{v}_{t-l})\boldsymbol{\epsilon}_t \tag{5.4}$$

$$\mathbf{a}_t = \mathbf{F}_X \mathbf{a}_{t-1} + \mathbf{g}_X \boldsymbol{\epsilon}_t / \mathbf{x}_t \tag{5.5}$$

For a more detailed view of all different forms that this model can take and their statistical properties, we point the interested reader towards Svetunkov (2017).

An advantage of this package is that it can automatically choose the best ETS model based on a number of information criteria and optimises required parameters. As such it does not require us to specify whether our time series experiences additive or multiplicative trend, seasonality and error terms. Instead the method calculates possible model specifications and compares them according to four information criteria: AIC, AICc, BIC and BICc. All four information criteria are used simultaneously in determining which model performs best overall.

Cubic smoothing splines smooth noisy data by estimating function values $f(x)$ though an optimisation approach which identifies that function $f(x)$ which minimises

$$\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda \int_a^b \{f''(t)\}^2 dt \tag{5.6}$$

with λ as a fixed constant which can be seen as a smoothing parameter and $a \leq x_1 \leq ... \leq x_n \leq b$ functioning as a boundary which has to include the data (Hastie & Tibshirani, 1990, p.27).

By minimising the criterion in Equation 5.6 the first part optimises closeness to the data and the second part penalises curvature of the function. Hastie and Tibshirani (1990) state that this equation has an explicit and unique minimiser which is a natural cubic spline with knots at the unique values of x_i. Furthermore, they state that cubic smoothing splines are approximately kernel smoothers with λ determining which shape the kernel takes. Choosing this parameter can be a challenge when using cubic splines if subjectively up to the researcher. It describes how closely the spline should follow the data versus create a smooth curve. In order to overcome this problem, we implement the approach by using the `smooth.spline` method from package `stats` (R Core Team, 2020). We make use of the package's generalised leave-one-out cross-validation for determining the optimal smoothing parameter λ, which is used if that parameter is not explicitly stated.

Lastly, for the **weekly averages** we calculate the average of each week of the time series. Compared to exponential smoothing and cubic smoothing splines, this presents an easy and direct method for smoothing variations in time series. Similar to the moving average method, which uses the same approach but moves the averaging-period along the series, this approach is a very intuitive approach. Compared to moving averages computing the weekly average also reduces the length of the time series, which can reduce computational expense when calculating distance matrices at a later stage. It should be noted, however, that a possible disadvantage is that by using a fixed window (7 days) we assume that we can represent the varying behaviour of the time series through a single value which relies on consistent behaviour of days of the week. Weekly averages also lag behind trends, as for example an increase occurring in the middle of a week would be averaged with the lower values in the beginning of the same week, dampening the apparent trend. We transform the time series into an explicit `ts` object and subsequently calculate the mean of each week using the `aggregate` function in package `stats` (R Core Team, 2020).

We now want to compare these three approaches to smoothing, exponential smoothing, cubic smoothing splines and weekly averages, through visualisation in order to identify differences between them. For this, four countries are selected to visualise here how the smoothing approaches handle different types of time series behaviour and to represent different geographic regions: Australia, Belgium, South Korea and Turkey.

The results for the three smoothing approaches are shown in Figure 5.2. It can be seen that the different smoothing methods have different advantages and disadvantages, and handle different problems of the data. The exponential smoothing is able to highlight the behaviour of Australia (blue) and Belgium (red) well, however, it "oversmooths" the two time series of South Korea (green) and Turkey (grey). Smoothing splines are able to capture the variations in the search behaviour of Turkey (grey), but have the tendency to smooth out much of the interesting behaviour of Australia (blue) and Belgium (red). The weekly averages deliver the best result for our specific application. The distinct behaviours of all four countries is preserved while still offering a much smoother finish compared to the raw time series in Figure 5.1. We thus decide to use weekly averages in place for the raw time series for the following analysis.

Using weekly averages we receive four matrices for the factor interest. Each contains a row wise time series for each country, spanning from 17.01.2020 - 9.10.2020 and 17.01.2019 - 9.10.2019 and for the search terms "Edinburgh" and "Scotland".

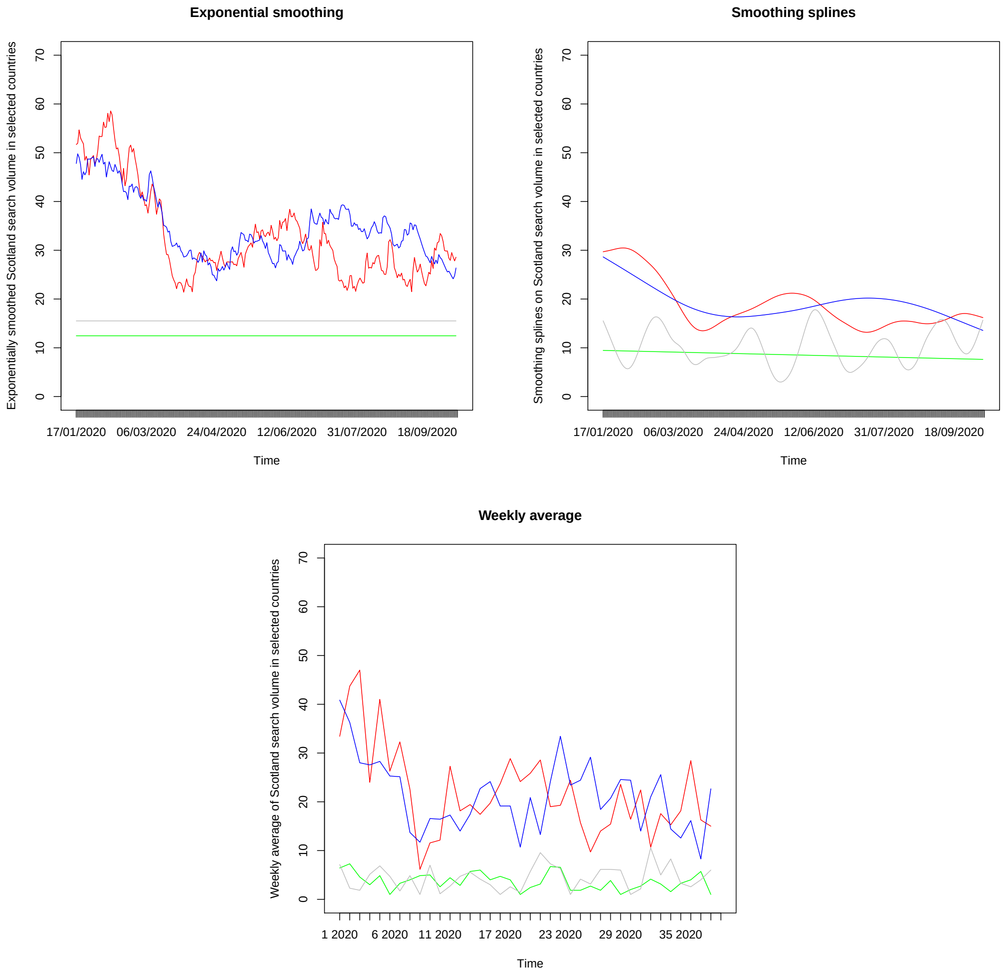

Figure 5.2: Google search behaviour for "Scotland" in the travel category. Shown for four selected countries: Australia (red), Belgium (blue), South Korea (green) and Turkey (grey). The values are smoothed using three different approaches: exponential smoothing (top left), smoothing cubic splines (top right), and weekly averages (bottom middle).

5.4.2 Affordability calculation

Based on the list of countries for which a CCI is available from the OECD as outlined in Table 5.1, we create a matrix of size 31 x 26 for each time period, one each for 2019 and 2020. The average CCI for the analysed time periods January to October 2019 and 2020 are shown in Table 5.2. The table shows that the average CCI for the time period in 2020 which occurred during the COVID-19 pandemic for all countries with the exception of Russia and Turkey is lower compared to the same months in 2019. Missing values are interpolated as the mean from the previous values to allow for subsequent analysis. We implement this approach using the mean imputation function na_mean from the R package imputeTS (S. Moritz & Bartz-Beielstein, 2017) for estimating the missing values. This imputation is only necessary for CCI 2020 values as all values for 2019 have been reported at the time of data collection. While necessary for our analysis, this imputation approach does carry the risk of overestimating CCI values for later months of 2020.

	CCI 2019	CCI 2020
Australia	99.826	98.848
Austria	100.516	99.567
Belgium	99.887	99.450
Brazil	100.370	98.625
Canada	100.868	98.994
China	105.125	103.383
Czech Republic	102.394	100.624
Denmark	100.668	100.005
Finland	99.188	98.125
France	99.798	99.063
Germany	101.044	99.703
Greece	100.835	100.268
Hungary	101.729	100.478
Italy	100.444	99.784
Ireland	101.166	99.237
Israel	100.984	98.547
Japan	99.076	96.770
Mexico	102.608	101.078
New Zealand	99.182	98.923
Poland	101.622	99.449
Portugal	100.620	98.567
Russia	99.287	100.256
South Africa	100.160	97.932
South Korea	99.485	97.351
Spain	101.042	97.943
Sweden	99.043	99.018
Switzerland	99.976	97.825
The Netherlands	99.877	99.101
Turkey	95.300	95.737
United Kingdom	99.841	98.691
United States	101.051	99.292

Table 5.2: Average Consumer Confidence Index in each country for the analysed time periods January to October 2019 and 2020.

5.4.3 Reachability calculation

We propose here the introduction of a reachability index. Our index takes into account the number and directness of flights between locations, thus more accurately representing how reachable a location is for someone compared to geographic distance. The factor also takes into account how external factors such as legal restrictions, which impact whether flights are allowed to happen between countries, influences reachability. Using the flight connection data described in Section 5.3, the reachability index is defined as

$$d_j^R = \begin{cases} 9999 & \text{for} f = 0 \\ \frac{1}{\sum_{t=1}^{T} f_{jt} * k_t} & \text{for} f \geq 0 \end{cases} \tag{5.7}$$

where d_j^R is the distance between Edinburgh and a country j, f_j is the number of flights between Edinburgh and j per day, and k is a binary indicator whether direct flights are available with $k = 2$ indicating availability and $k = 1$ indicating no availability. 9999 is an arbitrarily large number to indicate a very large distance due to no available flights for the given time period. This time period is $t = 1, ..., T$, with $T = 4$ months from July to October 2020.

It has to be noted that $k = 2$, indicating availability of direct flights on a given day, is simplifying a given issue: the data scraped only specifies *if* a direct flight is available, not the number of direct versus indirect flights. As an example, on the 13/09/2020 there were two flights available between Marseille and Edinburgh. The binary direct flight indicator in the dataset specifies that at least one of them was a direct flight, but it does not further specifies whether both were. We have therefore decided to use a binary k which takes the value 1 for no direct flights, meaning that the maximum distance a country can have to Edinburgh if any flights are available is 1 and the more flights and/or the more direct flights are available the distance shrinks.

Equation 5.7 gives us a distance vector $\mathbf{D}^R$ with elements d_j^R of size 31×1 for reachability between Edinburgh and each of the 31 countries as specified in Table 5.1. As mentioned before, London Heathrow is used in place for domestic flights leading to the distance between UK and Edinburgh being seemingly large. The values in distance vector $\mathbf{D}^R$ can be seen in Table 5.3. Not all countries from Table 5.1 can be seen reflected in Table 5.3, this is due to the fact that for some countries such as Turkey no connection can be found for the given time period. For these countries the arbitrarily large value 9999 is used in the analysis as outlined in Equation 5.7.

5.4.4 Clustering and combining distances

The objective is to create groups of similar countries based on their tourism behaviour. The objective of this is to allow for an overview of which countries behave similarity to each other

Country	Reachability of Edinburgh	Country	Reachability of Edinburgh
Australia	0.0084	Japan	0.0303
Austria	0.0039	New Zealand	0.0161
Belgium	0.0132	Poland	0.0024
Brazil	0.1111	Portugal	0.0023
Canada	0.0435	Russia	0.0200
China	0.0085	South Africa	0.1250
Czech Republic	0.0038	South Korea	0.0270
Denmark	0.0027	Spain	0.0015
Finland	0.0085	Sweden	0.0035
France	0.0022	Switzerland	0.0031
Germany	0.0024	The Netherlands	0.0037
Greece	0.0036	United Kingdom	0.0038
Hungary	0.0035	United States	0.0042
Italy	0.0017		

Table 5.3: List of key visitor countries for which a reachbility factor could be calculated as outlined in Equation 5.7. A small value indicates a small distance or a "good" reachability. For countries which are not listed no connection to Edinburgh could be found via Skyscanner for the chosen time period. This is the case for the following countries: Ireland, Israel, Mexico and Turkey.

and can thus be treated in a similar manner regarding marketing efforts. As our pre-existing knowledge about such groups is very limited, clustering allows us to form groups without the need for pre-labelling any data points. Specifically, to model this segmentation, we decide to use the "partition around medoids" algorithm (PAM) (Kaufman & Rousseeuw, 2009) as a straightforward algorithm suitable for handling multi-dimensional dissimilarity. The reason for choosing PAM is that the algorithm is able to cluster using any dissimilarity matrix as an input, enabling us to create a dissimilarity matrix using our three identified factors. We believe that these three dimensions best describe relevant tourism behaviour in our analysis and as the data comes in different formats, we first had to create a joined dissimilarity matrix which takes into account all dimensions. In contrast, popular clustering algorithm k-means requires distances calculated as Euclidean or Manhattan distances instead of, as we explain below, measures such as DTW. Using that approach would, however, leave out some important dynamic behaviour of the time series and we thus decide that PAM is the most suited algorithm here.

We are now interested in the dissimilarity between countries, which we define in the form of a distance matrix $\mathbf{D}$. Previously, we were calculating the distance of each country to Edinburgh in terms of interest, affordability, reachability. The dissimilarity of countries to each other now uses all three aspects described in Sections 5.4.1, 5.4.2 and 5.4.3. In this section we propose a dissimilarity metric which combines the information from these factors. We consider that all three dimensions carry equal importance in determining the number of tourists visiting or planning to visit Edinburgh.

In total, we are working with seven matrices: four matrices of Google search data covering “Edinburgh” and “Scotland” for the time period 17.01.2020 - 9.10.2020 and 17.01.2019 - 9.10.2019 (267 x 31 and 268 x 31 in size), two matrices with CCI values for 2019 and 2020 31 x 26, and one vector of size 31 x 1 with the reachability of Edinburgh for each country.

All matrices are first normalised. The dissimilarity matrices are then calculated as follows:

For **interest**, we make use of dynamic time warping (DTW), which was previously introduced in Chapter 3. This method aims to align two time series in such a way that it minimises the amount of warping and shifting that is required to do so and takes this minimum as a measure of dissimilarity. The main advantage of this method is that the overall time series shape is considered for calculating similarity, for example, if a peak occurs earlier in one time series than the other but the overall shape of the two is the same DTW would be able to detect that these two time series are similar. This is in contrast to lock-step methods such as Euclidean distances which calculate distances at fixed time stamps only. We begin by defining the warping curve $\phi(k)$ with $k = 1, ..., T$ for the two time series X and Y and their elements $X = (x_1, ..., x_N)$ and $Y = (y_1, ..., y_M)$ as

$$\phi(k) = (\phi_x(k), \phi_y(k)) \tag{5.8}$$

where $\phi_x(k) \in \{1...N\}$ and $\phi_y(k) \in \{1...M\}$ are the warping functions which re-map the elements of the two series X and Y respectively.

The average accumulated distortion between the two time series is then calculated as

$$d_\phi(X, Y) = \sum_{k=1}^{T} d(\phi_x(k), \phi_y(k)) m_\phi(k)/M_\phi \tag{5.9}$$

where $m_\phi(k)$ is a weighting coefficient and M_ϕ a normalisation constant.

The dissimilarity between the series X and Y is then calculated by detecting such an alignment that minimises the warping as seen in Equation 5.10.

$$D(X,Y)^{I(w,y)} = \min_{\phi} d_\phi(X,Y) \tag{5.10}$$

which describes the difference in the interest I of two countries X and Y for a year $y = \{2019, 2020\}$ and a search term $w = \{Edinburgh, Scotland\}$ (Giorgino et al., 2009). We implement this approach approach using the R package `TSdist` (Mori et al., 2016).

Affordability follows the same approach with

$$D(X,Y)^{A(y)} = \min_{\phi} d_\phi(X,Y) \tag{5.11}$$

with year $y = \{2019, 2020\}$ and $d_\phi(X,Y)$ as specified in Equation 5.9.

The vector of **reachability** indices D^R is calculated as described in Equation 5.7 and describes the distance between each country and Edinburgh. In order to determine the difference between two countries X and Y themselves their dissimilarity is now calculated as

$$D(X,Y)^{RD} = |d^R_X - d^R_Y|. \tag{5.12}$$

We now have seven distance matrices in total; four describing interest for Scotland and Edinburgh and the two time periods in 2019 and 2020, two for affordability for 2019 and 2020 which are valid for the whole region, and one for reachability:

Interest	Affordability	Reachability
$D^{I(Scot,2019)}$	$D^{A(2019)}$	D^{RD}
$D^{I(Edi,2019)}$	$D^{A(2020)}$	
$D^{I(Scot,2020)}$		
$D^{I(Edi,2020)}$		

A compound distance matrix is now calculated of these distances to derive one total measure of distance $\mathbf{D}$ through element-wise calculations as follows:

$$D(X,Y) = \sqrt{D^{I(Scot,2019)^2} + D^{I(Edi,2019)^2} + D^{I(Scot,2020)^2} + D^{I(Edi,2020)^2} + D^{A(2019)^2} + D^{A(2020)^2} + D^{RD^2}}. \tag{5.13}$$

The way that Equation 5.13 calculates each compound distance leads to a penalisation of extremely high distances in single dimensions compared to a simple average.

5.5 Results

5.5.1 The Interest-Affordability-Reachability model

Our proposed Interest-Affordability-Reachability (IAR) model uses three dimensions to describe tourism behaviour: intention in the form of 2019 and 2020 Google search trends time series which are smoothed as weekly averages, affordability in the form of 2019 and 2020 CCI values, and reachability in the form of a reachability index based on Skyscanner flight connections. These three dimensions are used to calculate a dissimilarity matrix **D** which describes the total dissimilarity between the countries in terms of how distant they are from Edinburgh. We then use PAM to cluster the countries with the number of clusters k being chosen by analysing the within-cluster sum of square (wss) compared to different possible values of k, as inbuilt in the method `fviz_nbclust` of R package `factoextra`. The PAM algorithm is chosen here because while its functionality is similar to that of k-means, forming stable and spherical groups of data points, it allows for the use of a distance matrix instead of raw data to be inputted into the algorithm.

Three clusters are identified which are shown in Table 5.4. Figure 5.3 depicts the average search volume for "Edinburgh" and "Scotland" in each cluster as well as the average CCI values in them. Table 5.5 shows the average reachability in each cluster.

In order to interpret the results each cluster is analysed regarding the average values of the three dimensions, allowing us to describe them in more detail and make recommendations to practitioners about them.

Cluster 1	Cluster 2	Cluster 3
Australia	Austria	Belgium
China	Brazil	Denmark
France	Canada	Greece
Germany	Czech Republic	Israel
Ireland	Finland	Italy
Poland	Hungary	Mexico
South Africa	Japan	Netherlands
Spain	New Zealand	Portugal
Turkey	Russia	Switzerland
United Kingdom	South Korea	
United States	Sweden	

Table 5.4: Result of the cluster analysis. Three clusters are identified.

Cluster 1	Cluster 2	Cluster 3
1818.01 / 0.018 *	0.025	2222.00 / 0.004 *

Table 5.5: Average reachability in each cluster, * indicates the average with and without two outlier countries with a value of 9999 indicating no detected connection between the country and Edinburgh for the analysed time period.

- **Cluster 1** : The cluster shows the highest search volume both for the term "Scotland" and "Edinburgh" (mean values 0.242 and 0.267 respectively). It shows a medium CCI which dips to the lowest level from the three clusters in the later terms of the analysed time period (mean value 99.046). The average reachability is in the middle both with and without the two not included countries Turkey and Ireland (mean values 1818.01 and 0.018). This implies that tourists from countries in Cluster 1 show high interest in visiting Edinburgh. The economic uncertainty and ability to reach Scotland might impact how quickly they are able to do so, but the high search volumes imply planning taking place.

- **Cluster 2** : This cluster shows the lowest search volume for both terms (mean value of 0.094 for Scotland and 0.074 for Edinburgh). The average CCI is around the lower level compared to the other clusters, in comparison with Cluster 1 it rises to the middle position (mean value 98.959). The reachability is the worst from all clusters with the largest distance (mean value 0.025). Tourists from these countries are unlikely to visit soon - they don't show a lot of interest as measured by search volume, nor do the CCI or reachability imply that they are able to make travel arrangements.

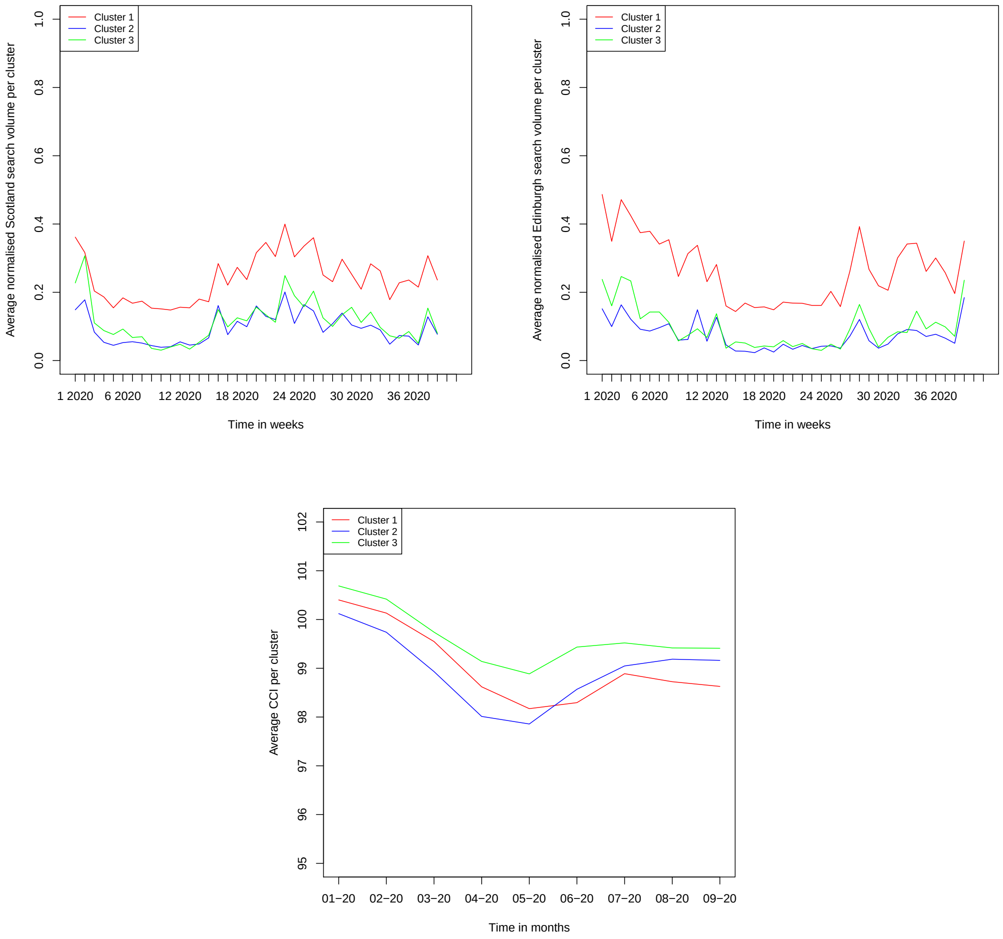

Figure 5.3: Average in clusters. Cluster 1 is shown in red, cluster 2 in blue and cluster 3 in green.

- **Cluster 3**: In this cluster we observe low search volumes (mean values 0.113 for Scotland and 0.097 for Edinburgh) but the highest average CCI from the three clusters (mean value 99.628). It shows the smallest distance when analysed without the two outlier countries Israel and Mexico, but the highest if analysed including them (mean values 2222.00 and 0.004). This cluster is sending mixed signals, while they are the most optimistic about their economic situation they show little interest in visiting Edinburgh or Scotland at the current time.

5.5.2 Comparison with other feature combinations

We compare the results of our model with a number of different approaches to demonstrate the advantages our approach has over them. The purpose of this comparison is to understand the role that each factor plays in the grouping of countries, as well as the impact of smoothing spiky time series and using reachability instead of geographical distance on the outcome.

- **Proposed IAL model**

  - Model 1: The proposed approach including smoothed Google Trends, CCI, reachability based on flight connections

- **Alternative multi-dimension models**

  These models present alternative approaches to our proposed selection of reachability factor (Model 2) and data pre-processin in the form of smoothing (Model 3).

  - Model 2: Smoothed Google Trends, CCI and reachability based on geographic distance between capital and Edinburgh
  - Model 3: Google Trends without smoothing, CCI and reachability based on flight connections

- **Single dimension model**

  These models show the impact individual features have on the clustering. Furthermore, a comparison of Model 6 and Model 7 demonstrates the isolated difference between the proposed reachability index and the geographic distance.

  - Model 4: Smoothed Google Trends only
  - Model 5: CCI only
  - Model 6: Reachability based on flight connections only
  - Model 7: Geo distance only

- **The impact of proposed reachability index**

  This model isolates impact that the proposed reachability index has by removing it from the IAL model and focusing on the remaining two dimensions.

– Model 8: Smoothed Google Trends and CCI

The geographic distances for models 2 and 7 are calculated using the `distGeo` function from package `geosphere` (Hijmans, 2019) which calculates the geodesic distance between two points on an ellipsoid. The countries capitals are chosen for that, in the two cases where there are multiple capitals Amsterdam is chosen for the Netherlands and Cape Town for South Africa.

Detailed plots for all alternative models can be found in the Appendix of this Chapter. Table 5.6 outlines a comparison of countries and their cluster membership. It should be noted that cluster labels such as "Cluster 1" are assigned by the algorithm depending on which data point it is visiting first. They do therefore not indicate that these clusters can be directly linked to each other. Cluster 1 in one model can be very different in terms of its values than Cluster 1 in another model. Instead, we are describing each cluster depending on their high or low average value in each feature. In the table, colours are henceforth assigned by using the same colours as the reference clustering result. This aims to make it possible for the reader to compare whether a specific country is always a member in a high interest cluster, for example. Red indicates high interest and at least some degree of good reachability and affordability. Blue represents low to medium low values in all categories. And green represents a medium ground or an unclear situation with mixed signals sent.

The comparison in Table 5.6 shows that depending on the features used in forming the clusters, each country becomes a member of a different group. The proposed approach has the most varied result with three resulting clusters of roughly equal size. The comparison shows how important it is to consider how variables are defined and treated before clustering.

In terms of the **alternative multi-dimension models**, models 2 and 3, they too show varied clustering results with three different groups compared to models with two or fewer dimensions which show only two clusters. This indicates that adding more dimensions might allow the model to identify more different clusters by picking up structures which are not present otherwise. A higher number of clusters can be useful for marketing segmentation efforts as it allows a more

Country	Proposed approach	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Australia	Red	Blue	Red	Blue	Red	Red	Red	Red
Austria	Blue	Green	Blue	Red	Red	Red	Red	Blue
Belgium	Green	Blue	Blue	Red	Blue	Red	Red	Blue
Brazil	Blue	Green	Blue	Red	Blue	Blue	Red	Blue
Canada	Blue	Green	Blue	Blue	Blue	Red	Red	Blue
Switzerland	Green	Red	Blue	Blue	Blue	NA	Red	Blue
China	Red	Red	Green	Blue	Blue	Red	Red	Red
Czechia	Blue	Green	Red	Red	Red	Red	Red	Red
Germany	Red	Blue	Red	Blue	Red	Red	Red	Red
Denmark	Green	Blue	Blue	Blue	Red	Red	Red	Blue
Spain	Red	Red	Red	Red	Red	Red	Red	Red
Finland	Blue	Blue	Red	Blue	Red	Red	Red	Red
France	Red	Blue	Red	Red	Red	Red	Red	Red
UK	Red	Blue	Red	Red	Red	Red	Red	Blue
Greece	Green	Red	Red	Blue	Red	Red	Red	Red
Hungary	Blue	Green	Red	Blue	Red	Red	Red	Red
Ireland	Red	Blue	Red	Red	Red	Blue	Red	Red
Israel	Green	Red	Red	Blue	Red	Red	Red	Red
Italy	Green	Red	Green	Blue	Blue	Red	Red	Blue
Japan	Blue	Red	Green	Blue	Blue	Red	Red	Blue
South Korea	Blue	Green	Green	Blue	Blue	Red	Red	Red
Mexico	Green	Red	Red	Blue	Red	Blue	Red	Red
Netherlands	Green	Red	Green	Blue	Blue	Red	Red	Blue
New Zealand	Blue	Blue	Blue	Blue	Blue	Red	Blue	Blue
Poland	Red	Blue	Red	Blue	Red	Red	Red	Red
Portugal	Green	Red	Blue	Blue	Red	Red	Red	Blue
Russia	Blue	Blue	Red	Blue	Red	Red	Red	Red
Sweden	Blue	Green	Green	Blue	Blue	NA	Red	Red
Turkey	Red	Blue	Red	Blue	Red	Blue	Red	Red
United States	Red	Blue	Red	Blue	Red	Blue	Red	Red
South Africa	Red	Blue	Red	Blue	Red	Red	Red	Red

Table 5.6: Model comparison of eight different models. The colours indicate three different interpretations of the respective cluster, with red indicating high interest and/or affordability and/or reachability, blue indicating lower interest, affordability and reachability, and green indicating mixed results.

nuanced approach to developing strategies. Depending on the definition of the dimensions, the results for countries can vary a lot. In the example of France, this country is a member of the high visitor chance cluster in all but one case. In the case of model 2, which uses search data, CCI and geographic distance, it is a member of the blue cluster which indicates low chance of visiting.

Looking at **single dimension models** next, Brazil demonstrates the differences that can occur between models. While the proposed approach in Model 1 assigns it membership in low chance cluster "Blue", Model 4 which clusters by interest only and Model 7 which uses geographic distance instead of reachability indicate a higher chance of visit for this country. Indeed, the results for Model 7 which clusters exclusively using geographic distance assign a "Red" (high chance) membership for all countries except New Zealand. This seems to be a highly unlikely outcome. Meanwhile, Model 4 which uses search data only paints a much more pessimistic picture, assigning membership to the "Blue" (low chance) cluster to all but eight countries.

Lastly, looking at Model 8 to identify the **impact of the proposed reachability index**, we can see that all countries which were identified as high visiting chance "Red" with the exception of UK are also identified as such in Model 8. However, by including the reachability index more variability seems to be found. All countries classified as low chance (blue) are also identified as low with the proposed approach, however a number of them were idenified having mixed results (green). This might indicate that while interest and affordability have a lowering effect, the reachability via flight connections might make the outlook for those countries more hopeful for the nearer future.

Overall, assigning these simplifying labels of "Red", "Blue" and "Green" to these countries was a challenging piece of analysis. A review of the modelling outputs attached in the Appendix shows that for many model outputs the difference between the clusters was minimal. Based on this a further analysis of the quality of the clustering results was conducted, the results of which are in Table 5.7.

		Size	Max member medoid diss	Avg member medoid diss	Cluster diameter	Min between cluster separation
Model 1	Cluster 1	11	1.3588	0.7375	10.7477	0.6082
	Cluster 2	11	1.0935	0.6609	2.4853	0.5774
	Cluster 3	9	1.1087	0.6741	3.4485	0.5774
Model 2	Cluster 1	14	1.3670	0.8585	10.7544	0.5586
	Cluster 2	7	1.3320	0.6599	2.3196	0.6565
	Cluster 3	10	1.1895	0.7609	3.9553	0.5586
Model 3	Cluster 1	17	1.3538	0.8032	9.9934	0.6146
	Cluster 2	8	0.8081	0.6164	2.1036	0.6146
	Cluster 3	6	0.7684	0.5573	1.8400	0.6150
Model 4	Cluster 1	23	0.9667	0.3054	1.7551	0.1018
	Cluster 2	8	0.5124	0.2964	10.6156	0.1018
Model 5	Cluster 1	20	1.1940	0.5842	4.7189	0.0340
	Cluster 2	11	0.8218	0.4092	3.6665	0.0340
Model 6	Cluster 1	24	0.3338	0.3198	0.3338	0.3337
	Cluster 2	5	0.3338	0.2670	2.8928	0.3337
Model 7	Cluster 1	30	14986016	6509169	14986016	15498621
	Cluster 2	1	0	0	0	15498621
Model 8	Cluster 1	19	1.4956	0.7762	4.7534	0.4067
	Cluster 2	12	1.0555	0.6228	10.6835	0.4067

Table 5.7: Quality measures for all eight models. Shown are: number of members per cluster, maximum dissimilarity between a cluster member and its medoid, average dissimilarity between a cluster member and its medoid, cluster diameter (maximal dissimilarity between two objects of the same cluster), and between cluster separation as the minimum dissimilarity between members of two different clusters.

Table 5.7 summarises the difference between the different models. Regarding cluster size, Model 1 produces the most even result with three clusters of similar size. Model 7 shows an extreme example for a result that groups together all but one observation into a group, resulting in very high dissimilarity within the clusters. Both Models 1 and 3 show good separation between the clusters, likely due to the inclusion of multiple variables. However, Model 3 which does not utilise smoothing for the interest data, forms one very large cluster, a possible interpretation of which could be an overemphasis of short spiky effects. Model 6 shows the most compact groups, but with very unequal sized clusters. The results from Model 1 and 2, which differ only in the use of airport connections instead of geographic distance, show that Model 1 produces clusters of slightly more even size, which are more compact with a slightly smaller average member to medoid dissimilarity. Overall, it can be said that the results from our proposed approach in Model 1 perform very well compared to other configurations.

5.6 Conclusion

The COVID-19 pandemic had profound effects on the way tourists travel. It affects not only the tourists' travel intentions but also their optimism about their future economic situation and travel connections between countries. For the recovery of the tourism sector, which is of importance for the Scottish economy, understanding tourists' behaviour is useful as it enables companies to plan ahead. Our findings not only introduce a method to capture all three important dimensions simultaneously, but the proposed model is also able to communicate findings in the form of groups of countries with a similar behaviour. Communicating these groups to local businesses and policy makers enables them to understand the dynamics of tourist behaviour in broad strokes and target clusters of countries with marketing efforts.

The main objective of this paper is to contribute to this understanding of tourism behaviour in different countries as well as the use of clustering for this purpose. Tourism behaviour depends on a number of factors, which we have included as interest, affordability and reachability in this analysis. As such, our contribution is two-fold in providing practical insights in the form of country groups and introducing a novel approach to including reachability based on number and directness of flight connections. A comparison with other model configurations demonstrates that using a reachability index instead of geographical distance changes the categorisation of several countries. Using a reachability factor instead of the distance has the advantage of identifying countries which are unreachable even though they are close, while taking into account external factors such as regulation impacting routes and the way routes change over time.

Our findings come in the form of groups of countries with a similar tourist behaviour regarding the three dimensions, which we derive using clustering. These groups can be targeted by companies and policy makers with marketing strategies, such as targeting those countries first which are the most likely to visit soon. Three clusters are identified which are of roughly equal size but show differences regarding interest, affordability and reachability. Cluster 1 consists of countries with optimistic tourists, showing high interest in visiting, medium to high affordability and a medium level reachability. Cluster 2 countries are characterised through not interested tourists, with low interest and reachability values while the affordability is at a medium level. Finally, cluster 3 countries experience high affordability levels, but interest is low and reachability values are mixed among the countries.

We compare our results to seven alternative approaches in an effort to understand the impact of the newly introduced reachability index, the individual parameters as well as a smoothing of the Google search data. Overall, the comparison demonstrates that our proposed approach is performing well in grouping together countries and forming well separated clusters of equal size. It can also be seen that including more variables in the analysis leads to a higher number of resulting clusters, indicating that more variations are being picked up. This is an advantage of these models as in marketing contexts segmentation in more than two groups can add more nuance to the targeted strategies.

However, there are a number of limitations to take into account, both from the nature of COVID-19 disruptions and from the limitations in the data. The inclusion of reachbility adds an interesting perspective into the analysis, however due to the exceptional circumstances of the pandemic not all countries showed any connection to Edinburgh for the time period. Those without a connection are included in the analysis using an arbitrarily large distance of 9999, which makes their analysis possible but might not reflect a realistic view on the connection between countries. For example, visitors might organise independent flight connections by first flying to a different destination and then connecting to Edinburgh, which is not taken into account if not sold as one unit on the Skyscanner website. Another limitation of our study is present in our use of keywords. We used Google Trends data for the keywords "Edinburgh" and "Scotland". While searches on Google were conducted for the travel category, this might include non-travel related search queries of these phrases especially for Baidu data. This limitation was necessary, however, due to the incorporation of Baidu data and the limitation of accurate keyword translation into Chinese. Word combinations (e.g. "Hotel Edinburgh") would have made these translation challenging, as noted previously by Vaughan and Chen (2015). We further justify our use of

"Edinburgh" and "Scotland" to indicate early and general interest in these locations as well as more concrete planning steps.

While the research carries limitations, the implications of the findings offer important insights for businesses in the tourism sector as well as policy makers. They uncover differences between countries as well as similarities and thus allow businesses to plan ahead depending on their knowledge about their customer base or target market. If a company is aware that the majority of their customers come from countries in Cluster 1 this would indicate that tourists from these countries are more optimistic about visiting again soon. On the contrary, relying on customers from countries in Cluster 2 or 3 can prove more difficult. The cluster descriptions also indicate the most problematic aspect that customers from a country perceive, for example customers in cluster 2 show low interest and less reachability, indicating that more structural factors are impacting their willingness to travel. All of these considerations can be taken by companies based on the findings in this chapter.

The presented approach also allows for the findings to be continuously updated. Currently, time series from January to October 2020 and their corresponding time periods in 2019 form the base of our understanding. However, all factors in the analysis will likely change, in some cases quite rapidly. Political decisions in the future such as vaccination passports will have a profound effect on the way tourists plan their trips. The advantage of using Google Trends data as indicator for tourism demand is clear, as it reflects these immediate changes and is also an indicator for hopefulness in the sense that tourists making abstract plans for travel will be included. CCI and flight connections meanwhile deliver a more facts-based perspective based on economic situation and regulations. As such our analysis provides a balanced view of hope and reality. The COVID-19 tourism recovery route will likely be a long and challenging one. These companies form an important part of not only Scotland's economy, and providing them with the tools necessary for understanding and working with the situation over time will be crucial for a successful recovery.

Finally, considering the introduced methodology we have also demonstrated two important points for the academic audience. Cluster analysis allowed us to form groups of countries with very limited pre-knowledge about their behaviour. We were able to use data from various sources and combine them into a joined dissimilarity matrix, which was used as an input for the clustering algorithm. While all three dimensions were considered with the same importance in this analysis, weighting the dimensions would allow the researcher to incorporate expert knowledge about the importance of one factor over the other. The challenges presented by using real-life data were numerous and we have presented some ways of overcoming them. Google search data showed erratic behaviour and smoothing methods had to be employed in order to not overemphasise small discrepancies. The fact that Google data and Baidu data was measured on different scales demonstrated the importance of understanding and handling different data formats to make them work together in forming the overall picture. Finally, our introduction of the reachability index showcases that distance, while possibly on the most important concepts in clustering, is relative. While theoretically based indices such as the Euclidean distance is often used, in reality it can vary and change over time and depends on external factors, such as the idea that people do not travel in straight lines but are reliant on modes of transportation. This Chapter shows that context matters when deciding what defines closeness.

Appendix to Chapter 5

Collection of model outputs for Section 5.5.2.

Model 2: Smooth search data, CCI and geographic distance

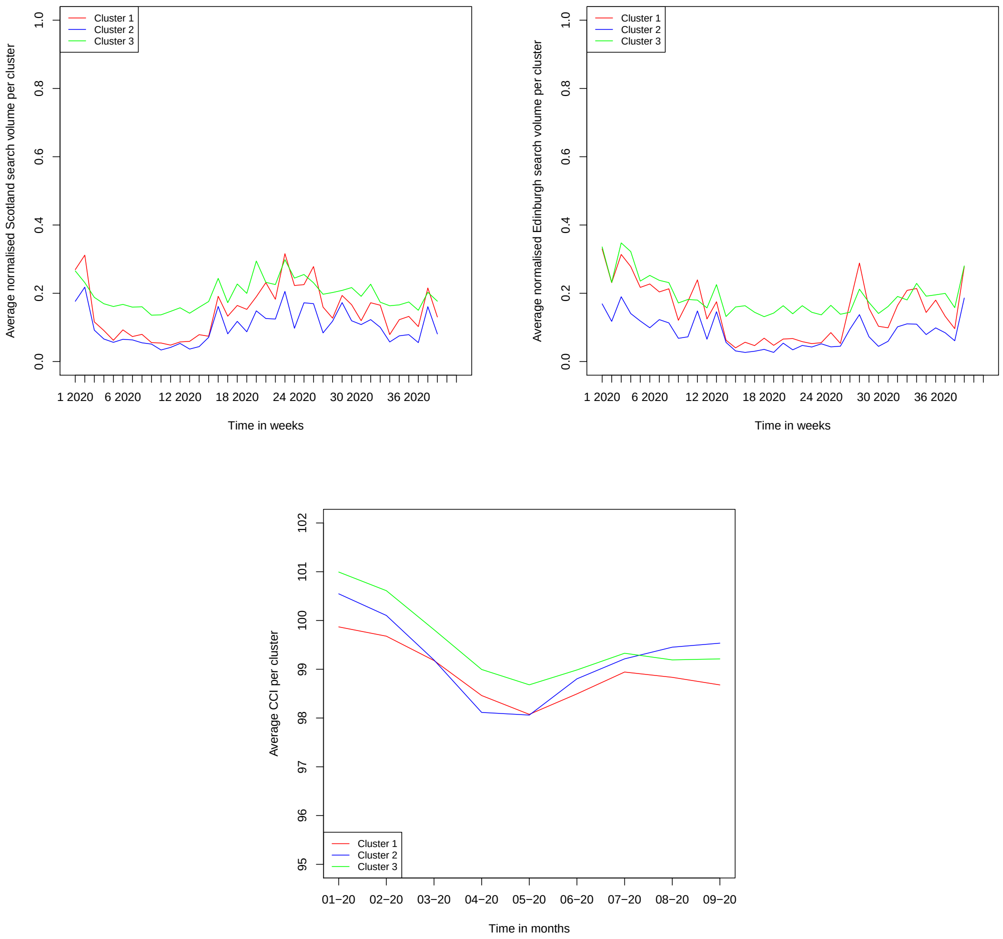

Figure 5.4: Model 2: Average search volume and CCI per cluster.

Model 3: Non-smooth search data, CCI, and reachability/flight connections

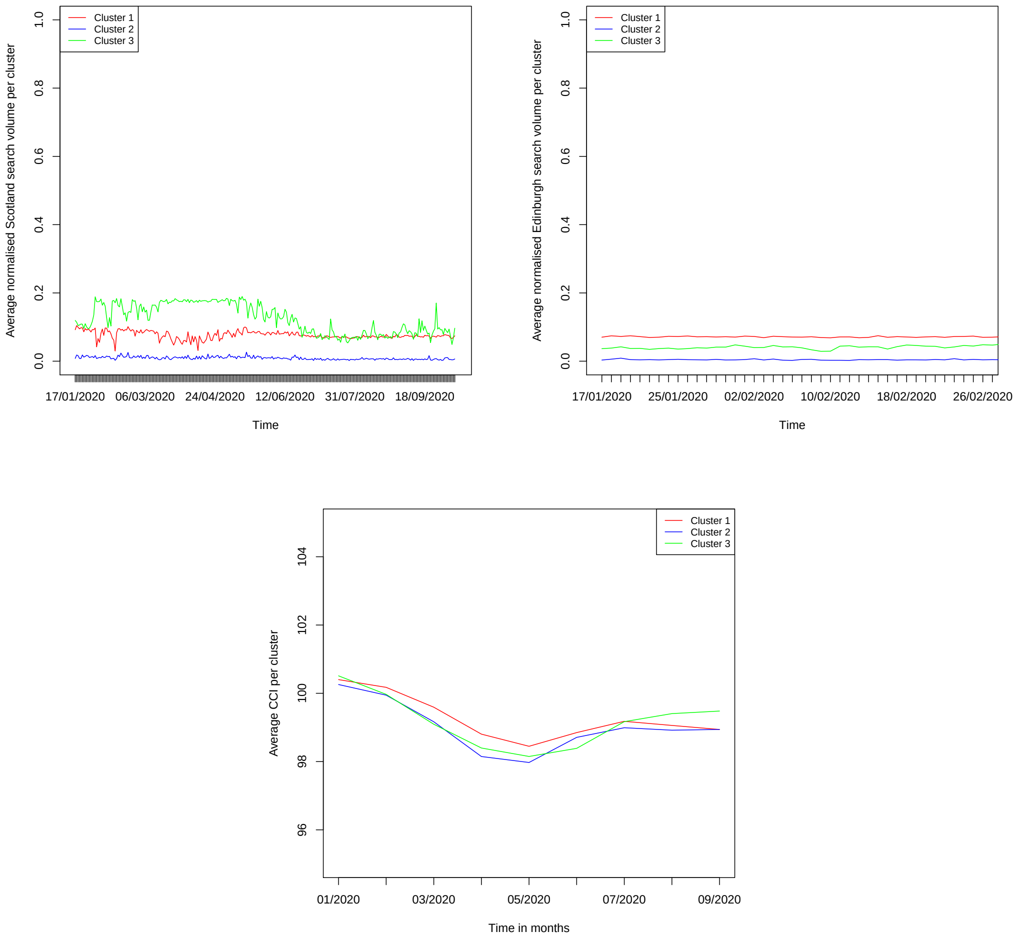

Figure 5.5: Model 3: Average search volume and CCI per cluster.

Cluster	Mean	Variance
1	-1764.5160	15438138
2	0.0245	0.0014
3	0.0125	0.0002

Table 5.8: Model 3: Average flight connection derived distance and variance per cluster.

Model 4: Smooth search data only

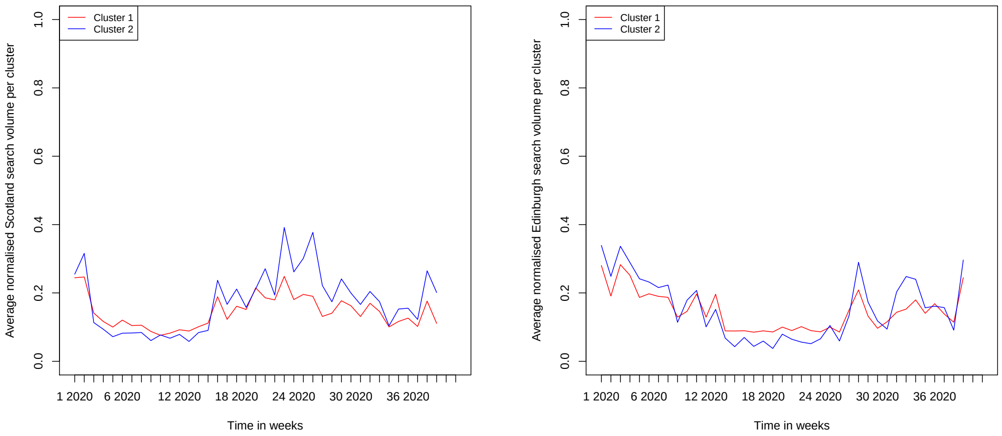

Figure 5.6: Model 4: Average search volume per cluster.

Model 5: CCI only

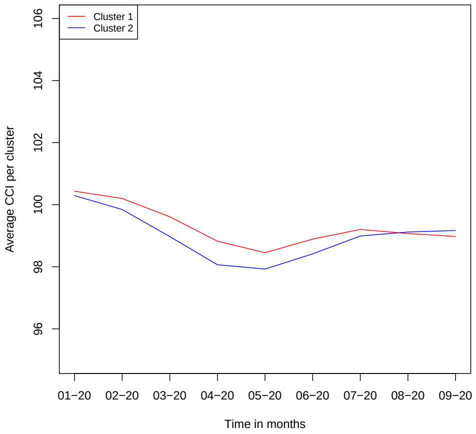

Figure 5.7: Model 5: Average CCI value per cluster.

Model 6: Reachability only

Cluster	Mean	Variance
1	0.01562	0.0007
2	-5999.3770	29994346

Table 5.9: Model 6: Average flight connection derived distance and variance per cluster.

Model 7: Geodistance only

Cluster	Mean	Variance
1	6979756	NA
2	15498621	9.118483e+12

Table 5.10: Model 7: Average flight connection derived distance and variance per cluster.

Model 8: Smooth search data and CCI

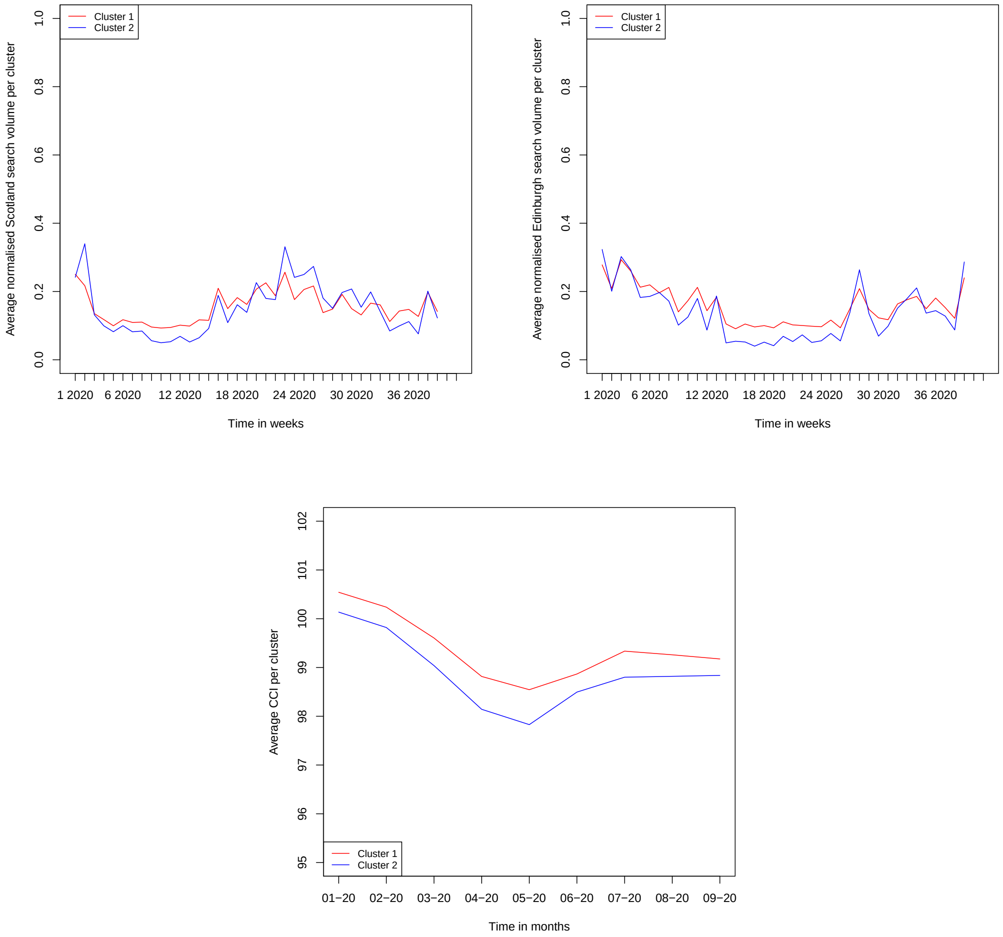

Figure 5.8: Model 8: Average search volume and CCI per cluster.

Chapter 6

Concluding remarks

This thesis investigated the use of clustering in three application areas and the methodological advancements which are necessary to make the methods useful in real-world contexts. The objective of the thesis motivating this approach has been to bridge a gap between theory and practice, by adjusting cluster methodologies as driven by the given context and available data. After a literature review of clustering with a focus on spatial, temporal and spatio-temporal methods in Chapter 2, we presented three use cases in public health, SME financing and tourism recovery in Chapters 3, 4 and 5 within which we introduced empirical contributions to the literature of the respective application area as well as methodological contributions. While presented in the context of these three scenarios, the relevance of our findings in a more generalised context has been discussed. Indeed, the three scenarios can be seen as case studies for a data and practice driven approach to method development.

Chapter 3, which resulted in an article discussing the topic in the context of risk analysis (Gieschen et al., 2021), used spatio-temporal clustering to form groups of Scottish GPs based on their geographical location and prescription behaviour over time. These groups can be used by the GPs themselves to compare themselves to their peers, and by policy makers to inform training and best practice activities. We introduced a novel method of taking into account spatially unequal distributed data by using a KDE based approach to modify the spatial distances between GPs. In doing so, we presented a way of overcoming the methodological limitation of the DBSCAN and ST-DBSCAN algorithm. Furthermore, we were the first to also use the DTW algorithm for the calculation of time series dissimilarity within the ST-DBSCAN algorithm. This is an important step for taking into account seasonality effects which are present in many application areas such as health care and prescriptions. In a more general sense, the combination of time series data and geospatial information for clustering as well as potential pitfalls with that have been presented. Real-life data seldom follow such distributions as expected from theory and method development, in many cases making adjustments to methodologies necessary on a case-to-case basis.

Chapter 4 made use of clustering for the analysis of access to finance of SMEs in the UK. By allowing for contagion effects to affect only companies which were spatially close and similar to each other, we were able to show that local level policy making which takes into account these effects can be useful as clearly there are effects connecting these SMEs in a spatially bound way. For this purpose, we introduce a novel approach of using a two-step process to spatial regression in which the **W** matrix which describes the spatial connectedness of SMEs is constructed with the help of clustering. This approach allows for the use of a sparse matrix in the regression model, which is not only computationally important in large scale analyses but also re-defines connectedness between companies based on not only their geographic closeness but also their similarity and thus their likely network structure. Among SMEs informal and formal networks through which information is shared are likely to exist for example through local entrepreneurial networking events and organisations connecting local business leaders within sectors. This Chapter also presented a way of utilising the power of cluster analysis when combining it with other existing methods, in this case regression. Clustering as an exploratory data analysis approach has been dominant in much of the literature, however, it also offers a way of representing networks and groups within which further analysis can take place.

Lastly, Chapter 5 explored the applicability of clustering in the context of market segmentation. Clustering was used to create groups of countries with a similar tourist behaviour in terms of interest, affordability and reachability. Companies can use these groups to formulate marketing strategies targeting them at different stages, which is of especially high importance during their recovery efforts after the COVID-19 pandemic. The introduction of a reachability index put into question whether geographical distance is always the deciding factor for behaviour or whether actual ability to reach something is a more important factor in times of modern travel. A comparison of eight model configurations showed the impact that different dimension definitions had on the outcome, with our proposed IAD model taking into account reachability in such a way that it reflected the ability of tourists to visit and which can be impacted through different scenarios such as the presented pandemic case. We introduce a new reachability index which is calculated using the number of flight connections and their directness. From the perspective of clustering, this Chapter also presented an important view on the use of different dimensions and data sources. While clustering utilises the concept of closeness or similarity, how we define this similarity is dependent on available data as much as the real life context. Expert knowledge of a particular area can thus be a crucial part of deciding how we define and then calculate similarity more generally.

The increasing importance of machine learning in public organisations and businesses has opened up a discussion about the use of data science and ML related methods for gathering insights and making decisions. In research, clustering is a known method for explorative data analysis, but as we have seen it can also be used for comparisons between and within groups, as a layer in spatial regression models, or for informing marketing strategies. This thesis contributed to existing knowledge two-fold: we have demonstrated methodological advancements and novel uses of existing clustering methodologies, contributing to the literature on clustering as a method more generally. Concretely, we have demonstrated how limitations in the DBSCAN algorithm can be overcome by including a density factor, how DTW can be used for a more accurate representation of time series similarity in connection with DBSCAN, how clustering can be used in a two-step approach to logistic regression to model within group contagion effects, and how data from multiple sources and formats can be brought together in order to define similarity. While presented within the context of use cases, all approaches can be generalised to other contexts, as discussed in the individual Chapters. We have expanded the use of clustering beyond exploratory analysis and demonstrated the importance of considering the nature of the data as well as our understanding of similarity in context. As such, this thesis contributes to the knowledge within the academic communities of both practice near areas of research by expanding their arsenal of methods useful to their field, and more methodological focused communities by highlighting the value in research motivated by practice and how this perspective can drive method innovation. Secondly, we have shown the value that clustering as a method brings to organisations and policy makers. Our findings contribute to research and practice on health policy and training, SME financing, regional financial policy making and tourism research. But more generally speaking, we have also demonstrated that there exists a gap between the usefulness of a method in a theoretical sense and its usefulness when presented with real world data, as well as how changes to methods allow for this gap to be bridged. This thesis thus seeks to stand as an example of data-driven methodology use and development.

While adding important new developments, the presented research is not without its limitations. We highlight the specific limitations of each use case within the respective Chapter and therefore want to stress here in particular the role that the data played in these limitations. In Chapter 3 we focus our research on one prescription drug, limiting the amount of data in our analysis to one time series per GP instead of pursuing a multi dimensional approach capturing a multitude of different drugs. Computational restraints played a major role in this choice as well as limited expert knowledge in the role that different drugs play. Antibiotics were chosen as a use case due to the high interest of practitioners in their prescription dynamics, however, by focusing on one antibiotic drug we might have limited our findings in terms of effects between them. This demonstrates the value of incorporating the advice of experts of a field into the development of a research project. The case presented in Chapter 4 on the other hand showed another problem often encountered with real life data. A large amount of missing values limited the amount of data that could be used for the analysis. While still presenting a sample of over 3,000 companies, a larger sample might be able to uncover dynamics especially in less represented areas of the country which are so far hidden from view. The nature of how the data set is collected, which limits how companies can be analysed over time, further complicates this. It allows us to only take a snapshot in time instead of considering changes to company behaviour and, for example, whether a particular SME was able to access finance in the past.

Lastly, Chapter 5 saw the challenge of estimating the actual number of visiting tourists to be used with the relative numbers provided by Google Trends data based on the experience and expert knowledge of industry partners. This not only relied on these numbers being more or less constant over time, but also on the sample of the industry partner being representative for the tourism industry as a whole. More generally speaking, this thesis also has the limitation of somewhat subjective decisions made by the author throughout the research journey. Choosing a particular method over another, while taking into consideration factors which made a method more suitable or strengthened the argument in favour of one method over another, still limits the statements we can make based on our findings. Only a limited number of clustering methods has been used within this thesis, leaving a vast amount of other approaches unexplored in our context.

Based on these limitations and those explored within the individual chapters, we see the potential for further research in multiple directions. From a theoretical standpoint, the role of the smoothing parameter in Chapter 3 would be one such avenue. We know that the parameter plays an important role in the amount of smoothing accounted for, but so far no detailed sensitivity analysis where this parameter is adjusted in different directions has been undertaken. We have seen the importance of choosing data when pursuing a data-driven approach to methodology as demonstrated in this thesis. By exploring the suggested methodological developments in other contexts, their value to practice and theory could be strengthened further. Furthermore, additional data sources could support potential explanations for our findings. For example, in Chapter 4 a possible explanation for contagion effects was the presence of formal or informal networks among SMEs which allowed for knowledge transfer between them. However, this was merely hypothesised as one possible explanation, a different one being the way that lenders make decisions on who to approve for a loan or overdraft. By investigating whether contagion effects correlate with a membership in an SME network, this explanation could be supported empirically. The observed differences in prescription behaviour in Chapter 3 could be elaborated on by incorporating data on demographic or socio-economic data in the different spatial regions. Further information on the GPs themselves could also be added to the analysis in an effort to understand why some GPs prescribe differently than their colleagues. Due to the functionality of clustering in discovering groups within data without the need for pre-labelling, there is a lot of potential in using it when there is limited pre-existing knowledge of structures. The methods' ability to include spatial and temporal dimensions is an important feature because data of such nature allows us to understand structures of a complex nature which experiences regional differences or varies over time. However, we have also explored whether our understanding of spatial data always accurately reflects our notion of distance or reachability. While in some cases the two might be synonymous, the idea of distribution of information and knowledge between two locations through digital channels or available travel via planes has come up as a question for future research. The spatially bound networks we have identified in Chapter 4 might change through the advancements in digital communication channels which connect companies irrespective of their geographic closeness to each other or to local banks. We have already included the concept of reachability of a tourist location in Chapter 5, but as travel becomes more affordable in the future and we are simultaneously discussing the sustainability of air travel this remains an area which is on the move. As such, future research might investigate what distance and reachability mean in context-dependent circumstances and how different definitions impact the use of spatial and spatio-temporal clustering methods. While the role of geographic distance will likely remain an important one in social sciences, technology will change our understanding of distance as we know it.

References

Agarwal, P., & Skupin, A. (2008). *Self-organising maps: Applications in geographic information science*. John Wiley & Sons.

Agarwal, S., & Hauswald, R. (2010). Distance and Private Information in Lending. *The Review of Financial Studies*, *23*, 2757-2788.

Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering–A decade review. *Information Systems*, *53*, 16–38. doi: 10.1016/j.is.2015.04.007

Agostino, M., & Trivieri, F. (2018). Who benefits from longer lending relationships? An analysis on European SMEs. *Journal of Small Business Management*, *56*(2), 274–293.

Anbaroglu, B., Heydecker, B., & Cheng, T. (2014). Spatio-temporal clustering for non-recurrent traffic congestion detection on urban road networks. *Transportation Research Part C: Emerging Technologies*, *48*, 47–65. doi: 10.1016/j.trc.2014.08.002

Anderson, C., Lee, D., & Dean, N. (2016). Bayesian cluster detection via adjacency modelling. *Spatial and Spatio-temporal Epidemiology*, *16*, 11–20. doi: 10.1016/j.sste.2015.11.005

Anderson, T. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, *41*(3), 359–364. doi: 10.1016/j.aap.2008.12.014

Andrienko, G., Andrienko, N., Bak, P., Bremm, S., Keim, D., von Landesberger, T., . . . 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), 200–221.

Andrienko, G., Andrienko, N., Rinzivillo, S., Nanni, M., Pedreschi, D., & Giannotti, F. (2009). Interactive visual clustering of large collections of trajectories. In *2009 ieee symposium on visual analytics science and technology* (p. 3-10). doi: 10.1109/VAST.2009.5332584

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). OPTICS: Ordering Points to Identify the Clustering Structure. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data* (pp. 49–60). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/304182.304187` doi: 10.1145/304182.304187

Armstrong, C., Craig, B., Jackson III, W. E., & Thomson, J. B. (2014). The moderating influence of financial market development on the relationship between loan guarantees for SMEs and local market employment rates. *Journal of Small Business Management*, *52*(1), 126–140.

Arribas-Gil, A., & Müller, H.-G. (2014). Pairwise dynamic time warping for event data. *Computational Statistics & Data Analysis*, *69*, 255–268. doi: 10.1016/j.csda.2013.08.011

Arthur, D., & Vassilvitskii, S. (2006, June). *k-means++: The advantages of careful seeding* (Technical Report No. 2006-13). Stanford InfoLab. Retrieved from `http://ilpubs.stanford.edu:8090/778/`

Artola, C., & Genre, V. (2011). Euro area SMEs under financial constraints: Belief or reality?

Asero, V., Gozzo, S., & Tomaselli, V. (2016). Building tourism networks through tourist mobility. *Journal of Travel Research*, *55*(6), 751-763. doi: 10.1177/0047287515569777

Bachem, O., Lucic, M., Hassani, H., & Krause, A. (2016b). Fast and provably good seedings for k-means. In *Advances in neural information processing systems* (pp. 55–63).

Bachem, O., Lucic, M., Hassani, S. H., & Krause, A. (2016a). Approximate k-means++ in sublinear time. In *Aaai* (pp. 1459–1467).

Bangwayo-Skeete, P. F., & Skeete, R. W. (2015). Can google data improve the forecasting performance of tourist arrivals? mixed-data sampling approach. *Tourism Management*, *46*, 454 - 464. doi: 10.1016/j.tourman.2014.07.014

Barro, D., & Basso, A. (2010). Credit contagion in a network of firms with spatial interaction. *European Journal of Operational Research*, *205*, 459-468.

Basel Committee on Banking Supervision. (2005). Studies on the validation of internal rating systems. *Working Paper*.

BDRC Continental. (2018). *Small- and Medium-Sized Enterprise Finance Monitor, 2011-2017.* UK Data Service, 19th Edition. (Accessed: 3 June 2019)

Beck, T., Demirgüç-Kunt, A., Laeven, L., & Maksimovic, V. (2006). The determinants of financing obstacles. *Journal of International Money and Finance*, *25*(6), 932–952.

Beck, T., Demirgüç-Kunt, A., & Maksimovic, V. (2008). Financing patterns around the world: Are small firms different? *Journal of Financial Economics*, *89*(3), 467–487.

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., & Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in brief*, *29*, 105340.

Bernacchia, A., & Pigolotti, S. (2011). Self-consistent method for density estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), 407–422. doi: 10.1111/j.1467-9868.2011.00772.x

Biçici, E., & Yuret, D. (2007). Locally scaled density based clustering. In *International Conference on Adaptive and Natural Computing Algorithms* (pp. 739–748). doi: 10.1007/978-3-540-71618-1_82

Bie, R., Mehmood, R., Ruan, S., Sun, Y., & Dawood, H. (2016). Adaptive fuzzy clustering by fast search and find of density peaks. *Personal and Ubiquitous Computing*, *20*(5), 785–793. doi: 10.1007/s00779-016-0954-4

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, *60*(1), 208–221. doi: 10.1016/j.datak.2006.01.013

Bishop, C. M. (2006). *Pattern recognition and machine learning.* springer.

Blangiardo, M., Finazzi, F., & Cameletti, M. (2016). Two-stage Bayesian model to evaluate the effect of air pollution on chronic respiratory diseases using drug prescriptions. *Spatial and Spatio-temporal Epidemiology*, *18*, 1–12. doi: 10.1016/j.sste.2016.03.001

Bokelmann, B., & Lessmann, S. (2019). Spurious patterns in google trends data - an analysis of the effects on tourism demand forecasting in germany. *Tourism Management*, *75*, 1-12. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0261517719300779` doi: https://doi.org/10.1016/j.tourman.2019.04.015

Borah, B., & Bhattacharyya, D. (2004). An improved sampling-based DBSCAN for large spatial databases. In *Proceedings of International Conference on Intelligent Sensing and Information Processing* (pp. 92–96). doi: 10.1109/ICISIP.2004.1287631

Box, G. E. P. (1970). *Time series analysis : forecasting and control / [by] george e.p. box and gwilym m. jenkins.* San Francisco: Holden-Day.

British Business Bank. (2018). *Annual Report and Accounts 2018.* Retrieved from `https://annualreport2018.british-business-bank.co.uk` (Accessed: 26 September 2019)

Brown, M., Ongena, S., Popov, A., & Yeşin, P. (2011). Who needs credit and who gets credit in Eastern Europe? *Economic Policy*, *26*(65), 93–130.

Brunner, M. I., Furrer, R., & Favre, A.-C. (2019). Modeling the spatial dependence of floods using the Fisher copula. *Hydrology and Earth System Sciences*, *23*(1), 107–124.

Bulchand-Gidumal, J., & Melián-González, S. (2021). Post-COVID-19 behavior change in purchase of air tickets. *Annals of Tourism Research*, 103129. doi: https://doi.org/10.1016/j.annals.2020.103129

Cairns, K. J., Marshall, A. H., & Kee, F. (2011). Using simulation to assess cardiac first-responder schemes exhibiting stochastic and spatial complexities. *Journal of the Operational Research Society*, *62*(6), 982–991. doi: 10.1057/jors.2010.27

Calabrese, R., Andreeva, G., & Ansell, J. (2019). "Birds of a Feather" Fail Together: Exploring the Nature of Dependency in SME Defaults. *Risk Analysis*, *39*(1), 71-84.

Campello, M., Graham, J. R., & Harvey, C. R. (2010). The real effects of financial constraints: Evidence from a financial crisis. *Journal of Financial Economics*, *97*(3), 470–487.

Carbó-Valverde, S., Rodrıguez-Fernández, F., & Udell, G. F. (2008). *Bank lending, financing constraints and SME investment* (Tech. Rep.). working paper.

Carbó-Valverde, S., Rodrıguez-Fernández, F., & Udell, G. F. (2016). Trade credit, the financial crisis, and SME access to finance. *Journal of Money, Credit and Banking*, *48*(1), 113–143.

Casey, E., & O'Toole, C. M. (2014). Bank lending constraints, trade credit and alternative financing during the financial crisis: Evidence from European SMEs. *Journal of Corporate Finance*, *27*, 173–193.

Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, *40*(1), 200 - 210. Retrieved from http://www.sciencedirect.com/science/article/pii/S0957417412008767 doi: 10.1016/j.eswa.2012.07.021

Celeux, G., Forbes, F., & Peyrard, N. (2003). EM procedures using mean field-like approximations for Markov model-based image segmentation. *Pattern Recognition*, *36*(1), 131–144. doi: 10.1016/S0031-3203(02)00027-4

Chakraborty, A., & Hu, C. X. (2006). Lending relationships in line-of-credit and nonline-of-credit loans: Evidence from collateral use in small business. *Journal of Financial Intermediation*, *15*(1), 86–107.

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*(6), 1–36. Retrieved from `http://www.jstatsoft.org/v61/i06/`

Chavis, L. W., Klapper, L. F., & Love, I. (2011). The impact of the business environment on young firm financing. *The World Bank Economic Review*, *25*(3), 486–507.

Chen, L., Özsu, M. T., & Oria, V. (2005). Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD International conference on management of data* (pp. 491–502). doi: 10.1145/1066157.1066213

Chen, M.-H. (2015). Understanding the impact of changes in consumer confidence on hotel stock performance in Taiwan. *International Journal of Hospitality Management*, *50*, 55-65. doi: https://doi.org/10.1016/j.ijhm.2015.07.010

Chen, Y.-T., Sun, E. W., & Lin, Y.-B. (2020). Merging anomalous data usage in wireless mobile telecommunications: Business analytics with a strategy-focused data-driven approach for sustainability. *European Journal of Operational Research*, *281*(3), 687-705. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0377221719301948` (Featured Cluster: Business Analytics: Defining the field and identifying a research agenda) doi: https://doi.org/10.1016/j.ejor.2019.02.046

Choi, H., & Varian, H. (2012). Predicting the present with google trends. *Economic record*, *88*, 2–9. doi: 10.1111/j.1475-4932.2012.00809.x

Cole, R. A. (2008). Who needs credit and who gets credit? Evidence from the surveys of small business finances.

Coll, B., Moutari, S., & Marshall, A. (2014). Pattern Recognition Approach for Road Collision Hotspot Analysis: Case Study of Northern Ireland. *Proceedings of the ITRN2014, University of Limerick.*

Collineau, L., Carmo, L. P., Endimiani, A., Magouras, I., Müntener, C., Schüpbach-Regula, G., & Stärk, K. D. C. (2018). Risk Ranking of Antimicrobial-Resistant Hazards Found in Meat in Switzerland. *Risk Analysis*, *38*(5), 1070-1084. doi: 10.1111/risa.12901

Columbus, L. (2021, Jan 17). 76% Of Enterprises Prioritize AI & Machine Learning In 2021 IT Budgets. *Forbes*. Retrieved from `https://www.forbes.com/sites/louiscolumbus/2021/01/17/76-of-enterprises-prioritize-ai--machine-learning-in-2021-it-budgets` (Accessed: 10 June 2021)

Cosh, A., Cumming, D., & Hughes, A. (2009). Outside enterpreneurial capital. *The Economic Journal*, *119*(540), 1494–1533.

Craig, B. R., Jackson III, W. E., & Thomson, J. B. (2007). Small firm finance, credit rationing, and the impact of SBA-guaranteed lending on local economic growth. *Journal of Small Business Management*, *45*(1), 116–132.

Cressie, N. (1992). Statistics for spatial data. *Terra Nova*, *4*(5), 613–617.

De Angelis, L., & Dias, J. G. (2014). Mining categorical sequences from data using a hybrid clustering method. *European Journal of Operational Research*, *234*(3), 720–730. doi: 10.1016/j.ejor.2013.11.002

Deloitte. (2020). *COVID-19 global mobility update.* `file:///tmp/deloitte-ch-covid-19-global-mobility-update-07-14-october-2020-2.pdf`.

Demir, E., & Ersan, O. (2018). The impact of economic policy uncertainty on stock returns of Turkish tourism companies. *Current Issues in Tourism*, *21*(8), 847-855. doi: 10.1080/13683500.2016.1217195

Department of Health and Social Care. (2016). *Government response to the Review on Antimicrobial Resistance* (Tech. Rep.). Retrieved from `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\_data/file/553471/Gov\_response\_AMR\_Review.pdf`

Dias, J. G., Vermunt, J. K., & Ramos, S. (2015). Clustering financial time series: New insights from an extended hidden Markov model. *European Journal of Operational Research*, *243*(3), 852–864. doi: 10.1016/j.ejor.2014.12.041

Di Lascio, F. M. L., Durante, F., & Pappada, R. (2017). Copula–based clustering methods. In *Copulas and dependence models with applications* (pp. 49–67). Springer. doi: "https://doi.org/10.1007/978-3-319-64221-5_4"

Disegna, M., D'Urso, P., & Durante, F. (2017). Copula-based fuzzy clustering of spatial time series. *Spatial Statistics*, *21*, 209–225. doi: 10.1016/j.spasta.2017.07.002

Durkin, M. J., Jafarzadeh, S. R., Hsueh, K., Sallah, Y. H., Munshi, K. D., Henderson, R. R., & Fraser, V. J. (2018). Outpatient antibiotic prescription trends in the United States: a national cohort study. *Infection Control & Hospital Epidemiology*, *39*(5), 584–589. doi: 10.1017/ice.2018.26

D'Urso, P., De Giovanni, L., Disegna, M., & Massari, R. (2019). Fuzzy clustering with spatial–temporal information. *Spatial Statistics*, *30*, 71-102. Retrieved from `https://www.sciencedirect.com/science/article/pii/S2211675318301994` doi: https://doi.org/10.1016/j.spasta.2019.03.002

Edachery, J., Sen, A., & Brandenburg, F. J. (1999). Graph clustering using distance-k cliques. In *International Symposium on Graph Drawing* (pp. 98–106).

Ertöz, L., Steinbach, M., & Kumar, V. (2003). Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In *Proceedings of the 2003 SIAM International Conference on Data Mining* (pp. 47–58). doi: 10.1137/1.9781611972733.5

Eslinger, R., & Morgan, J. D. (2017). Spatial Cluster Analysis of High-Density Vehicle–Bear Collisions and Bridge Locations. *Papers in Applied Geography*, *3*(2), 171–181. doi: 10.1080/23754931.2017.1299633

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In (Vol. 96(34), pp. 226–231).

Eurocontrol. (2021). *COVID-19 impact on the european air traffic network.* `https://www.eurocontrol.int/covid19`. (Accessed: 1 March 2021)

Fernandes, G. B., & Artes, R. (2016). Spatial dependence in credit risk and its improvement in credit scoring. *European Journal of Operational Research*, *249*, 517-524.

Ferrando, A., & Griesshaber, N. (2011, February). *Financing Obstacles Among Euro Area Firms: Who Suffers the Most?* ECB Working Paper No. 1293. (Available at SSRN: https://ssrn.com/abstract=1757728)

Ferretti, V., & Montibeller, G. (2019). An Integrated Framework for Environmental Multi-Impact Spatial Risk Analysis. *Risk Analysis*, *39*(1), 257-273. doi: 10.1111/risa.12942

Foreign & Commonwealth Office. (2020). *Foreign Secretary advises all British travellers to return* *to the UK now [Press release].* Retrieved from `"https://www.gov.uk/government/news/foreign-secretary-advises-all-british-travellers-to-return-to-the-uk-now"` (Accessed: 7 July 2020)

Fränti, P., & Sieranoja, S. (2019). How much can k-means be improved by using better initialization and repeats? *Pattern Recognition*, *93*, 95–112.

Gabor, M. R., Conţiu, L. C., & Oltean, F. D. (2012). A comparative analysis regarding European tourism competitiveness: emerging versus developed markets. *Procedia Economics and Finance*, *3*, 361–366.

Gallego, I., & Font, X. (2020). Changes in air passenger demand as a result of the COVID-19 crisis: using Big Data to inform tourism policy. *Journal of Sustainable Tourism*, 1-20. doi: 10.1080/09669582.2020.1773476

Gama, A. P. M., & Van Auken, H. (2015). The interdependence between trade credit and bank lending: commitment in intermediary firm relationships. *Journal of Small Business Management*, *53*(4), 886–904.

George, A. N., Stewart, J. R., Evans, J. C., & Gibson, J. M. (2020). Risk of Antibiotic-Resistant Staphylococcus aureus Dispersion from Hog Farms: A Critical Review. *Risk Analysis*, *40*(8), 1645-1665. doi: 10.1111/risa.13495

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, *61*, 115–125. doi: 10.1016/j.dss.2014.02.003

Giannetti, M., Burkart, M., & Ellingsen, T. (2011). What you sell is what you lend? Explaining trade credit contracts. *The Review of Financial Studies*, *24*(4), 1261–1298.

Gieschen, A., Ansell, J., Calabrese, R., & Martin-Barragan, B. (2021). Modeling Antimicrobial Prescriptions in Scotland: A Spatiotemporal Clustering Approach. *Risk Analysis*. (Advance online publication) doi: 10.1111/risa.13795

Giesecke, K., & Weber, S. (2006). Credit contagion and aggregate losses. *Journal of Economic Dynamics & Control*, *30*, 741-767.

Gilje, E. P. (2017). Does local access to finance matter? evidence from US oil and natural gas shale booms. *Management Science*, *65*(1), 1–18.

Giorgino, T., et al. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of statistical Software*, *31*(7), 1–24.

Gomide, J., Veloso, A., Meira, W., Almeida, V., Benevenuto, F., Ferraz, F., & Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. In *Proceedings of the 3rd International Web Science Conference.* doi: 10.1145/2527031.2527049

González-Olabarria, J. R., Mola-Yudego, B., & Coll, L. (2015). Different Factors for Different Causes: Analysis of the Spatial Aggregations of Fire Ignitions in Catalonia (Spain). *Risk Analysis*, *35*(7), 1197-1209. doi: 10.1111/risa.12339

Google. (2020). *FAQ about Google Trends data.* Retrieved from `"https://support.google.com/trends/answer/4365533\?hl=en-GB\&ref\_topic=6248052"` (Accessed: 23 July 2020)

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857–871.

Graham, A., Kremarik, F., & Kruse, W. (2020). Attitudes of ageing passengers to air travel since the coronavirus pandemic. *Journal of Air Transport Management*, *87*, 101865. Retrieved from `https://www.sciencedirect.com/science/article/pii/S096969972030449X` doi: https://doi.org/10.1016/j.jairtraman.2020.101865

Gursoy, D., & Chi, C. G. (2020). Effects of covid-19 pandemic on hospitality industry: review of the current situations and a research agenda. *Journal of Hospitality Marketing & Management*, *29*(5), 527-529. doi: 10.1080/19368623.2020.1788231

Gössling, S., Scott, D., & Hall, C. M. (2020). Pandemics, tourism and global change: a rapid assessment of covid-19. *Journal of Sustainable Tourism*, *0*(0), 1-20. doi: 10.1080/09669582.2020.1758708

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Elsevier Science.

Hardin, J. W., Hardin, J. W., Hilbe, J. M., & Hilbe, J. (2007). *Generalized linear models and extensions.* Stata press.

Hartuv, E., & Shamir, R. (2000). A clustering algorithm based on graph connectivity. *Information Processing Letters*, *76*(4), 175-181. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0020019000001423` doi: https://doi.org/10.1016/S0020-0190(00)00142-3

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models* (Vol. 43). CRC press.

Havranek, T., & Zeynalov, A. (2021). Forecasting tourist arrivals: Google Trends meets mixed-frequency data. *Tourism Economics*, *27*(1), 129-148. doi: 10.1177/1354816619879584

Healthcare Quality and Improvement Directorate. (2018). *Practising Realistic Medicine: Chief Medical Officer for Scotland annual report* (Tech. Rep.). Retrieved from `https://www.gov.scot/publications/practising-realistic-medicine/`

Hijmans, R. J. (2017). geosphere: Spherical Trigonometry [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=geosphere` (R package version 1.5-7)

Hijmans, R. J. (2019). geosphere: Spherical trigonometry [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=geosphere` (R package version 1.5-10)

Hinneburg, A., & Gabriel, H.-H. (2007). DENCLUE 2.0: Fast Clustering based on Kernel Density Estimation. In *International symposium on intelligent data analysis* (pp. 70–80). doi: 10.1007/978-3-540-74825-0_7

Hosany, S., & Prayag, G. (2013). Patterns of tourists' emotional responses, satisfaction, and intention to recommend. *Journal of Business Research*, *66*(6), 730-737. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0148296311003201` (International Tourism Behavior in Turbulent Times) doi: https://doi.org/10.1016/j.jbusres.2011.09.011

Hu, X.-B., Li, H., Guo, X., van Gelder, P. H. A. J. M., & Shi, P. (2019). Spatial Vulnerability of Network Systems under Spatially Local Hazards. *Risk Analysis*, *39*(1), 162-179. doi: 10.1111/risa.12986

Huang, X., Zhang, L., & Ding, Y. (2017). The Baidu Index: Uses in predicting tourism flows – a case study of the Forbidden City. *Tourism Management*, *58*, 301-306. doi: https://doi.org/10.1016/j.tourman.2016.03.015

Hwang, Y.-H., Gretzel, U., & Fesenmaier, D. R. (2006). Multicity trip patterns: Tourists to the United States. *Annals of Tourism Research*, *33*(4), 1057-1078. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0160738306000569` doi: https://doi.org/10.1016/j.annals.2006.04.004

Hyndman, R., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach.* Springer Science & Business Media.

Inniss, T. R. (2006). Seasonal clustering technique for time series data. *European Journal of Operational Research*, *175*(1), 376–384. doi: 10.1016/j.ejor.2005.03.049

Izakian, H., Pedrycz, W., & Jamal, I. (2015). Fuzzy clustering of time series data using dynamic time warping distance. *Engineering Applications of Artificial Intelligence*, *39*, 235–244. doi: 10.1016/j.engappai.2014.12.015

Jiménez, G., Ongena, S., Peydró, J.-L., & Saurina, J. (2012). Credit supply and monetary policy: Identifying the bank balance-sheet channel with loan applications. *American Economic Review*, *102*(5), 2301–26.

Johns Hopkins University. (2020). *COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.* Retrieved from `"https://coronavirus.jhu.edu"` (Accessed: 4 March 2021)

Juhro, S. M., & Iyke, B. N. (2020). Consumer confidence and consumption expenditure in Indonesia. *Economic Modelling*, *89*, 367-377. doi: https://doi.org/10.1016/j.econmod.2019.11.001

Katz, N., Panas, L., Kim, M., Audet, A. D., Bilansky, A., Eadie, J., … Carrow, G. (2010). Usefulness of prescription monitoring programs for surveillance–analysis of Schedule ii opioid prescription data in Massachusetts, 1996–2006. *Pharmacoepidemiology and Drug Safety*, *19*(2), 115–123. doi: 10.1002/pds.1878

Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. In *Proceedings of the Statistical Data Analysis Based on the L1 Norm Conference, Neuchatel, Switzerland* (pp. 405–416).

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis* (Vol. 3). John Wiley & Sons.

Kim, J., Veremyev, A., Boginski, V., & Prokopyev, O. A. (2020). On the maximum small-world subgraph problem. *European Journal of Operational Research*, *280*(3), 818-831. doi: https://doi.org/10.1016/j.ejor.2019.07.042

Kisilevich, S., Mansmann, F., Nanni, M., & Rinzivillo, S. (2009). Spatio-temporal clustering. In *Data mining and knowledge discovery handbook* (pp. 855–874). Springer.

Kjærulff, T. M., Ersbøll, A. K., Gislason, G., & Schipperijn, J. (2016). Geographical clustering of incident acute myocardial infarction in Denmark: A spatial analysis approach. *Spatial and spatio-temporal epidemiology*, *19*, 46–59. doi: 10.1016/j.sste.2016.05.001

Klier, T., & McMillen, D. P. (2008). Clustering of Auto Supplier Plants in the United States. *Journal of Business & Economic Statistics*, *26*(4), 460-471.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, *43*(1), 59–69.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480.

Kourgiantakis, M., Apostolakis, A., & Dimou, I. (2020). COVID-19 and holiday intentions: the case of Crete, Greece. *Anatolia*, *0*(0), 1-4. doi: 10.1080/13032917.2020.1781221

Kriegel, H.-P., Kröger, P., Sander, J., & Zimek, A. (2011). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(3), 231–240. doi: 10.1002/widm.30

Kufel, T., et al. (2020). ARIMA-based forecasting of the dynamics of confirmed Covid-19 cases for selected European countries. *Equilibrium. Quarterly Journal of Economics and Economic Policy*, *15*(2), 181–204.

Lascu, D.-N., Manrai, L. A., Manrai, A. K., & Gan, A. (2018). A cluster analysis of tourist attractions in Spain. *European Journal of Management and Business Economics*.

Lee, N., & Brown, R. (2017). Innovation, SMEs and the liability of distance: the demand and supply of bank funding in UK peripheral regions. *Journal of Economic Geography*, *17*(1), 233–260.

Lee, N., & Luca, D. (2019). The big-city bias in access to finance: evidence from firm perceptions in almost 100 countries. *Journal of Economic Geography*, *19*(1), 199–224.

LeSage, J. P. (2009). *Introduction to spatial econometrics.* Boca Raton: CRC Press.

Liao, T. W. (2005). Clustering of time series data – A survey. *Pattern Recognition*, *38*(11), 1857–1874. doi: 10.1016/j.patcog.2005.01.025

Lin, F.-J., & Lin, Y.-H. (2016). The effect of network relationship on the performance of SMEs. *Journal of Business Research*, *69*(5), 1780 - 1784. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0148296315004786` (Designing implementable innovative realities) doi: https://doi.org/10.1016/j.jbusres.2015.10.055

Lin, J., Khade, R., & Li, Y. (2012). Rotation-invariant similarity in time series using bag-of-patterns representation. *Journal of Intelligent Information Systems*, *39*(2), 287–315.

Liu, A., Vici, L., Ramos, V., Giannoni, S., & Blake, A. (2021). Visitor arrivals forecasts amid COVID-19: A perspective from the Europe team. *Annals of Tourism Research*, *88*, 103182. doi: https://doi.org/10.1016/j.annals.2021.103182

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2018). cluster: Cluster analysis basics and extensions [Computer software manual]. (R package version 2.0.7-1 — For new features, see the 'Changelog' file (in the package source))

Mai, F., Fry, M. J., & Ohlmann, J. W. (2018). Model-based capacitated clustering with posterior regularization. *European Journal of Operational Research*, *271*(2), 594–605. doi: 10.1016/j.ejor.2018.04.048

Manolova, T. S., Manev, I. M., & Gyoshev, B. S. (2014). Friends with money? Owner's financial network and new venture internationalization in a transition economy. *International Small Business Journal*, *32*(8), 944–966.

Marbac, M., Biernacki, C., & Vandewalle, V. (2017). Model-based clustering of Gaussian copulas for mixed data. *Communications in Statistics-Theory and Methods*, *46*(23), 11635–11656. doi: 10.1080/03610926.2016.1277753

Matioli, L. C., Santos, S. R., Kleina, M., & Leite, E. A. (2018). A new algorithm for clustering based on kernel density estimation. *Journal of Applied Statistics*, *45*(2), 347–366. doi: 10.1080/02664763.2016.1277191

Matiza, T. (2020). Post-covid-19 crisis travel behaviour: towards mitigating the effects of perceived risk. *Journal of Tourism Futures*.

Mazanec, J. A. (2010). Tourism-Receiving Countries in Connotative Google Space. *Journal of Travel Research*, *49*(4), 501-512. doi: 10.1177/0047287509349269

McMillen, D., & McMillen, M. D. (2013). Package 'McSpatial'. *Nonparametric Spatial Data Analysis. August*, *4*.

Mehmood, R., Zhang, G., Bie, R., Dawood, H., & Ahmad, H. (2016). Clustering by fast search and find of density peaks via heat diffusion. *Neurocomputing*, *208*, 210–217. doi: 10.1016/j.neucom.2016.01.102

Moews, B., Argueta, J. R., & Gieschen, A. (2021). Filaments of crime: Informing policing via thresholded ridge estimation. *Decision Support Systems*, *144*, 113518. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167923621000282 doi: https://doi.org/10.1016/j.dss.2021.113518

Mölter, A., Belmonte, M., Palin, V., Mistry, C., Sperrin, M., White, A., … Van Staa, T. (2018). Antibiotic prescribing patterns in general medical practices in England: Does area matter? *Health & Place*, *53*, 10–16. doi: 10.1016/j.healthplace.2018.07.004

Montani, S., Portinale, L., Leonardi, G., Bellazzi, R., & Bellazzi, R. (2006). Case-based retrieval to support the treatment of end stage renal failure patients. *Artificial Intelligence in Medicine*, *37*(1), 31–42. doi: 10.1016/j.artmed.2005.06.003

Mori, U., Mendiburu, A., & Lozano, J. A. (2016). Distance measures for Time Series in R: The TSdist package. *R Journal*, *8*(2), 451–459.

Moritz, A., Block, J. H., & Heinz, A. (2016). Financing patterns of European SMEs–an empirical taxonomy. *Venture Capital*, *18*(2), 115–148.

Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time Series Missing Value Imputation in R. *The R Journal*, *9*(1), 207–218. doi: 10.32614/RJ-2017-009

Murphy, K. P. (2012). *Machine learning a probabilistic perspective.* Cambridge, MA: MIT Press.

Musgrove, D., Hughes, J., & Eberly, L. (2016). Hierarchical copula regression models for areal data. *Spatial Statistics*, *17*, 38-49. doi: https://doi.org/10.1016/j.spasta.2016.04.006

Mwaura, S., & Carter, S. (2017). No joy with banks' round here: the geography of the usage of bank financing among UK SMEs. In *RSA Annual Conference.*

Nanni, M., & Pedreschi, D. (2006). Time-focused clustering of trajectories of moving objects. *Journal of Intelligent Information Systems*, *27*(3), 267–289.

Ng, R. T., & Han, J. (2002). CLARANS: A Method for Clustering Objects for Spatial Data Mining. *IEEE Transactions on Knowledge & Data Engineering*, *14*, 1003–1016. doi: 10.1109/TKDE.2002.1033770

NHS Education for Scotland. (2020). *Scottish Reduction in Antimicrobial Prescribing (ScRAP).* Retrieved from https://www.nes.scot.nhs.uk/education-and-training/by-theme-initiative/healthcare-associated-infections/training-resources/scottish-reduction-in-antimicrobial-prescribing-(scrap).aspx (Accessed: 15 April 2020)

NHS ISD. (2018). *General Practice - GP Workforce and practice list sizes 2008-2018* (Tech. Rep.). Retrieved from `http://www.isdscotland.org/Health-Topics/General-Practice/Publications/2018-12-11/2018-12-11-GPWorkforce2018-Summary.pdf`

NHS Scotland. (2013). *Healthboard Areas of NHS Scotland.* Retrieved from `https://www.scot.nhs.uk/mapofscotlandshowversion-2/`

OECD. (2020). *Consumer confidence index (CCI) (indicator).* (Accessed: 28 July 2020) doi: 10.1787/46434d78-en

Owen, R., Botelho, T., & Anwar, O. (2016). Exploring the success and barriers to SME access to finance and its potential role in achieving growth. *ERC Research Paper*, *53*.

Pan, B., Wu, D. C., & Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology*, *3*(3), 196-210.

Paparrizos, J., & Gravano, L. (2017). Fast and Accurate Time-Series Clustering. *ACM Transactions on Database Systems (TODS)*, *42*(2), 8. doi: 10.1145/3044711

Pei, T., Jasra, A., Hand, D. J., Zhu, A.-X., & Zhou, C. (2009). DECODE: a new method for discovering clusters of different densities in spatial data. *Data Mining and Knowledge Discovery*, *18*, 337–369. doi: 10.1007/s10618-008-0120-3

Perles-Ribes, J. F., Ivars-Baidal, J. A., Ramón-Rodríguez, A. B., & Vera-Rebollo, J. F. (2020). The typological classification of tourist destinations: The region of Valencia, a case study. *Tourism Economics*, *26*(5), 764-773. doi: 10.1177/1354816619838413

Petersen, M. A., & Rajan, R. G. (1997). Trade Credit: Theories and Evidence. *The Review of Financial Studies*, *10*(3), 661–691.

Pinkse, J., & Slade, M. E. (1998). Contracting in space: An application of spatial statistics to discrete-choice models. *Journal of Econometrics*, *85*(1), 125–154.

Polyzos, S., Samitas, A., & Spyridou, A. E. (2020). Tourism demand and the covid-19 pandemic: an lstm approach. *Tourism Recreation Research*, 1-13. doi: 10.1080/02508281.2020.1777053

Popov, A., & Udell, G. F. (2012). Cross-border banking, credit access, and the financial crisis. *Journal of International Economics*, *87*(1), 147–161.

Povinelli, R. J., Johnson, M. T., Lindgren, A. C., & Ye, J. (2004). Time series classification using Gaussian mixture models of reconstructed phase spaces. *IEEE Transactions on Knowledge and Data Engineering*, *16*(6), 779–783. doi: 10.1109/TKDE.2004.17

Presbitero, A. F., Udell, G. F., & Zazzaro, A. (2014). The Home Bias and the Credit Crunch: A Regional Perspective. *Journal of Money, Credit and Banking*, *46*(1), 53-85.

Psillaki, M., & Eleftheriou, K. (2015). Trade Credit, Bank Credit, and Flight to Quality: Evidence from French SMEs. *Journal of Small Business Management*, *53*(4), 1219–1240.

Public Health England. (2019, November). *Antibiotic awareness resources.* Retrieved from `https://www.gov.uk/government/collections/european-antibiotic-awareness-day-resources` (Accessed: 15 April 2020)

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ramires, A., Brandão, F., & Sousa, A. C. (2018). Motivation-based cluster analysis of international tourists visiting a World Heritage City: The case of Porto, Portugal. *Journal of Destination Marketing & Management*, *8*, 49-60. Retrieved from `https://www.sciencedirect.com/science/article/pii/S2212571X16300579` doi: https://doi.org/10.1016/j.jdmm.2016.12.001

Räsänen, T., & Kolehmainen, M. (2009). Feature-based clustering for electricity use time series data. In M. Kolehmainen, P. Toivanen, & B. Beliczynski (Eds.), *Adaptive and natural computing algorithms* (pp. 401–412). Berlin, Heidelberg: Springer Berlin Heidelberg.

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science*, *344*(6191), 1492–1496. doi: 10.1126/science.1242072

Romagosa, F. (2020). The COVID-19 crisis: Opportunities for sustainable and proximity tourism. *Tourism Geographies*, *22*(3), 690-694. doi: 10.1080/14616688.2020.1763447

Romano, T., Cambini, C., Fumagalli, E., & Rondi, L. (2021). Setting network tariffs with heterogeneous firms: The case of natural gas distribution. *European Journal of Operational Research.* Retrieved from `https://www.sciencedirect.com/science/article/pii/S0377221721004331` doi: https://doi.org/10.1016/j.ejor.2021.05.019

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65. Retrieved from https://www.sciencedirect.com/science/article/pii/0377042787901257 doi: https://doi.org/10.1016/0377-0427(87)90125-7

Ruiz, C., Spiliopoulou, M., & Menasalvas, E. (2007). C-DBSCAN: Density-Based Cclustering with Constraints. In *International workshop on rough sets, fuzzy sets, data mining, and granular-soft computing* (pp. 216–223). doi: 10.1007/978-3-540-72530-5_25

Ryan, C., & Huyton, J. (2000). Who is interested in Aboriginal tourism in the Northern territory, Australia? a cluster analysis. *Journal of Sustainable Tourism*, *8*(1), 53-88. doi: 10.1080/09669580008667349

Ryan, R. M., O'Toole, C. M., & McCann, F. (2014). Does bank market power affect SME financing constraints? *Journal of Banking & Finance*, *49*, 495–505.

Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, *1*(1), 27–64. doi: 10.1016/j.cosrev.2007.05.001

Scotland's AI Strategy. (2021). Retrieved from https://www.scotlandaistrategy.com (Accessed: 10 June 2021)

Scott, D. W. (1992). *Multivariate Density Estimation. Theory, Practice and Visualization.* New York: Wiley.

Scott, D. W. (2009). Sturges' rule. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*(3), 303–306. doi: 10.1002/wics.35

Scottish Antimicrobial Prescribing Group. (2018). *Scottish Reduction in Antimicrobial Prescribing (ScRAP) Programme V2 Support Pack.* Retrieved from "https://www.nes.scot.nhs.uk/media/4081795/scrap_support_pack_-_feb_2018.pdf"

Scottish Antimicrobial Prescribing Group. (2020a). *Antibiotic awareness.* Retrieved from https://www.sapg.scot/antibiotic-awareness/ (Accessed: 15 April 2020)

Scottish Antimicrobial Prescribing Group. (2020b). *Primary care.* Retrieved from https://www.sapg.scot/quality-improvement/primary-care/ (Accessed: 15 April 2020)

Scottish Enterprise. (2019). *Building Scotland's Future Today.* (Retrieved from "https://www.scottish-enterprise.com/media/3109/scottish-enterprise-building-scotlands-future-today.pdf", Accessed: 26 Sept 2019)

Serra, J., & Arcos, J. L. (2014). An empirical evaluation of similarity measures for time series classification. *Knowledge-Based Systems*, *67*, 305–314. doi: 10.1016/j.knosys.2014.04.035

Seyfi, S., Hall, C. M., & Shabani, B. (2020). COVID-19 and international travel restrictions: the geopolitics of health and tourism. *Tourism Geographies*, *0*(0), 1-17. doi: 10.1080/14616688.2020.1833972

Siliverstovs, B., & Wochner, D. S. (2018). Google trends and reality: Do the proportions match?: Appraising the informational value of online search behavior: Evidence from swiss tourism regions. *Journal of Economic Behavior & Organization*, *145*, 1-23. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167268117302937 doi: https://doi.org/10.1016/j.jebo.2017.10.011

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis* (Vol. 26). Chapman & Hall.

Skyscanner. (n.d.). *Skyscanner: About us.* https://www.skyscanner.es/es/en-gb/eur/about-us. (Accessed: 01 May 2021)

Smith, C. M., Lessells, R., Grant, A. D., Herbst, K., & Tanser, F. (2018). Spatial clustering of drug-resistant tuberculosis in Hlabisa subdistrict, KwaZulu-Natal, 2011–2015. *The International Journal of Tuberculosis and Lung Disease*, *22*(3), 287–293. doi: 10.5588/ijtld.17.0457

Sokal, R., & Michener, C. (1958). A Statistical Method for Evaluating Systematic Relationships. In *University of Kansas Science Bulletin.* University of Kansas, 38(2), 1409–1438.

Spatial patterns of cultural tourism in Portugal. (2015). *Tourism Management Perspectives*, *16*, 107-115. doi: https://doi.org/10.1016/j.tmp.2015.07.010

Svetunkov, I. (2017, February 28). *Statistical models underlying functions of 'smooth' package for R* (Working Paper). Lancaster University Management School.

Svetunkov, I. (2020). smooth: Forecasting using state space models [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=smooth (R package version 2.6.0)

Terrell, G. R., & Scott, D. W. (1992). Variable Kernel Density Estimation. *The Annals of Statistics*, *20*(3), 1236–1265. doi: 10.1214/aos/1176348768

The Federal Government of Germany. (2020). *Cabinet extends travel warning.* Retrieved from "https://www.bundesregierung.de/breg-en/news/reisewarnung-verlaengert-1749424" (Accessed: 7 July 2020)

The Scottish Government. (2021, March). *Scotland's artificial intelligence strategy.* Retrieved from "https://www.scotlandaistrategy.com/s/Scotlands_AI_Strategy_Web_updated_single_page_aps.pdf" (Accessed: 10 June 2021)

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Tiwasing, P., Gorton, M., Phillipson, J., Maioli, S., & Newbery, R. (2019). Spatial Variations in SME Productivity. *ESRC Productivity Insights Network.*

Tormene, P., Giorgino, T., Quaglini, S., & Stefanelli, M. (2009). Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation. *Artificial Intelligence in Medicine*, *45*(1), 11–34. doi: 10.1016/j.artmed.2008.11.007

Tourism and Events Division. (2020). *Policy: Tourism and events.* Retrieved from "https://www.gov.scot/policies/tourism-and-events/" (Accessed: 22 November 2020)

Van Puyenbroeck, T., Montalto, V., & Saisana, M. (2021). Benchmarking culture in Europe: A data envelopment analysis approach to identify city-specific strengths. *European Journal of Operational Research*, *288*(2), 584-597. Retrieved from https://www.sciencedirect.com/science/article/pii/S0377221720305191 doi: https://doi.org/10.1016/j.ejor.2020.05.058

Vaughan, L., & Chen, Y. (2015). Data mining from web search queries: A comparison of google trends and baidu index. *Journal of the Association for Information Science and Technology*, *66*(1), 13–22. doi: 10.1002/asi.23201

Verma, A., & Butenko, S. (2013). Network clustering via clique relaxations: A community based. *Graph Partitioning and Graph Clustering*, *588*, 129.

Vos, E., Yeh, A. J.-Y., Carter, S., & Tagg, S. (2007). The happy story of small business financing. *Journal of Banking & Finance*, *31*(9), 2648–2672.

Wang, K., Gasser, T., et al. (1997). Alignment of curves by dynamic time warping. *The Annals of Statistics*, *25*(3), 1251–1276.

Wang, M., Wang, A., & Li, A. (2006). Mining spatial-temporal clusters from geo-databases. In *International Conference on Advanced Data Mining and Applications* (pp. 263–270).

Wang, X., Smith, K., & Hyndman, R. (2006). Characteristic-based clustering for time series data. *Data mining and knowledge Discovery*, *13*(3), 335–364.

Wehinger, G. (2012). Bank deleveraging, the move from bank to market-based financing, and SME financing. *OECD Journal: Financial Market Trends*, *2012*(1), 65–79.

Wilhelm, S., & de Matos, M. G. (2013). Estimating spatial probit models in R. *The R Journal*, *5*(1), 130–143.

Williams, E. (2011). Aviation Formulary V1. 42. *Aviation*, *1*, 42. Retrieved from `ftp://ftp.bartol.udel.edu/anita/amir/My_thesis/Figures4Thesis/CRC_plots/Aviation\%20Formulary\%20V1.46.pdf`

World Health Organization. (2020). *WHO Director-General's statement on IHR Emergency Committee on Novel Coronavirus (2019-nCoV).* Retrieved from `"https://www.who.int/dg/speeches/detail/who-director-general-s-statement-on-ihr-emergency-committee-on-novel-coronavirus-(2019-ncov)"` (Accessed: 7 July 2020)

Xie, G., Qian, Y., & Wang, S. (2021). Forecasting chinese cruise tourism demand with big data: An optimized machine learning approach. *Tourism Management*, *82*, 104208. doi: https://doi.org/10.1016/j.tourman.2020.104208

Xing, Z., Pei, J., & Philip, S. Y. (2012). Early classification on time series. *Knowledge and Information Systems*, *31*(1), 105–127. doi: 10.1007/s10115-011-0400-x

Xu, R., & Wunsch, D. (2008). *Clustering* (Vol. 10). John Wiley & Sons.

Yang, X., Pan, B., Evans, J. A., & Lv, B. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management*, *46*, 386-397. doi: https://doi.org/10.1016/j.tourman.2014.07.019

Yang, Y., Altschuler, B., Liang, Z., & Li, X. R. (2020). Monitoring the global COVID-19 impact on tourism: The COVID19tourism index. *Annals of Tourism Research*, 103120. doi: https://doi.org/10.1016/j.annals.2020.103120

Zagmutt, F. J., Schoenbaum, M. A., & Hill, A. E. (2016). The Impact of Population, Contact, and Spatial Heterogeneity on Epidemic Model Predictions. *Risk Analysis*, *36*(5), 939-953. doi: 10.1111/risa.12482

Zahn, C. (1971). Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters. *IEEE Transactions on Computers*, *C-20*(1), 68-86. doi: 10.1109/T-C.1971.223083

Zelnik-Manor, L., & Perona, P. (2004). Self-tuning spectral clustering. In (Vol. 17, pp. 1601–1608).

Zenker, S., & Kock, F. (2020). The coronavirus pandemic–a critical discussion of a tourism research agenda. *Tourism management*, *81*, 104164.

Zhang, J. (1992). The mean field theory in EM procedures for Markov random fields. *IEEE Transactions on signal processing*, *40*(10), 2570–2583.

Önder, I. (2017). Forecasting tourism demand with google trends: Accuracy comparison of countries versus cities. *International Journal of Tourism Research*, *19*(6), 648-660. doi: https://doi.org/10.1002/jtr.2137