國立臺灣大學電機資訊學院暨中央研究院
資料科學學位學程
碩士論文

Data Science Degree Program
College of Electrical Engineering and Computer Science
National Taiwan University and Academia Sinica
Master Thesis

以時間序列分群法結合 R/Shiny 資料視覺化探討 COVID-19 傳播

An Investigation to the Spread of COVID-19 via Time Series Clustering and its Data Visualization via R/Shiny App

李奕宏

Yi-Hung Li

指導教授：潘建興 博士、吳沛遠 博士

Advisor: Frederick Kin Hing Phoa Ph.D., Pei-Yuan Wu Ph.D.

中華民國 112 年 01 月

January, 2023

doi:10.6342/NTU202300086

國立臺灣大學碩士學位論文
口試委員會審定書

以時間序列分群法結合 R/Shiny 資料視覺化探討 COVID-19 傳播

An Investigation to the Spread of COVID-19 via Time Series Clustering and its Data Visualization via R-Shiny App

本論文係 李奕宏（學號 R10946001）在國立臺灣大學資料科學學位學程研究所完成之碩士學位論文，於民國 112 年 1 月 10 日承下列考試委員審查通過及口試及格，特此證明

口試委員：吳沛遠 （簽名）

（指導教授）

（簽名）　（簽名）

（簽名）　

學程主任：謝宏昀 （簽名）

Acknowledgements

I am very grateful to one of my thesis advisors, Dr. Frederick Kin Hing Phoa, for what he had brought to me through this unforgettable journey. To be honest, I have thought for thousands of times that I might be unsuitable for pursuing a career of data scientist after I joined the Data Science Degree Program of NTU. Just during the darkest period that I lost my direction for numerous times, Dr. Phoa gave me a lot of recognition and took me into his field of study. No matter what trouble I stuck into in 2022, Dr. Phoa always encouraged me to keep fighting and suggested me on various kinds of solutions. I would say that being advised by Dr. Phoa is one of my greatest luck in these two years. Besides, I would like to also thank Dr. Pei-Yuan Wu for his recognition and approval to join his team. Though he was always busy on advising lots of students from his department during the whole year, he still expressed his consideration to my status and let me to report my progress on research to him. He is also thoughtful for me on my work load, so he even didn't assign any task for me to take care of in this year. And that allowed me to dedicate my whole self to my thesis. There are still a lot of people I met in NTU worth to be mentioned, but I'd like to keep them in my heart since I can't express my gratitude enough for them on lighting up my master student life. I wish all of them to be filled with happiness and achieve their goals in the rest of their lives.

doi:10.6342/NTU202300086

摘要

隨著 2019 年新型冠狀病毒的傳播帶來全球性的災害，科學家們開始意識到適合流行疾病分析的機器學習方法論之重要性。而在眾多相關演算法的發展之中，此研究著重在一個適合時間序列資料分析的分群法—「基於模型的遞迴分割法」之時間序列版。在這篇研究中，我快速地探討了幾篇和這個演算法的發展還有一些與本篇發想的方法論有關的文獻，嘗試優化前篇研究所提出之 R/Shiny 視覺化工具，並且提出了兩大優化這個演算法的方法論，其中包含了基於人口的標準化技巧，以及鄰近域相關特徵的衍生性概念。以台灣一年半的新型冠狀病毒每日確診數資料為研究對象，我不僅在應用此演算法及上述方法論的實驗中得到了一些流行相關的洞見，也透過了統計方法驗證了我的方法論的有效性。完整地分析過實驗後，我針對這個演算法的分群任務和預測任務都個別提出了相對應的建議參數設定。最後，針對開發過程和實驗過程中的限制，我列舉了幾項本研究未來可能可以優化的方向。

關鍵字：基於模型的遞迴分割法、新型冠狀病毒、視覺化工具、時間序列

doi:10.6342/NTU202300086

Abstract

The outbreak of COVID-19 raised awareness of the need to develop machine learning methodology suitable for epidemiological pattern recognition. This study focuses on time-series model-based recursive partitioning and one of its derivative applications for visualization. Several literature reviews on related studies including the evolution of this methodology are provided. This study documents the improved functionality enhanced user-friendliness of an R/Shiny application for visualizing time-series model-based recursive partitioning and proposes several methodologies to strengthen the computational results from this algorithm. With Taiwan township-level COVID-19 spread data, several experiments were conducted to gain insights on the spread patterns and validate the effectiveness of this study's main contributions, which include the population-based scaling technique and the derivative concept of adjacent domain-relevant attributes. Finally, various option configurations for clustering and forecasting tasks based on this epidemic-related algorithm are suggested. After considering the limitation of this work, several promising future directions are provided

doi:10.6342/NTU202300086

at the end of this study.

Keywords: model-based recursive partitioning, time series, Shiny, COVID-19

doi:10.6342/NTU202300086

Contents

doi:10.6342/NTU202300086

doi:10.6342/NTU202300086

List of Figures

doi:10.6342/NTU202300086

List of Tables

doi:10.6342/NTU202300086

Chapter 1 Introduction

1.1 Motivation

In late December 2019, a pneumonia of unknown cause characterized by the clinical manifestations of fever, cough, nasal congestion, fatigue emerged in Wuhan, China ([1, 2]). By January 7, 2020, Chinese scientists had isolated a novel coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2; previously known as 2019-nCoV), from these patients with virus-infected pneumonia, which was later designated coronavirus disease 2019 (COVID-19) in February, 2020, by WHO [3]. Several weeks later, a global epidemic of COVID-19 loomed and therefore led WHO to declare a Public Health Emergency of International Concern on 30 January, 2020. Since then, several critical fields including education, economy, and politics began to suffer from the great impacts of the virus. After nearly three years, the global environment is still unstable and undergoing unforeseen structural changes day by day. As of October, 2022, the world has experienced a large burden of morbidity and

doi:10.6342/NTU202300086

mortality. There were 7.1 million reported deaths and 17.6 million estimated deaths from COVID-19, as reported by the Institute of Health Metrics and Evaluation [4] and 63 million reported cases, as reported by the Coronavirus Research Center of Johns Hopkins University [5].

Although expected to be a country of high risk due to the proximity to mainland China, Taiwan rapidly responded in the very early period to establish an effective strategy against the spread of COVID-19. This early response successfully lowered the number of confirmed cases reported, which in turn significantly alleviated the harm brought by the pandemic in the comparison with other countries. For example, in the first five weeks of reporting the first confirmed case, the CECC, Central Epidemic Command Center, implemented a guide of approximately 120 action items including border control from the air and sea, technology-enabled case identification, and quarantine of suspicious cases, etc. [6]. However, as the virus continued to evolve, Taiwan finally encountered two surges of cases. The first one was in mid-May 2021, which mainly was linked to exposures in the Wan-Hua District of Taipei City. The cause of the surge was widely believed to be due to a change in border control policy in early April, as well as a subsequent COVID-19 outbreak at an airport hotel in Taoyuan City, the international airport city [7]. The second surge was from mid-April 2022, with the spread of the Omicron variant and the associated link with the strategy changing from

doi:10.6342/NTU202300086

Zero-COVID to Living-with-COVID. The cumulative reported confirmed cases in mid-November 2022 were about 8 million, with 13,645 reported deaths [5]. Taiwanese are still experiencing the long tail of COVID-19.

With the aim of lessening or eliminating the further influences from not only COVID-19, but from as yet unknown future epidemics, novel machine learning methodologies for investigating COVID-19 spread patterns need to be developed. Hence, in this study, we focus on a modern algorithm for time series pattern clustering and forecasting and attempt to improve the effectiveness of this methodology using the spread of COVID-19 in Taiwan as a case study.

1.2 Background

Since the global outbreak of COVID-19, data scientists have focused on developing machine learning methodology for recognizing COVID-19 spread patterns. As various studies have pointed out, the outcome of these models or analyses can help to construct more effective healthcare or political strategies to control the spread in time [8-13]. We review the general ideas presented in these studies below.

Several supervised forecasting methods, including Autoregressive Integrated Moving Average (ARIMA), have been applied. [14] conducted a relatively early study on the application of a simple econometric ARIMA model to predict the global

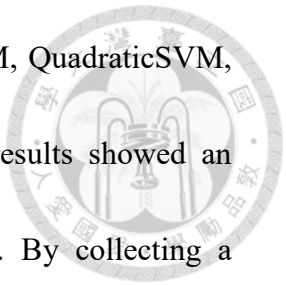
doi:10.6342/NTU202300086

prevalence and incidence of COVID-19. It depicted the potential trend with the assumptions that the virus would not develop new mutations. On the other hand, [15] narrowed down their scope to India and applied a univariate ARIMA to forecast daily cases for the next 50 days, while assuming no additional interventions. They also built a Nonlinear Autoregressive (NAR) neural network to provide a performance comparison. The results showed two incredible R^2 values of 0.95 and 0.97, respectively.

Random forest (RF) is also a widely-used traditional supervised algorithm in this field of study. To forecast the COVID-19 cumulative confirmed cases for the next one, three, and six days, [12] compared random forest with many other models, such as cubist regression (CUBIST), support vector regression (SVR), and stacking-ensemble learning, etc. Using evaluation metrics like symmetric mean absolute percentage error, the experiments revealed that RF was not the best choice but still effective. To assist medical decision-making, [16] used a semi-labeled large patient dataset from 146 countries and applied machine learning models such as RF to categorize patients with different levels of priority to be hospitalized. This study proposed several novel AI-based solutions to overcrowded health care systems around the world during the COVID-19 pandemic.

Another algorithm worth mentioning is Support Vector Machine (SVM). Not only did [16] applied SVM on their tasks, but [17] proposed a methodology that combines

doi:10.6342/NTU202300086

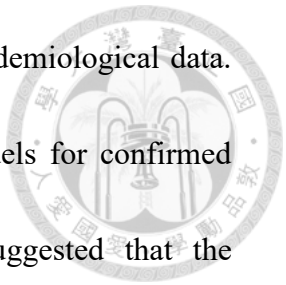


several machine learning classification models, such as LinearSVM, QuadraticSVM, etc., with digital signal processing for genome analyses. Their results showed an amazing 100% accuracy classifying COVID-19 virus sequences. By collecting a dataset of X-ray images, [18] integrated a deep learning model, ResNet50, with SVM to build an effective classification model that can help medical practitioners in diagnosing COVID-19 patients. The resulting classification model performed better than other models.

Exploratory clustering analyses using K-Means were conducted as well to examine the internal patterns of COVID-19 spread. In [19], K-Means clustering was applied to check the correlation of temperature and three stages of COVID-19, including suspended, confirmed, and death. The results validated the hypothesis that temperature was a significant factor and the addition of more potential attributes were expected in future work. [20] classified 155 countries with several attributes, such as metrics of air pollution and socio-economic status, using the K-Means algorithm. The study identified several COVID-19-related metrics by which countries could not be stratified, such as the number of confirmed cases, and the number of deaths and case fatality rate, respectively.

Except for the above-mentioned data-driven approaches, Long Short-Term Memory (LSTM) [21] has also been commonly used for prediction since it is designed

doi:10.6342/NTU202300086



to handle sequential data and thus is well-suited for modelling epidemiological data. [13] conducted assessments with several common prediction models for confirmed cases, deaths, and recoveries in ten major countries. They suggested that the Bidirectional LSTM (Bi-LSTM) and LSTM had better performance than the Gated Recurrent Network (GRU), SVR, or ARIMA. [22] focused on Canada and used LSTM to predict not only the trend but also the first stopping time point of COVID-19. A recent study [23] utilized LSTM, Bi-LSTM, and encoder-decoder LSTM (ED-LSTM) to do multi-step infection prediction in India, while pointing out challenges in forecasting due to the lack of reliable data.

Among these algorithms, model-based recursive partitioning [24] is attractive because it offers explainable results with key features useful for further analyses. Forecasting daily confirmed cases may also help in preparing hospital resources and finding a suitable time to implement strategies for the next stage of the pandemic. Furthermore, clustering target observations for further exploratory analyses allows experts from different domains to clarify the possible causes of the status quo and better prepare for similar pandemics.

1.3 Objectives and Organization

This research follows [8] in developing an interactive R/Shiny application for

doi:10.6342/NTU202300086

COVID-19 pattern recognition utilizing model-based recursive partitioning. The primary contributions of this study are:

(1) Doing a literature review on related studies and suggesting potential directions of improvement,

(2) Strengthening the proposed application by developing several useful functions and optimizing the interpretability of the statistical results,

(3) Modifying the composition of domain-relevant attributes used in the former studies based on research identifying features highly correlated with COVID-19 spread,

(4) Conducting a series of experiments on spatial summary features to check their effectiveness,

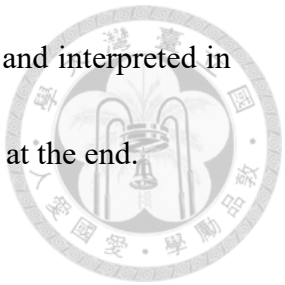
(5) Comparing the original standardization method with a new scaling method of population,

(6) Giving suggestions on option configuration of epidemic-related Model-based recursive partitioning methodology.

The remainder of the paper is organized as follows: Chapter 2 reviews several literature streams directly associated with the proposed methodology. Chapter 3 integrates and improves the proposed R/Shiny app for use in the experiments. Chapter 4 introduces experimental setups including data collection, feature selection and

doi:10.6342/NTU202300086

preprocessing, and experimental designs, and the results are shown and interpreted in Chapter 5. We conclude and present findings and suggest extensions at the end.



doi:10.6342/NTU202300086

Chapter 2 Literature Review

The approach analyzed here is based on the algorithm proposed by [24], which is called model-based recursive partitioning (MOB). We further adapt the method of single-step MOB tree, suggested by [25], that applies a specially designed form of linear regression as the fitting function for time series data. Moreover, we modify the interactive application from [8] to improve its user-friendliness. We then followed their study on high volume Taiwan COVID-19 spread data to conduct more extensive experiments to examine the possibility of better results, while also provide suggestions for improving the methodology from both above-mentioned studies. Due to the detailed experimental setups and the content of datasets, we review studies suggesting how to conduct data removal and missing value imputation.

2.1 Model-based Recursive Partitioning

We begin with the model-based recursive partitioning algorithm. This clustering

doi:10.6342/NTU202300086

algorithm is based on a greedy forward search to find an unstable variable with a split point that divides a group of data into two smaller, homogenous subgroups. This process repeats until a tree is formed that either reaches a maximum depth specified or cannot make any further improvement in evaluation metrics. In the resulting tree, all observations are separated into terminal nodes, where one terminal node represents a parametric statistical model that fits all observations in the corresponding subset. The advantages offered by this algorithm are: [25-27]

- Improved interpretability of classical trees
- Provides an exploratory way to assess variables
- Lower computation time versus other forecasting methods
- Stabler predictions compared to classical trees and more flexible predictions as compared to single models
- Ability to assess the extent to which model parameters are more stable across the levels of one or more covariates

The MOB algorithm consists of four main steps:

(1) Fit a parametric model to all observations of the current node by estimating the parameters through minimizing the objective function; either the method of ordinary least square (OLS) or maximum likelihood (ML) can be applied.

(2) Perform a generalized M-fluctuation test [28] on all partitioning variables to

doi:10.6342/NTU202300086

inspect overall instability and select the variable associated with the highest value; otherwise stop.

(3) Compute the split point(s), with respect to the selected partitioning variable in the last step, that locally optimizes the objective function, either for a fixed or an adaptive number of splits.

(4) Split the current node, and repeat the above steps for each resulting node.

To sum up, the model-based recursive partitioning fits a local statistical model to different subgroups of observations, then splits them based on parametric instability to create smaller homogeneous subgroups. The algorithm is a special type of classification and regression tree and can be therefore easily visualized for further analysis.

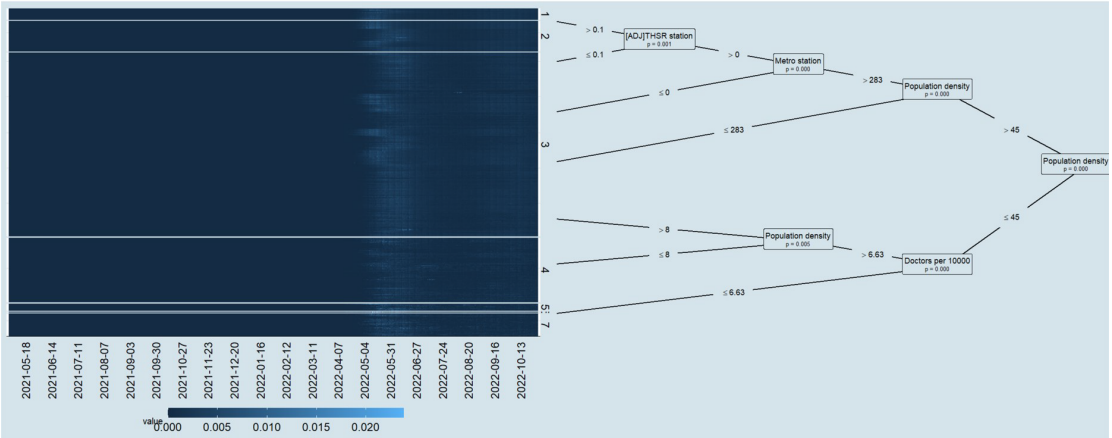


Figure 2-1: Example of MOB tree for time series data

doi:10.6342/NTU202300086

2.2 Time Series Clustering using Domain-Relevant Attributes

Domain-relevant attributes (DRA) are often described as cross-sectional features of time series data that conceptually link them into subgroups. For example, when dealing with a dataset of highway sections to forecast traffic flow, domain-relevant attributes may include the length of sections, or information about whether the sections pass a service area, etc. After discovering that most studies analyze domain-relevant attributes only in a post-hoc fashion, and identifying the limitations of applying ARIMA to MOB trees [29], [25] proposed a solution that captures linear trend, additive seasonality, and autocorrelation as well as incorporating information provided by domain-relevant attributes into the clustering process by fitting linear regression in MOB trees. The model is shown as:

$$y_t = \alpha_0 + \alpha_1 trend + \beta_1 Season_{1t} + \beta_2 Season_{2t} + \cdots + \beta_m Season_{mt} + \gamma_1 y_{t-1} + \gamma_2 y_{t-2} + \cdots + \gamma_m y_{t-m} + \varepsilon_t \quad 2\text{-}1$$

where the subscript $\boldsymbol{t}$ means time and subscript $\boldsymbol{j}$ (j = 1,2,…,m) means season, the y_t represents the series value at time t, $Season_{jt}$ is 1 when time t is in season j, and y_{t-k} therefore implies the kth lagged value. While trying to capture multiplicative

doi:10.6342/NTU202300086

trend and seasonality, we can change the left-hand side y_t to $\log(y_t)$. After we fit this model into a MOB tree with domain-relevant attributes as partitioning variables, we obtain several subgroups of series data, which differ by temporal behavior and attribute profile. See Figure 2-1 for an example.

2.3 An Interactive Approach to Time Series Clustering

Using Taiwan COVID-19 spread patterns, [8] developed an interactive tool based on the proposed algorithm in the last section with the goal of providing more analytical insights for domain experts. In this research, the main algorithmic solution proceeds as follows:

(1) Separate outlier time series, which includes “all zero” values

(2) Normalize the series by group by subtracting the mean and dividing the standard deviation

(3) Run the MOB tree on the series with Eq. 2-1

(4) Stop the tree by best improvement on mean square error, tree simplicity, AIC [30], or BIC [31]

The collected data is composed of 183 daily series involving the level of city, township, and districts with a date range from January 1 to June 10 of 2021. Features

doi:10.6342/NTU202300086

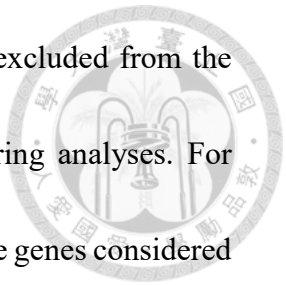
include region, administrative level, population, and two Boolean variables, imported, and airport. The model used is a linear trend with six seasonal dummies (Sunday to Friday), and lags 1 to 7.

Since domain-relevant attributes can be used for further valuable analyses, we would like to find other features potentially associated with COVID-19 spread in Taiwan. However, standardization with mean and standard deviation may sometimes lead to data misrepresentation. As [32-34] suggested, normalization of COVID-19 series by population can avoid the potential differences in the absolute number of cases reported among communities and make them comparable. Other studies related to clustering methods or statistical/ML methods on COVID-19 also used infection rates instead of total number of cases or standardized number of cases [19, 35-37]. Therefore, in this study, we replace the original scaling method with the infection rate. Based on these two points of potential improvement, we expect to gain better resulting MOB trees that provide more interpretable and valuable insights.

2.4 Data Removal and Missing Value Imputation

Removing unimportant regions or clusters has always been challenging for data clustering algorithms [38]. By deleting several observations, machine learning

doi:10.6342/NTU202300086



approaches are expected to perform better since noisy signals are excluded from the training process, allowing focus on the individuals of interest during analyses. For example, in [39], they compared different filtering methods to remove genes considered irrelevant from the set to be analyzed in several genetic network analyses. They even showed the negative effects of not filtering for comparison, revealing an obvious bias in the resulting cluster analysis. On the other hand, just as [20] mentioned in one of their limitations, finding consistent and compatible data for all individuals is not always easy due to a wide variety of reasons. Hence, under reasonable assumptions, we remove some of the observations that were either not strongly linked with the regions we focused on or lacked compatible and reliable information for domain-relevant attributes. These pre-processing steps provide a less biased and more insightful analysis.

According to the discussion above, missing values are also challenging for machine learning methods. In this study, some weather-related domain-relevant attributes can exhibit missing values due to temporary shutdown or maintenance of weather stations. A well-studied geo-statistical approach called Kriging Interpolation [40] is suitable in dealing with this kind of problem. [41] presented a concrete example of applying Ordinary Kriging to predict time-series of air temperature data for the purpose of filling in missing values. The result was that this method is capable of imputing missing values. Recently, [42] proposed a novel methodology for spatio-

doi:10.6342/NTU202300086

temporal data imputation, which is called CUTOFF. In their study, a procedure of cross-validation was applied to optimize the parameters. Then, in a comparison of analyses of rainfall data, they proved their method was significantly better than other four well-known imputation methods, including K-Nearest Neighbors, Singular Value Decomposition, Multiple Imputation, and Random Forest. Though there are too many effective imputation methods to review here, we decided to apply a relatively intuitive approach, which is to impute the mean values of all neighboring townships. While there is likely a slight bias compared to other more complex methods, we save the costs of having to provide proof of the assumptions for the different means. This is in fact a more conservative choice as long as there are not too many missing values and one cannot completely check whether one's data fit the assumptions of each imputation method.

doi:10.6342/NTU202300086

Chapter 3 Interactive Programming by R/Shiny Apps

3.1 Introduction to R/Shiny Apps

When it comes to the ability of explaining exploratory analyses, a powerful visualization can concisely illustrate computational results and offer additional opportunities for discovering valuable insights. R/Shiny [43] enables developers with limited coding experience to design interactive and powerful web applications that easily integrate R's statistical graphs. In past decades, numerous studies have used Shiny to provide user-friendly solutions for domain experts and non-experts to interpret analytical insights efficiently. As an example of several previous works, [44], as a contribution to the bioinformatics field, created shinyCircos for the circular visualization of genomic data. In the field of education, [45] developed a Shiny application as teaching tools for statistics. Finally, in the field of research methodology, [46] proposed ROBVIS for visualizing risk-of-bias assessments, which helps researchers to perform systematic reviews.

doi:10.6342/NTU202300086

3.2 The Use of R/Shiny Apps in Previous Work on COVID-19 Research

There are many valuable applications based on R/Shiny worth mentioning, and thus it is important to apply an interactive visualization framework to data mining associated with COVID-19. To list just a few examples, several related tools were proposed such as COVID-19 tracker [47], the metaCOVID [48], and COVID19-world [49], etc. Among them, [8] developed a tool featuring great visualizations for presenting time-series MOB tree structures and related subgroup information with the use of the R package proposed by [50]. For more details on this Shiny application, the user interface is shown below. Once the time series data is uploaded and several parametric settings are selected via the interactive side tab, the tool instantly initiates partitioning computation and provides one MOB heatmap of the original series order plus its branch plot, another one showing the MOB heatmap of seasonality plus its branch plot, a series plot with average line for each cluster, and a coefficient plot comparing OLS models to check the similarities and differences in terms of parameters. These figures are referred to Figures 1-4 in [8]. Although this application enables us to easily manipulate data and see the corresponding results, we aim to optimize the user interface, user experience and add new functions to the tool, with the objective of maximizing interpretability.

doi:10.6342/NTU202300086

3.3 The Improved Version - Interpretable Pattern Recognition Software

The proposed version was named “Interpretable Pattern Recognition Software,” and a screenshot is given in Figure 3-2. The first improvement was to design an external JSON file called “variable_dict.json” to save information about domain-relevant attributes. Since data scientists or engineers may use shorter aliases for column names (names of domain-relevant attributes) or attribute values, the saved information would be used in two places to avoid misunderstanding results:

(1) The displayed attribute names on the side tab for user selection would be replaced by the real whole variable names

(2) The categorical values displayed on the branch plot would be replaced by their real value names.

(3) For ease of comparison, we took the same data and parametric settings as with [8] (see Figure 1 in [8] and Figure 3-1).

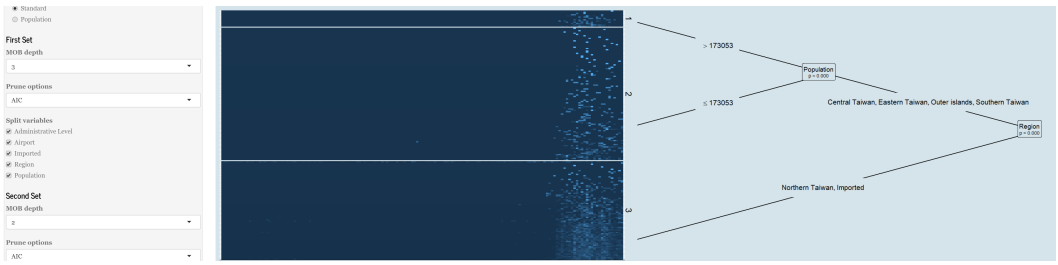


Figure 3-1: UI and branch plot with data from [8]

doi:10.6342/NTU202300086

Secondly, to provide more useful information with the part of series plots, instead of all lines and an average line, we used the suggestions from [25] and constructed a red average line, a blue median line, and shaded inter-quartile range since daily outliers may sometimes destroy the shape of series and cause hurt the interpretation of results.

Last, to offer a modern-style design, we applied the extension package "shinythemes" [51] to the original UI. Adding a journal-style and Economist-style themes could make the visualization even more satisfying. See Figure 3-2 for details.

doi:10.6342/NTU202300086

Figure 3-2: UI of Interpretable Pattern Recognition Software

doi:10.6342/NTU202300086

Chapter 4 Experimental Studies

In this study, we collected the Taiwan township-level COVID-19 data from May of 2021 to October of 2022 from [52]. For spatial summary features, we obtained a township borders dataset from [53]. To include population information for the infection rate, we also gathered a dataset from [54]. Unless specified, all datasets in this study stem from December 2021. Please refer to Table 4-1 for more details on other attributes.

4.1 Origin of Data Feature

[33, 55, 56] suggested that both daily temperature and relative humidity play key roles in predicting COVID-19 daily cases or mortality rates in several communities of Germany/China. [32] also discovered that during a multi-prefecture study in Japan, temperature, humidity, and population density all have great influences on COVID-19 spread. [11], on the other hand, proposed the concept that age structure of population is important for COVID-19 transmission. For mobility, [57] concluded that reductions in

doi:10.6342/NTU202300086

travel rate between communities can play a decisive role in stopping the outbreak, and [58] suggested that all connected communities, no matter directly or indirectly, have great risk of outbreak. Lastly, in several other studies that are systematically reviewed in [59], hospital resources have certain kinds of association with COVID-19 spread pattern.

As Tobler [60] proposed the first law of geography -- "Near things are more related than distant things," machine learning researchers have been making efforts on the field of spatial feature engineering, which is defined as the process of developing additional information from raw data using geographic knowledge [61] on various mining tasks. Among lots of methods, spatial lag is of interest and was described as the mean value of all neighboring areas. To the best of our knowledge, there are few studies on epidemic pattern recognition using spatial lags of original dependent attributes as derivative attributes. Therefore, this study examined the effectiveness of a modified version of spatial lags of domain-relevant attributes on clustering and forecasting with time-series-related MOB tree, using Taiwan COVID-19 spread pattern as a case study. To integrate all neighborhood information and their own information at a time, the proposed spatial features calculated a mean of their own values and all neighborhood values. For example, if Town A (attribute Z_1 = 5) was adjacent to Town B (attribute Z_1 = 11) and C (attribute Z_1 = 2), then a proposed feature ${}_{adj}Z_1$ would be 6 in Town A. Since the nature

doi:10.6342/NTU202300086

of this spatial summary feature is based on adjacent townships in this study, we named it “adjacent domain-relevant attribute” (ADRA).

4.2 Data Preprocessing

Among the 368 townships in Taiwan, we removed 11 (listed in Table 4-2) for two main reasons. (1) The lack of manned or automatic weather stations and physically connected neighboring townships increased the difficulty of imputation in terms of weather attributes. (2) The proximity to other islands that do not belong to Taiwan is greater than to Taiwan, which may bring about unconsidered interaction effects that affect the spread pattern.

For temperature and relative humidity, we refer to the weather station data in Taiwan [62]. If there is more than one station in the same area, we take the mean of non-null values to represent the monthly average. Also, when dealing with lack of data (which may be due to the absence of station or station temporary shutdown), we gathered all the non-null values from neighboring townships of the same period to calculate the mean for representing the monthly average. This simple interpolation approach reduced the numbers of townships with missing data from 59 to 4.

For the number of doctors, all 3 districts in Hsinchu City were merged into only

doi:10.6342/NTU202300086

one regional unit; a similar process was used in all 2 districts in Chiayi City. Such a merge originated from their shared same zip codes and their short separation in 1990 only. To properly address this issue, we decided to share the values for each district involved since the demographical characteristics are similar.

For the airport part, we separated the observations into 3 levels: without an airport, with a domestic airport (i.e., no international flights), and with an international airport. The reason is that we considered airports with any domestic flights to be an important transmission path for COVID-19 within Taiwan, but airports with any international flights bring greater threats since they involve inter-country transmission.

For the metro part, although there are 7 systems in Taiwan, to focus on the effect of mobility, we only considered 4 main systems that have higher daily passenger capacities, which are the Taipei Metro Network (TRTC), Taoyuan Metro Network (TYMC), Taichung Metro Network (TMRT), and Kaohsiung Metro Network (KRTC).

4.3 Experimental Designs

In this study, the experiments were mainly conducted to examine 3 aspects:

(1) The effectiveness of adjacent domain-relevant attributes

(2) The effectiveness of population-based scaling methods

doi:10.6342/NTU202300086

(3) The differences between AIC/BIC pruning criteria

Therefore, in this $2 \times 2 \times 2$ primary experiment, we adopt every set of settings to create 9 MOB trees from depth 1 to 9 and calculate the clustering MSE and forecasting adjusted MAE. We expect to examine the differences in either the tree structure (i.e., the partitioning variables used or the splitting points in each node) or the performance metrics (i.e., the MSE or MAE). To confirm the statistical significance, several paired t-tests using clustering MSE were applied on trees under same parametric configurations, except the one that aims for comparison. Besides analyzing a single parametric setting at a time, we found the interaction effectiveness between the different proposed settings and thereafter suggested various option configurations for different tasks of this algorithm. We also created township cluster distribution maps on Taiwan to further visualize the cluster distribution.

doi:10.6342/NTU202300086

Table 4-1: Attributes in experiments

Feature	Type	Description	Values Type (Values)	Sources
Population density	Demography	Population density of the district (people/km2)	Numerical	[54]
Dependency rate	Demography	Population aged between 15-64 divided by aged beyond 65	Numerical	[63]
Aging index	Demography	Population aged beyond 65 divided by aged below 15	Numerical	[63]
Temperature	Weather	Mean of monthly average temperature from 2021/05-2022/10	Numerical	[62]
Relative humidity	Weather	Mean of monthly average relative humidity from 2021/05-2022/10	Numerical	[62]
Number of doctors	Hospital resources	Number of doctors per 10 thousand people	Numerical	[64]
International airports	Transportation	Whether the township has an international airport	Categorical (0:No; 1:Yes)	[65]
Domestic airports	Transportation	Whether the township has a domestic airport	Categorical (0:No; 1:Yes)	[65]
Metro	Transportation	Whether the township has a metro station	Categorical (0:No; 1:Yes)	[66]
THSR	Transportation	Whether the township has a Taiwan HSR station	Categorical (0:No; 1:Yes)	[66]
Transfer station	Transportation	Whether the township has a public transport interchange	Categorical (0:No; 1:Yes)	[67]

doi:10.6342/NTU202300086

Table 4-2: Townships removed from this study

County	Township
Penghu County	Cimei Township
Kinmen County	Jincheng Township
Kinmen County	Jinhu Township
Kinmen County	Jinsha Township
Kinmen County	Jinning Township
Kinmen County	Lieyu Township
Kinmen County	Wuqin Township
Lienchiang County	Beigan Township
Lienchiang County	Nangan Township
Lienchiang County	Juguang Township
Lienchiang County	Dongyin Township

In the forecasting experiments, we performed forecasting for the next 7 days (2022/11/01-2022/11/07) to compare the prediction adjusted MAE, so as to validate whether our proposed scaling method is better. Since the MAE scales would be different for standardization daily cases with regional mean/standard deviation and for that with regional population, we adjusted them by scaling back to the original series scale (i.e., the same as the raw daily confirmed case reports). Like the clustering tasks, to validate the statistical significance, several paired t-tests using adjusted forecasting MAE were applied on trees under the same parametric configurations, except the one that aims for comparison. For all tests included in this study, the confidence levels were set to 0.95.

doi:10.6342/NTU202300086

Also, to in order to apply paired t-tests to check to the significance of difference in trees under settings of interest, we assume the distribution of differences to be normal.

Finally, to further check the differences in effectiveness with respect to the previous study, [8], we extended the clustering and forecasting experiments to compare with their methods. To be fair, we modified the data by removing the observations outside the 357 selected townships, retaining the features excluded “Imported”, and collecting the data that suit our training and test periods (2021/05/01-2022/10/31 and 2022/11/01-2022/11/07 respectively). Tree depths were picked at 6 and 7, which were reasonable based on our experience in this study.

doi:10.6342/NTU202300086

Chapter 5 Results and Interpretations

The experiment results are organized in Table 5-1 and Table 5-2. All tables related to clustering on population-based scaling show the MSE value scaled by 10^{10} for a better display. For detailed experiment results, please refer to Appendix A. To simplify the interpretation of the trees, we selected several MOB trees that performed well enough with suitable complexity and are often mentioned. The aliases for these trees

Table 5-1: MSE of Clustering Experiments

Scaling Method	Adjacent domain-relevant attributes			
	Without		With	
	Pruning Option		Pruning option	
	AIC	BIC	AIC	BIC
Standard	0.1329	0.1330	0.1324	0.1329
Population	1066.3775	1068.0249	1061.9860	1063.6214

Table 5-2: Adjusted MAE of Forecasting Experiments

Scaling Method	Adjacent domain-relevant attributes			
	Without		With	
	Pruning Option		Pruning option	
	AIC	BIC	AIC	BIC
Standard	36.953	35.239	35.320	34.214
Population	13.396	13.439	13.214	13.216

doi:10.6342/NTU202300086

Table 5-3: Parametric details of selected MOB trees

Tree alias	Depth	Pruning Option	Scaling method	Inclusion of adjacent DRA
Tree 1	5	AIC	Standard	No
Tree 2	5	AIC	Standard	Yes
Tree 3	9	AIC	Standard	No
Tree 4	9	AIC	Standard	Yes
Tree 5	9	AIC	Population	No
Tree 6	9	AIC	Population	Yes
Tree 7	5	AIC	Population	Yes

and their parametric configurations are provided in Table 5-3 and a list of numbers of observations in each cluster with respect to selected tree is provided in Table 5-4.

5.1 Effectiveness of Adjacent Domain-Relevant Attributes

Starting with one of the most important findings, when focusing on either scaling method, the adjacent domain-relevant attributes were used at an early stage of depth 2, and due to the evolutionary characteristics of the MOB trees led by their greedy forward search nature, the produced trees afterwards were totally different in structure. Taking

doi:10.6342/NTU202300086

the selected trees at depth 5 as an example, we observe the differences in Figure 5-1 and Figure 5-2, where spatial summary features on population density

Table 5-4: Number of townships in each cluster of several selected trees

Cluster number	Tree 1	Tree 2	Tree 7
1	3	47	1
2	25	66	2
3	30	30	58
4	49	12	91
5	25	1	59
6	22	3	72
7	12	1	7
8	14	7	8
9	59	10	2
10	69	73	1
11	14	83	21
12	35	1	5
13	--	4	23
14	--	9	14

([ADJ]Population density) were considered rather than regular features on population density (Population density) at depth 1, thus creating different clustering results (See Table 5-4, Figure 5-3). Taking a deeper look at the branch plots, we observe that several critical domain-relevant attributes were included in both trees, regardless of whether they were spatial features. To further investigate whether the top 3 attributes changed

doi:10.6342/NTU202300086

with respect to the involvement of adjacent domain-relevant attributes, we created Table 5-7 based on depth 9 of different scaling methods (Trees 3 to 6) to see the changes in most-used splitting variables. In this table, spatial information was excluded, e.g., adjacent population density was counted as regular population density, since we only focused on the essential domain-relevant attributes. As we can see, though compositions of feature on Tree 3 and Tree 4 were similar, those on Tree 5 and Tree 6 were quite different. While we can observe that population density and features related to age structure were always critical to Taiwan COVID-19 spread clustering, several features such as those related to weather or medical resources were taken into consideration when adjacent domain-relevant attributes and population-based scaling were involved. We concluded that the inclusion of geographical information affected the following splitting rules and thus the partitioning results we cared about.

In addition to the tree structure, Table 5-4 suggests that trees splitting with adjacent domain-relevant attributes are likely to have unequal-sized clusters. This may be a great benefit if some precise regional strategies against epidemic can be implemented in specific townships, no matter what the sizes of subsets are. Also, to further support this assumption, we notice that the epidemic is often transmitted through township borders. Similar prevention or control policies in the neighborhood regions can jointly stop the spread of disease more effectively. As shown noticed in Figure 5-3, due to the nature of

doi:10.6342/NTU202300086

adjacency, trees splitting with adjacent domain-relevant attributes seem to have more neighborhood townships in each cluster while the distribution map of the tree without them seem to be more fragile, somehow making it harder for further regional analyses and decision making.

Last, from the MSE values given in Table A-1 and Table A-2, we discovered that, although there were nearly no differences in small trees, adjacent domain-relevant attributes helped MOB sub-models to fit better in larger MOB trees. Also, the AIC values in Table 5-5 reveal stability towards significant difference on involvement of adjacent domain-relevant attributes. Hence, we conclude that integrating geographical information improves partitioning with the AIC pruning option.

Table 5-5: P-values from paired t-test for clustering comparison of involvement of ADRA

Scaling Method			
Standard		Population	
Pruning option		Pruning Option	
AIC	BIC	AIC	BIC
0.0774**	0.4147	0.0849**	0.0806**

Significant at the * 15% level; ** 10% level; *** 5% level; **** 1% level.

Table 5-6: P-values from paired t-test for forecasting comparison of involvement of ADRA

Scaling Method			
Standard		Population	
Pruning option		Pruning Option	
AIC	BIC	AIC	BIC
0.1494*	0.3370	0.1572	0.03838**

Significant at the * 15% level; ** 10% level; *** 5% level; **** 1% level.

doi:10.6342/NTU202300086

Regarding forecasting, we see a slightly lower average MAE in trees with adjacent domain-relevant attributes (See Table 5-2). Moreover, the differences in tree pairs under same settings are clear when the tree depth increases (See Table A-3, Table A-4). Referring to Table 5-6 for results on comparative tests, we noticed that standard scaling with BIC pruning again showed no differences in paired performance. We thus further suggest other parametric settings (i.e., AIC) to improve forecasting results with the addition of adjacent domain-relevant attributes, especially for deep trees.

Table 5-7: DRA frequency and ranking in Trees 3~6

Tree	Top 1 attribute	Top 2 attribute	Top 3 attribute
Tree 3	Aging index (9)	Population density (7)	Dependency rate (6)
Tree 4	Population density (19)	Aging index (8)	Dependency rate (7)
Tree 5	Population density (15)	Aging index (10)	Dependency rate (10)
Tree 6	Population density (22)	Average relative humidity (8)	Aging index/Number of doctors per 10000 (8)

5.2 Effectiveness of Population-Based Scaling Methods

We observe Tree 2 and Tree 7 to verify the differences between scaling methods. As the set of branches showed in Figure 5-4, we noticed that it was different from what

doi:10.6342/NTU202300086

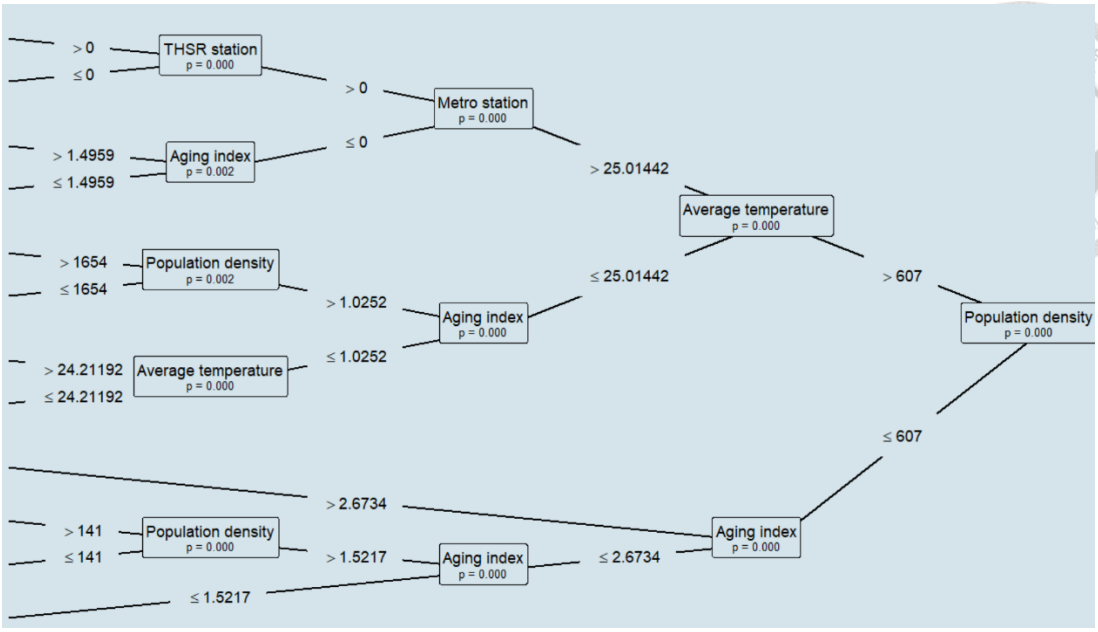


Figure 5-1: Branch plot of Tree 1

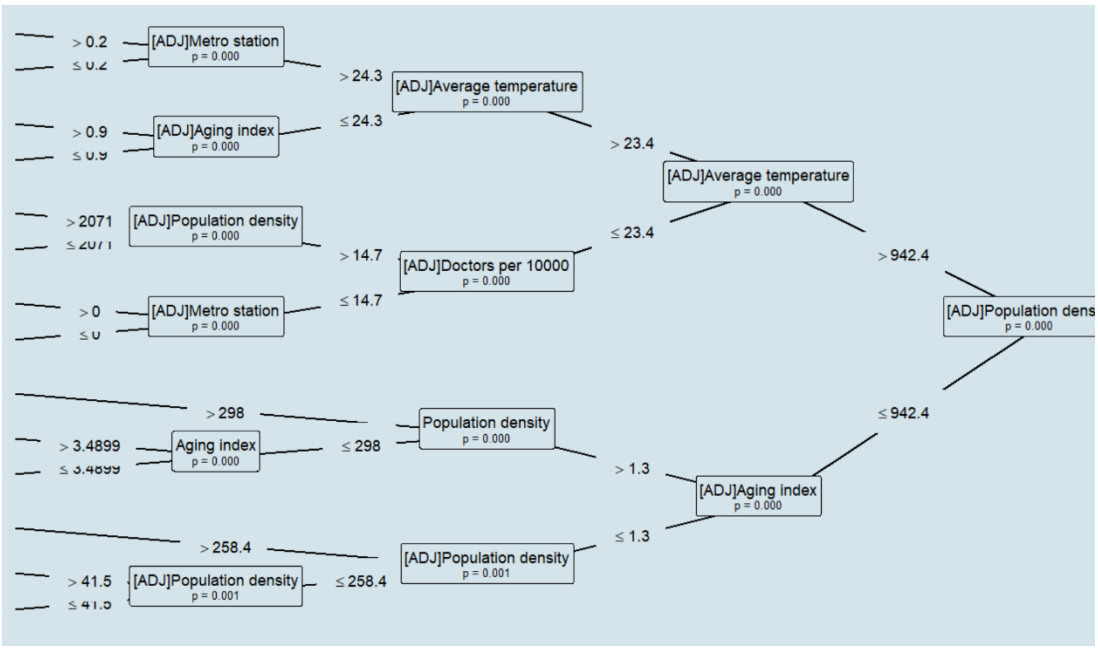


Figure 5-2: Branch plot of Tree 2

was shown in Figure 5-2. While Tree 2 emphasized mix of various types of adjacent domain-relevant attributes, including population density, age structure, transportation, medical resources, and weather, Tree 7 has adjacent population density as a stem attribute, as the first two depth relied only on it and there were multiple existences in the whole tree (nearly 50% of all splitting variables), and several other attributes such

doi:10.6342/NTU202300086

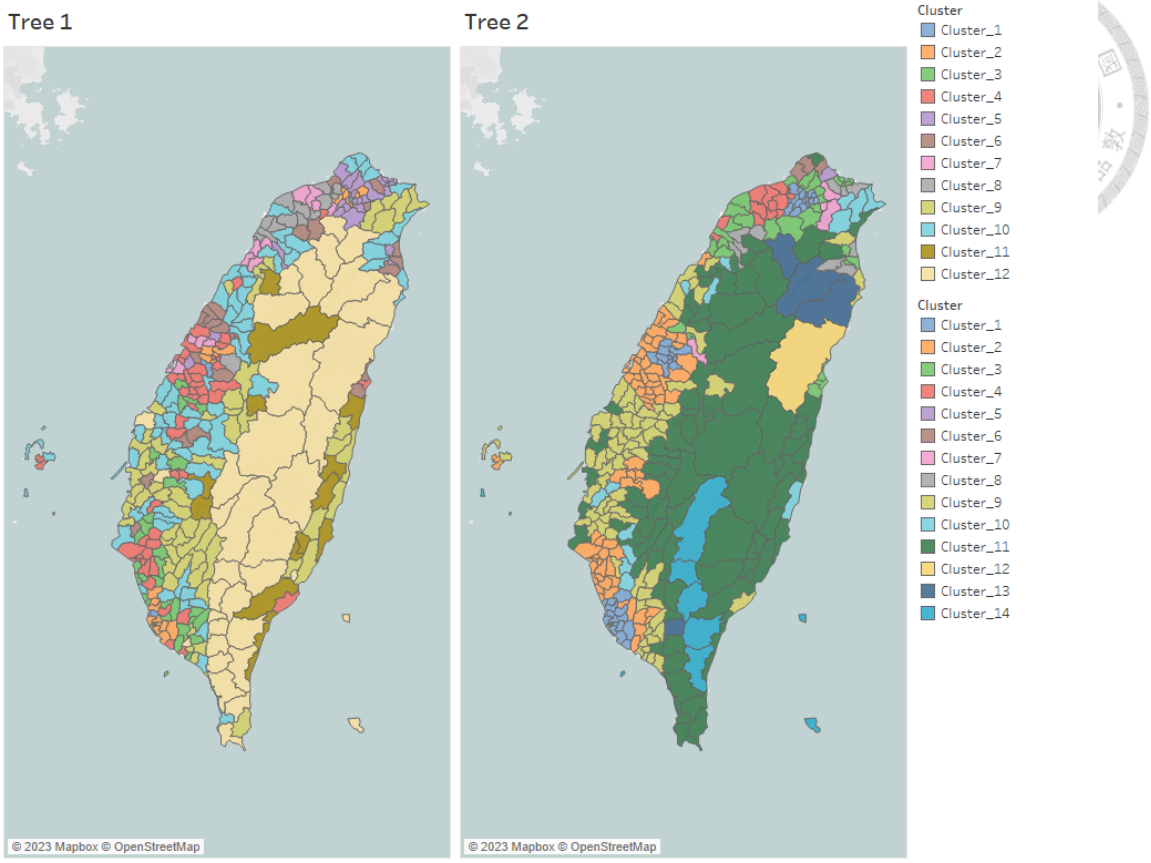


Figure 5-3: Distribution maps for Tree1 versus Tree2

as age structure (2 times), temperature (1 time), and, the second most important, the transportation (30% of all splitting variables). This phenomenon could also be discovered in Table 5-7, where Tree 6 has reached 22 times in existence of attributes related to population density, which was nearly 30% of all splitting variables. Also, the composition of the list of top 3 attributes was distinguishable from others due to the involvement of attributes related to weather and medical resources. To sum up, the population-based scaling method not only generated a different set of branching rules, but the branching rules were highly associated with population-density-related domain-

doi:10.6342/NTU202300086

relevant attributes, as well as their integration with attribute types other than demography.

From Table 5-4, we cannot see any difference in cluster sizes between Tree 2 and Tree 7, since they were all unevenly distributed. However, after drawing the distribution maps (Figure 5-5), we noticed that the clusters in population-based method MOB Trees were easier to identify because they fit the regional characteristics of Taiwan. For example, Cluster 1 contained only a region with one small domestic airport, while Cluster 2 involved two regions with the second and third largest international airports. Cluster 3 focused on several high-risk areas, such as most of Taipei City, that were around the first two clusters, and Cluster 4, 5, and 6 surrounded the third cluster with different regional features. Rather than generating fragile patterns in some metropolitan areas, the diffusion effects were clearly initiated from several critical points, then finally expanded to all Taiwan townships. This diffusion phenomenon helped policy makers to take measures such as hierarchical containment controls.

Table 5-8: P-values from paired t-test for forecasting comparison of scaling method

Adjacent domain-relevant attributes			
Without		With	
Pruning option		Pruning Option	
AIC	BIC	AIC	BIC
<0.0001****	<0.0001****	<0.0001****	<0.0001****

Significant at the * 15% level; ** 10% level; *** 5% level; **** 1% level.

doi:10.6342/NTU202300086

To compare the results MOB trees built on different scales, we refer to the adjusted MAE in Table 5-2, since they were on the same scales. Here, key facts are the great differences in predictability between these two scaling methods. Population-based scaling exhibited greater forecasting power than standard scaling with statistical significance (Also check Table 5-8). Hence, we suggest the application of population-based scaling rather than mean-based scaling method in forecasting tasks.

Table 5-9: P-values from paired t-test for clustering comparison of pruning options

Adjacent domain-relevant attributes			
Without		With	
Scaling Method		Scaling Method	
Standard	Population	Standard	Population
0.0727**	0.0783**	0.0420***	0.1140*

Significant at the * 15% level; ** 10% level; *** 5% level; **** 1% level.

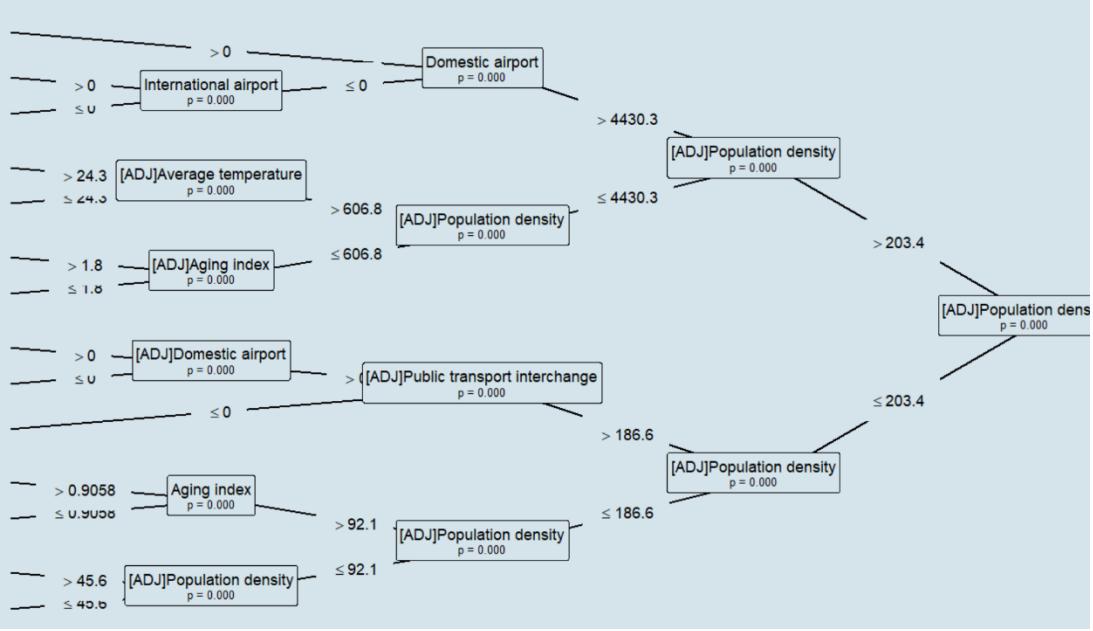


Figure 5-4: Branch plot of Tree 7

doi:10.6342/NTU202300086

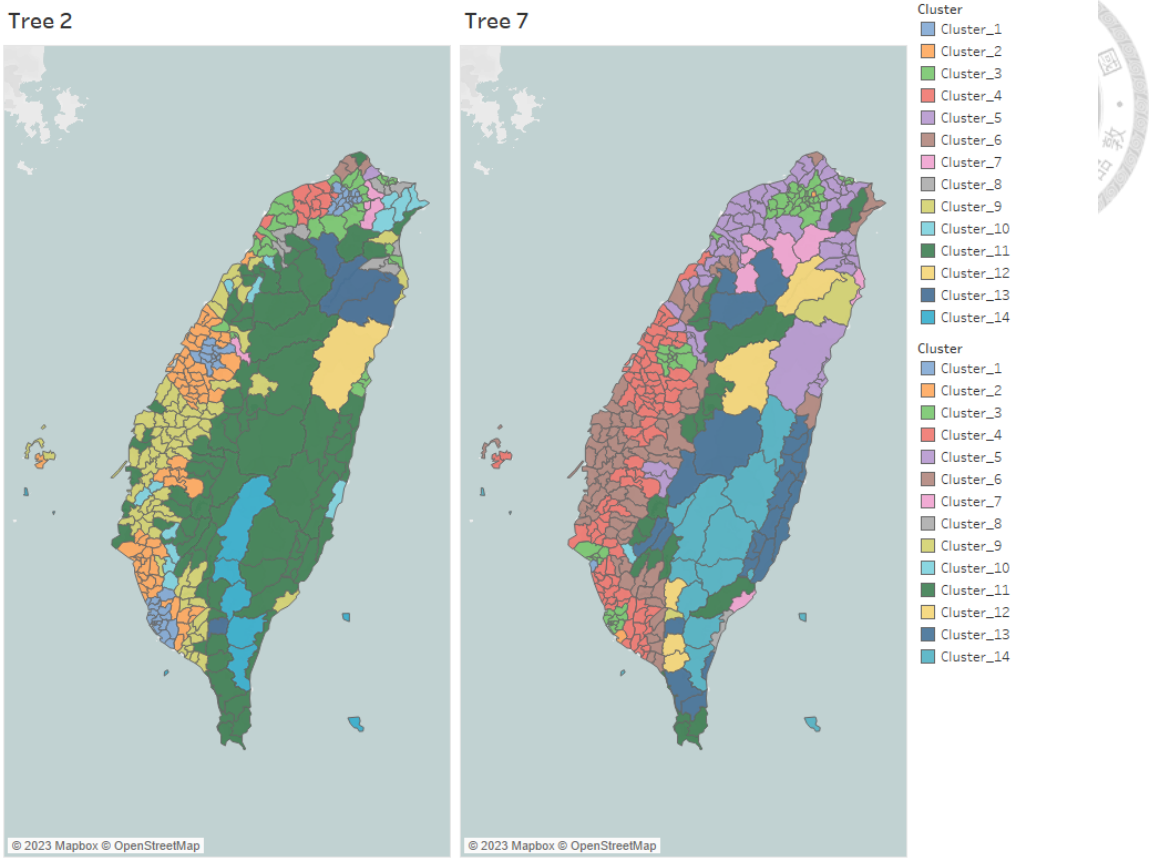


Figure 5-5: Distribution maps of Tree 2 versus Tree 7

5.3 Differences between Different Pruning Criteria

As we observe in Table A-1 and Table A-2, a tree with the AIC pruning option always produced an identical or lower clustering MSE value than another with the BIC pruning option under same settings. After seeing Table 5-9 to confirm the statistical significance, we therefore suggest that AIC did a better job in paired clustering MSE, so we take only AIC as the controlled pruning option for this chapter. In addition to the above-mentioned property, we also noticed that the BIC pruning option stopped MOB

doi:10.6342/NTU202300086

Table 5-10: P-values from paired t-test for forecasting comparison of pruning options

Adjacent domain-relevant attributes			
Without		With	
Scaling Method		Scaling Method	
Standard	Population	Standard	Population
0.0861**	0.1221*	0.3985	0.9865

Significant at the * 15% level; ** 10% level; *** 5% level; **** 1% level.

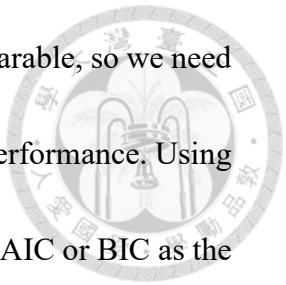
trees from growing after depth 6 under the standard scaling method with adjacent domain-relevant attributes. After considering the BIC pruning option mentioned in the comparative test results in Sec. 5.1, we conclude that AIC is more stable in generating clusters.

In the forecasting experiments (Table 5-2), we observed that trees with the BIC pruning option showed slightly better performance than those with the AIC pruning option under the standard scaling method. When the population-based scaling method was applied, trees showed nearly no differences in predictive performance between pruning options. Table 5-10 displays the statistical significance. Therefore, we recommend using the population-based scaling method to avoid the issue of deciding which pruning option to apply when performing clustering and/or forecasting tasks.

5.4 Suggested Option Configuration

According to the above analyses of experimental results, we propose the following suggested option configurations for using this methodology in different application

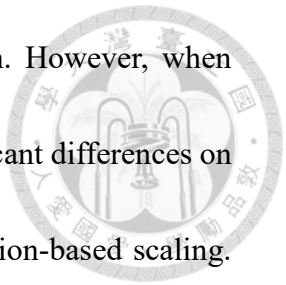
doi:10.6342/NTU202300086



scenarios. For clustering tasks, different scaling methods are incomparable, so we need to find a stable option configuration that shows potential for better performance. Using the information in Table 5-9, we encounter the dilemma of choosing AIC or BIC as the pruning option. Yet the configuration of population-based scaling with adjacent domain-relevant attributes showed relatively not significant difference in performance between AIC and BIC, so we can further overcome the trouble of choosing pruning option by using this configuration. Secondly, from Table 5-1 and Table 5-5, we verify that under population-based scaling, involvement of adjacent domain-relevant attributes performs better, and the differences in results from lack of adjacent domain-relevant attributes are roughly equally important for either pruning option of AIC or BIC. Therefore, we conclude that the configuration of population-based scaling with adjacent domain-relevant attributes and either pruning option, AIC or BIC, is suitable for clustering using our proposed methodology.

On the other hand, for the task of forecasting, we need to recall that population-based scaling performed better than standard scaling (See Table 5-2, Table 5-8). Using the same logic from the clustering tasks, we initially preferred a configuration that generated more stable performance without considering any pruning option. From Table 5-10, we see the involvement of adjacent domain-relevant attributes did not lead to any statistically significant difference between the pruning options, so we intend to

doi:10.6342/NTU202300086



further use this property to build a desirable option configuration. However, when moving to Table 5-6, we noticed that BIC can stably produce significant differences on the utilization of adjacent domain-relevant attributes under population-based scaling. Therefore, we suggest using a configuration of population-based scaling with adjacent domain-relevant attributes and the BIC pruning option.

5.5 Comparison with the Previous Study

By conducting clustering and forecasting experiments using the methodology of the previous study, we can compare their performance with our methodology and obtain the following results. Note that since the values of clustering MSE from different scaling methods are incomparable, we do not use the suggested configuration from the last section. Instead, a similar configuration of standard scaling with adjacent domain-relevant attributes and the pruning option of AIC was applied. As we can see in Table 5-11, their clustering performance was worse than ours using the standard scaling method. In Table 5-12, we use our suggested configuration for forecasting. Obviously, the application of population-based scaling had better predictive ability again the methodology of the previous study. Therefore, we conclude that our methodology yields better results.

doi:10.6342/NTU202300086

Table 5-11: Comparison of clustering MSE with previous study

Depth	Previous study	Our methodology
6	0.1322	0.1319
7	0.1322	0.1311

Table 5-12: Comparison of forecasting adjusted MAE with previous study

Depth	Previous study	Our methodology
6	32.789	13.174
7	32.789	13.419

doi:10.6342/NTU202300086

Chapter 6 Discussion and Conclusion

In the development process of the improved R/Shiny application, we focused on creating reliable visualization details. There exists an unsolved problem related to the separation of generation on branch plots and their corresponding MOB heatmap. In specific, the ends of branches could not be precisely located at the corresponding sections of MOB heatmap that represented the clusters belonging to the branches. Such problem is originated from the original design in "partykit" that does not share the same x-axis (date), and it scales the leaf plot (sections/clusters of the heatmap) only with its height but not the width at the same time. These could destroy the heatmap visualization and make it unreadable, and thus the previous version did not apply the original package "partykit" to draw a heatmap and its branch plot simultaneously. Although we do not solve this problem directly, we suggest a simple remedy with the following procedure.

1. For each cluster in the MOB heatmap, locate its center row denoted as $\bar{y}_i$ for $i = 1, \dots, k$, k is the number of clusters.
2. Standardize these row numbers as $\bar{c}_i = \bar{y}_i/m$, where m is total number of rows.
3. One-to-one manually draw lines from all leftmost nodes of branch plots to all rightmost boundary of the MOB heatmap at height $\bar{c}_i$, $i = 1, \dots, k$.

doi:10.6342/NTU202300086

There are some limitations of the plots via “Shiny” that remain unsolved yet. Since the number of clusters is dynamic, the series plot section lacks the flexibility to generate plots with different sizes each time. We suggest the developers to offer an easier-to-use way to dynamically adjust the space of the “tabPanel”. In addition, the evolutionary nature of MOB tree implies that the branches grow as the maximum depth changes, but when the maximum depth was set to 10, the “mob” function of “partykit” package always produced trees with depth of 2. We suspect some undetermined errors occurred in the original source code.

There are some limitations of our approach to be noticed. First, we assume that the nature behind the data is unchanged. For our case, we assume no further mutations occur in our COVID-19 virus. Such invariant assumption is essential to Eq. 2-1, because our approach of using past data is incapable of handling trend as the epidemic continue to evolve. If some newer interference exists at some time point of analysis, the clustering rules and accuracy on trend prediction are no longer statistically sound. To determine when our approach fails, the simplest method is to check if our forecasted responses within the validation period exceeds the confidence band at a certain acceptable α. Second, our study only focuses on the case of Taiwan COVID-19 spread, so the validity of applying this approach to similar cases in foreign countries remains unknown, so as to other pandemic spreads other than COVID-19. We do not have a

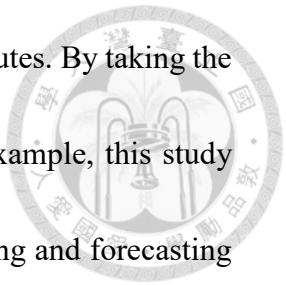
doi:10.6342/NTU202300086

consistency check for our method under a general framework, and it will be a main topic in the future work.

We suggest four improvements from our work. Firstly, the addition of other domain-relevant attributes which are potentially associated with COVID-19 spread may produce more effective clustering and forecasting results and thus provide more insights for domain experts to construct further analyses. For example, PM 2.5 pollution or the median household annual income, etc. Secondly, the only assumption we made regarding weather-related domain-relevant attributes is based on the neighboring effect. We expect that, after carefully checking the assumptions, more factors can be considered since they may affect the means we take to impute the missing data. For example, a weighted average may be applied based on the numbers or situations of township weather stations. Moreover, as packages in R grow more and more powerful, we plan to improve our R/Shiny application to provide more effective visualizations and more efficient functions. The concept of adjacent domain-relevant attributes may be refined or optimized to offer a novel methodology for generating derivative attributes in future work.

In this study, we provided an improved version of R/Shiny visualization tool for time-series model-based recursive partitioning and proposed several novel methodologies for applying this algorithm, which involved the population-based

doi:10.6342/NTU202300086



scaling technique and the concept of adjacent domain-relevant attributes. By taking the analysis on township-level Taiwan COVID-19 spread data as an example, this study showed that our ideas can improve both the performance in clustering and forecasting tasks. We also suggested various option configurations for different tasks using this algorithm. Lastly, we discussed promising future directions in developing this methodology. Though new diseases will always afflict the human race, data scientists can play a key role in developing new machine learning methodologies to mitigate their future impact.

doi:10.6342/NTU202300086

References

[1] Huang, C., et al., *Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China.* The lancet, 2020. **395**(10223): p. 497-506.

[2] Velavan, T.P. and C.G. Meyer, *The COVID-19 epidemic.* Tropical medicine & international health, 2020. **25**(3): p. 278.

[3] Zhou, F., et al., *Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study.* The lancet, 2020. **395**(10229): p. 1054-1062.

[4] Institute for Health Metrics and Evaluation. *IHME: COVID-19 Projections*. [cited 2022 November 1]; Available from: https://covid19.healthdata.org/global?view=cumulative-deaths&tab=trend.

[5] Coronavirus Research Center of Johns Hopkins University. *Covid-19 map*. [cited 2022 November 16]; Available from: https://coronavirus.jhu.edu/map.html.

[6] Wang, C.J., C.Y. Ng, and R.H. Brook, *Response to COVID-19 in Taiwan: big data analytics, new technology, and proactive testing.* Jama, 2020. **323**(14): p. 1341-1342.

[7] Huang, J.-H., et al., *Rapid response of a medical center upon the surge of COVID-19 epidemic in Taiwan.* Journal of Microbiology, Immunology and Infection, 2022. **55**(1): p. 1-5.

[8] Ashouri, M. and F.K.H. Phoa, *Interactive tool for clustering and forecasting patterns of Taiwan COVID-19 spread.* Plos one, 2022. **17**(6): p. e0265477.

[9] Shinde, G.R., et al., *Forecasting models for coronavirus disease (COVID-19): a survey of the state-of-the-art.* SN Computer Science, 2020. **1**(4): p. 1-15.

[10] Malik, Y.S., et al., *How artificial intelligence may help the Covid-19 pandemic: Pitfalls and lessons for the future.* Reviews in Medical Virology, 2021. **31**(5): p. 1-11.

[11] Dowd, J.B., et al., *Demographic science aids in understanding the spread and fatality rates of COVID-19.* Proceedings of the National Academy of Sciences,

doi:10.6342/NTU202300086

2020. **117**(18): p. 9696-9698.

[12] Ribeiro, M.H.D.M., et al., *Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil.* Chaos, Solitons & Fractals, 2020. **135**: p. 109853.

[13] Shahid, F., A. Zameer, and M. Muneeb, *Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM.* Chaos, Solitons & Fractals, 2020. **140**: p. 110212.

[14] Benvenuto, D., et al., *Application of the ARIMA model on the COVID-2019 epidemic dataset.* Data in brief, 2020. **29**: p. 105340.

[15] Khan, F.M. and R. Gupta, *ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India.* Journal of Safety Science and Resilience, 2020. **1**(1): p. 12-18.

[16] Pourhomayoun, M. and M. Shakibi, *Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making.* Smart Health, 2021. **20**: p. 100178.

[17] Randhawa, G.S., et al., *Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study.* Plos one, 2020. **15**(4): p. e0232391.

[18] Sethy, P.K. and S.K. Behera, *Detection of coronavirus disease (covid-19) based on deep features.* 2020.

[19] Siddiqui, M.K., et al., *Correlation between temperature and COVID-19 (suspected, confirmed and death) cases based on machine learning analysis.* J Pure Appl Microbiol, 2020. **14**(suppl 1): p. 1017-1024.

[20] Carrillo-Larco, R.M. and M. Castillo-Cara, *Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach.* Wellcome open research, 2020. **5**.

[21] Hochreiter, S. and J. Schmidhuber, *Long short-term memory.* Neural computation, 1997. **9**(8): p. 1735-1780.

[22] Chimmula, V.K.R. and L. Zhang, *Time series forecasting of COVID-19 transmission in Canada using LSTM networks.* Chaos, Solitons & Fractals, 2020. **135**: p. 109864.

[23] Chandra, R., A. Jain, and D. Singh Chauhan, *Deep learning via LSTM models for COVID-19 infection forecasting in India.* PloS one, 2022. **17**(1): p. e0262708.

[24] Zeileis, A., T. Hothorn, and K. Hornik, *Model-based recursive partitioning.* Journal of Computational and Graphical Statistics, 2008. **17**(2): p. 492-514.

[25] Ashouri, M., G. Shmueli, and C.-Y. Sin, *Tree-based methods for clustering time series using domain-relevant attributes.* Journal of Business Analytics,

doi:10.6342/NTU202300086

2019. **2**(1): p. 1-23.

[26] Finch, W.H., *Structural equation modelling trees for invariance assessment.* International Journal of Quantitative Research in Education, 2017. **4**(1-2): p. 72-93.

[27] Rusch, T., A. Zeileis, and K. Hornik. *Logistic regression trees for ordinal and preference data*. in *BOOK OF ABSTRACTS*. 2016.

[28] Zeileis, A. and K. Hornik, *Generalized M-fluctuation tests for parameter instability.* Statistica Neerlandica, 2007. **61**(4): p. 488-508.

[29] Hyndman, R.J. and G. Athanasopoulos, *Forecasting: principles and practice*. 2018: OTexts.

[30] Akaike, H., *Information theory and an extension of the maximum likelihood principle*, in *Selected papers of hirotugu akaike*. 1998, Springer. p. 199-213.

[31] Schwarz, G., *Estimating the dimension of a model.* The annals of statistics, 1978: p. 461-464.

[32] Rashed, E.A., et al., *Influence of absolute humidity, temperature and population density on COVID-19 spread and decay durations: multi-prefecture study in Japan.* International journal of environmental research and public health, 2020. **17**(15): p. 5354.

[33] Ganegoda, N.C., et al., *Interrelationship between daily COVID-19 cases and average temperature as well as relative humidity in Germany.* Scientific reports, 2021. **11**(1): p. 1-16.

[34] Adams, A., et al., *The disguised pandemic: The importance of data normalization in COVID-19 web mapping.* Public Health, 2020. **183**: p. 36.

[35] Kumar, J. and K. Hembram, *Epidemiological study of novel coronavirus (COVID-19).* arXiv preprint arXiv:2003.11376, 2020.

[36] Sameni, R., *Mathematical modeling of epidemic diseases; a case study of the COVID-19 coronavirus.* arXiv preprint arXiv:2003.11371, 2020.

[37] Tian, Y., I. Luthra, and X. Zhang, *Forecasting COVID-19 cases using Machine Learning models.* MedRxiv, 2020.

[38] Perret, B., et al., *Removing non-significant regions in hierarchical clustering and segmentation.* Pattern Recognition Letters, 2019. **128**: p. 433-439.

[39] Tritchler, D., E. Parkhomenko, and J. Beyene, *Filtering genes for cluster and network analysis.* BMC bioinformatics, 2009. **10**(1): p. 1-9.

[40] Oliver, M.A. and R. Webster, *Kriging: a method of interpolation for geographical information systems.* International Journal of Geographical Information System, 1990. **4**(3): p. 313-332.

[41] Shtiliyanova, A., et al., *Kriging-based approach to predict missing air temperature data.* Computers and Electronics in Agriculture, 2017. **142**: p.

doi:10.6342/NTU202300086

440-449.

[42] Feng, L., et al., *CUTOFF: A spatio-temporal imputation method.* Journal of Hydrology, 2014. **519**: p. 3591-3605.

[43] Winston Chang, J.C., JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen , Jonathan McPherson, Alan Dipert and Barbara Borges, *shiny: Web Application Framework for R*. 2021.

[44] Yu, Y., Y. Ouyang, and W. Yao, *shinyCircos: an R/Shiny application for interactive creation of Circos plot.* Bioinformatics, 2018. **34**(7): p. 1229-1231.

[45] Potter, G., et al., *Web application teaching tools for statistics using R and shiny.* Technology Innovations in Statistics Education, 2016. **9**(1).

[46] McGuinness, L.A. and J.P. Higgins, *Risk-of-bias VISualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments.* Research synthesis methods, 2021. **12**(1): p. 55-61.

[47] Shrotri, M., et al., *An interactive website tracking COVID-19 vaccine development.* The Lancet Global Health, 2021. **9**(5): p. e590-e592.

[48] Evrenoglou, T., I. Boutron, and A. Chaimani, *metaCOVID: An R-Shiny application for living meta-analyses of COVID-19 trials.* medRxiv, 2021.

[49] Tebé, C., et al., *COVID19-world: a shiny application to perform comprehensive country-specific data visualization for SARS-CoV-2 epidemic.* BMC Medical Research Methodology, 2020. **20**(1): p. 1-7.

[50] Zeileis, A., T. Hothorn, and K. Hornik, *party with the mob: Model-Based Recursive Partitioning in R.* R package version 0.9-9999, 2010.

[51] Chang, W., *shinythemes: Themes for Shiny*. 2021.

[52] National Center for High-Performance Computing. *Taiwan Reports of the COVID-19 Pandemic*. [cited 2022 October]; Available from: https://covid-19.nchc.org.tw/dt_005-covidTable_taiwan.php

[53] Government Open Data Platform. *Township Borders Dataset (TWD97 Coordinates)*. [cited 2022 September]; Available from: https://data.gov.tw/dataset/7441.

[54] Government Open Data Platform. *Township Population Density Dataset*. [cited 2022 September]; Available from: https://data.gov.tw/dataset/8410.

[55] Qi, H., et al., *COVID-19 transmission in Mainland China is associated with temperature and humidity: A time-series analysis.* Science of the total environment, 2020. **728**: p. 138778.

[56] Ma, Y., et al., *Effects of temperature variation and humidity on the mortality of COVID-19 in Wuhan.* medRxiv, 2020.

[57] Siegenfeld, A.F. and Y. Bar-Yam, *Eliminating covid-19: A community-based analysis.* arXiv preprint arXiv:2003.10086, 2020.

doi:10.6342/NTU202300086

[58] Hossain, M.P., et al., *The effects of border control and quarantine measures on the spread of COVID-19.* Epidemics, 2020. **32**: p. 100397.

[59] Klein, M.G., et al., *COVID-19 models for hospital surge capacity planning: A systematic review.* Disaster medicine and public health preparedness, 2022. **16**(1): p. 390-397.

[60] Tobler, W.R., *A computer movie simulating urban growth in the Detroit region.* Economic geography, 1970. **46**(sup1): p. 234-240.

[61] Sergio J. Rey, D.A.-B., Levi J. Wolf, *Geographic Data Science with Python*. 2020.

[62] CWB Observation Data Inquire System. *Observation Data of CWB's Manned and Automatic Weather Stations*. [cited 2022 September]; Available from: https://e-service.cwb.gov.tw/HistoryDataQuery/index.jsp.

[63] Society and Economics Geographic Information System. *Statistics for Districts*. [cited 2022 September]; Available from: https://segis.moi.gov.tw/STAT/Web/Platform/QueryInterface/STAT_QueryTopProduct.aspx.

[64] Taiwan Medical Association. *2019 Statistics for Taiwan Medical Practitioners and Medical Care Institutions*. [cited 2022 September]; Available from: https://www.tma.tw/stats/index_NYearInfo.asp?/2019.html.

[65] Civil Aeronautics Administration of MOTC. *The Diagram of Airports*. [cited 2022 September]; Available from: https://www.caa.gov.tw/Article.aspx?a=982&lang=2.

[66] Institute of Transportation of MOTC. *MOTC Transport API V2*. [cited 2022 September]; Available from: https://ptx.transportdata.tw/MOTC/?urls.primaryName=%E8%BB%8C%E9%81%93V2#/Metro/MetroApi_Station_2092.

[67] Wikipedia. *List of ROC Public Transfer Interchanges*. [cited 2022 September]; Available from: https://zh.m.wikipedia.org/zh-tw/%E4%B8%AD%E8%8F%AF%E6%B0%91%E5%9C%8B%E6%B1%BD%E8%BB%8A%E5%AE%A2%E9%81%8B%E8%BB%8A%E7%AB%99%E5%88%97%E8%A1%A8.

doi:10.6342/NTU202300086

Appendix A — Experiment Results

Table A-1: Clustering MSE of standard scale without/with adjacent domain-relevant attributes

Depth	Adjacent domain-relevant attributes			
	Without		With	
	Pruning option		Pruning option	
	AIC	BIC	AIC	BIC
1	0.1347	0.1347	0.1347	0.1347
2	0.1337	0.1337	0.1338	0.1338
3	0.1333	0.1333	0.1335	0.1335
4	0.1330	0.1330	0.1331	0.1331
5	0.1326	0.1326	0.1326	0.1328
6	0.1325	0.1326	0.1319	0.1321
7	0.1323	0.1325	0.1311	0.1321
8	0.1321	0.1325	0.1307	0.1321
9	0.1319	0.1324	0.1306	0.1321
Average	0.1329	0.1330	0.1324	0.1329

doi:10.6342/NTU202300086

Table A-2: Clustering MSE of population scale without/with adjacent domain-relevant attributes

Depth	Adjacent domain-relevant attributes			
	Without		With	
	Pruning option		Pruning option	
	AIC	BIC	AIC	BIC
1	1089.0511	1089.0511	1089.0511	1089.0511
2	1084.0704	1084.0704	1083.9985	1083.9985
3	1074.2400	1074.2400	1073.4363	1073.4363
4	1067.5645	1067.5645	1069.2587	1069.3804
5	1062.0057	1062.2173	1063.3706	1063.7132
6	1059.3889	1060.1166	1056.0253	1056.3569
7	1057.1062	1059.8747	1048.3567	1050.1937
8	1054.0129	1058.4821	1040.1674	1044.1332
9	1049.9577	1056.6073	1034.2098	1042.3289
Average	1066.3775	1068.0249	1061.9860	1063.6214

Table A-3: Forecasting MAE of standard scale without/with adjacent domain-relevant attributes

Depth	Adjacent domain-relevant attributes			
	Without		With	
	Pruning option		Pruning option	
	AIC	BIC	AIC	BIC
1	34.228	34.228	34.228	34.228
2	32.296	32.296	32.620	32.620
3	31.877	31.877	33.738	33.738
4	34.951	34.951	36.906	36.906
5	36.889	34.803	35.783	34.815
6	40.168	35.671	32.976	33.904
7	41.517	35.671	39.194	33.904
8	40.209	35.671	36.150	33.904
9	40.442	41.981	36.283	33.904
Average	36.953	35.239	35.320	34.214

doi:10.6342/NTU202300086

Table A-4: Forecasting MAE of population scale without/with adjacent domain-relevant attributes

Depth	Adjacent domain-relevant attributes			
	Without		With	
	Pruning option		Pruning option	
	AIC	BIC	AIC	BIC
1	13.039	13.039	13.039	13.039
2	12.919	12.919	13.032	13.032
3	13.138	13.138	13.168	13.168
4	13.142	13.142	12.848	12.932
5	13.216	13.205	12.847	13.078
6	13.549	13.858	12.712	13.174
7	13.696	13.892	13.784	13.419
8	13.678	13.764	13.857	13.596
9	14.188	13.991	13.640	13.502
Average	13.396	13.439	13.214	13.216

doi:10.6342/NTU202300086