Time-Series Cluster Analysis for Irish Covid-19 Data at County Level

Patrick Garrett (01571095)
School of Computer Science
National University of Ireland, Galway

Supervisors
Dr. Matthias Nickles

In partial fulfillment of the requirements for the degree of
MSc in Computer Science (Artificial Intellgence)
September, 2021

DECLARATION I, Patrick Garrett, do hereby declare that this thesis entitled Time-Series Cluster Analysis for Irish Covid-19 Data at County Level is a bonafide record of research work done by me for the award of MSc in Computer Science (Artificial Intelligence) from National University of Ireland, Galway. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

Signature

Patrick Garrett

Acknowledgement A word of thanks to my supervisor Dr. Matthias Nickles for his guidance throughout this project. I would also like to thank my family, in particular my mother, for their support and encouragement over the past year.

Abstract

It has been a year-and-a-half since the Covid-19 virus was declared a pandemic by the World Health Organisation. At time of writing there have been over 219 million recorded cases of Covid-19 globally, with over 357,000 of those cases occurring in Ireland. Governments and medical researchers worldwide have conducted numerous studies and acquired massive amounts of data to track the spread of the disease with the intention of implementing measures to control it. In this report I analyse the spread of Covid-19 within Ireland by clustering time-series data representing the number of infections recorded in each county since the outbreak of the virus. Different configurations of common clustering algorithms are applied to publicly available Covid-19 data with the intention of identifying the most coherent clusters. Nine clusters of counties are identified. Statistical tests are then performed to identify socioeconomic and demographic factors that may correlate with the clusters.

Keywords: time-series, clustering, Ireland, Irish, counties, Covid-19, infectious disease

Contents

List of Figures

Chapter 1

Introduction

In this introductory section, I outline the motivation behind this project with reference to the Irish context of the Covid-19 pandemic.

1.1 Overview

In the project I utilise unsupervised time-series clustering to group counties of the Republic of Ireland based on the similarity of their Covid-19 infection rates. I then compare the composition of each cluster with publicly available socio-economic, demographic, and environmental data. Finally, I report if a relationship exists between clusters of counties and external societal factors.

1.2 Covid-19 in Ireland

The island of Ireland is composed of two political entities: Republic of Ireland, which contains traditional 26 counties, and Northern Ireland, composed of six traditional counties. The population of the Republic as recorded in the most recent census is 4.75 million [17]. The most populous county by far with 1.35

million people is County Dublin in the east of the country, location of the capital city Dublin. Northern Ireland has a population of approximately 1.881 million [58].

The first confirmed case of Covid-19 on the island of Ireland was diagnosed in a woman on February 27th, 2020 [12]. She had travelled to Northern Ireland from Italy, the earliest European hotspot of the disease. Two days later, the first case was diagnosed in the Republic of Ireland in a school student who had also recently arrived in the country from Italy [56]. On March 12th, the Irish government launched an action plan to fight the spread of the Covid-19 by shutting educational facilities and implementing restrictions on the number of people permitted at indoor and outdoor gatherings [42]. These restrictions did not curtail the spread of the disease and by March 22nd cases of Covid-19 had been confirmed in every county of the Republic of Ireland, prompting the government to impose a full lockdown, limiting all non-essential journeys among the population [56]. Restrictions were gradually eased over the following months as the daily number of cases decreased [32]. However, a second wave of the virus swept the country from June 2020 [23] onwards which forced the Irish government to reimpose a general lockdown with Level 5 Restrictions. An easing of those restrictions in the weeks leading up to Christmas preceded an even more aggressive third wave of the virus [46] with the daily number of cases reaching a peak in January 2021. Vaccinations for Covid-19 began to be administered to the population in December 2020, with the vaccination program currently in progress. The Covid-19 pandemic led to unprecedented disruption to Irish public life, triggered an economic recession [7], put massive strain on the nation's healthcare system [52], and, by time of writing, caused the deaths of 5112 people [73].

1.3 Motivation

In understanding the diffusion of the Covid-19 virus (or other infectious diseases) it may be helpful to government, policymakers, and epidemiologists to know if particular regions of the country display similar patterns of infection and if there are related societal factors which contribute to this spread. My main motivation in this report is to explore the possibility of using unsupervised machine learning, specifically time-series clustering, to identify clusters of counties with similar Covid-19 infection patterns before investigating if there are socio-economic, demographic, or environmental variables which correlate with these clusters.

To my knowledge no other report has been published that investigates the effectiveness of time-series clustering at modelling the spread of Covid-19 throughout the island of Ireland. I hope by conducting this study that I can contribute both to a greater understanding of the nature of the pandemic and to the usefulness of time-series clustering in modelling infectious diseases.

1.4 Research Questions

By conducting this study, I intend to answer the following questions:

1. Can time-series clustering be utilised to increase knowledge of the spread of Covid-19 in Ireland?

2. What does time-series clustering reveal about the spread of Covid-19 in Ireland?

3. What clusters of counties in Ireland experienced the spread of Covid-19 in similar ways?

4. Are there independent socio-economic, demographic, and environmental factors which correlate with the spread of Covid-19 in Ireland?

Chapter 2

Background and Related Work

In this section I provide an overview of existing literature relevant to this project. It includes descriptions of concepts which are relevant to this project.

2.1 Time-Series Clustering

2.1.1 Clustering

Clustering is a machine learning technique that involves placing similar unlabelled data into related or homogenous clusters. Data within a cluster should have maximum similarity to other objects in that cluster while having maximum dissimilarity with objects in other clusters.[72] Grouping unlabelled data objects in this way facilitates the discovery of structure and patterns in the data, which may provide insights for those analysing the data.

2.1.2 Time-Series

A time-series is a sequence of continuous, real-valued data points [9] ordered by chronological intervals. Time-series can be used to represent data variables that

change in value as a function of time [3], e.g. stock prices [2], political opinion poll numbers [57]. A time-series object included in the dataset of this project is plotted in Figure 2.1. In this report, the number of confirmed cases of Covid-19 recorded in each county of the Republic of Ireland will be represented as a time-series.

2.1.3 Clustering Applied to Time-Series

While each time-series consists of multiple datapoints, it can still be considered as a single object [61], thus it is possible to apply clustering to a time-series dataset.

A detailed decade review of time-series clustering is provided in [3].

In this project, I will employ whole time-series clustering, i.e. clustering of a dataset of time-series with respect to their similarity [3]. There are five major components in whole time-series clustering:

1. Time-series representation/dimensionality reduction:

   As time-series are high-dimensional [43] (for example, a time-series representing Covid infections for one Irish county will contain a datapoint for each day since the start of the pandemic, approximately 16 months ago), dimensionality reduction is usually performed on the dataset [24][48]. The representation of the time-series is transformed to a reduced dimensionality – if two time-series displayed similar shape in higher dimensions then they will also display similar shape in the transformation space[3]. Examples of dimensionality reduction methods include Discrete Fourier Transfer [5][26] and Discrete Wavelet Transform [19] . For this project, the time-series to be clustered are short enough that dimensionality reduction is unnecessary.

2. Similarity/dissimilarity measures in time-series clustering.

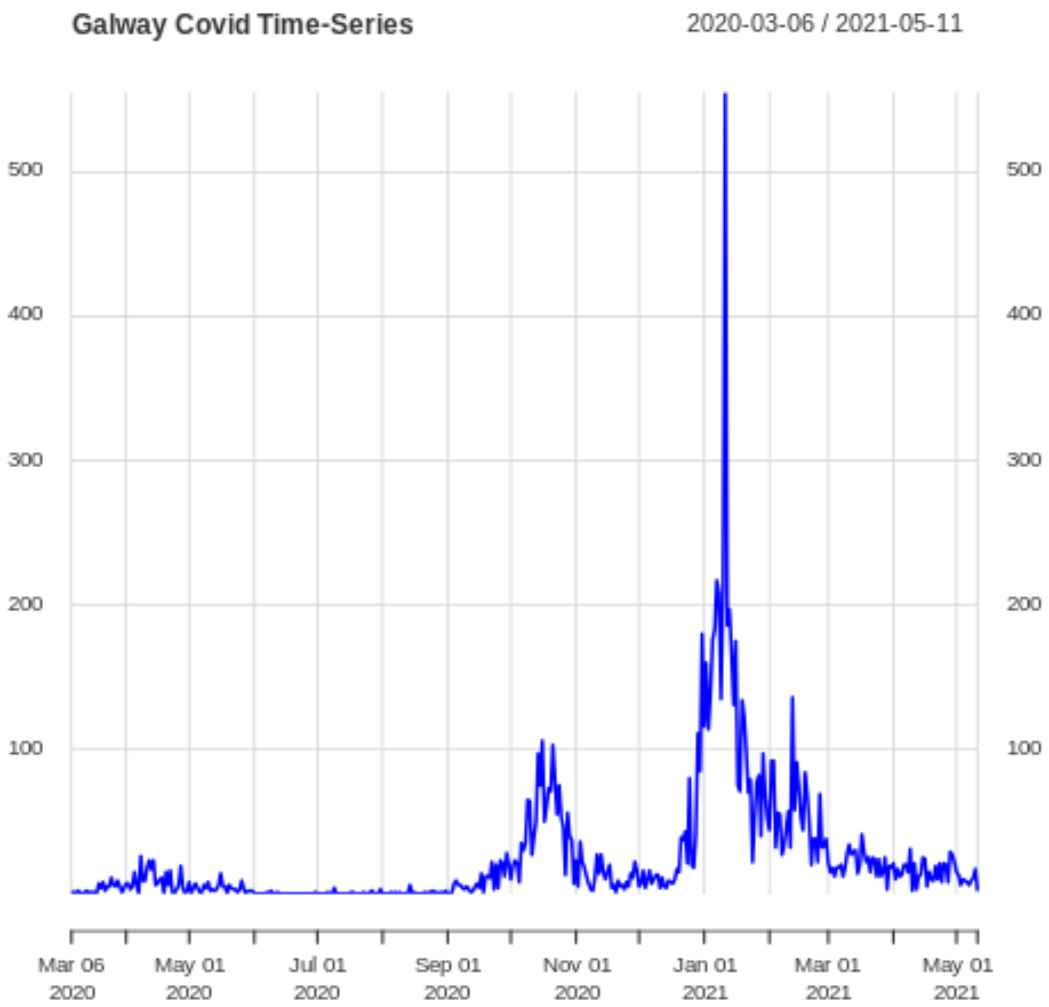

Figure 2.1: Galway County Covid Infection Time-Series

Similarity in time-series clustering is measured by distance, which is calculated approximately (unlike in traditional clustering, where distance is exactly measured) [3]. Choice of similarity measure depends on the characteristic of the time-series. Where the time-series being compared have correlating time intervals, similarity in time can be measured using, for example, Euclidean distance [26]. When time-series have asynchronised time intervals, similarity in shape is usually measured. Dynamic Time Warping is a common method for measuring similarity/dissimilarity in shape [22].

3. Time-series cluster prototypes

   An important factor in cluster quality is the accuracy of the cluster prototype (the object/time-series that best represents the cluster). Clustering algorithms like k-means require a good quality prototype in order to converge [3]. Three main approaches outlined in the survey are defining the prototype as the medoid sequence of the set [41], defining the prototype as the average sequence of the set [4][34], and using local search to identify the prototype [36].

4. Clustering Algorithms

   The two types of algorithm most used for time-series clustering are hierarchical clustering and partitional clustering [3]. Hierarchical clustering [41] makes a hierarchy of clusters using either an agglomerative approach – starting with every object as its own cluster and gradually merging them- or a divisive approach, where one begins with one large cluster and splits them in a divisive manner. Hierarchical clustering is computationally expensive but it does have the advantage of not requiring the final number of clusters to be pre-defined.

   In partitional clustering, the number of clusters the dataset will be divided

into must be pre-defined, for example the value of the k parameter in the k-means algorithm [51]. This can be a challenge in real-world datasets where prior knowledge is not available. However, partitional algorithms tend to be faster than the hierarchical type [13].

5. Evaluation methods for clustering

   Evaluation of the quality of clusters is difficult in the absence of data labels. Definition of clusters is subjective. When a ground truth is available, external index measures such as Rand Index [62] and purity [75] can be used to evaluate quality. When the data labels are unknown, internal index measures such as Calinski-Harabasz and Silhouette Index [3] and Sum of Squared Errors [49] can be used to assess the quality of the identified clusters. As there does not exist ground truth cluster values for the dataset in this project, a selection of internal cluster validity indices are deployed to evaluate clusters generated.

2.2 Time-Series Clustering for Covid-19 Analysis

In [74], the authors deploy a hierarchical approach to cluster countries based on similarity of the shape of the time-series representing their Covid-19 infections over the period January to April 2020. Country population and area are included as factors. The main challenge the authors faced was comparing time-series of countries of different populations and densities, and time-series with different (asynchronous) beginning and end points. Utilising shape-based similarity can help overcome this challenge although the authors claim to have success using Euclidean distance for the distance measure as opposed to DTW. This study was

conducted soon after the beginning of the pandemic and suffered from lack of data.

In [53] the authors performed clustering of seven countries based on their number of Covid infections. By comparing the time-series of each nation's infection rate and rescaling these values to account for differing population size, the authors were able to use fuzzy c-means clustering to group the countries. A membership matrix was deployed to determine the probability that a country was a member of any one cluster. Again, the study was conducted early in the pandemic and only assessed the numbers of a handful of countries and so was limited by lack of data.

In [8] the authors deploy elements of graph theory, specifically hierarchical trees and minimum spanning trees, to cluster nearly two hundred countries based on the similarity of the time-series representing their new Covid cases. Similarity of trends is calculated with Pearson's correlation coefficient. The authors were able to identify three main clusters of countries that exhibit similar Covid infection time-series. The conclusion contains some soft speculation about the broader societal factors which contribute to typical Covid spread among countries in each cluster.

In [39], the authors utilise clustering to group countries based on their time-series of new cases of Covid, and their Covid death rate time-series. After factoring in the offset representing the average number of days between Covid infection and Covid death, the authors calculated an anomaly score for each country which indicated if their Covid death rate was above or below the mean. This anomaly score is used to indicate how effective a country was at preventing the deaths of Covid patients. The authors deployed hierarchical clustering and k-means to form the clusters. To determine the appropriate number of pre-defined clusters, the authors used six different methods including Silhouette Score and Dunn Index.

In [54] the paper's authors analyse, but do not cluster, the Covid infection time-series of nine countries. They observed a power-law growth in the cumulative number of Covid cases in each country. The authors quantified the similarity between infection curves by calculating Distance Correlation between the countries of the study. Some prescriptive measures, such as increased social distancing, are suggested by the authors to flatten the infection. However, they do not explore the correlating societal factors which may contribute to Covid spread.

In [64], the states of the USA are clustered based on the similarity of their time-series of Covid-19 infections and deaths. The authors deploy a parametric metric based on Dynamic Time Warping to measure similarity of time-series. This approach proves useful as the states of the USA experienced their first cases at different times – DTW facilitates comparison despite this lack of synchronicity of time-series. A hierarchical approach is used to identify clusters, with Cophenetic Correlation Coefficient used to analyse the accuracy of those clusters. A total of nine clusters were identified – the Calinski-Harabascz criterion was used to obtain the optimal number of clusters. The authors also analysed Logistic, Gompertz and SIR(Susceptible-Infected-Recovered) mathematical models to allow for modelling and prediction of Covid time-series in different states.

The paper [55] also involves time-series clustering of Covid data in the United States, this time at county-level rather than state-level. The authors include an explanatory model of the clusters by correlating them with a selection of demographic, socio-economic and political variables. The authors conclude that exogenous variables such as the population density of a county, the age profile of its citizens, and even the political allegiance of its Governor have a correlation with the rate of infection the county experienced. To carry out their study, the author deployed k-means clustering, Euclidean distance as a similarity metric, and they rescaled the seven-day moving average of Covid infections so that the

value fell between 0 and 1.

The authors of [37] employed a different approach for using time-series clustering for Covid-19 analysis. Instead of clustering time-series representing Covid cases and deaths, the authors cluster the population of an area based on the amount of time people spent at home after the implementation of a post-pandemic lockdown in a particular urban area, specifically Atlanta, USA. The time-series for the study were extracted from mobility records available from SafeGraph[66], a data company that aggregates anonymised location data. By identifying the part of Atlanta where each cluster was located, the authors were able to compare those clusters with demographic and socio-economic data obtained from the U.S. Census Bureau. Their conclusions were that people from a lower income background were less likely to stay at home during the pandemic (likely because of the nature of their work) and thus are more likely to be exposed to infection. The authors utilised k-means to identify clusters with Euclidean distance as the chosen similarity metric.

2.3 Other Clustering Approaches for Covid-19 Analysis

In [45] the author uses agglomerative hierarchical clustering with Euclidean distance to group the states and union territories of India based on the similarity of their Covid infection, recovery, and death numbers. Distance between clusters is evaluated with analysis of variance (ANOVA) technique, while the elbow method was employed to identify the optimal number of clusters (in this case, six). The paper was published early in the pandemic and thus had access to limited Covid data.

In [16] the authors clustered 155 countries first by eight demographic, eco-

nomic and health related metrics, including GDP, healthcare coverage and the proportion of population that is male. These variables were chosen because of their connection with Covid-19 infections. The k-means clustering algorithm was deployed with the optimal number of clusters (six) being identified through use of the elbow method combined with prior knowledge used to assess if countries fell into expected clusters. Distance between clusters was evaluated with ANOVA. When the clusters were identified, the authors compared them to recorded Covid infection, death, and Covid death rate data of the countries within each cluster. The model identified a clear correlation between the selected external variables and the Covid infection rate of countries within each cluster. However, the model did not group countries well according to number of deaths or fatality rate. One observation made was that the cluster with largest infection rate also had the best healthcare coverage, suggesting that recorded infections in a country are strongly related to the ability of that country to mass-test its population.

In a manner similar to [16], the authors of [21] perform unsupervised k-means clustering on a dataset of 187 countries, based on 15 different national variables. They then compare the composition of each cluster with the Covid infection rates of that cluster's countries to detect if a correlation exists. After performing principle component analysis on the country dataset, the authors demonstrate that 8 of the 15 original variables account for 99% of the cumulative variance. This allows for a significant reduction in dimensionality of the dataset to be clustered. The elbow method is deployed to identify an optimal number of clusters. The paper does not include much technical detail of how the identified clusters correlate with the Covid infection rates of their constituent countries.

In [63] the authors utilise unsupervised k-means to cluster 79 countries based on 18 socio-economic and health-related variables. Then, using Pearson's correlation coefficient, the correlation, negative or positive, is calculated between

the variables with each cluster's Covid infection and death rates. Through these means, the authors highlight environmental factors which correlate strongly with Covid-19. According to this study respiratory conditions like asthma have a high correlation with Covid death while smoking correlates negatively with Covid infection. High health expenditure correlates positively increased Covid infection rates which again suggests Covid detection is related to the scale of a nation's testing programme.

2.4 Covid-19 Analysis in the Irish Context

The following papers study the spread of Covid in Ireland.

In [6] the authors use least squared regression to compare the Covid infection time-series for counties that border Northern Ireland with those of the rest of the state. They included in their analysis other societal factors such as population density of the region, age, and prevalence of circulatory diseases. The results of the study were that border counties experienced higher levels of Covid infection than the rest of the state. The authors conclude that increased number of cases in border counties is a result of the higher level of Covid infection in Northern Ireland, itself a consequence of a delayed Covid containment policy implemented by the UK.

In [27] the authors analyse the connection between demographic factors and the number of Covid cases confirmed within the electoral wards of county Dublin. Kernel PCA is performed to extract features from census data. These features are then used to cluster the electoral wards by means of an unsupervised neural network method. Both Silhouette index and Davies-Bouldin index are deployed to validate the final choice of seven clusters. After the clusters of electoral wards are identified the authors compare them with Covid infection rates in those ar-

eas. The results of the study indicate an association between high numbers of Covid cases and areas with low proportion of elderly people, a high percentage of employment, and a high percentage of private rent.

2.5 Research Gap Addressed

While time-series clustering has been performed on Covid-19 infection rates before, it has mainly been applied at international level for comparison of different nations [74], [53], [8], [39].

In researching the background literature for this project, I encountered two published reports of time-series clustering for Covid-19 data within one country, both involving the USA [64],[55].

Therefore, the application of time-series clustering of Covid-19 data at intra-national level is still underexplored.

While cluster analysis [27] and time-series analysis [6] have been applied within the Irish context, there does not appear to exist a study involving the nationwide application of time-series clustering within the Republic of Ireland.

Therefore, the main research gap I wish to address is: What insights can time-series clustering of Covid-19 provide when applied to Ireland?

Also, most of the works involving time-series clustering of Covid-19 data were published relatively early in the pandemic, and therefore were limited by lack of data. This project will involve analyse of Covid-19 time-series for the first 14 months of the pandemic, and therefore may provide knowledge of the different dynamics of the second and third waves of the virus.

Chapter 3

Methodology

3.1 Choice of Software Library

In this project, for the performance of clustering on the time-series dataset, I employed the dtwclust library[67] in R. The main advantage of using dtwclust is that it provides a common infrastructure for testing and comparing different clustering approaches to a common time-series dataset. Dtwclust facilitates the implementation of different configurations of the two most common types of time-series clustering, i.e. partitional clustering and hierarchical clustering, which are the methods I utilise for experiments in this project (other significant but less common clustering methods include density-based, grid-based, and model-based clustering, which are detailed in the most recent major survey of time-series clustering [3]).

A variety of distance measures, including Euclidean distance and dynamic time warping, can be defined using the dtwclust library, and by altering such parameters the user can implement the equivalent of such popular clustering algorithms as k-means and k-medoids.

The dtwclust package also enables the comparison of cluster validity indices

(CVIs) of different clustering algorithm configurations, providing the user with clear information related to comparative cluster quality.

Clustering algorithms, distance metrics, and cluster validity indices utilised in this project will be detailed in later sections.

3.2 Datasets

For this project I obtain publicly available data from the following sources:

- Republic of Ireland Covid-19 Data:

  The Irish government publishes Covid data on the Government of Ireland website [31]. Datasets, including the daily recorded number of Covid infections and Covid-related deaths, are available in a variety of formats, including csv which is the format used in the project.

  The daily number of recorded Covid infections and Covid-related deaths are listed for each of the 26 counties of the Republic of Ireland. A cyber-attack [30] committed on the I.T. system of Ireland's Health Service Executive on May 14th 2021 disrupted daily updates to publicly available Covid-19 datasets. Because of this, the Covid data detailing infections in the Republic of Ireland terminates on May 11th 2021.

- Northern Ireland Covid-19 Data:

  The British government makes available Covid data, including daily infections and deaths, for Northern Ireland at both regional level and local authority level [33].While local authority level would provide finer-grained detail about the location of Covid-19 infections and deaths within Northern Ireland, that data is only available dating back to August 2020. As such, it

does not provide detail of how Northern Irish local authorities experienced the first wave of the virus and thus cannot be compared with counties in the Republic of Ireland. Because of this, Northern Ireland Covid infection data is represented as one time-series, taking the total number of time-series clustered in this project to 27.

- Demographic Data - Republic of Ireland :

  Data relating to societal and demographic variables such as income and age levels was downloaded from the Census 2016 Open Data Site[29].

- Demographic Data - Northern Ireland :

  Northern Irish socio-economic and demographic datasets are provided by the Northern Ireland Statistics and Research Agency and are available from the website of Open Data NI [59].

3.3 Data Pre-Processing

To create time-series for this project, csv files containing Covid-19 data for Republic of Ireland and Northern Ireland were downloaded from their original sources mentioned in the previous section and merged using Microsoft Excel. The time-series for this project are of equal length in time, beginning on March 6th 2020 (the first day available in the Northern Irish Covid dataset) and ending on May 11th 2021 (when the previously mentioned cyber attack disrupted the I.T. systems of the Irish health service). This creates a dataset with dimensions 27 (counties) x 432 (days of recorded Covid infections).

The csv file containing Covid data is imported into R as a data frame. Before the data can be used with a clustering algorithm, it must be converted to an xts

time-series object. Xts (eXtensible Time Series [**?**]) is an extension or child of the zoo library in R. A time-series can be created using xts by selecting a column from the data frame which is of a time-based class. In this dataset that column is 'Date'. In an xts object, this column can be defined as a date index. The remaining columns, consisting of a matrix of data representing Covid infections for counties in the dataset, are now ordered by the date index.

3.3.1 Z-Normalisation

As mentioned in a previous section, there is significant variation in the population of Irish counties, and so the number of Covid infections recorded in each county has also varied greatly. For example, county Dublin experienced its worst day of Covid on January 3rd 2021, with 3654 infections recorded. Meanwhile, Ireland's least populous county Leitrim recorded 43 infections on its worst day at the peak of the third wave of the virus. To meaningfully cluster counties using typical distance measures like Euclidean distance, their Covid infection time-series must be normalised to a common scale.

To normalise each county's time-series, I convert raw Covid infection numbers to their standard score or z-score. The z-score describes how many standard deviations a measured raw data point is above or below the mean value of the collection or population of data in which the data point is contained. In the context of this project, the target data point is infections recorded in a county for a particular day and the data collection in which it is contained is the complete time-series. The z-score scales values to a range of between -1 and +1, negative if the original data point is below the mean, and positive if it is above. The formula for the z-score is shown in Figure 3.1 where x = the observed data point, μ = the mean value for the collection/population, and σ = the standard deviation of the

$$Z = \frac{x - \mu}{\sigma}$$

Figure 3.1: Z-normalisation Formula

collection/population.

In this project I perform z-normalisation on the dataset using the zscore function in the R language. The zscore function performs normalisation on each row (time-series) of the dataset discretely, thus allowing each time-series to maintain its original shape.

3.4 Clustering

For this project, I conduct clustering experiments using both hierarchical and partitional clustering methods.

3.4.1 Hierarchical Clustering

In hierarchical clustering, the algorithm either treats each object as a single cluster and iteratively merges them into larger clusters (agglomerative or 'bottom-up' clustering) or begins with one large cluster and gradually splits it up until each object forms its own cluster (divisive or 'top-down' clustering)[3]. The dtwclust library utilises the hclust function native to R which is an agglomerative approach.

Hierarchical clustering has several advantages. Firstly, it can display it's results in the form of dendrogram (for example, see Figure 4.2) which can provide a clear visualisation of the generated clusters. Also, the number of clusters to be created does not need to be pre-defined before operation, unlike partitional clustering. Finally, hierarchical clustering is deterministic, always returning the same

results depending on the distance metric/linkage control used, and thus avoids the ambiguity involved in methods with a stochastic element, e.g. partitional clustering [67].

A major disadvantage of hierarchical clustering is computational complexity[67][3]. Hierarchical clustering requires the calculation of a distance matrix for the entire dataset – with N being the number of objects in the dataset, time and memory complexity is O(N2)[67]. While this makes hierarchical clustering unsuited to large datasets, it is not impractical for use in this project.

3.4.1.1 Linkage Controls

In agglomerative hierarchical clustering, the merging of clusters is based on a dissimilarity metric known as a linkage. The dtwclust library includes the hclust implementation of hierarchical clustering with which several different linkage methods can be deployed. In this project I perform hierarchical clustering with the following linkage methods:

- Average Linkage :

  In average linkage[69], the distance between two clusters is the average distance between all pairs of objects in both clusters.

- Ward Linkage:

  The objective of Ward's method for linkage is to minimise within-cluster variance[71]. It achieves this by, at each step, merging clusters with minimum cluster distance [20]. In dtwclust, the original Ward's criterion is implemented as "ward.d2".

- Centroid:

  In centroid linkage, the inter-cluster distance is based on the Euclidean distance between the centroid or mean values of each cluster. [69]

- Single

  In Single linkage, distance is calculated as the minimum distance between any pair objects in opposing clusters[68].

3.4.2 Partitional Clustering

In partitional clustering, each object is assigned to one of a pre-defined (k) number of clusters. The objective of the method is to minimise the intra-cluster distance, by minimising the total distance between all data points in the cluster from the centre point (centroid) or prototype of the cluster, while also maximising inter-cluster distance [67][3].

The process begins by randomly choosing k number of data points in the set to act as centroids. After the distance between all data points is calculated, each object is assigned to the cluster around its nearest centroid/prototype. Over multiple iterations, the centroids of each cluster are updated using a prototyping function and all data points are re-assigned to the cluster of the nearest updated centroid [67]. This process is repeated until either a fixed number of iterations are completed or there is no further change in the composition of the clusters.

A disadvantage of partitional clustering is that it is partially stochastic in nature because of the random initialisation of centroids. As such, different executions of a partitional clustering method will return different results, with convergence to local optima possible. Another disadvantage of partitional clustering

is that the k number of clusters must be defined in advance. Techniques for identifying the optimal number of clusters exist, including the Silhouette method [65] (detailed in a later section). However, partitional clustering is faster than hierarchical clustering when applied to time-series data.

The most popular partitional clustering algorithm is k-means [50] where the centroid of each cluster is the mean value of its component objects. In the k-medoids (PAM – Partition Around Medoids) algorithm the centroid or medoid is the object from the cluster that is closest to the centre of the cluster. Both k-means and k-medoids can be implemented using the dtwclust library, and I deploy both approaches in clustering experiments in this project.

3.4.3 Distance/Dissimilarity Metric

To cluster time-series it is necessary to quantify the similarity or dissimilarity between each time-series object. This similarity/dissimilarity can be measured using a distance metric. In this project, I deploy two distance metrics for comparing and clustering time series, Euclidean distance [26] and Dynamic Time Warping [22].

3.4.3.1 Euclidean Distance

Euclidean distance is calculated as the square root of the sum of the square differences between two vectors [20](in the context of comparing time-series, the difference in value between series at each time step).

As the time-series in this project are univariate time-series of equal length, Euclidean distance is a suitable distance metric for clustering.

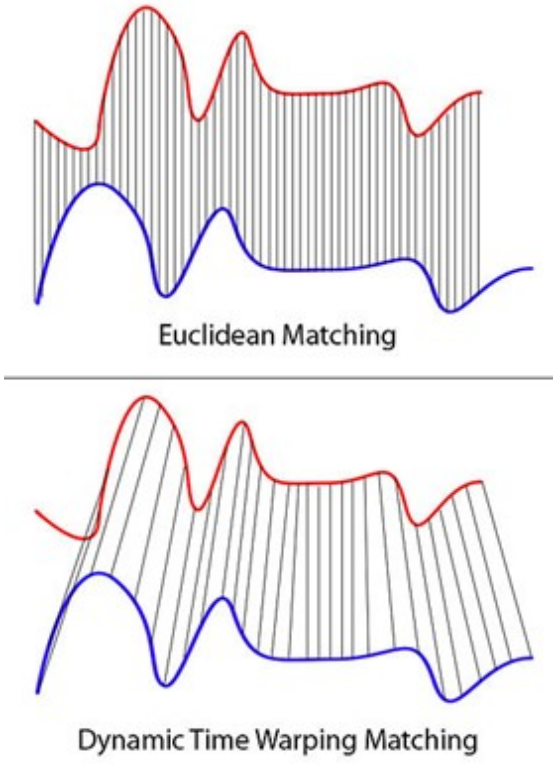

Figure 3.2: Dynamic Time Warping (image:Creative Commons)

3.4.3.2 Dynamic Time Warping (DTW)

Dynamic time warping is type of distance metric algorithm which can be used to compress or stretch a series to make it resemble another series of different length [28]. Performing dynamic time warping on time-series can facilitate the comparison of multiple series based on shape rather than purely on raw distance in space. This is useful when investigating the similarity of behaviour of time-series even if they vary in length.

Figure 3.2 presents an intuitive impression of how dtw operates. The time-series pictured are of different length. Calculating Euclidean distance between corresponding points in time will not allow an accurate comparison of the shape/behaviour of the two series. Dynamic time warping, on the other hand, seeks an optimal

warping path between the peaks and troughs of both series, which when used in clustering can help identify series of similar shape, independent of their position in time [3].

Although dynamic time warping is more complex and computationally expensive than Euclidean distance, in these experiments I include clustering configurations that utilise dtw, for the purposes of comparison and verification of performance.

3.4.4 Cluster Validity Indices

The application of time-series clustering to the dataset used in this project is an unsupervised process. There does not exist a ground truth with which to verify the composition of the clusters generated. Evaluating the quality of the clusters, therefore, is somewhat subjective. However, the dtwclust library includes a variety of cluster validation indices (CVIs) which can be used to compare and evaluate the results of any clustering configurations deployed.

In this project, I utilise four CVIs that can be applied to gauge the "internal" quality of clusters. Internal CVIs evaluate cluster purity in the absence of predefined cluster labels, whereas "external" CVIs require the presence of a ground truth.

Below is a brief description of the four internal CVIs used in these experiments, based on a recent review of cluster validity indices [10]. The indices operate by estimating both intra-cluster distance and inter-cluster distance, and then calculate a quality score by either summing the estimates or performing division of the two values. Cluster quality is indicated by either how low or how high the score is, depending on the index deployed.

3.4.4.1 Silhouette index (Sil)

The silhouette index [65] is the summation of the intra-cluster distance, estimated by the distance between all objects in the cluster, and the inter-cluster distance, which is based on nearest neighbour distance. A higher score indicates better cluster quality.

3.4.4.2 Calinski-Harabasz index (CH)

In Calinski-Harabasz [14], intra-cluster distance is based on the distance from all objects of the cluster to the cluster centroid. Inter-cluster distance is estimated by the distance between cluster centroids and the global centroid (the centroid that uses the entire dataset). Score is calculated by a division of the two distance values. A higher indicates better cluster quality.

3.4.4.3 Dunn index

Dunn index [25] involves using nearest neighbour distance to estimate intra-cluster distance, and maximum cluster diameter to estimate inter-cluster distance. The final score, of which a higher value indicates better cluster purity, is calculated by a division of the two distance estimates.

3.4.4.4 COP index

COP index [35] estimates intra-cluster distance based on the distance from cluster objects to the centroid. Inter-cluster distance is based on furthest neighbour distance. Final score is calculated by division of the two distances. A lower score indicates better cluster quality.

3.4.5 Cluster Evaluation

To identify the best performing configuration of clustering algorithm, the four CVIs are recorded and majority vote is applied. The configuration with the highest score in most of the indices is chosen to perform final clustering on the dataset.

3.4.6 Defining Number of Clusters

As previously stated, some clustering algorithms such as k-means and k-medoids require the pre-definition of final cluster numbers before they can implemented. The silhouette method [65] (as detailed in the previous section) is a widely used approach for defining the optimal number of clusters.

Utilising the fviz_nbclust function in R (from the factoextra package [40], an optimal number of final clusters can be obtained for clustering algorithms utilising Euclidean distance including k-means, PAM (k-medoids) and hierarchical clustering.

After applying the silhouette method to the dataset (an upper limit of 15 possible clusters was set to prevent the formation of trivially small clusters), the following optimal number of clusters for selected Euclidean distance clustering algorithms were obtained:

K-Means: 11 Clusters, see Figure 3.3

K-Medoids: 11 Clusters, see Figure 3.4

Hierarchical Clustering: 9 Clusters, see Figure 3.5

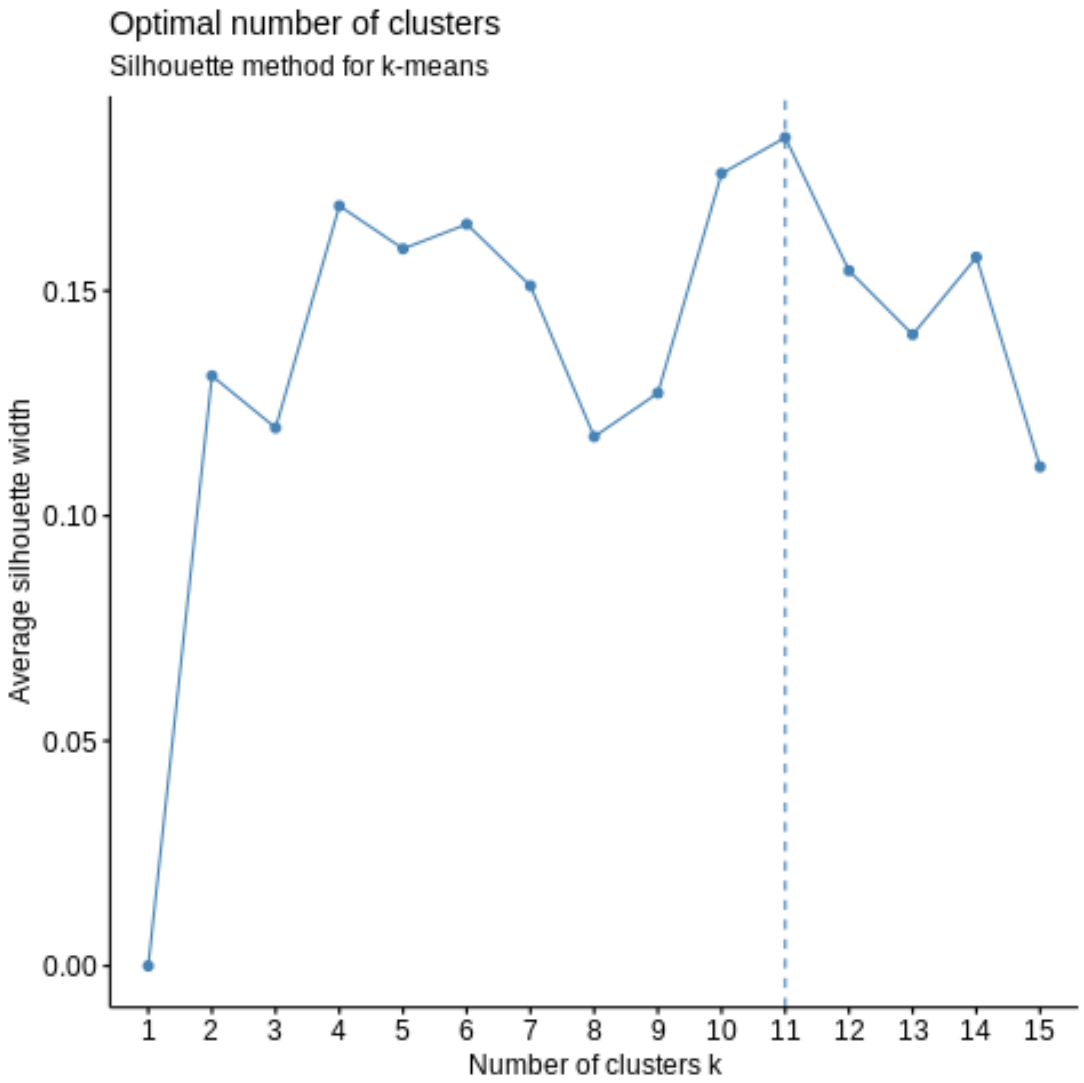

Figure 3.3: The optimal k value for k-means using Silhouette method is 11 clusters

3.4.7 Final Configurations of Clustering Algorithms for Clustering Experiments

3.4.7.1 Partitional Configurations Utilising Euclidean Distance

For Euclidean distance-based partitional configurations with a stochastic element, such as k-means and k-medoids, multiple iterations of clustering were performed

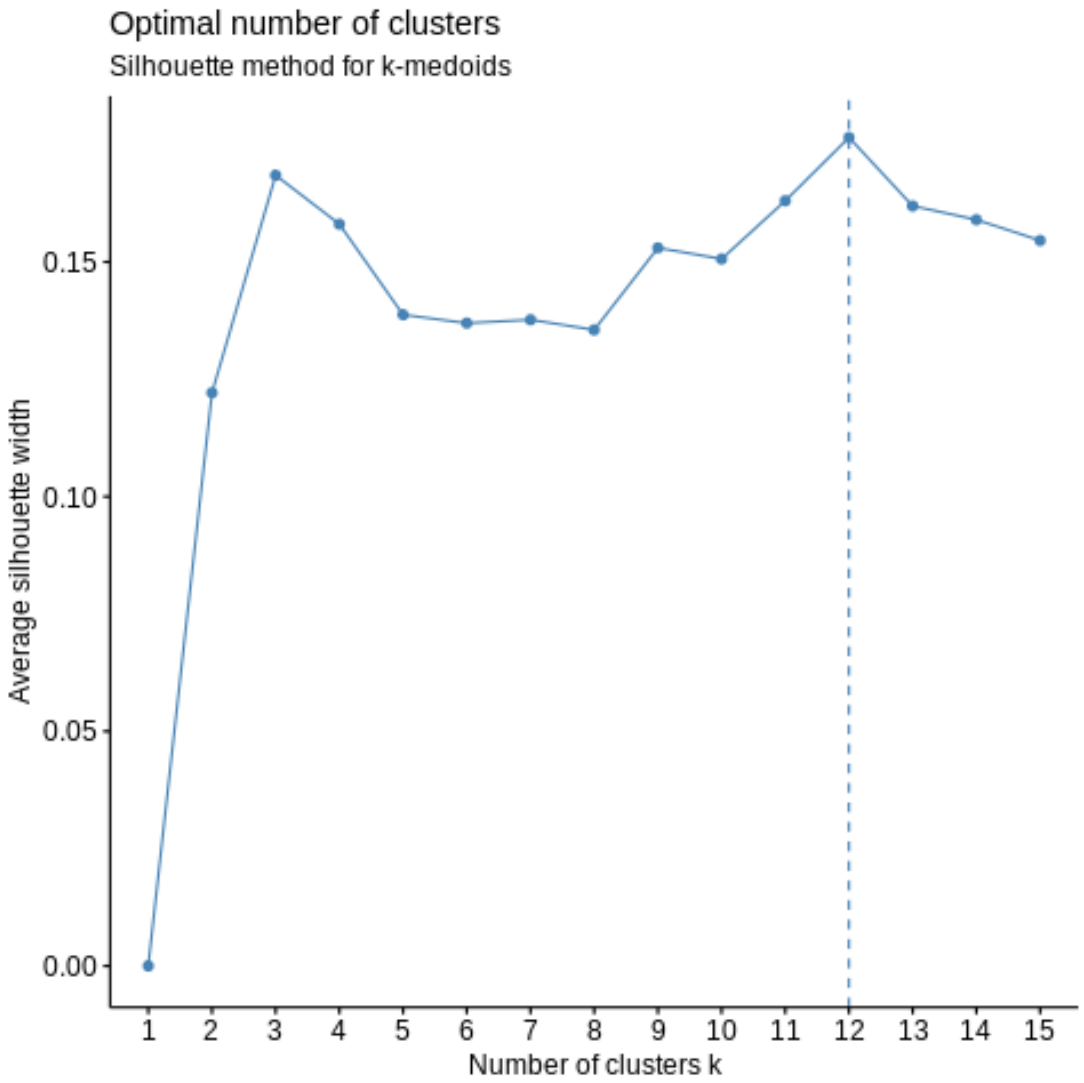

Figure 3.4: The optimal k value for k-medoids using Silhouette method is 12 clusters

on the dataset using the pre-defined number of clusters detailed in the previous section. Each iteration uses a different seed - in this experiment I chose seeds 1 to 20 for the purpose of simplicity. The CVI values for each run were recorded and then averaged to find a representative value for each index. If the average CVI value for a particular configuration compares well with other configurations, the seed of that configuration with the best CVI values is chosen as the final

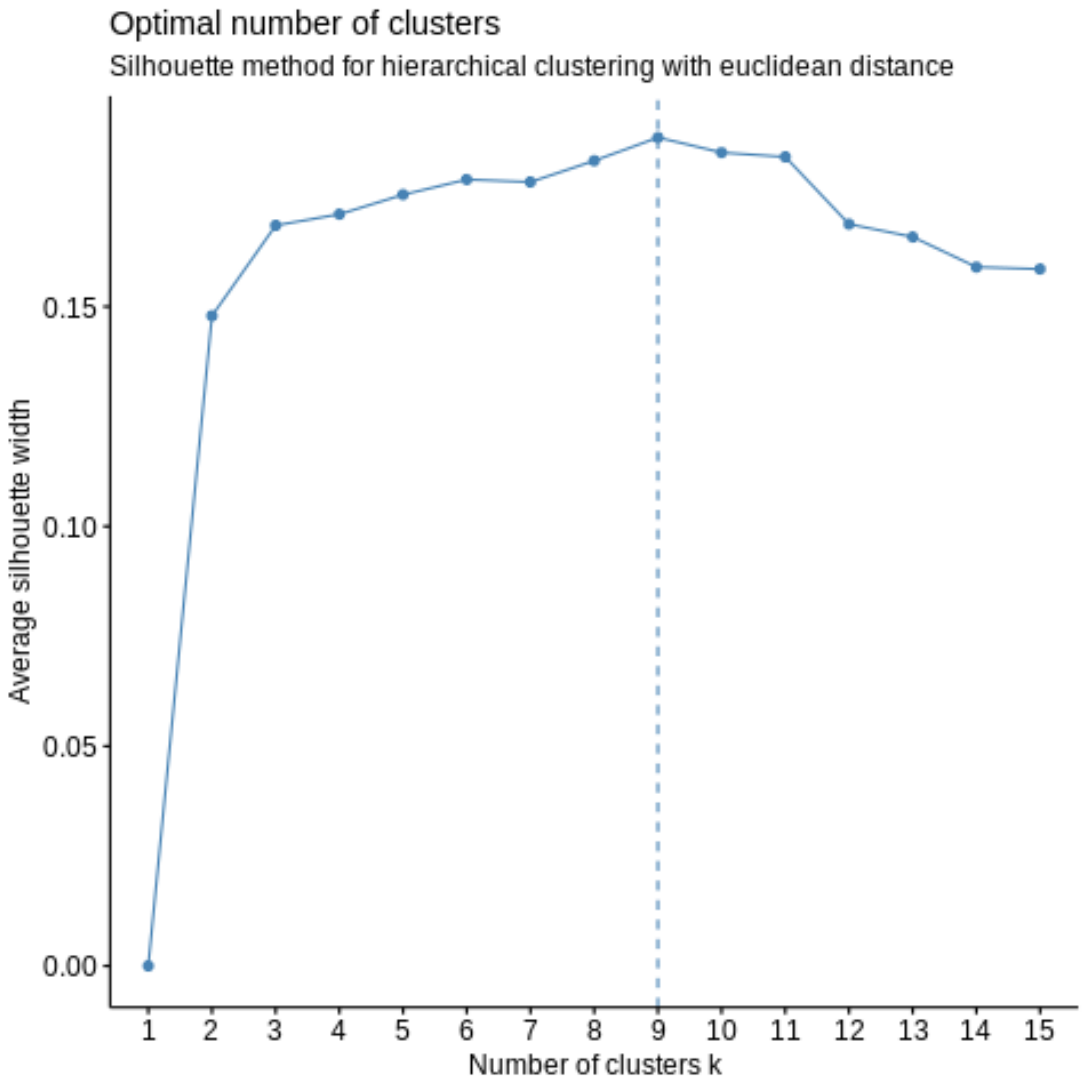

Figure 3.5: The optimal k value for hierarchical clustering using Silhouette method is 9 clusters

clustering configuration.

Configuration #1 is an implementation of k-medoids (PAM - Partitioning Around Medoids), with the centroid of each cluster represented being the object closest to the centre.

Configuration #2 is an implementation of k-means, with the centroid of each cluster being the mean value.

3.4.7.2 Hierarchical Configurations Utilising Euclidean Distance

For deterministic hierarchical configurations, the four cluster validity index values were obtained after performing clustering on the dataset.

Configurations #3 to #6 of this experiment perform hierarchical clustering with Euclidean distance. They utilise average, Ward, centroid, and single linkage respectively.

3.4.7.3 Partitional Clustering Utilising Dynamic Time Warping

For configurations using dynamic time warping (of which the application of the silhouette method did not return a value for optimal number of clusters) the average CVI values for seeds 1 through 20 are obtained for k number of clusters 2 to 15. For the value of k with the best scoring CVIs, if those values compare favourably with CVI scores from other configurations, then the seed with the best cluster index scores is chosen as the final cluster configuration for application on the dataset.

Configuration #7 of this experiment uses dtw_basic distance, an optimised version of dynamic time warping included in dtwclust[67]. This configuration uses DBA (DTW barycenter averaging) [60] to define cluster centroids. DBA is a prototyping function developed for dynamic time warping, which involves using a time-series object as a reference centroid. As in DTW multiple time points from each series object may map to a single point on the centroid series, in DBA the values from these time points are collected into groups and the mean value for each point on the centroid is calculated based on these groups [67].

After performing partitional clustering using dtw_basic distance for twenty seeds, for each cluster number value (k) from 2 to 15, the following average CVI scores were obtained (see Figure 3.6) .

Average CVI Values for DTW Basic

K.Value	Silhouette_Score	CH	Dunn	COP
2	**0.08215967**	**17.06687142**	0.50630371	0.66393067
3	0.04648292	9.28306523	0.45400622	0.61299563
4	0.04809333	6.4497263	0.48759063	0.58004786
5	0.04332572	5.21788742	0.47805752	0.5503818
6	0.03460754	4.17883449	0.47316835	0.52358655
7	0.03870448	3.58902231	0.48321881	0.4965889
8	0.0343068	3.19103722	0.47064897	0.47064718
9	0.03752143	2.82569425	0.49458316	0.44249785
10	0.03571778	2.54249112	0.4899587	0.41822313
11	0.03395086	2.33180809	0.48881022	0.39257111
12	0.03351369	2.1689697	0.50429792	0.36618375
13	0.03677144	1.99096056	0.51461916	0.34159112
14	0.04109594	1.85042678	0.51474637	0.31893011
15	0.03874117	1.72327221	**0.52620303**	**0.2950083**

Figure 3.6: Average CVI scores for Configuration # 7

The best average score for each CVI is highlighted in bold. The configurations involving the generation of two clusters and fifteen clusters respectively attain the best score in two CVIs each. As the configuration with a k value of 2 performs best by far in the silhouette and Calinski-Harabasz indices, and also generates a comparably good score in Dunn index, I have chosen it as the best configuration of dtw_basic for this dataset.

Configuration #8 uses a dtwclust implementation of dynamic time warping distance, dtw_lb, that leverages lower bounding, specifically the lb_improved method of lower bounding [47]. As DTW is so computationally expensive, with distances between all points on time-series pairs requiring calculation, lower bounding can be introduced as a method to limit the number of calculations between points which are excessively distant (and thus certain not to be correct matches in the time-series shape) thus assisting in reducing demands on CPU time. [44]. Configuration #8 also uses DTW barycenter averaging as a centroid function. Dynamic time warping with lower bounds is only suitable for clustering time-series of equal length [67]. The dataset in this project satisfies that condition.

After performing partitional clustering using dtw_lb distance for twenty seeds, for each cluster number value (k) from 2 to 15, the average CVI scores shown in Figure 3.7 were obtained.

The configuration with a k value of 2 attained the best score in the silhouette and Calinski-Harabasz indices. Because it also attains a comparatively high score in Dunn, it is designated the best configuration for dtw with lower bounding.

Average CVI Values for dtw_lb

K.Value	Silhouette_Score	CH	Dunn	COP
2	**0.17460193**	**17.0668701**	0.14990342	1.3994214
3	0.04604803	9.28311699	0.10884583	1.16089172
4	0.03120179	6.44972511	0.12150872	1.04736903
5	-0.01354749	5.2178863	0.12151859	0.97057531
6	-0.04756921	4.17884197	0.11605458	0.91977303
7	-0.05608156	3.58902281	0.12611887	0.87105677
8	-0.08241206	3.1910351	0.11532734	0.820366
9	-0.07238761	2.8256878	0.12642421	0.76809175
10	-0.08512415	2.54249348	0.13264635	0.72258137
11	-0.08252679	2.33180833	0.13821335	0.67479541
12	-0.09564696	2.16897098	0.14161038	0.62778504
13	-0.09135907	1.99096061	0.14979431	0.57983107
14	-0.07486173	1.85042534	0.15841121	0.53887917
15	-0.07057375	1.72327173	**0.16509254**	**0.49554355**

Figure 3.7: Average CVI scores for Configuration # 8

Summary of Clustering Configurations and CVI Values (bold indicates best scoring CVI)

Configuration.ID	Clustering.Type	Distance.Measure	K.Value	Linkage	Centroid	Silhouette	CH.Score	Dunn	COP
1	Partitional (K-Medoids)	Euclidean	12	NA	PAM	0.07184315	2.78188941	0.46667181	0.31207351
2	Partitional (K-Means)	Euclidean	11	NA	Mean	0.07949016	2.2968503	0.45765088	**0.30795052**
3	Hierarchical	Euclidean	9	Average	NA	0.1745238	4.35726823	**0.67698177**	0.37350417
4	Hierarchical	Euclidean	9	Ward	NA	**0.18847315**	3.58857173	0.63068802	0.35851105
5	Hierarchical	Euclidean	9	Centroid	NA	0.01172029	1.24704656	0.52736282	0.46068612
6	Hierarchical	Euclidean	9	Single	NA	0.08762357	3.02567054	0.65374791	0.43031043
7	Partitional (DTW)	DTW Basic	2	NA	DBA	0.08215967	**17.06687142**	0.50630371	0.66393067
8	Partitional (DTW)	DTW LB	2	NA	DBA	0.17460193	17.0668701	0.14990342	1.3994214

Figure 3.8: Summary of clustering configurations and CVI scores

3.4.7.4 Summary of Clustering Configurations and CVI Values

Figure 3.8 provides a summary of each configuration and the score it attained for each cluster validity index. The best score for each CVI is highlighted in bold.

No single configuration achieves a majority of highest CVI scores. Configurations #1 and #2, whose CVI values were averaged over 20 runs, achieve the best (lowest) scores for COP index. However, their scores for other indices range from poor to medium.

Configurations #7 and #8 using dynamic time warping and whose CVI values were also averaged over 20 runs, achieve high scores for Calinski-Harabasz index, but mediocre/poor scores for Dunn and COP indices.

To get a clearer idea of which configuration performs best overall, I have ranked each configuration for all four validity indices (see Figure 3.8) . After summing the CVI ranking scores for all configurations, the configuration with the lowest score is deemed to be the best performer.

The configurations that achieved the best scores overall, based on all cluster validity indices applied to this dataset, are configurations #3 -hierarchical clus-

Ranking of Configurations for each Cluster Validity Indices

Configuration	Silhouette	CH	Dunn	COP	Total
1	7	6	6	2	21
2	6	7	7	1	21
3	3	3	1	4	11
4	1	4	3	3	11
5	8	8	4	6	26
6	4	5	2	5	16
7	5	1	5	7	18
8	2	2	8	8	20

Figure 3.9: Ranking of clustering configurations based on CVI values

tering with Euclidean distance and average linkage - and #4, hierarchical with Euclidean distance and Ward criterion linkage.

After performing clustering on the dataset with configurations #3 and #4, the clusters shown in Figure 3.10 and Figure 3.11 were formed.

Although there are similarities between the composition of both sets of clusters, configuration #4 using Ward's criterion appears to have produced clusters with a more even distribution of size, with fewer clusters with size 1. This will assist in performing one way ANOVA tests in the section of the project and for that reason I designate it as the final clustering configuration. More analysis of the final generated clusters will be provided in the results section of this document.

Clusters created by Hierarchical Clustering with Average Linkage

Cluster.Number	Counties	Cluster.Size
1	Carlow,Kilkenny,Waterford,Wexford	4
2	Cavan,Meath	2
3	Clare,Cork,Dublin,Kerry,Kildare,Limerick,Louth,Monaghan,Tipperary,Wicklow	10
4	Donegal,Leitrim,Sligo	3
5	Galway,Mayo,Roscommon	3
6	Laois,Offaly	2
7	Longford	1
8	Westmeath	1
9	Northern.Ireland	1

Figure 3.10: Clusters generated by hierarchical clustering with average linkage

3.5 Statistical Analysis of Final Clusters

In the previous section, time-series clustering was performed on the dataset. In this section, tests are performed to assess if there exists a statistically significant relationship between selected socio-economic variables and the clusters of counties previously identified.

To assess the statistical significance of demographic variables, one-way ANOVA (Analysis of Variance) is applied. If ANOVA returns a p-value of less than 0.05 when analysing the variance of a particular variable relative to the cluster set, then a post-hoc Tukey test is performed to identify clusters which differ significantly from one another.

Clusters created by Hierarchical Clustering with Ward Linkage

Cluster.Number	Counties	Cluster.Size
1	Carlow,Kilkenny,Waterford,Wexford	4
2	Cavan,Meath	2
3	Clare,Cork,Kerry,Limerick,Louth,Monaghan,Tipperary	7
4	Donegal,Leitrim,Sligo	3
5	Dublin,Kildare,Wicklow	3
6	Galway,Mayo,Roscommon	3
7	Laois,Offaly	2
8	Longford,Westmeath	2
9	Northern.Ireland	1

Figure 3.11: Clusters generated by hierarchical clustering with Ward linkage

3.5.1 ANOVA (Analysis of Variance)

A one-way ANOVA test is a method of examining if a particular common measurement (dependent variable) differs significantly between different groups of a population, or if there is no significant difference between the groups (i.e. the null hypothesis holds) [11] .

In one-way ANOVA, variance is analysed by calculating the between-group variance and then dividing that value by the within-group variance. Between-group variance is calculated by summing the squared differences between the mean value of each group and the overall mean value. This value is then divided by the degrees of freedom (DF). The DF is calculated as the number of observations or groups minus one[11].

Identifying within-group variance begins by calculating the sum of squared differences between each observation value and the mean of its group. This value is then divided by within-group DF.

The f-ratio is the value returned by dividing the between-group variance by within-group variance. If this value is large, it indicates that the differences between groups is particularly larger than differences within groups. In the context of this project, a high f-ratio indicates that a particular demographic variable or measurement differs significantly between at least two clusters of counties, suggesting a possible correlation. From the f-ratio, the p-value is calculated, indicating the probability that the null hypothesis is true. If the p-value is less than the alpha value of 0.05, then it is probable that there is a correlation between the value of the dependent variable and the groups in the experiment.

For one-way ANOVA to operate successfully, some assumptions about the dataset must be satisfied[15]:

1. Independence of observations: the value of one observation within a group cannot be linked to the value of another observation. This condition is true for the dataset in this project.

2. Homogeneity of variance: population of each group must have similar variance. In this project, the Levene Test is R is deployed to assess if a dependent variable satisfies this condition.

3. Normal Distribution: values for each variable are distributed normally around the mean[15]. The R implementation for Shapiro-Wilks test is utilised to verify if a dependent variable is of normal distribution.

3.5.2 Post-hoc Tukey Test

While the ANOVA test reveals if there exists a disparity between groups, it does not identify between which groups the significant difference lies. The post-hoc Tukey test[1], also known Tukey's test of Honestly Significant Difference, can perform this task.

The Tukey test performs comparisons between pairs of group means identified by the ANOVA test and computes the honestly significant difference between groups using Student's q distribution. The TukeyHSD function in R is deployed to perform this function.

3.5.3 Demographic Variables

Demographic Variables

One-way ANOVA and Tukey tests are applied to the following variables, taken from the Irish census of 2016 and the website of the Central Statistics Office, to assess if a correlation exists with the counties clustered in the previous section:

- Percentage of county population in managerial class
- Percentage high professionals
- Percentage low professionals
- Percentage non-manual workers
- Percentage manual workers
- Percentage semi-skilled workers
- Percentage unskilled workers
- Percentage workers of own account
- Percentage of population who are farmers
- Percentage agricultural workers
- Percentage of county population aged 20 - 24

- Percentage aged 25 - 29
- Percentage aged 30 - 34
- Percentage aged 35 - 39
- Percentage aged 40 - 44
- Percentage 45 - 49
- Percentage 50 - 54
- Percentage 55 - 59
- Percentage 60 - 64
- Percentage aged 65+
- Population of county divided by number of Covid cases
- Percentage of workers in manufacturing industry
- Percentage of workers in transport and communications
- Percentage of workers in the building industry

I was unable to find corresponding data in several of these categories for Northern Ireland. For this reason, combined with the fact that Northern Ireland has formed its own singleton cluster in our model, I have not applied one-way ANOVA testing to Northern Ireland in this section.

Chapter 4

Results

4.1 Results of Clustering

After performing hierarchical clustering with Ward linkage on the dataset, nine clusters were generated, as shown in Figure 4.1

What is immediately evident on viewing the clusters is their geographical consistency. Apart from the border counties of Louth and Monaghan, who were grouped together with several southern Munster counties, no cluster contains counties that do not border one another. This suggests that there is a coherence to time-series clustering of Irish counties. It also makes intuitive sense that neighbouring counties will have experienced an infectious disease in similar ways.

4.2 Results of One-way ANOVA Tests

To successfully apply one-way ANOVA tests, the dependent variable must satisfy assumptions that are mentioned in a previous section, namely homogeneity of

Clusters created by Hierarchical Clustering with Ward Linkage

Cluster.Number	Counties	Cluster.Size
1	Carlow,Kilkenny,Waterford,Wexford	4
2	Cavan,Meath	2
3	Clare,Cork,Kerry,Limerick,Louth,Monaghan,Tipperary	7
4	Donegal,Leitrim,Sligo	3
5	Dublin,Kildare,Wicklow	3
6	Galway,Mayo,Roscommon	3
7	Laois,Offaly	2
8	Longford,Westmeath	2
9	Northern.Ireland	1

Figure 4.1: Final clusters generated by hierarchical clustering with Ward linkage

variance and normal distribution. To assess these two qualities, the Shapiro-Wilks test and the Levene test can be applied respectively.

After performing these tests on the demographic variables, the results shown in Figure 4.3 were generated.

Values > 0.05 are highlighted in bold and thus pass the threshold of the test. Any variable that does not pass both the Shapiro-Wilks test and the Levene test does not satisfy the assumptions for one-way ANOVA testing and is discarded.

The one-way ANOVA test is performed on the remaining variables and the results shown in Figure 4.4 are returned.

A p-value < 0.05 (highlighted in bold) suggests that there exists a statistically

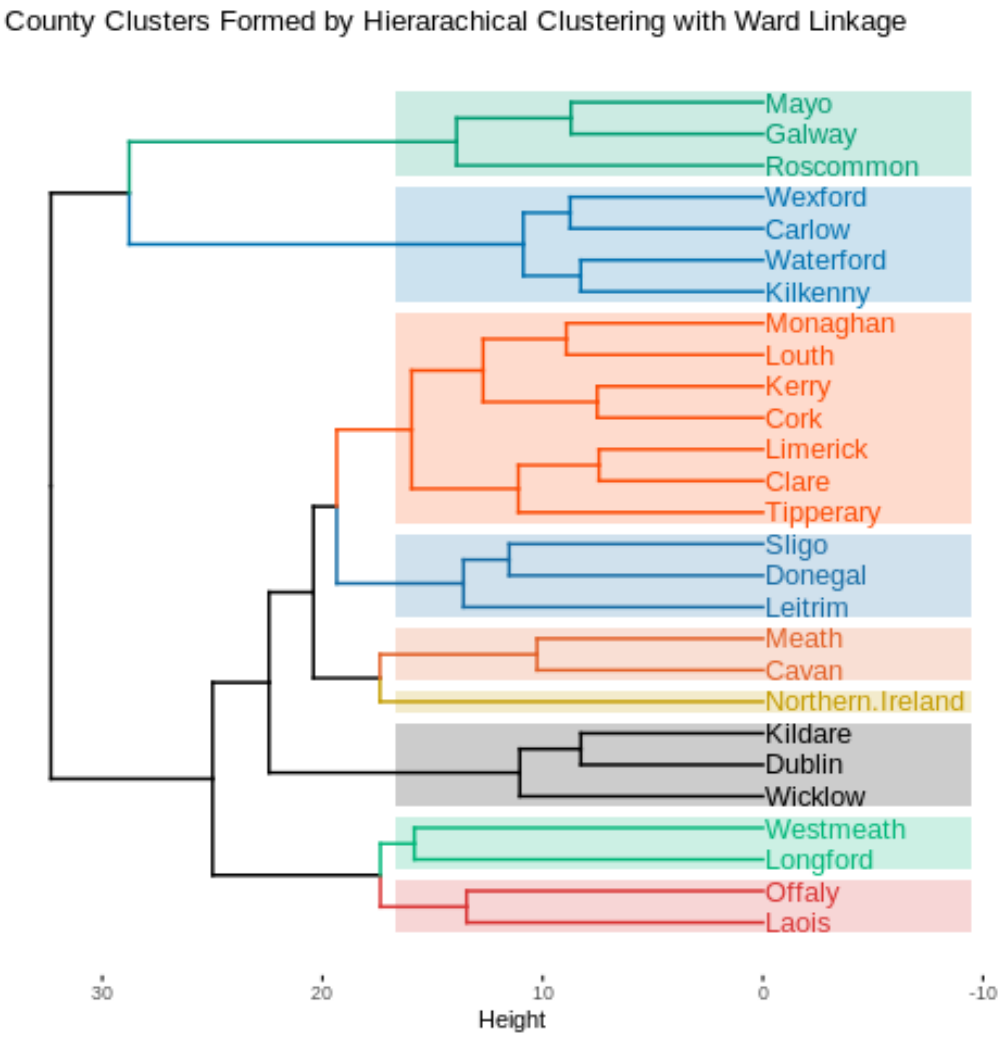

Figure 4.2: Dendrogram highlighting the clusters formed by hierarchical clustering with Ward Linkage

significant difference between at least two groups in the model.

4.3 Results of Tukey's Test

Six of the demographic variables (outlined in Demographic Variables) attain a p-value of less than 0.05. Post-hoc Tukey's test (Tukey's Honestly Significant

Scores for tests to check ANOVA assumptions (values > 0.05 in bold)

Variable	Shapiro-Wilks Value (Normal Distribution)	Levene Value (Homogeneity of Variance
% Managers	0.000123	4.3e-05
% High Professionals	**0.093148**	**0.948913**
% Low Professionals	**0.694792**	**0.598153**
% Non-Manual Workers	**0.921570**	**0.368764**
% Manual Workers	**0.436919**	**0.998727**
% Semi Skilled	**0.256467**	**0.557731**
% Unskilled	**0.127105**	**0.783905**
% Workers of own account	**0.166329**	**0.208377**
% Farmers	**0.469606**	**0.259999**
% Agricultural Workers	**0.353159**	**0.951645**
% Aged 20-24	**0.075127**	**0.919636**
% 25-29	0.000149	**0.602015**
% 30-34	0.005237	**0.427865**
% 35-39	**0.657155**	**0.136713**
% 40-44	0.008632	0.023352
% 45-49	**0.927559**	**0.722599**
% 50-54	**0.280949**	**0.701063**
% 55-59	**0.199681**	**0.865841**
% 60-64	**0.201038**	**0.910051**
% over 65	**0.892621**	**0.857505**
Populations Divided by Cases	**0.581440**	**0.502146**
% Manufacturing	**0.357283**	**0.513768**
% Transport/Communications	0.002656	**0.583977**
% Builders	**0.792302**	**0.850374**

Figure 4.3: ANOVA Assumption Test Results

Difference test) is now performed on the selected variables to identify which clusters differ in terms of that variables mean value. Those tests generated the Tukey Pairwise Confidence Interval Plots, shown in Figure 4.5,Figure 4.6,Figure 4.7,Figure 4.8,Figure 4.9, and Figure 4.10. Comparisons with an asterisk indicate that

P-Values for different dependent variables after ANOVA test

Variable	P_Value
% High Professionals	0.072362
% Low Professionals	0.116118
% Non-Manual Workers	0.348479
% Manual Workers	**0.048735**
% Semi Skilled	**0.046377**
% Unskilled	0.132399
% Workers of own account	0.470482
% Farmers	**0.048692**
% Agricultural Workers	0.398997
% Aged 20-24	0.717288
% 35-39	**0.023702**
% 45-49	0.929792
% 50-54	0.813037
% 55-59	0.072538
% 60-64	**0.018446**
% over 65	**0.011381**
Populations Divided by Cases	0.814790
% Manufacturing	0.134477
% Builders	0.251471

Figure 4.4: ANOVA Test Results

the adjusted p-value for those two clusters is < 0.05, indicating a significant difference between those two clusters:

- Percentage Semi-Skilled Workers (Figure 4.5)

  In this Tukey plot, the cluster pair 5 and 3 (marked with an asterisk) show significant difference.

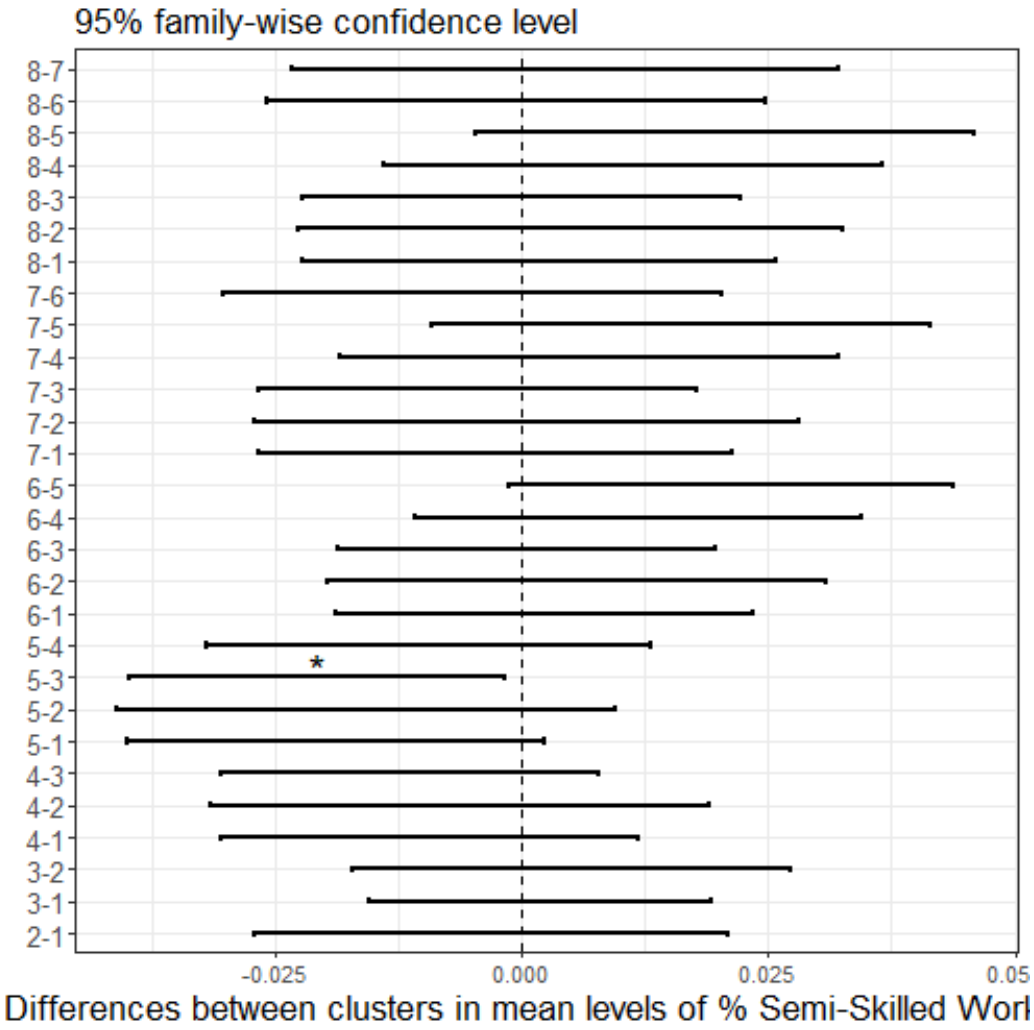

Figure 4.5: Tukey Test: Percentage Semi-Skilled Workers

- Percentage of Farmers (Figure 4.6)

The Tukey plot for this variable shows a significant difference (marked with an asterisk) between mean percentage of farmers in the counties of clusters 6 and 5.

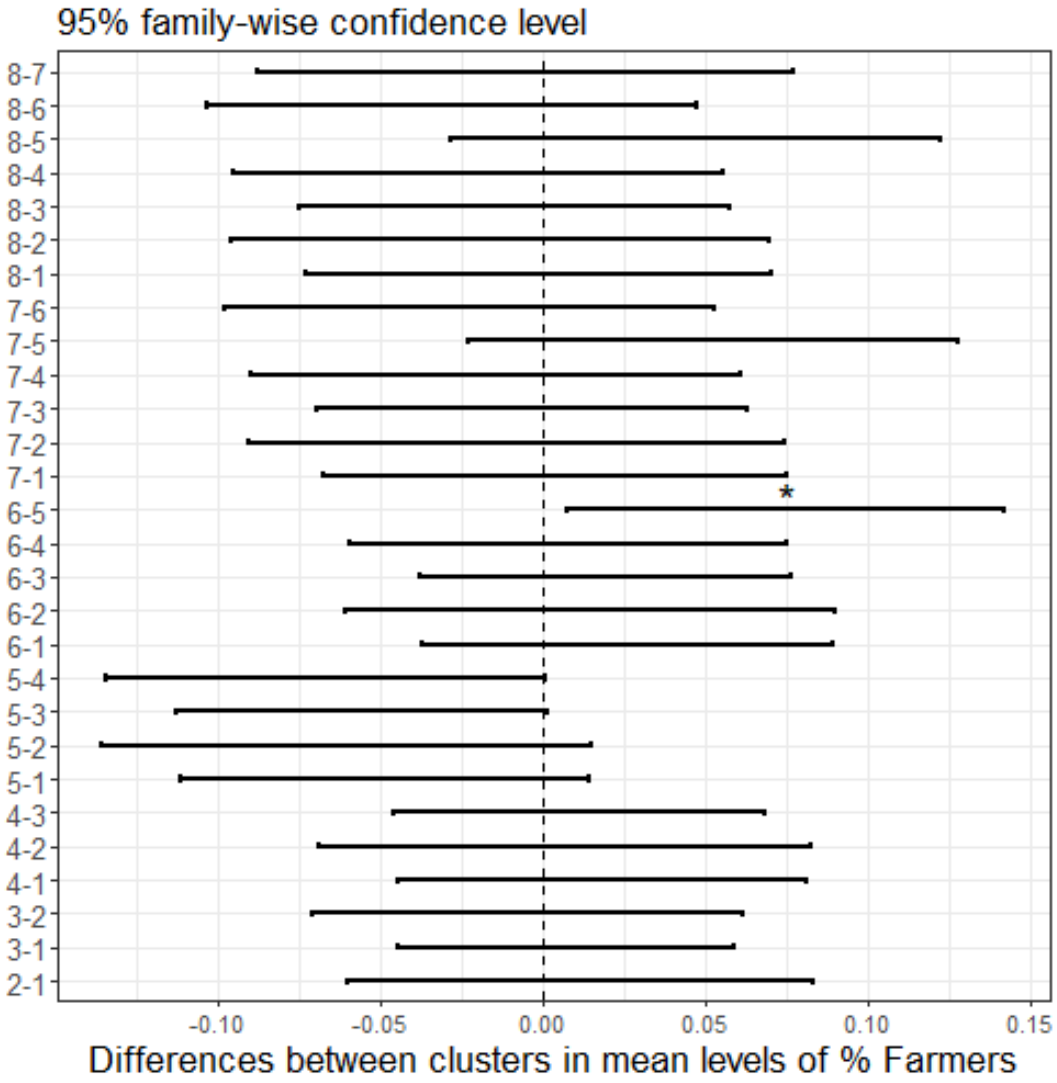

Figure 4.6: Tukey Test: Percentage Farmers

- Percentage of Population Aged 35-39 (Figure 4.7)

  There is significant pairwise difference (marked with an asterisk) between clusters 5 and 6 , and clusters 5 and 4.

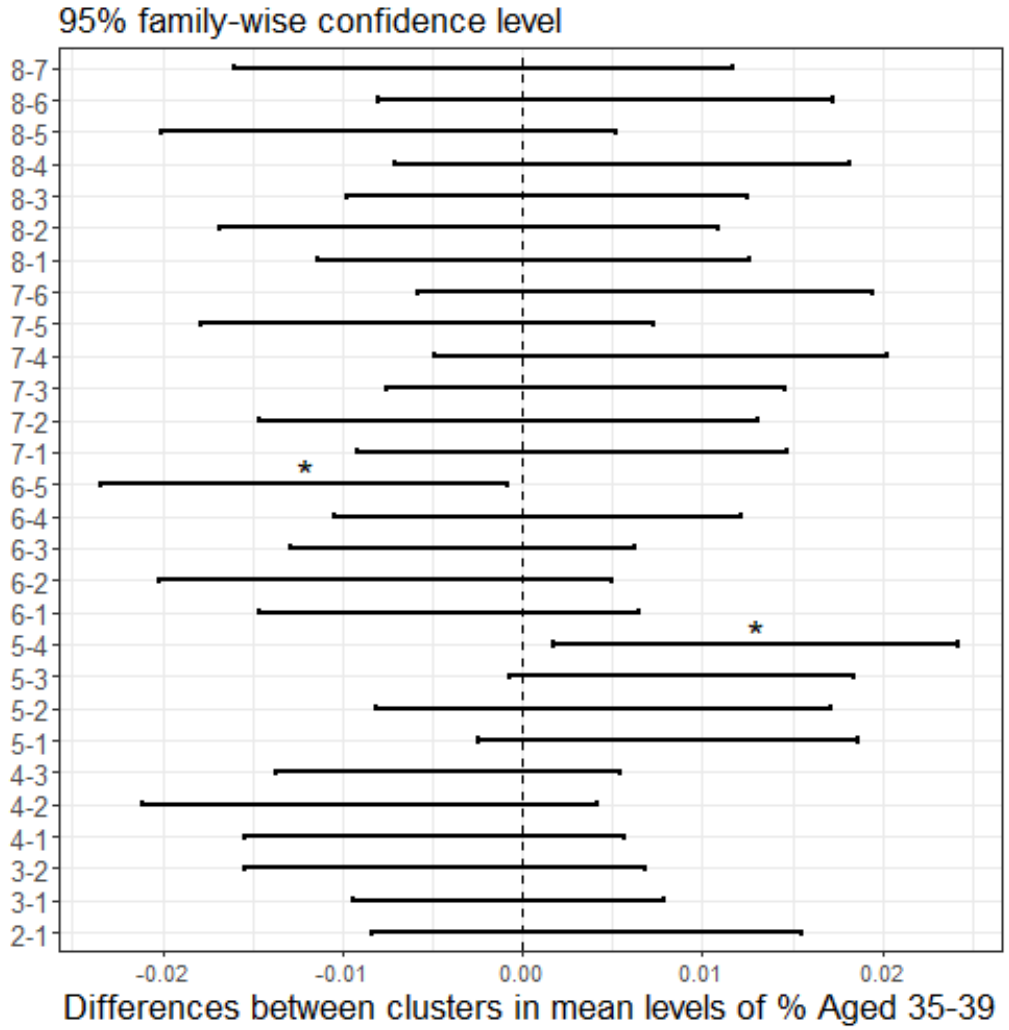

Figure 4.7: Tukey Test: Percentage Aged 35-39

- Percentage of Population Aged 60-64 (Figure 4.8)

  This Tukey plot shows a significant pairwise difference (marked with an asterisk) between clusters 4 and 5, and clusters 5 and 6

- Percentage of Population Aged 65+ (Figure 4.9)

  There is significant pairwise difference (marked with an asterisk) between clusters 5 and 6 , and clusters 5 and 4.

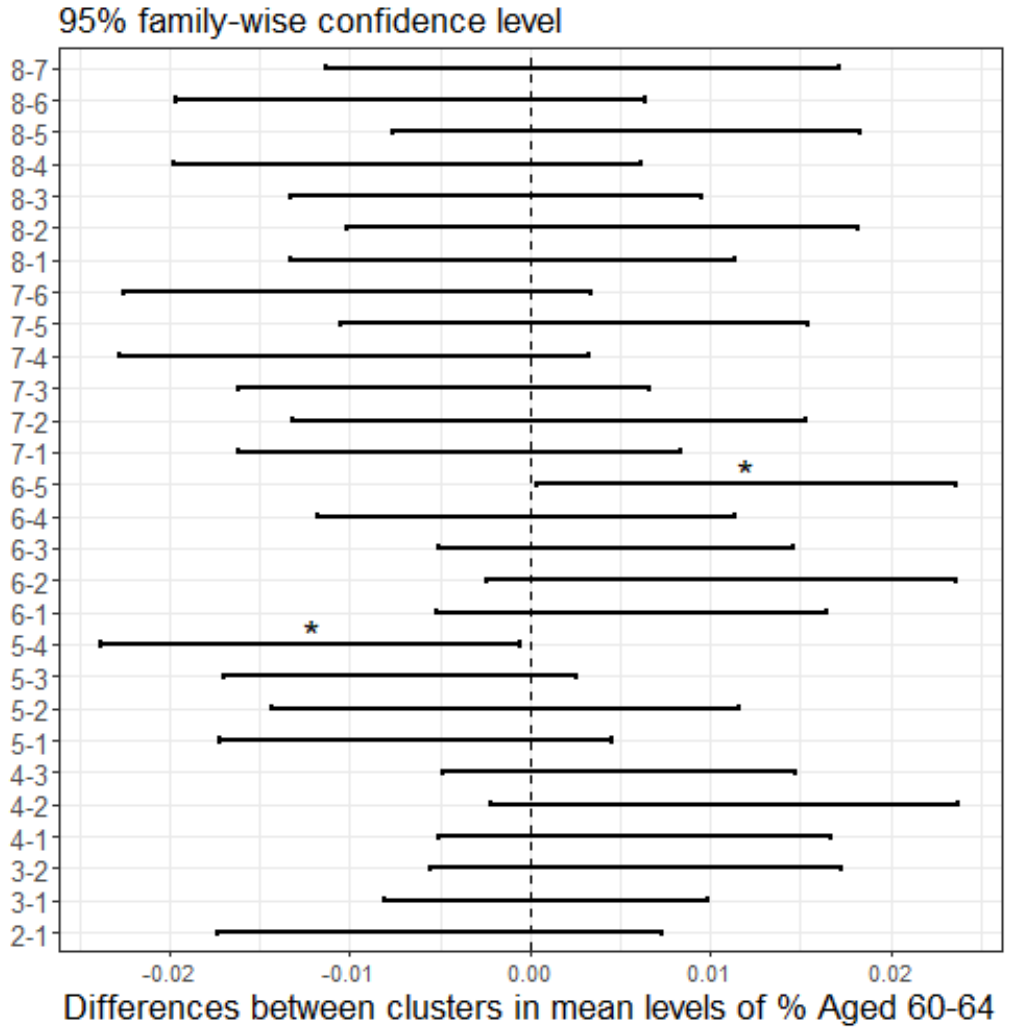

Figure 4.8: Tukey Test: Percentage Aged 60-64

- Percentage of Population in Manual Work (Figure 4.10)

  The Tukey test reveals that no pair of clusters exhibits a significant difference in mean levels of the percentage of population who are manual workers.

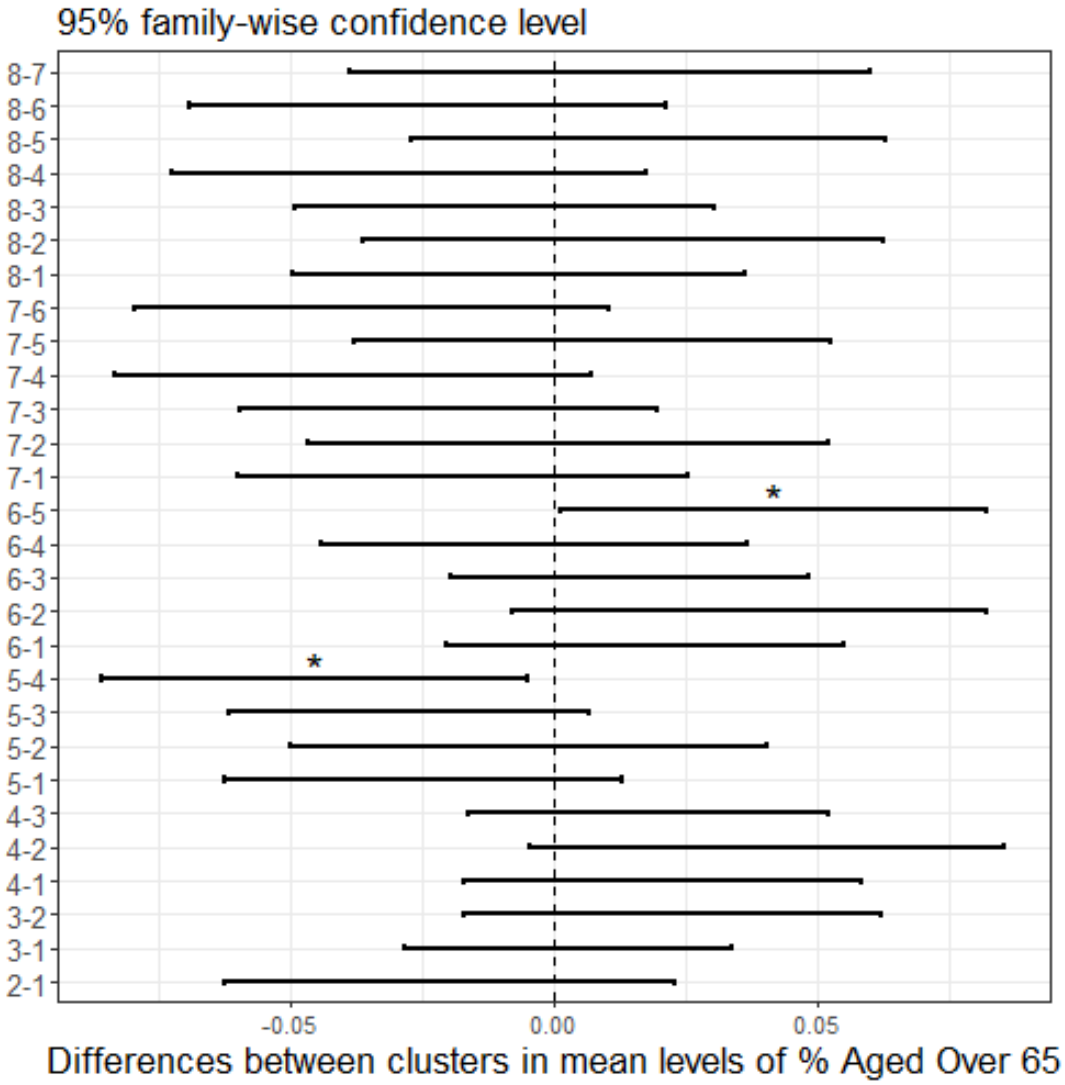

Figure 4.9: Tukey Test: Percentage Aged 65+

4.4 Discussion of Results of Tukey Tests

Firstly, despite the results of the ANOVA test returning a p-value of less than 0.05, the Tukey test for percentage manual workers indicates there is no significant difference between any two clusters for that variable.

There are repeated differences between cluster #5 (Dublin, Kildare, and Wick-

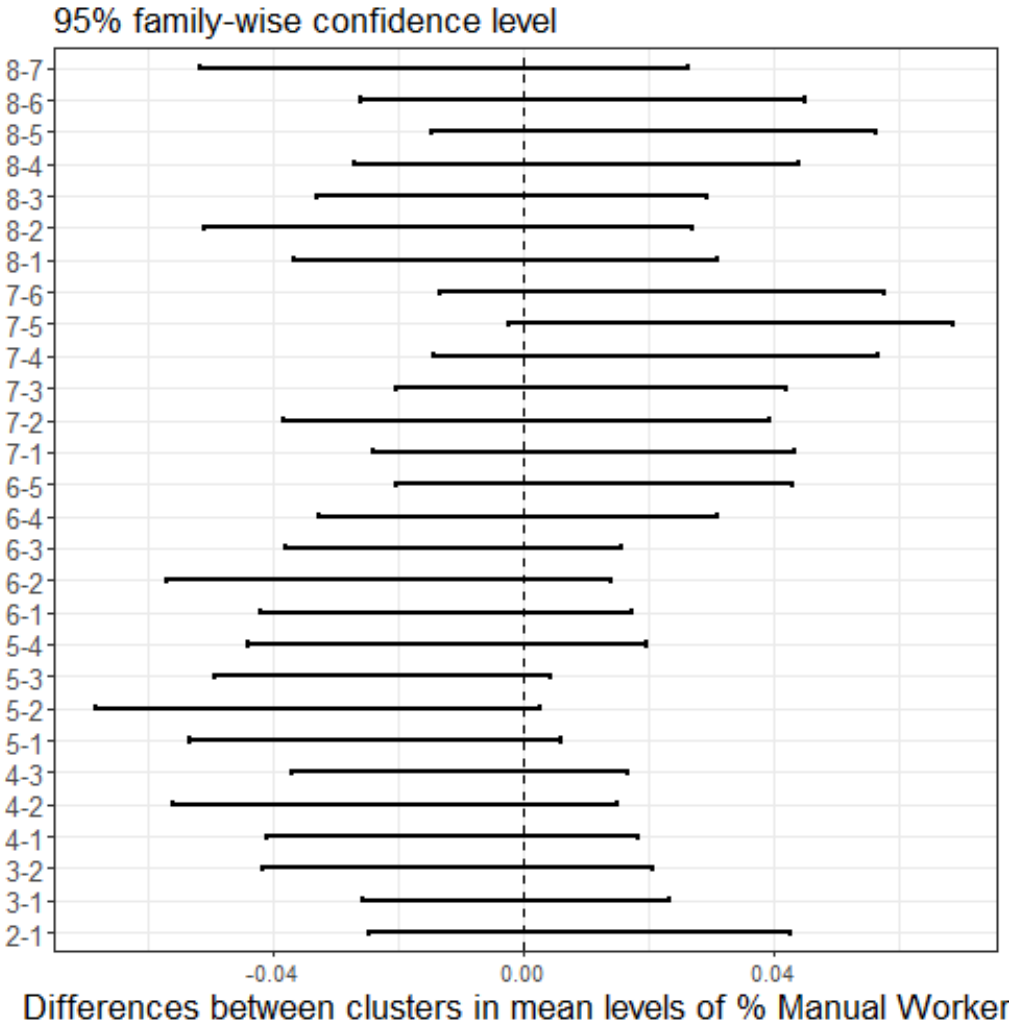

Figure 4.10: Tukey Test: Percentage Manual Workers

low) and clusters #4 (Donegal, Leitrim, and Sligo) and #6 (Galway, Mayo, and Roscommon). Three of these variables are related to age, with the percentage of farmers also showing a variation. The composition of each of these clusters shows a clear urban/rural contrast. Cluster #5 contains densely populated urban and suburban areas built around Dublin City. Clusters #4 and #6, however, include much more sparsely populated counties in the west and north-west of the country. While it may be worth investigating if the Covid infection rate is related to these

variables, one may also conclude their variation is more likely indicative of the urban-rural divide between the counties in those respective clusters, for example the tendency of younger people in rural areas to migrate to urban areas and cities for work. [18]

While the Tukey tests indicate that in some variables there does exist a significant inter-cluster difference, there is never more than one pair of clusters that differ significantly in any one demographic variable. This suggests that none of the examined demographic variables conclusively correlate with the shape of Covid infection rate time-series. The results also serve to highlight the limitation of performing ANOVA and post-Tukey tests on such a small dataset: any clusters that differ are only composed of a small sample of two or three counties, and thus do not provide a definitive pattern of behaviour.

The results of statistical tests suggest that none of the selected variables conclusively correlate with the clusters of counties generated in this experiment.

4.5 Interpretation of Clusters

While the statistical tests detailed in the previous section did not identify any explanatory or correlating variable, informal observations of the clusters may be possible by examining their time-series and centroids:

- Cluster # 1 (Figure 4.11)(Figure 4.12)
  The counties did not experience any significant spike in Covid infections until the third wave of the virus which struck in January 2021. This suggests that Covid restrictions implemented by the Irish government were effective in this region.

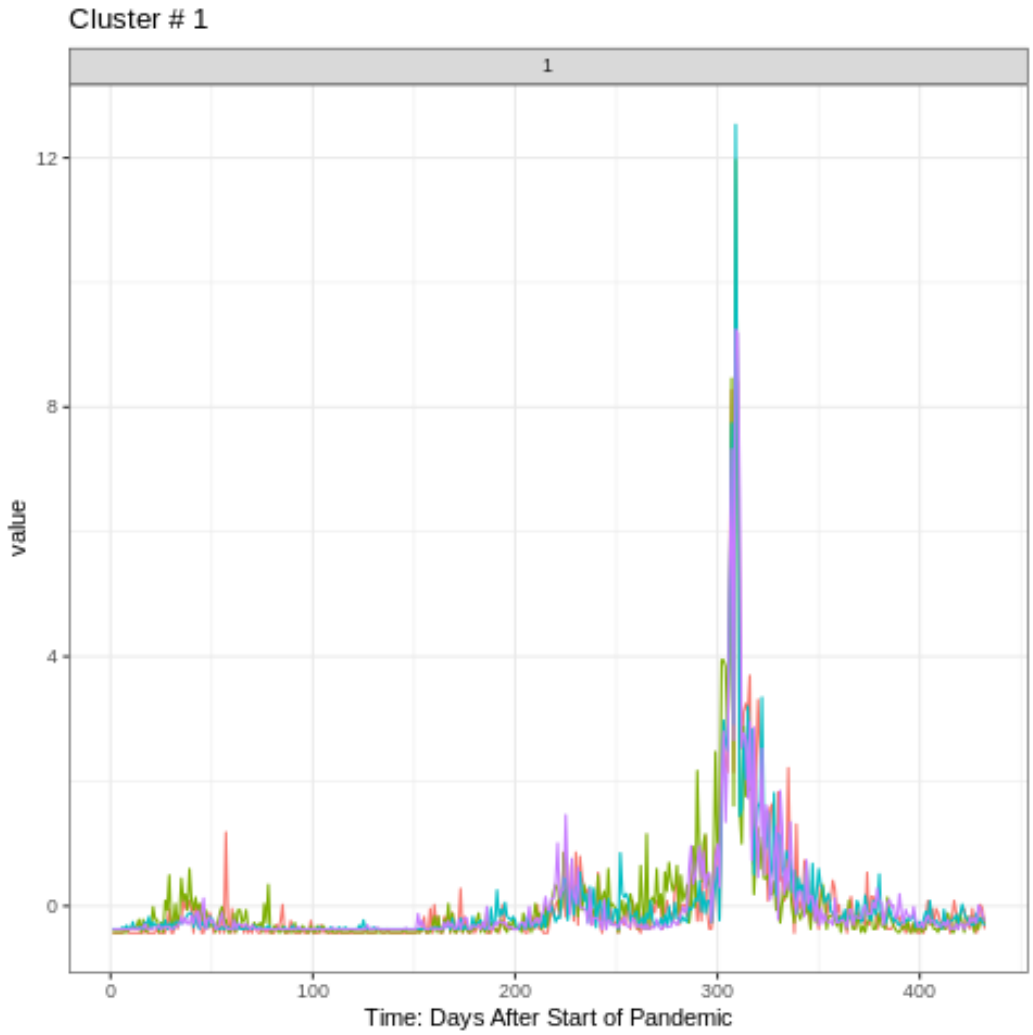

Figure 4.11: Time-Series of Cluster #1

- Cluster #2 (Figure 4.13)(Figure 4.14)

  Both counties in this cluster experienced drastic spikes in Covid infection cases during the second wave of the virus autumn 2020. Outbreaks of Covid during the second wave of the virus were linked to private gatherings, with a GP in Cavan linking the increase in infections to large post-match celebrations during the GAA championship final season in September and October [70].

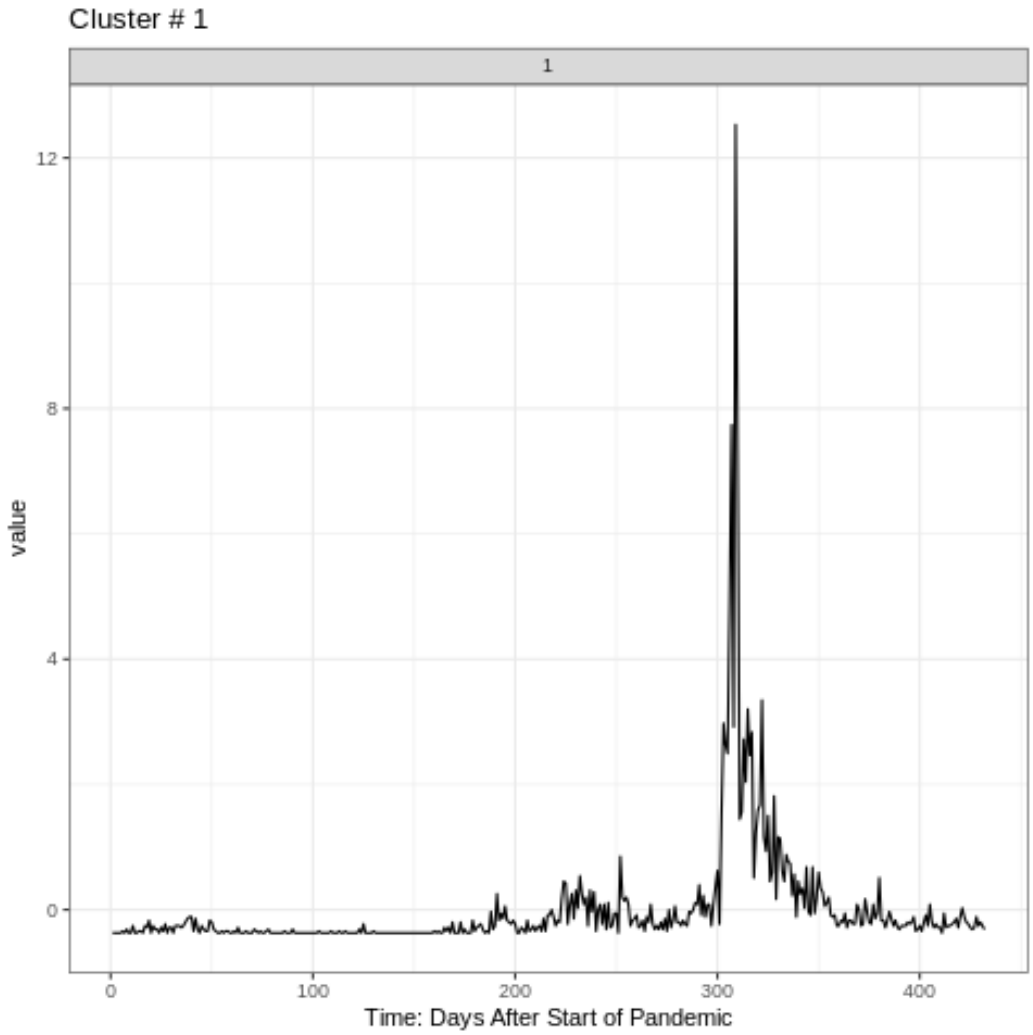

Figure 4.12: Centroid of Cluster #1

- Cluster #3 (Figure 4.15)(Figure 4.16)

  Like Cluster #1, the counties in this cluster did not appear to experience any drastic spikes in Covid cases until the third wave. And like Cluster #1, these counties are also mainly located in the south of the country, with the obvious exception of Louth and Monaghan which are located on the Northern Irish border.

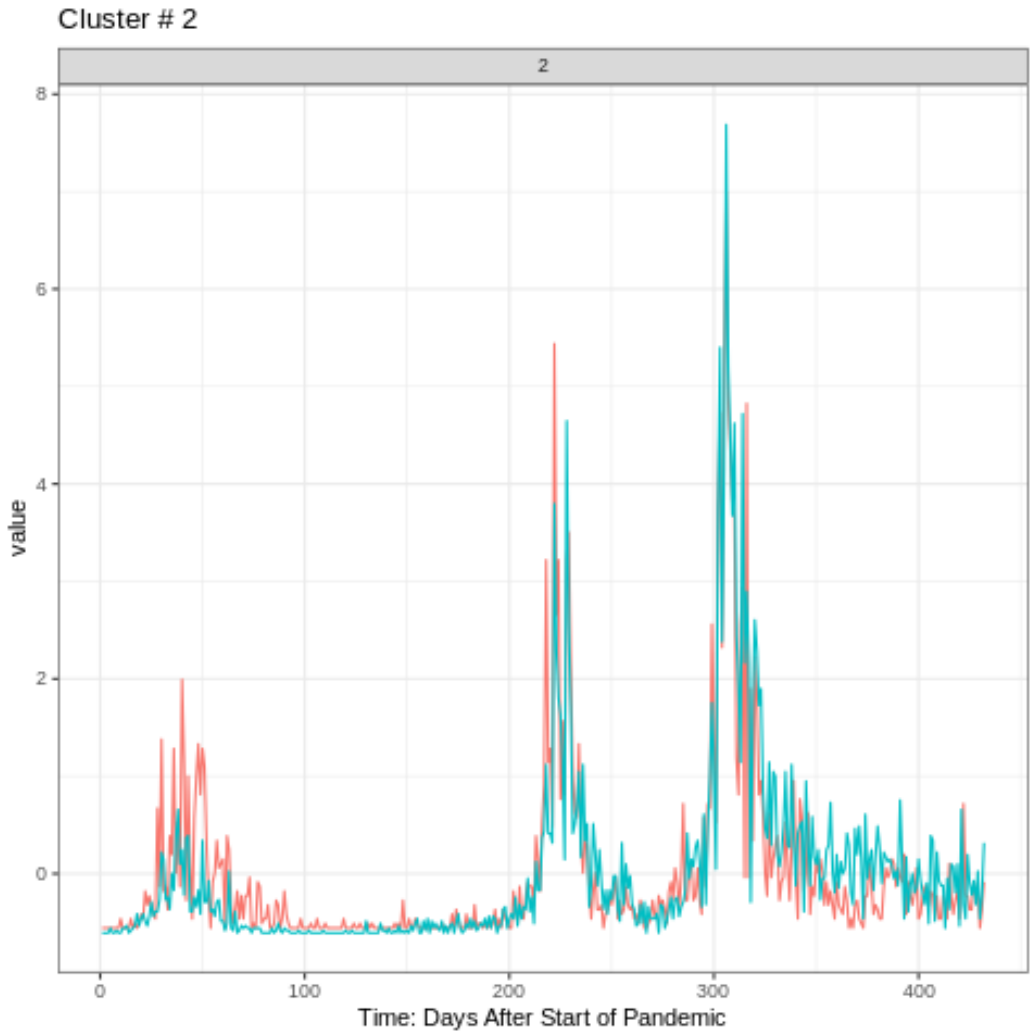

Figure 4.13: Time-Series of Cluster #2

- Cluster #4 (Figure 4.17)(Figure 4.18)

  These counties appear to have experienced small spikes of infection during the first wave, followed by moderate spikes during the second.

- Cluster #5 (Figure 4.19)(Figure 4.20)

  Apart from their geographical proximity, there does not appear to be a similarity between these counties based on the shape of the time-series.

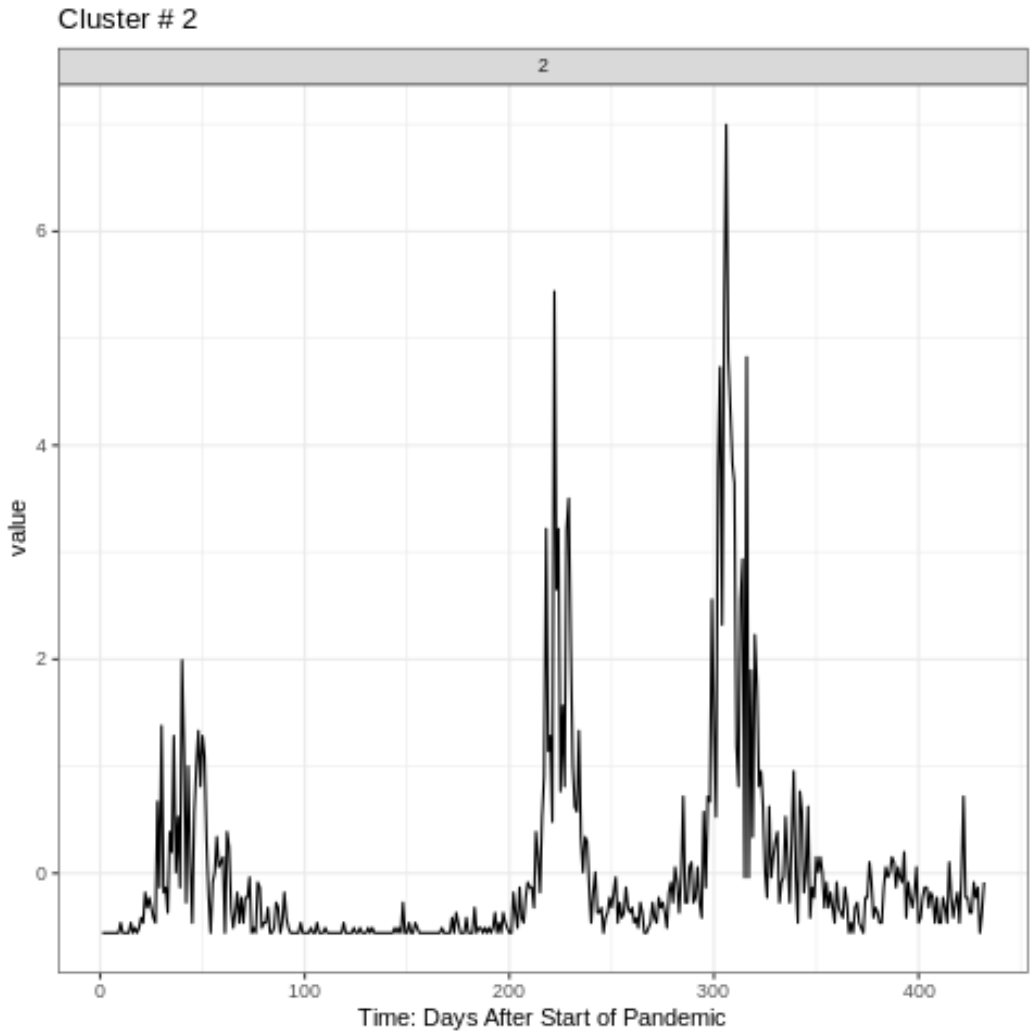

Figure 4.14: Centroid of Cluster #2

- Cluster #6 (Figure 4.21)(Figure 4.22)

  Counties Galway and Mayo appear to have experienced only gentle spikes in the first and second waves, before similarly drastic spikes during the third.

- Cluster #7 (Figure 4.23)(Figure 4.24)

  Laois and Offaly, although neighbouring counties, display differing shapes in their Covid infection time-series, with Offaly experiencing a more drastic

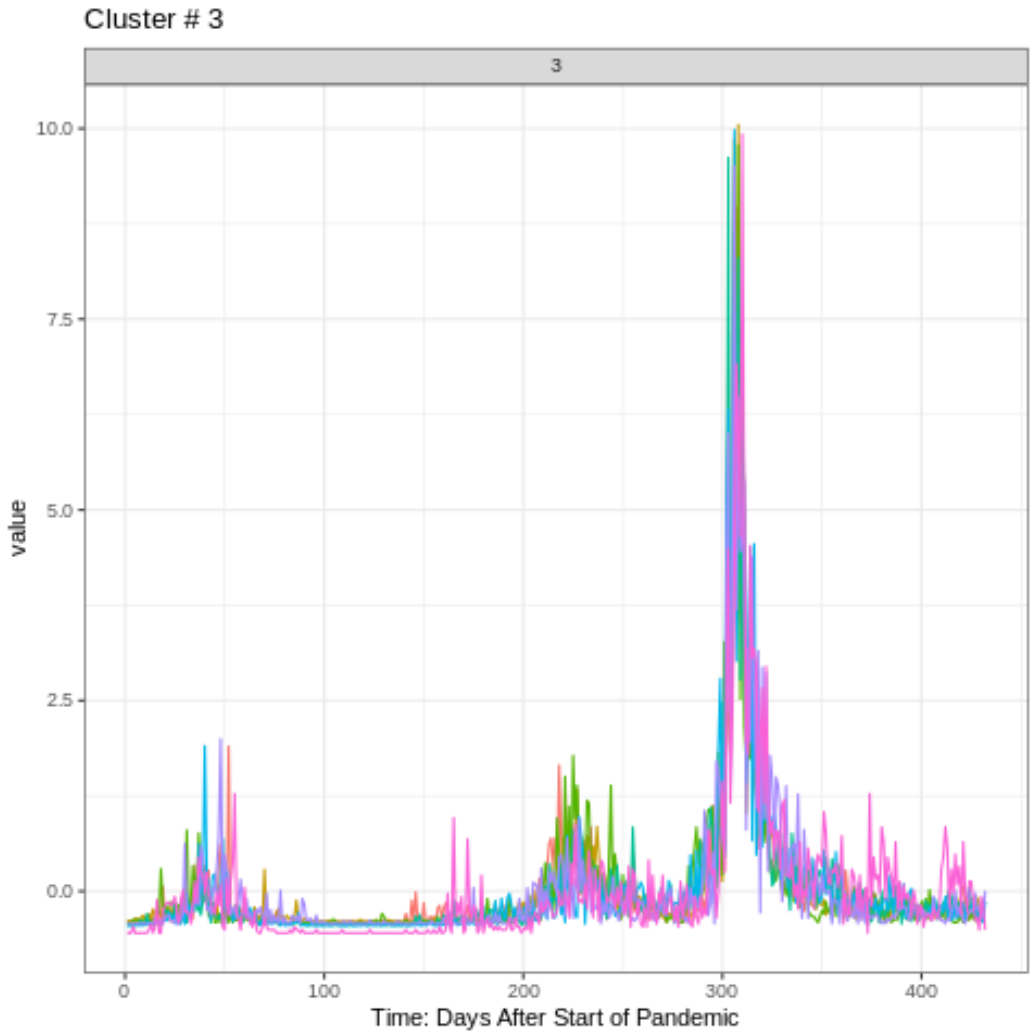

Figure 4.15: Time-Series of Cluster #3

spike during the first wave.

- Cluster #8 (Figure 4.25)(Figure 4.26)
  Both Longford and Westmeath experienced drastic spikes of Covid-19 infections during the first wave. The spike in cases in Longford can be linked to an outbreak of Covid-19 in a meat plant in the town of Ballymahon in May 2020[38].

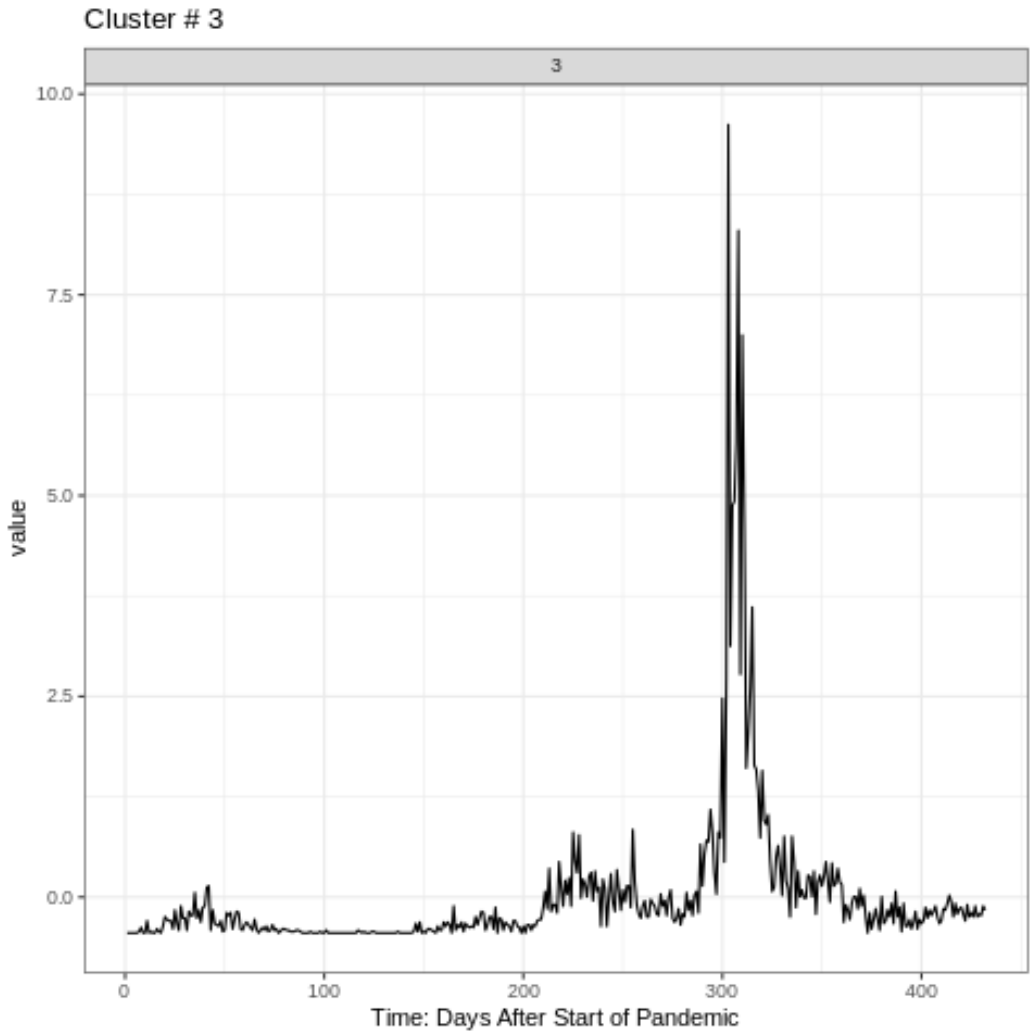

Figure 4.16: Centroid of Cluster #3

As Ballymahon is just three kilometres from the Westmeath border, the spike in Westmeath cases may be attributed to infected staff of the meat plant whose Covid infection was recorded over the border.

- Cluster # 9 (Figure 4.27)(Figure 4.28)
  Northern Ireland has formed its own cluster, while counties bordering Northern Ireland are spread between different clusters. This suggests that while

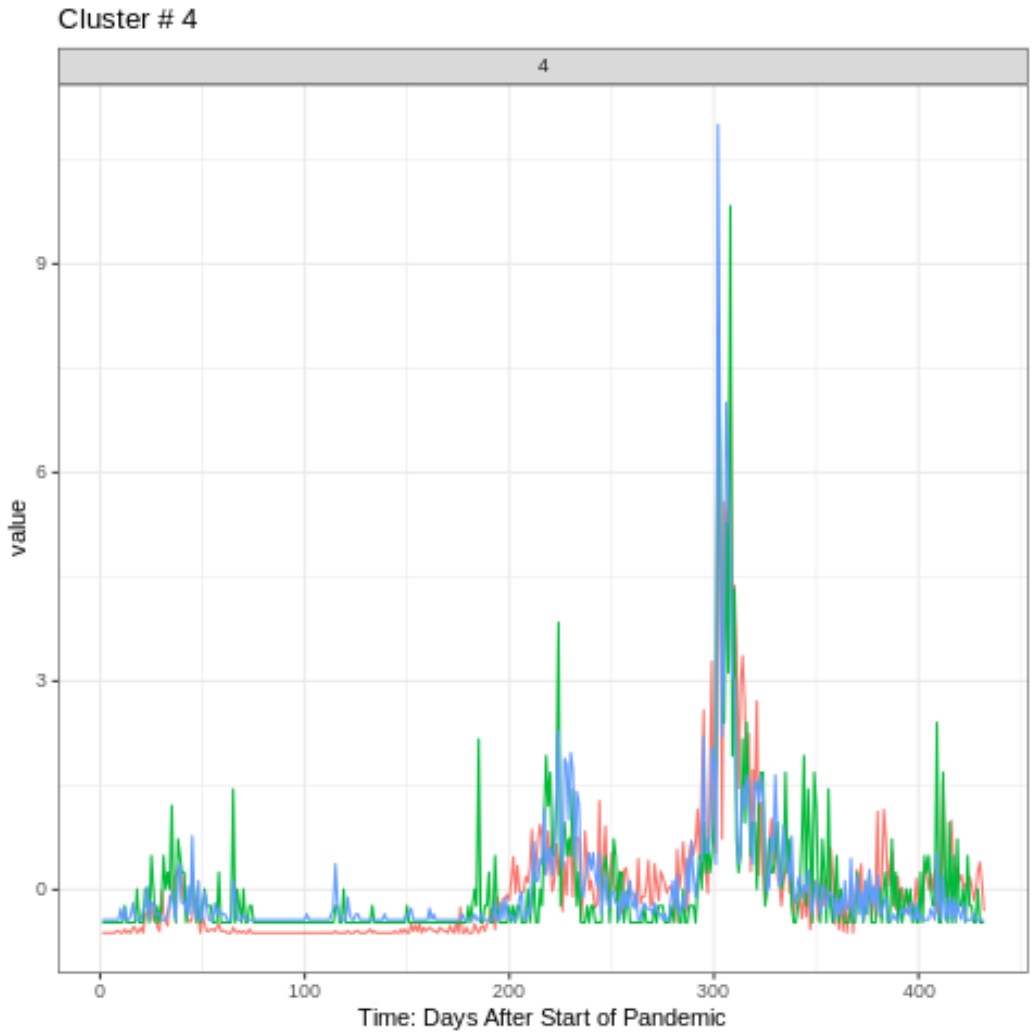

Figure 4.17: Time-Series of Cluster #4

higher Covid rates in border counties are linked to their proximity to Northern Ireland [6], this closeness of location does not affect the shape of the Covid infection time-series.

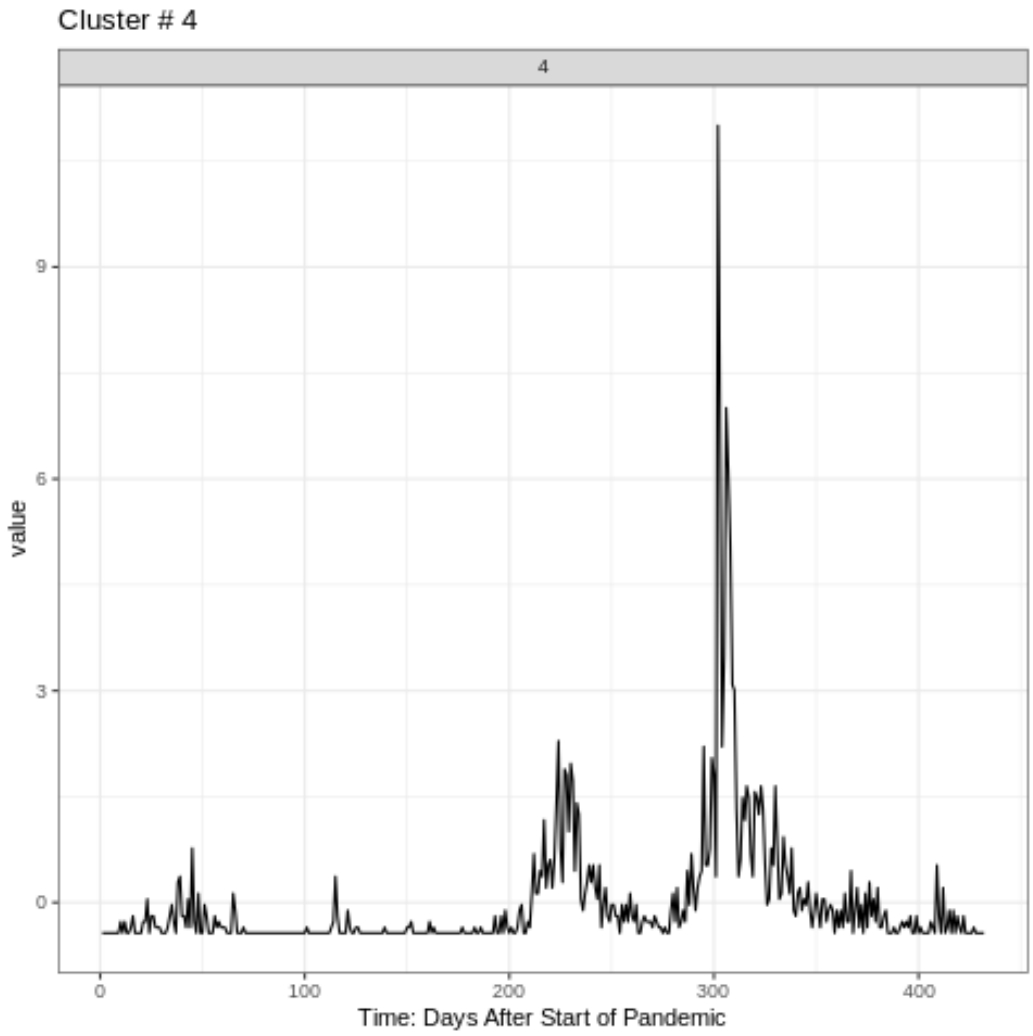

Figure 4.18: Centroid of Cluster #4

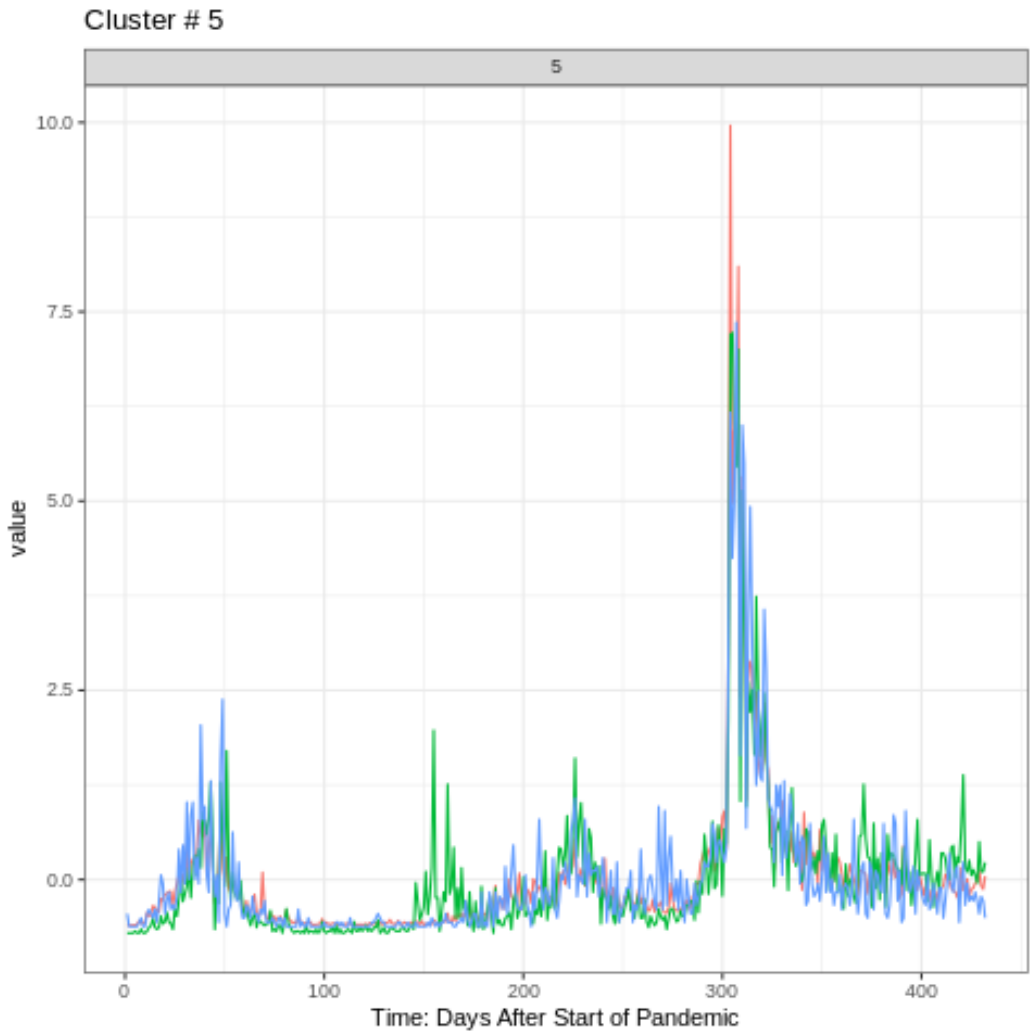

Figure 4.19: Time-Series of Cluster #5

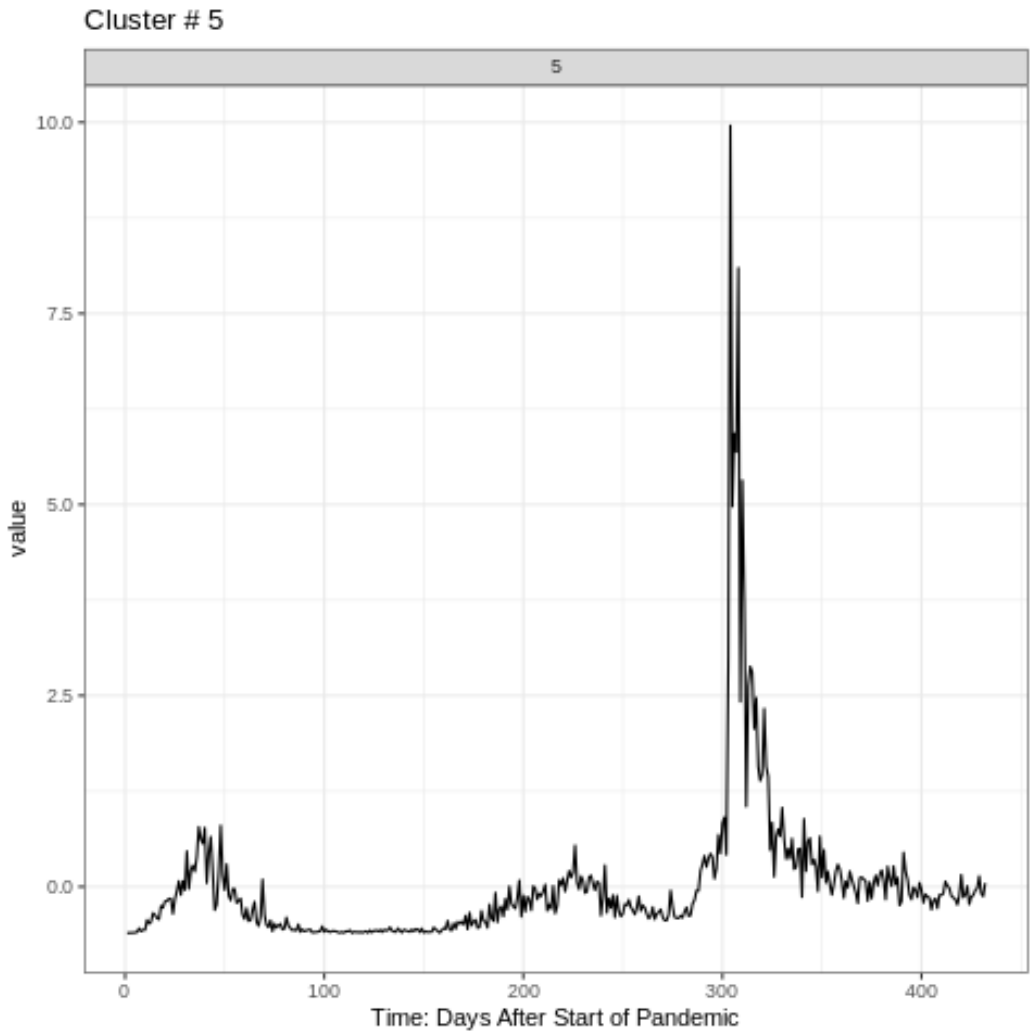

Figure 4.20: Centroid of Cluster #5

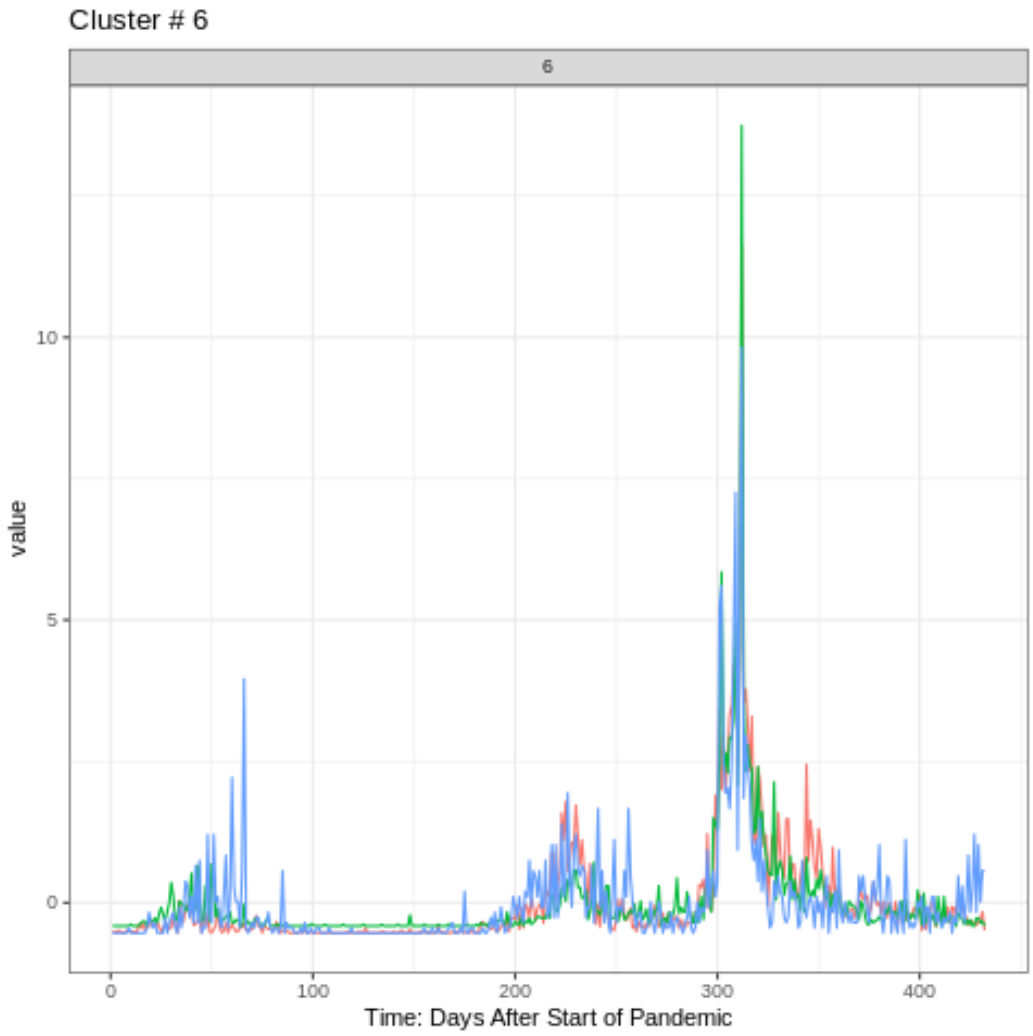

Figure 4.21: Time-Series of Cluster #6

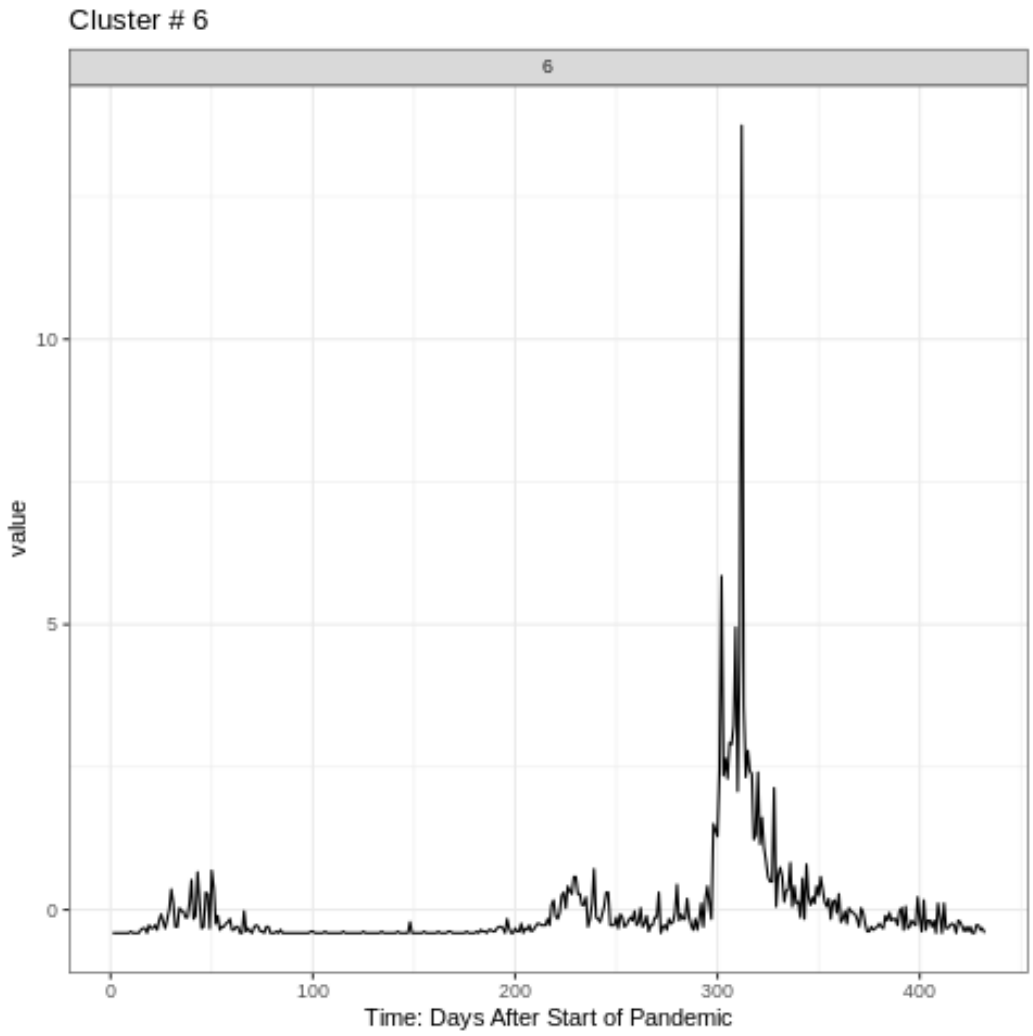

Figure 4.22: Centroid of Cluster #6

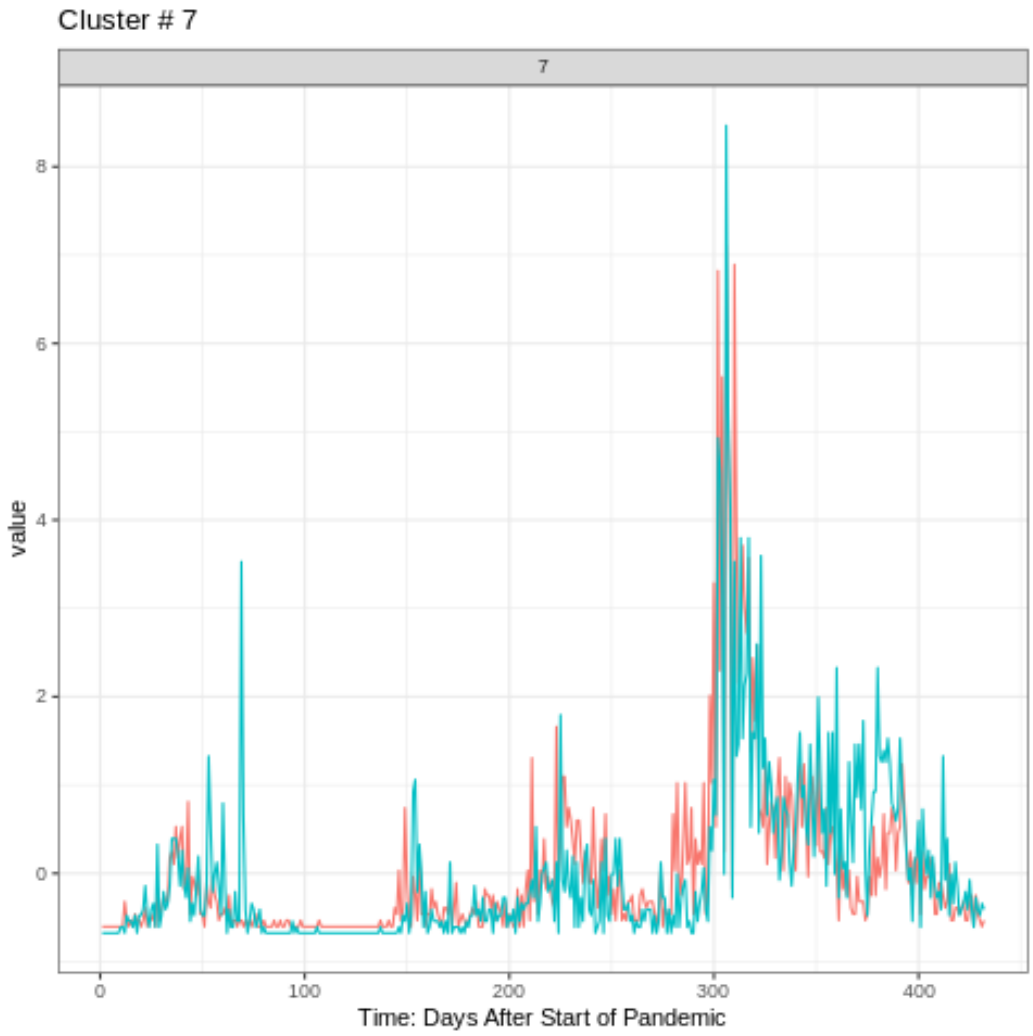

Figure 4.23: Time-Series of Cluster #7

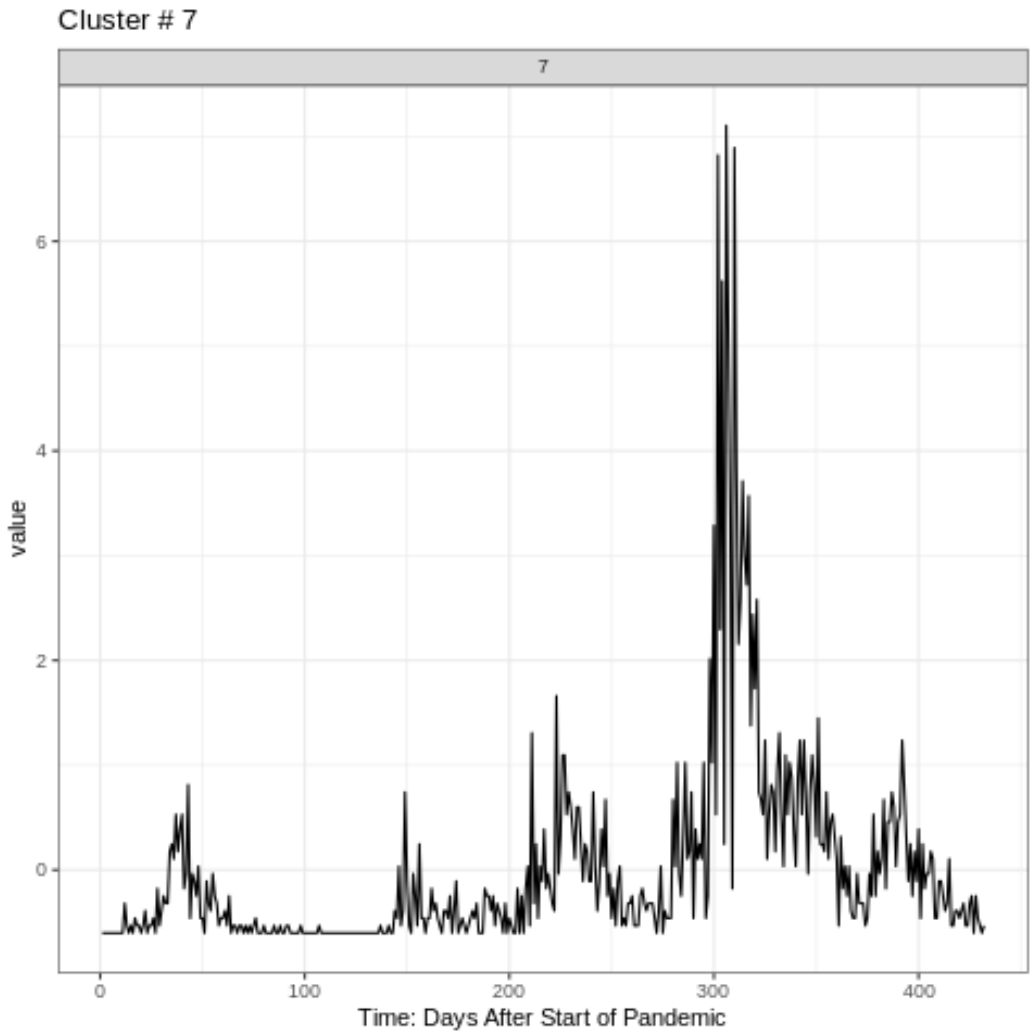

Figure 4.24: Centroid of Cluster #7

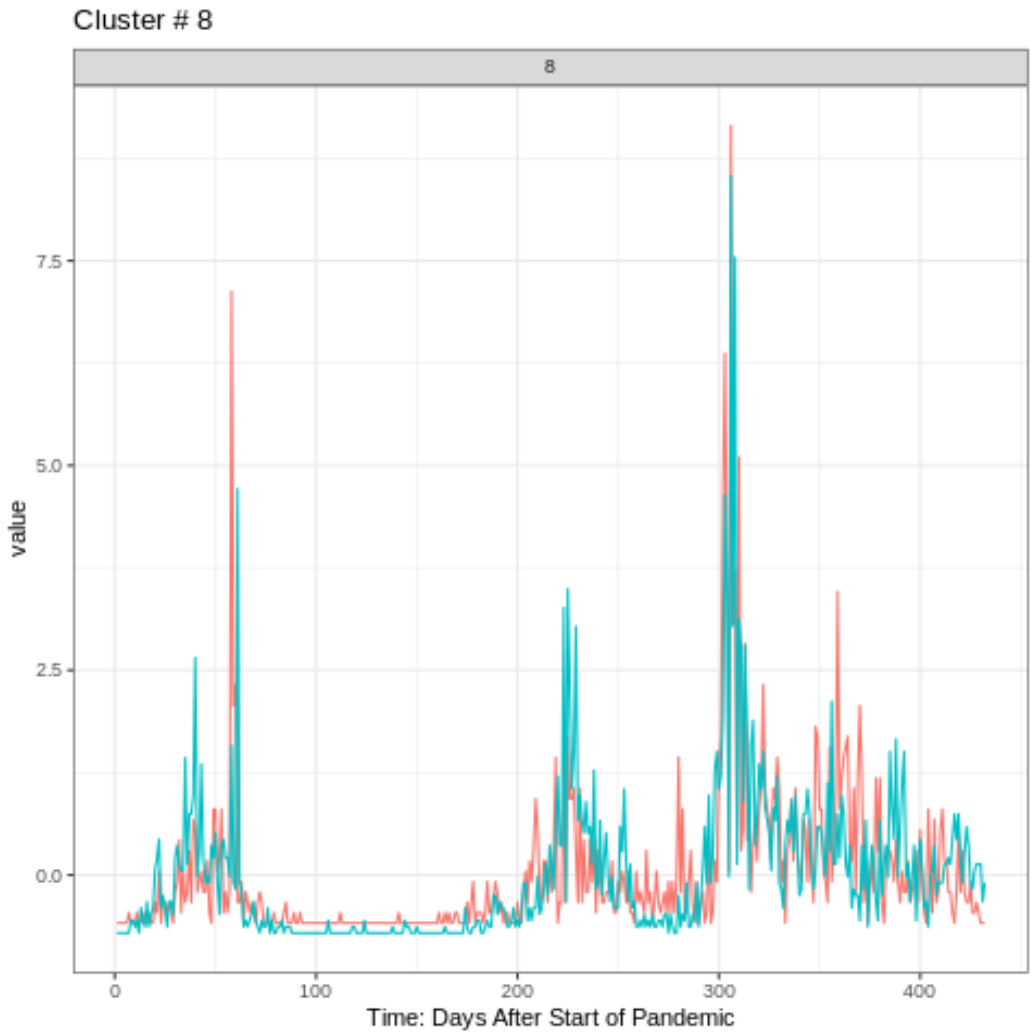

Figure 4.25: Time-Series of Cluster #8

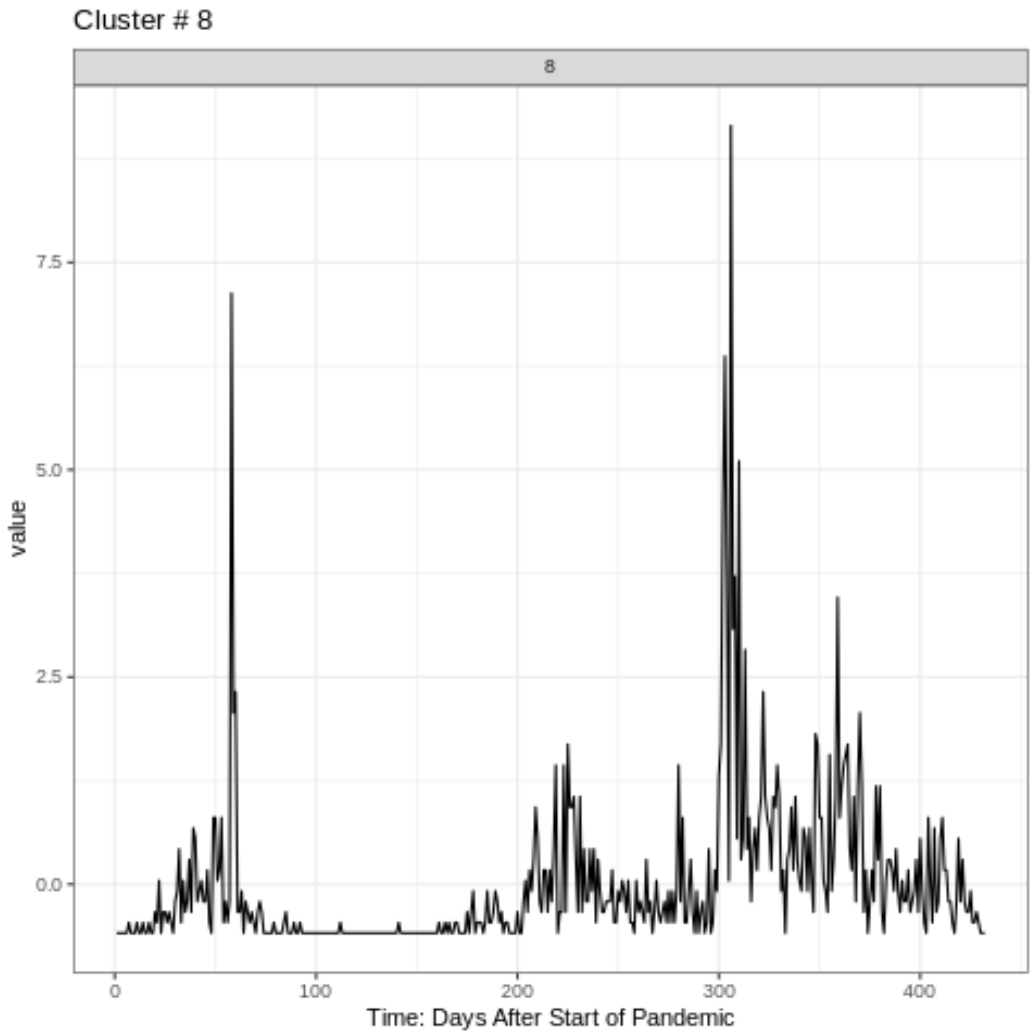

Figure 4.26: Centroid of Cluster #8

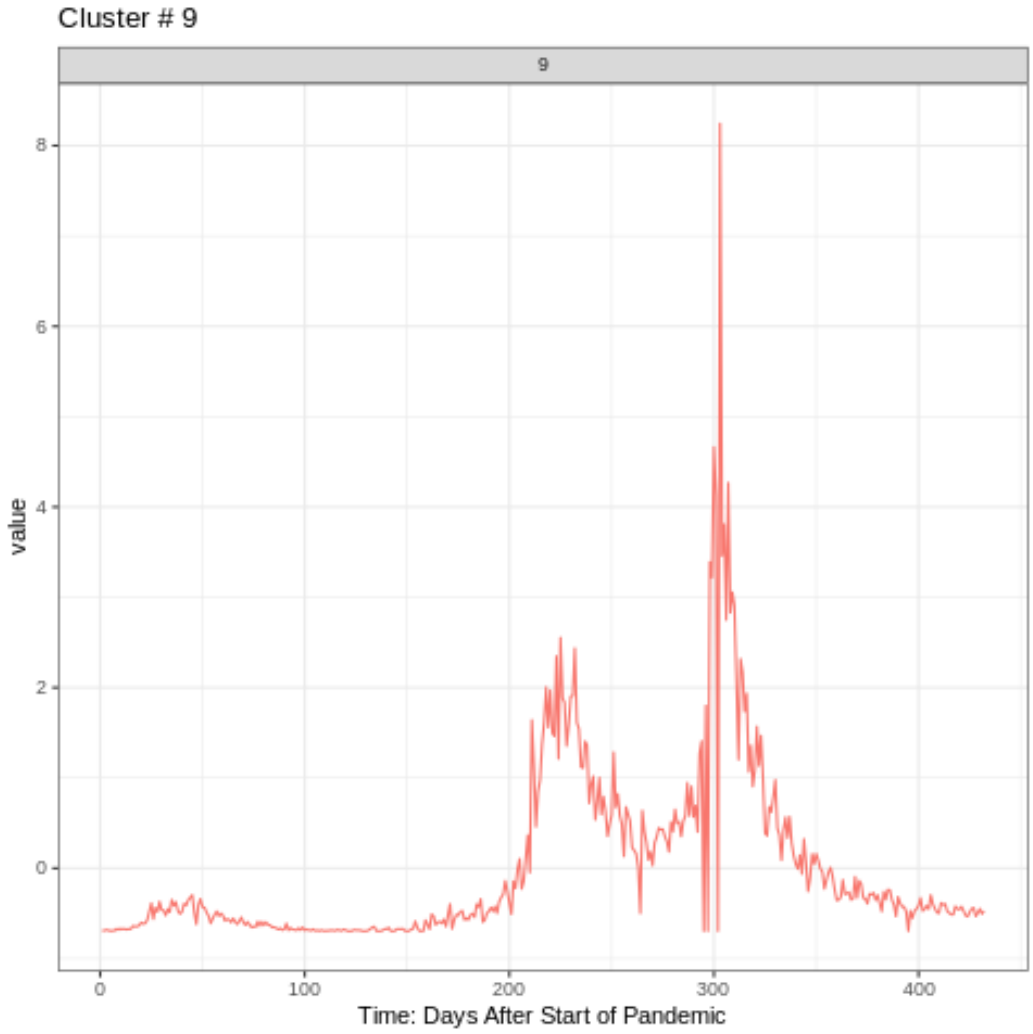

Figure 4.27: Time-Series of Cluster #9

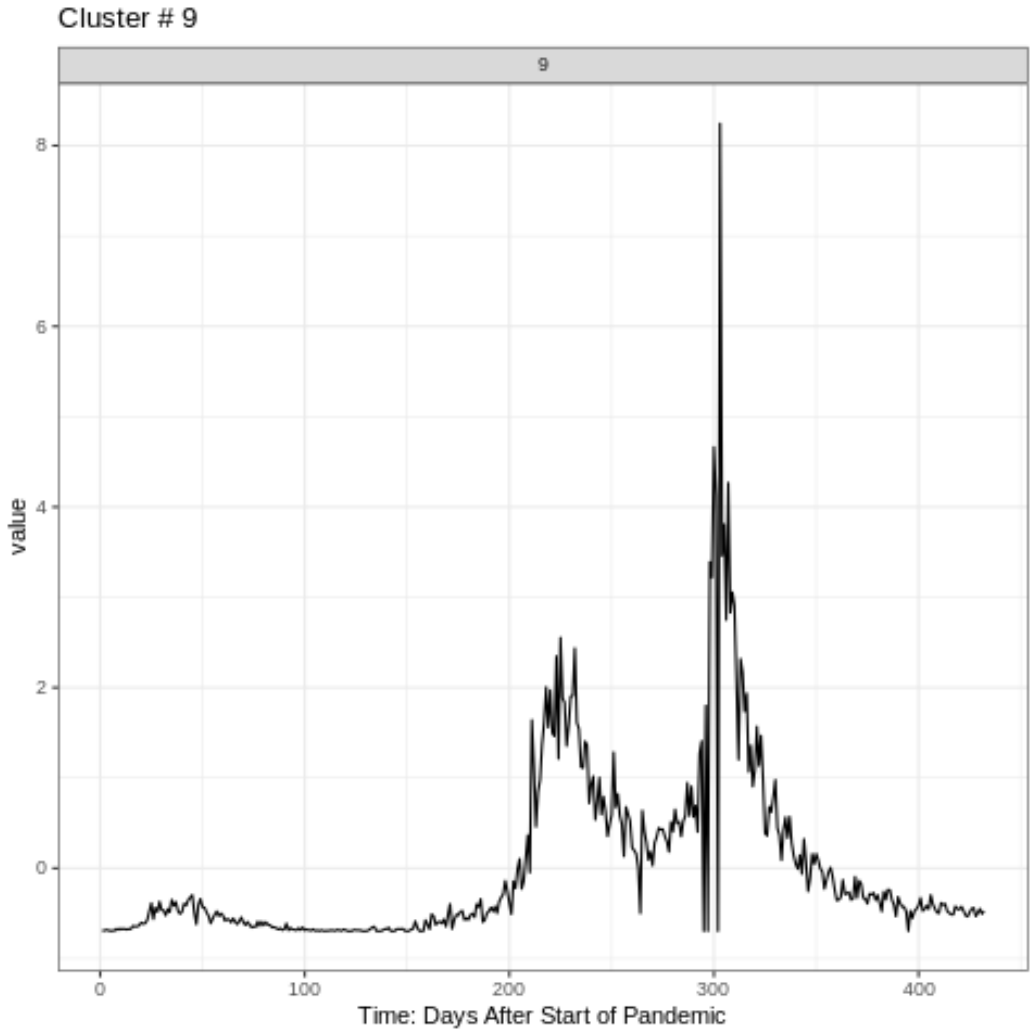

Figure 4.28: Centroid of Cluster #9

Chapter 5

Conclusions

5.1 Research Questions Answered

1. Can time-series clustering be utilised to increase knowledge of the spread of Covid-19 in Ireland?

   The results of this project, including the geographic coherence of the clusters formed, suggest that time-series clustering can reveal information about the spread of Covid-19 at a high regional level. However, the use of a dataset composed of the Covid infection time-series of Irish counties is perhaps not fine-grained or detailed enough to reveal any specific information about what factors affect the spread of the disease.
   A larger dataset, consisting of more focused local time-series combined with more specific local demographic data, would likely return more interesting results. An extension of this project may involve time-series clustering of a dataset of Covid infection rates of the 166 Local Electoral Areas of Ireland. Such a dataset is publicly available from the Government of Ireland website, although the updating of this dataset was disrupted soon after its

establishment by the previously mentioned ransomware attack on the Irish Health Service Executive, and so it was not possible to use it in this project.

2. What does time-series clustering reveal about the spread of Covid-19 in Ireland?

   Time-series clustering at county level reveals that neighbouring counties experienced the spread of the virus in similar ways, even if their relative infection numbers differed significantly.

3. What clusters of counties in Ireland experienced the spread of Covid-19 in similar ways?

   Time-series clustering identified nine clusters of Irish counties (including Northern Ireland, which is represented as a single entity) with similar time-series (see Figure 4.1).

4. Are there independent socio-economic, demographic, and environmental factors which correlate with the spread of Covid-19 in Ireland?

   Time-series clustering of Covid infection rates of Irish counties did not identify any such demographic factor that conclusively correlates with the spread of the disease.

5.2 Conclusion

In this project I performed clustering on a dataset of time-series objects representing the Covid-19 infection rates of Irish counties. I deployed a range of different configurations of common clustering algorithms to generate clusters. Although clustering is a highly subjective discipline, I used four cluster validity indices to help identify the final choice of clusters. The best final group of clusters was generated by hierarchical clustering with Euclidean distance and using Ward's criterion as linkage control. I then performed ANOVA and post-hoc Tukey tests to investigate if there exists a possible correlation between the clusters and a selection of county-level demographic variables recorded in the last Irish census of 2016. These tests revealed that there was no demographic variable that conclusively correlated with the final clusters of counties.

5.3 Future Work

As previously mentioned, the clustering of a larger, more locally-focused dataset of time-series objects (for example the Covid-19 infection rates of Irish Local Electoral Authorities) could possibly reveal more interesting and specific information related to the spread of the disease in Ireland. It would be particularly interesting to explore time-series clustering to assess if local urban and rural areas experience infectious diseases in consistently different ways

Clustering of Irish Covid death rates could also provide interesting insights. If the clusters formed by time-series clustering of Covid death rates differs significantly from clusters of similar infection rates, this will raise questions of how regional branches of the Health Service Executive managed the treatment of Covid

patients. It could also shed light on which regions provided most or least access to Covid treatment and/or testing. A similar approach was used by the authors of [39], although at international level.

The field of clustering is vast and, in this project, time limitations meant I could only test a relatively small selection of clustering configurations. More exploration of this discipline may allow me to build on and improve the work done in this project.

References

[1] H. Abdi and L. J. Williams. Tukey's honestly significant difference (hsd) test. *Encyclopedia of research design*, 3(1):1–5, 2010. 40

[2] S. Aghabozorgi and Y. W. Teh. Stock market co-movement assessment using a three-phase clustering method. *Expert Systems with Applications*, 41(4, Part 1):1301–1314, 2014. doi: https://doi.org/10.1016/j.eswa.2013.08.028. 6

[3] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah. Time-series clustering – A decade review. *Information Systems*, 53:16–38, 2015. doi: https://doi.org/10.1016/j.is.2015.04.007. 6, 8, 9, 16, 20, 21, 22, 25

[4] S. R. Aghabozorgi, T. Y. Wah, A. Amini, and M. R. Saybani. A new approach to present prototypes in clustering of time series. 2011. 8

[5] R. Agrawal, C. Faloutsos, and A. Swami. Efficient similarity search in sequence databases. In D. B. Lomet, editor, *Foundations of Data Organization and Algorithms*, pages 69–84, Berlin, Heidelberg, 1993. Springer Berlin Heidelberg. ISBN 978-3-540-48047-1. 6

[6] R. Ahmed and P. May. Does high COVID-19 spread impact neighbouring countries? Evidence from Ireland. *HRB open research.*, 4(56):56, 2021. 14, 15, 60

[7] M. Allen-Coghlan and P. Varthalitis. Comparing two recessions in ireland: Global financial crisis vs covid-19. *QUARTERLY ECONOMIC COMMENTARY*, page 87, 2020. 2

[8] E. Alvarez, J. Gabriel Brida, and E. Limas. Comparisons of COVID-19 dynamics in the different countries of the World using Time-Series clustering. *medRxiv*, 2020. doi: 10.1101/2020.08.18.20177261. 10, 15

[9] C. Antunes and A. L. Oliveira. Temporal Data Mining: an overview. 2001. 5

[10] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, 2013. 25

[11] R. A. Armstrong, S. Slade, and F. Eperjesi. An introduction to analysis of variance (anova) with special reference to data from clinical experiments in optometry. *Ophthalmic and Physiological Optics*, 20(3):235–241, 2000. 38

[12] BBC. Coronavirus: First case confirmed in Northern Ireland. URL `https://www.bbc.com/news/uk-northern-ireland-51665704`. Accessed on 08.07.2021. 2

[13] P. S. Bradley, U. Fayyad, C. Reina, et al. Scaling EM (expectation-maximization) clustering to large databases. *Microsoft Research*, pages 0–25, 1998. 9

[14] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974. 26

[15] R. N. Cardinal and M. R. Aitken. *ANOVA for the behavioral sciences researcher*. Psychology Press, 2013. 39

[16] R. Carrillo-Larco. Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach. 12, 13

[17] Central Statistics Office. Census of Population 2016 - Preliminary Results, . URL https://www.cso.ie/en/releasesandpublications/ep/p-cpr/censusofpopulation2016-preliminaryresults/geochan/. Accessed on 08.07.2021. 1

[18] Central Statistics Office. Urban, Rural, Regional - CSO - Central Statistics Office , . URL https://www.cso.ie/en/releasesandpublications/ep/p-cp3oy/cp3/urr/. Accessed on 07.09.2021. 53

[19] K.-P. Chan and A. W.-C. Fu. Efficient time series matching by wavelets. In *Proceedings 15th International Conference on Data Engineering (Cat. No.99CB36337)*, pages 126–133, 1999. doi: 10.1109/ICDE.1999.754915. 6

[20] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs. Nbclust: an r package for determining the relevant number of clusters in a data set. *Journal of statistical software*, 61(1):1–36, 2014. 21, 23

[21] L. Chaudhary and B. Singh. Community detection using unsupervised machine learning techniques on COVID-19 dataset. *Social network analysis and mining.*, 11(1):1, 2021. 13

[22] S. Chu, E. Keogh, and D. Hart. Iterative Deepening Dynamic Time Warping for Time Series. *Proceedings of the 2nd SIAM International Conference on Data Mining*, 01 2002. 8, 23

[23] C. Comiskey, A. Snel, and P. Banka. The second wave: estimating the hidden asymptomatic prevalence of covid-19 in ireland as we plan for imminent immunisation. *HRB Open Research*, 4(19):19, 2021. 2

[24] G. Duan, Y. Suzuki, and K. Kawagoe. Grid Representation for Efficient Similarity Search in Time Series Databases. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages x123–x123, 2006. doi: 10.1109/ICDEW.2006.63. 6

[25] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. 1973. 26

[26] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. *SIGMOD record*, 23(2):419–429, 1994. 6, 8, 23

[27] M. Ghahramani and F. Pilla. Leveraging artificial intelligence to analyze the COVID-19 distribution pattern based on socio-economic determinants. *Sustainable cities and society.*, 69:102848, 2021. 14, 15

[28] T. Giorgino. Computing and visualizing dynamic time warping alignments in r: the dtw package. *Journal of statistical Software*, 31(1):1–24, 2009. 24

[29] Government of Ireland. OSi Census 2016 OpenData, . URL `https://census2016.geohive.ie`. Accessed on 09.09.2021. 18

[30] Government of Ireland. gov.ie - Cyber attack on HSE systems, . URL `https://www.gov.ie/en/news/ebbb8-cyber-attack-on-hse-systems`. Accessed on 07.09.2021. 17

[31] Government of Ireland. DATA.GOV.IE - Datasets, . URL `https://data.gov.ie/dataset`. Accessed on 06.07.2021. 17

[32] Government of Ireland. Government publishes roadmap to ease COVID 19 restrictions and reopen Ireland's society and economy, . URL `https://www.gov.ie/en/press-release/`

e5e599-government-publishes-roadmap-to-ease-covid-19-restrictions-and-reope/ Accessed on 08.07.2021. 2

[33] Government of the United Kingdom. Coronavirus in the UK - Download Data. URL https://coronavirus.data.gov.uk/details/download. Accessed on 08.07.2021. 17

[34] L. Gupta, D. L. Molfese, R. Tammana, and P. G. Simos. Nonlinear alignment and averaging for estimating the evoked potential. *IEEE transactions on biomedical engineering*, 43(4):348–356, 1996. 8

[35] I. Gurrutxaga, I. Albisua, O. Arbelaitz, J. I. Martín, J. Muguerza, J. M. Pérez, and I. Perona. Sep/cop: An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index. *Pattern Recognition*, 43(10):3364–3373, 2010. 26

[36] V. Hautamaki, P. Nykanen, and P. Franti. Time-series clustering by approximate prototypes. In *2008 19th International conference on pattern recognition*, pages 1–4. IEEE, 2008. 8

[37] X. Huang, Z. Li, J. Lu, S. Wang, H. Wei, and B. Chen. Time-Series Clustering for Home Dwell Time during COVID-19: What Can We Learn from It? *ISPRS International Journal of Geo-Information*, 9(11), 2020. doi: 10.3390/ijgi9110675. 12

[38] independent.ie. Gardai give final warning to staff at factory where half are infected with Covid-19. URL https://www.independent.ie/business/farming/news/farming-news/gardai-give-final-warning-to-staff-at-factory-where-half-are-infected-with-html. Accessed on 07.09.2021. 58

[39] N. James and M. Menzies. Cluster-based dual evolution for multivariate time series: Analyzing COVID-19. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(6):061108, 2020. doi: 10.1063/5.0013156. 10, 15, 75

[40] A. Kassambara and F. Mundt. Package 'factoextra'. *Extract and visualize the results of multivariate data analyses*, 76, 2017. 27

[41] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009. 8

[42] B. Kennelly, M. O'Callaghan, D. Coughlan, J. Cullinan, E. Doherty, L. Glynn, E. Moloney, and M. Queally. The covid-19 pandemic in ireland: An overview of the health service and economic policy response. *Health Policy and Technology*, 9(4):419–429, 2020. 2

[43] E. Keogh and S. Kasetty. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data mining and knowledge discovery*, 7(4):349–371, 2003. 6

[44] E. Keogh and C. A. Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and information systems*, 7(3):358–386, 2005. 33

[45] S. Kumar. Monitoring Novel Corona Virus (COVID-19) Infections in India by Cluster Analysis. *Annals of data science.*, 7(3):417–425, 2020. 12

[46] L. F. Lacey. Characterization of ireland's third wave of covid-19 infections using an exponential function, with a time-dependent rate coefficient and its associated information entropy. 2021. 2

[47] D. Lemire. Faster retrieval with a two-pass dynamic-time-warping lower bound. *Pattern recognition*, 42(9):2169–2180, 2009. 33

[48] J. Lin, E. Keogh, S. Lonardi, and B. Chiu. A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, DMKD '03, page 2–11, New York, NY, USA, 2003. Association for Computing Machinery. doi: 10.1145/882082.882086. 6

[49] J. Lin, M. Vlachos, E. Keogh, and D. Gunopulos. Iterative incremental clustering of time series. In *International Conference on Extending Database Technology*, pages 106–122. Springer, 2004. 9

[50] S. Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982. 23

[51] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967. 9

[52] C. Magner, N. Greenberg, F. Timmins, V. O'Doherty, and B. Lyons. The psychological impact of covid-19 on frontline healthcare workers 'from heartbreak to hope'. *Journal of Clinical Nursing*, 30(13-14):e53, 2021. 2

[53] M. R. Mahmoudi, D. Baleanu, Z. Mansor, B. A. Tuan, and K.-H. Pho. Fuzzy clustering method to compare the spread rate of Covid-19 in the high risks countries. *Chaos, Solitons & Fractals*, 140:110230, 2020. doi: https://doi.org/10.1016/j.chaos.2020.110230. 10, 15

[54] C. Manchein, E. L. Brugnago, R. M. da Silva, C. F. O. Mendes, and M. W. Beims. Strong correlations between power-law growth of COVID-19 in four continents and the inefficiency of soft quarantine strategies. *Chaos: An*

Interdisciplinary Journal of Nonlinear Science, 30(4):041102, 2020. doi: 10.1063/5.0009454. 11

[55] F. M. Megahed, L. A. Jones-Farmer, and S. E. Rigdon. A retrospective cluster analysis of COVID-19 cases by county. *bioRxiv*, 2020. doi: 10.1101/2020.11.12.379537. 11, 15

[56] G. Mills, W. Cullen, N. Moore, and R. Foley. Making sense of publicly available data on COVID-19 in Ireland. *medRxiv*, 2020. 2

[57] T. Mirowski, S. Roychoudhury, F. Zhou, and Z. Obradovic. Predicting Poll Trends Using Twitter and Multivariate Time-Series Classification. In *Social Informatics*, pages 273–289, Cham, 2016. Springer International Publishing. ISBN 978-3-319-47880-7. 6

[58] Northern Ireland Statistics and Research Agency. 2011 Census. URL `https://www.nisra.gov.uk/statistics/census/2011-census`. Accessed on 08.07.2021. 2

[59] OpenDataNI. URL `https://www.opendatani.gov.uk/`. Accessed on 08.07.2021. 18

[60] F. Petitjean, A. Ketterlin, and P. Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern recognition*, 44(3):678–693, 2011. 31

[61] R. Pradeep Kumar and P.Nagabhushan. Time Series as a Point - A Novel Approach for Time Series Cluster Visualization. *Proceedings of the Conference on Data Mining, 2006 ,pp.24–29.* 6

[62] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971. 9

[63] S. A. Rizvi, M. Umair, and M. A. Cheema. Clustering of Countries for COVID-19 Cases based on Disease Prevalence, Health Systems and Environmental Indicators. *medRxiv*, 2021. doi: 10.1101/2021.02.15.21251762. 13

[64] F. Rojas, O. Valenzuela, and I. Rojas. Estimation of COVID-19 dynamics in the different states of the United States using Time-Series Clustering. *medRxiv*, 2020. doi: 10.1101/2020.06.29.20142364. 11, 15

[65] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. 23, 26, 27

[66] SafeGraph Inc. SafeGraph - Places, Data & Foot Traffic Insights. URL `https://www.safegraph.com`. Accessed on 08.07.2021. 12

[67] A. Sarda-Espinosa. *dtwclust: Time Series Clustering Along with Optimizations for the Dynamic Time Warping Distance*, 2019. URL `https://CRAN.R-project.org/package=dtwclust`. R package version 5.5.6. 16, 21, 22, 31, 33

[68] P. H. Sneath. The application of computers to taxonomy. *Microbiology*, 17 (1):201–226, 1957. 22

[69] R. R. Sokal. A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.*, 38:1409–1438, 1958. 21, 22

[70] thejournal.ie. Post-match parties and house gatherings: Where GPs notice Covid outbreaks originating. URL `https://www.thejournal.ie/article.php?id=5234054`. Accessed on 07.09.2021. 54

[71] J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963. 21

[72] T. Warren Liao. Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874, 2005. 5

[73] Wikipedia. COVID-19 pandemic data. Accessed on 08.07.2021. 2

[74] V. Zarikas, S. G. Poulopoulos, Z. Gareiou, and E. Zervas. Clustering analysis of countries using the COVID-19 cases dataset. *Data in Brief*, 31:105787, 2020. doi: https://doi.org/10.1016/j.dib.2020.105787. 9, 15

[75] Y. Zhao and G. Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine learning*, 55(3):311–331, 2004. 9

Appendix A

Appendix

Code for this project : https://github.com/paddy-garrett/Covid_Clustering