VizTS – Visualisation of Time Series Clustering

Arief Sulistio, Goh I-Vy, Lim Si Ling Evelyn

Abstract — Time series datasets contain valuable information that can be obtained through pattern discovery. Time series clustering is an unsupervised technique commonly performed to partition time series data into groups based on similarity or distance to uncover interesting patterns with respect to time. Time series clustering has a wide variety of strategies and a series specific to Dynamic Time Warping (DTW) distance. dtwclust is a package built on statistical software R and has many algorithms implemented specifically tailored to DTW. A great amount of effort went into optimising the efficiency in the implementation of algorithms, and the functions were designed with flexibility and extensibility in mind. As such, dtwclust is a package with up-to-date and robust time series clustering algorithms which are comparable to, if not more superior than the expensive commercial-of-the-shelves analytical toolkit such as SAS Enterprise Miner. However, till date, the usage of dtwclust package tends to be confined within academic research as it requires intermediate R programming skill.

The project aims to provide a user-friendly interface to dtwclust package by using R Shiny framework. Application is designed to allow casual users to import data, manage, explore, calibrate, visualize and evaluate cluster results without having to type a single line of code. Representing time series cluster structures as visual images (visualization of time series data) can help users quickly understand the structure of data, clusters, anomalies, and other regularities in datasets. In addition to that, the application aims to empower users by providing flexibility in data selection and data exploration and incorporated interactive graph visualisation to enhance data exploration, to aid in the interpretability of the outputs of the clusters and to investigate the similarities or dissimilarities within the cluster.

Index Terms—Time Series, dynamic time warping, clustering, hierarchical, partitional, dtwclust, distance measure, cluster evaluation matrix

1. Introduction

With higher computation power, powerful data storage and processors, real-world data have taken the chance to store and keep data for a longer time. As such, there is also an increasing focus on looking time series data related analysis.

Time series data is classified as a dynamic data because its feature value change as a function of time, is high dimensional and large in data size. It is of interest due to their ubiquity in various area. Clustering such data is advantageous because it leads to discovery of interesting patterns in time series datasets. There are many real-world applications in time series clustering. For example, in biology; functional clustering time series gene expression data, in climate; analysing PM_{10} and $PM_{2.5}$ concentrations in New Zealand, in finance; finding seasonality patterns in retail and in psychology; analysing human behaviour in psychological domain. [1]

2. Literature Review

Time series clustering is a type of clustering algorithm made to handle dynamic data. It is a non-supervised technique to group similar time series with similar trend by minimizing intra-cluster distance. Elements to consider are the (dis)similarity or distance measure, time series prototype (centroids), the clustering algorithm itself, and cluster evaluation [2]. Although the Euclidean and Manhattan distance measure are the most commonly used distance measure, it is defined for series of equal length and sensitive to noise, scale and time shifts.

As such, the Dynamic Time Warping (DTW) algorithm, which has been proposed around 1970 in the context of speech recognition [3], is one of the most used measure of the similarity between two time series. The DTW aims to find the best alignment between two time series by constructing a warping matrix and search for the optimal warping path. In Figure 1 below, the coloured solid squares depict the best alignment between the time series Q and time series C.

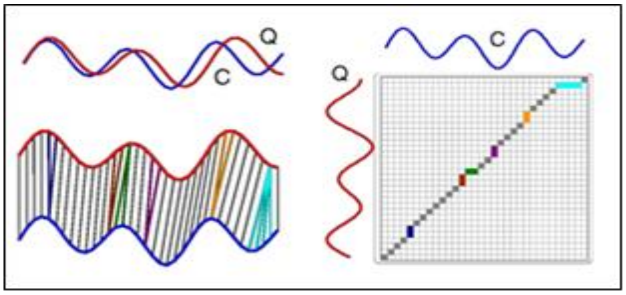


Figure 1: Dynamic Time Warping between Time Series Q and C [4]

There are several available R packages for data clustering. For example, flexclust and cluster packages which implements many partitional and hierarchical clustering respectively. However, neither are specifically targeted at time series data. There are also packages such as Tdist and TSclust which are more time series oriented but focuses on dissimilarity measures for time series. The dtw package on the other hand does not include the lower bound technique, although it implements extensive functionality with respect to DTW. Therefore, dtwclust R package which incorporates both classic and new clustering algorithm such as, k-Shape and TADPole which were originally implemented in MATLAB, and caters for time series data, is employed in this paper [2].

3. Motivation and Objective

3.1. Motivation

The current dtwclust package provides comprehensive functions which incorporates both classic time series clustering approaches and improvised techniques which were introduced in the past few years. It is intended to provide a consistent and user-friendly way for users to apply time series clustering algorithms with different distance metrics and centroid algorithms, taking into consideration the nuances of time-series data.

However, existing dtwclust package does not offer data preparation related functions for time series clustering and conversion of raw transactional data to time aggregated series data requires some data transformation with functions in other R packages. In addition, output plots are based on default R base which can be further improved in terms of visualisation and interactivity. Also, the clustering result do not offer insights on the characteristics of the cluster and it require users to merge the clustering result with other attributes.

3.2 Objective

This project aims integrate dtwclust package with packages meant for data manipulation and those which provide interactivity with the objective to build a hassle-free interface for user to perform time series clustering without the need to code and to visualise the clustering result in a more interactive and visual manner in order to uncover pattern which have potential use case in the respective domains.

3.2.1 Visual Data Exploratory

In order to have a better understanding of the data, data exploratory is a crucial process prior to perform any form of analysis. This application aims provide users with multiple time aggregation options and the frequency of the selected time aggregation will be plotted accordingly. Also, filters will be put in place to provide users with greater flexibility in data exploration.

3.2.2 Interactive Clustering Result

Even though dtwclust package is mathematically robust and relatively computational efficient, the output plot uses default R base plot function which can be further improve in terms of visualisation and interactivity. Visualisation of dendrogram from hierarchical clustering will be enhanced and time series charts for each cluster will be made interactive.

3.2.3 Visualisation of Cluster Features

As with all types of clustering, it is important for users to understand the characteristics of the clusters through the overlaying of clustering result with other attributes. The application also strives to achieve this objective by providing placeholders for users to select variables of their interest so that they can visually comprehend the clustering result.

4. Application Design

Functionalities of the VizTS: Clustering Edition application were designed and put in place to fulfill the objectives which were set in the previous section with the design framework in Figure 2.

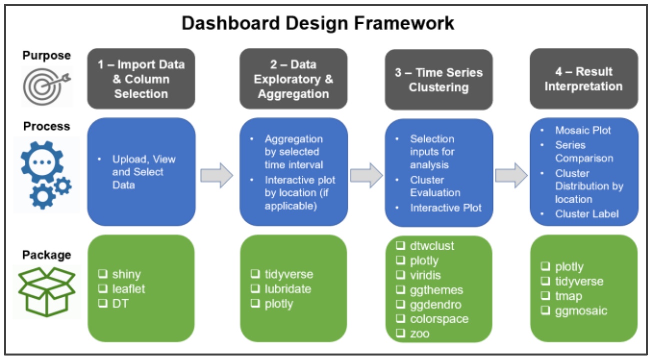
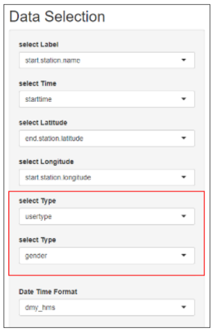

Dashboard Design Framework
Purpose
Process
Package
1 – Import Data & Column Selection
2 – Data Exploratory & Aggregation
3 – Time Series Clustering
4 – Result Interpretation
• Upload, View and Select Data
• Aggregation by selected time interval
• Interactive plot by location (if applicable)
• Selection inputs for analysis
• Cluster Evaluation
• Interactive Plot
• Mosaic Plot
• Series Comparison
• Cluster Distribution by location
• Cluster Label
❑ shiny
❑ leaflet
❑ DT
❑ tidyverse
❑ lubridate
❑ plotly
❑ dtwclust
❑ plotly
❑ viridis
❑ ggthemes
❑ ggdendro
❑ colorspace
❑ zoo
❑ plotly
❑ tidyverse
❑ tmap
❑ ggmosaic

Figure 2: Dashboard Design Framework

4.1 Data Import and Data Selection

Application provides the options to import comma, semicolon and tab separated data with or without header and for various text qualified data (None, double quote, single quote). Data table is used for users to view the imported data and to search for field names.



Data Selection

select Label
start.station.name

select Time
starttime

select Latitude
end.station.latitude

select Longitude
start.station.longitude

select Type
usertype

select Type
gender

Date Time Format
dmy_hms

From Figure 3, there is flexibility to choose columns from the imported data to perform Time Series Clustering Analysis and data table is used to display the selected columns. There are 2 extra variables called 'Type' for users to put in attributes of interest for more meaningful interpretation of clustering results.

Figure 3: Data selection options

4.2 Data Exploratory

In the Data Exploratory Tab, the VizTS: Clustering Edition application also incorporates the map feature if the latitude and longitude data is available. Refer to Figure 4 for snapshot of the map feature.

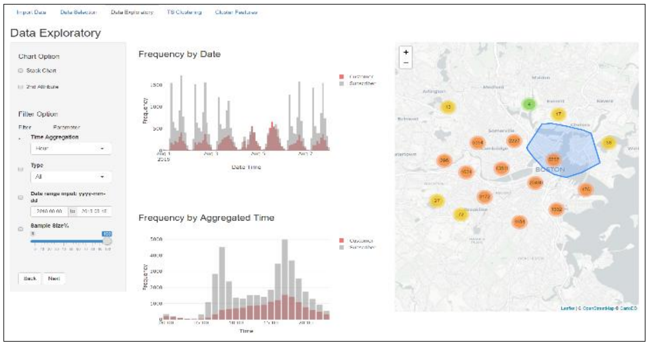
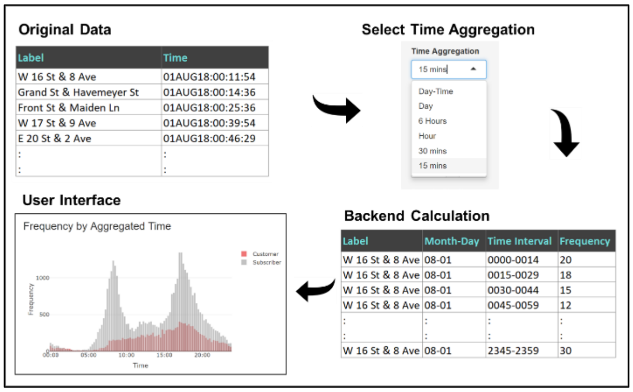


Figure 4: Snapshot of Data Exploratory Tab

One of the key features under data exploratory is the ability for user to examine the dataset at selected time aggregation which is an important data preparation step in any time series analysis. Application has an in-built aggregation function which transforms transactional data into time aggregated data as part of data pre-processing. As shown in Figure 5, application provides a drop-down list of time aggregation by 15 minutes, 30 minutes, 1-Hour, 6-hour, Day and Day-time intervals for user to select and view the aggregated data. Time aggregation under Day-Time is set at 1-Hour interval.



Original Data

Label	Time
W 16 St & 8 Ave	01AUG18:00:11:54
Grand St & Havemeyer St	01AUG18:00:14:36
Front St & Maiden Ln	01AUG18:00:25:36
W 17 St & 9 Ave	01AUG18:00:39:54
E 20 St & 2 Ave	01AUG18:00:46:29
:	:
:	:

Select Time Aggregation

Time Aggregation
15 mins

Backend Calculation

Label	Month-Day	Time Interval	Frequency
W 16 St & 8 Ave	08-01	0000-0014	20
W 16 St & 8 Ave	08-01	0015-0029	18
W 16 St & 8 Ave	08-01	0030-0044	15
W 16 St & 8 Ave	08-01	0045-0059	12
:	:	:	:
:	:	:	:
W 16 St & 8 Ave	08-01	2345-2359	30

User Interface

Figure 5: Work flow of Time Aggregation Function within the application

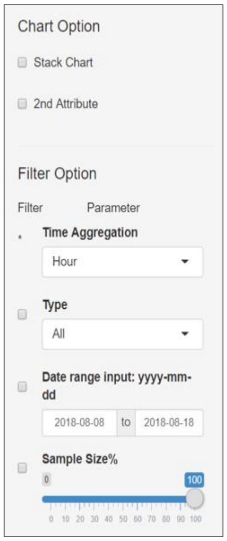


Chart Option

□ Stack Chart

□ 2nd Attribute

Filter Option

Filter Parameter

Time Aggregation: Hour

□ Type: All

□ Date range input: yyyy-mm-dd: 2018-08-08 to 2018-08-18

□ Sample Size%: 0 – 100

Figure 6 shows the options available for users in Data Exploratory. The default plot provides overlaying of histogram by levels for the selected variable and users have the option of stacked histogram chart. Users can also view the distribution by second attribute by selecting ‘2nd Attribute’ option illustrated in Figure 6.

Also, there is a drop down list for ‘Type’ reflects the categories for the selected attribute (Type1/2) and date range can be selected. Users can also random sample a percentage of the original data if a large dataset is used. Checkbox on the left of each parameter provides users the flexibility to filter the data based on the selected parameter of which will be used in time series clustering analysis in the following tab.

Figure 6: Illustration of various data filter that are available for user to select under data exploratory

Figure 7 below provides a comparison of the default view – where first attribute and overlaying chart and optional view – Stack Chart and 2nd Attribute are checked in the check box in Chart Option.

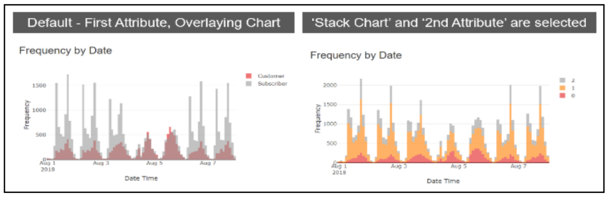


Figure 7: Plots for the different selections.

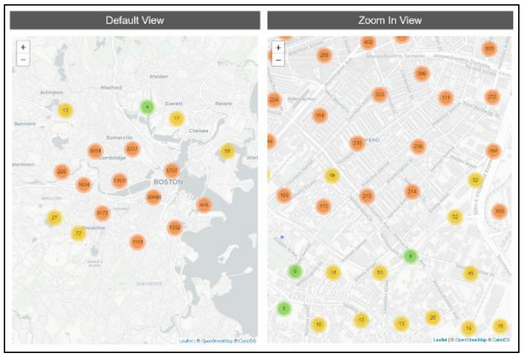


Figure 8: Interactive Geospatial Distribution plots

For datasets with latitude and longitude, interactive geospatial distribution plot provides users with a neat overview of the frequency distribution with respect to location (using location clusters) and it also allows users to zoom into specific area of interest as illustrated in Figure 8.

4.3 Time Series Clustering

As the dtwclust package incorporates a series of different clustering technique, the application provides users with the options to select between various types (clustering), different options for distance computation, centroid algorithm and method algorithm. Figure 9 below illustrates the user-friendly drop-down menu that built into the application to allow user to select the different clustering technique.

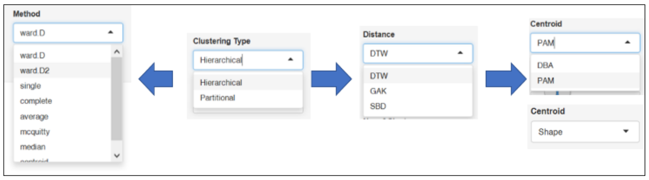
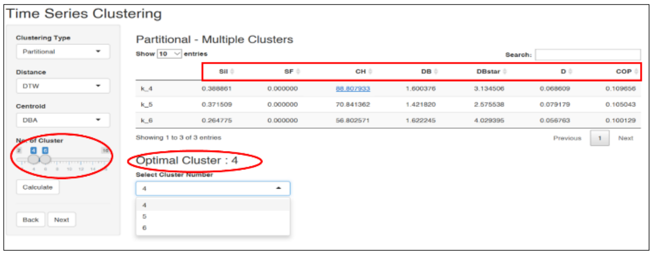


Figure 9: User-friendly drop-down menu for clustering calibration

In cases of partitional clustering, where k cluster must first be determined by user, the application also provides user with the option to select a range of cluster for computation. When multiple cluster is selected, the application will employ a standardized Cluster Evaluation Metrics, where cluster validity indices (CVI) using various method are utilized and compare to each other. The majority vote is used to decide on the optimal cluster.



Time Series Clustering

Clustering Type: Partitional

Distance: DTW

Centroid: DBA

No. of Cluster: 4 – 6

Calculate

Back Next

Partitional - Multiple Clusters

Show 10 entries Search:

	Sil	SF	CH	DB	DBstar	D	COP
k_4	0.388861	0.000000	88.807933	1.600376	3.134506	0.068609	0.109656
k_5	0.371509	0.000000	70.841362	1.421820	2.575538	0.079179	0.105043
k_6	0.264775	0.000000	56.802571	1.622245	4.029395	0.056763	0.100129

Showing 1 to 3 of 3 entries Previous 1 Next

Optimal Cluster : 4

Select Cluster Number: 4

4
5
6

Figure 10: Clustering evaluation can be performed.

Figure 10 provides a snapshot of the application. The red box in the snapshot are the various method of CVIs. The optimal cluster from the range of cluster selected, is recommended based on the majority vote of the CVIs. Users are able to select either the optimal cluster recommended or their choice of cluster to be visualized, to examine the characteristics of each cluster. Figure 11 below is a snapshot of the generated line graph. As the application uses plotly, the plot is an interactive line graph, which therefore further enhance investigation of cluster’s characteristics.

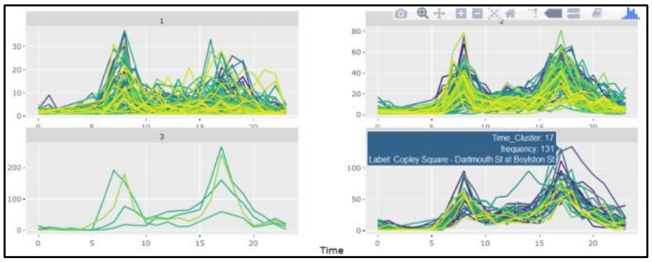


Figure 11: Time Series Clustering using Optimal Cluster of 4

The agglomerative hierarchical clusering is commonly displayed as a tree diagram called a dendrogram. A dendrogram begin with each object in a separate cluster and at each step, the two cluster that are most similar are joined into a single new cluster [5]. When the user performs a Hierarchial Clustering, the VizTS: Clustering Edition leverages on ggplot2 and plotly to build a dendrogram plot to enhance the current base plot. The dendrogram in the application is colored based on cluster and is an interactive plot as well. Refer to Figure 12 for illustration.

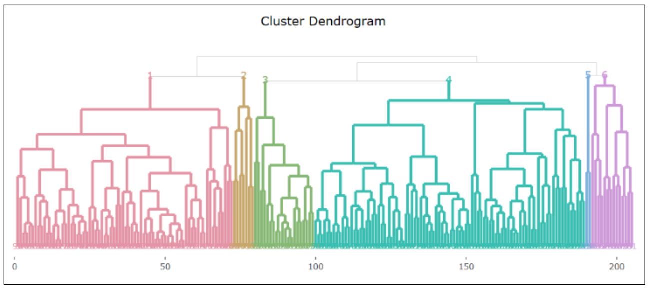


Figure 12: Cluster Dendrogram using Hierarchical Clustering with cluster of 6.

4.4 Review of Clustering Features

One of the key features of the VizTS: Clustering Edition application is the capability to incorporate the map feature when the dataset consists of latitude and longitude data. This will therefore enable users in discovery and exploratory to further understand the cluster's characteristics and idiosyncrasies. Figure 13 illustrates four different cluster that can be seen on the map. Each cluster is different in colour and the size of the point represents the frequency or volume of the data.

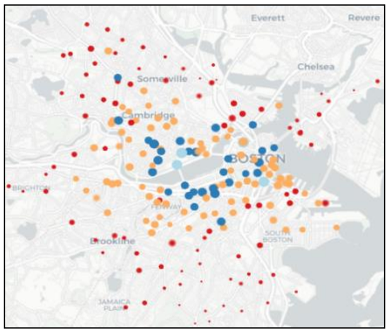
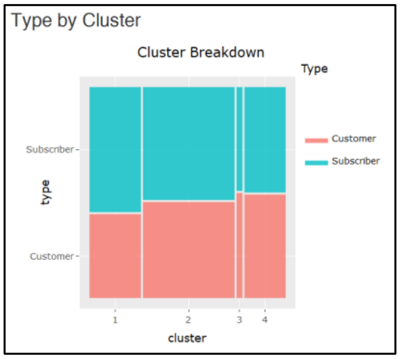


Figure 13: Map feature with clusters differentiated by colour and size of each point represents the frequency of data.



From the initial data selection tab, users can select a type, which is a parameter of interest for analysing. In the Cluster Feature tab, the user can visualize the proportion of the type in each cluster by utilizing the mosaic plot in built in the application. Refer to Figure 14.

Figure 14: Analyzing parameter of interest by cluster.

Lastly, the Cluster Features tab also encourage users to further investigate the individual series within each cluster by providing the capability to compare up to three different series (preferably from the same cluster). This allows users to be able to visualise how similar the series are and why are they being considered as one cluster. The line graph also leverages on plotly to incorporate interactivity as can be seen in Figure 15.

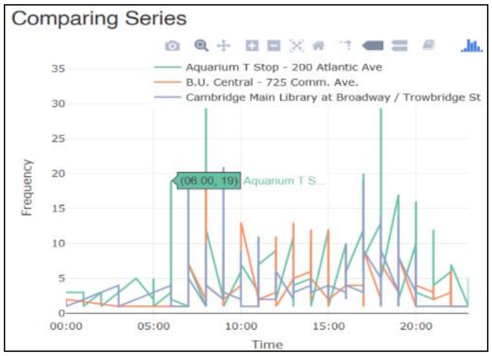


Figure 15: Compare and contrast individual series within the same cluster.

5. DISCUSSION

5.1 Boston BLUEbikes Case Study

The objective of this case study is to understand the demand of bicycles at the respective stations in Boston City [6] and to determine if there are any similar time series pattern for service provider to better manage the distribution of bicycles. BLUEbikes system data for one week in August 2018 (1 August – 7 August 2018) data was used for this analysis.

Figure 16 shows the frequency of bicycle usage based on start date and time of the trip for subscriber and customers. Subscribers are users on annual or monthly pass while customers are users on single trip or day pass. Frequency of usage by date shows that for subscribers, there were 2 distinct peaks on weekdays which are 1 to 3, 6 and 7 Aug 2018. It was also observed that the peaks on Friday were lower than that of the other weekdays. It is interesting to note that there was slightly more usage from customers compared to subscribers on weekends and customers' usage usually peaked around early afternoon. 15 minutes interval was selected for 'Frequency by Aggregated Time' plot.

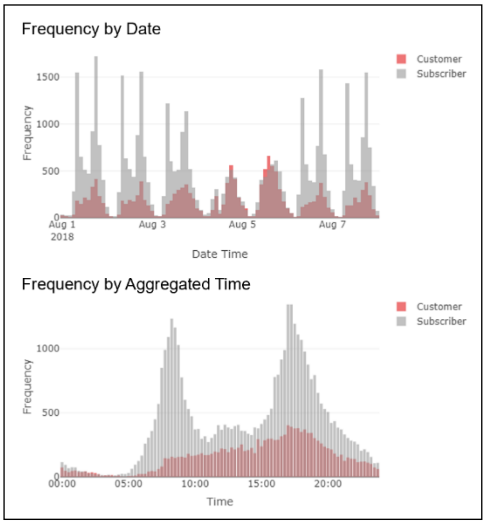
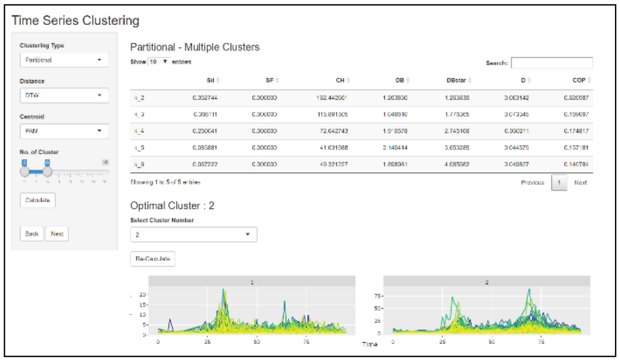


Figure 16: Histogram of Frequency plot by Date and Aggregated Time (15 minutes) for BLUEbikes data

Next, time series clustering was performed for a range of 2 to 6 clusters with Partional clustering, DTW distance and PAM centriod. Cluster evaluation result in Figure 17 shows that 2 is the optimal number of clusters. Time Series plots reveal that bike demand for Cluster 2 is almost 3 times of Cluster 1 and the position of the peaks is different for the two clusters.



Time Series Clustering

Clustering Type: Partitional

Distance: DTW

Centroid: PAM

No. of Cluster: 2 – 6

Partitional - Multiple Clusters

Show 10 entries

	Sil	SF	CH	DB	DBstar	D	COP
k_2	0.352744	0.000000	192.442061	1.263830	1.263830	0.063142	0.620987
k_3	0.386111	0.000000	115.891505	1.049810	1.773365	0.073345	0.169087
k_4	0.250041	0.000000	72.642743	1.910578	2.745106	0.060211	0.174817
k_5	0.085881	0.000000	41.631088	2.140414	3.653285	0.044575	0.157181
k_6	0.087222	0.000000	49.321227	1.888961	4.085082	0.040827	0.140784

Showing 1 to 5 of 5 entries

Optimal Cluster : 2

Select Cluster Number: 2

Figure 17: Time Series Clustering for BLUEbikes Data

Figure 18 below shows cluster features of the two clusters. From the mosaic plot, proportion of customer is slightly higher in Cluster 2 compared to Cluster 1. Since this dataset contains geographical coordinates, geographical distribution of the clusters was plotted with orange points representing Cluster 1 and blue points representing Cluster 2. Size of the points denote demand of bikes at the respective docking stations. Demand of bicycle from Cluster 2 is higher than that of Cluster 1 from mosaic plot and the map.

Three stations in Cluster 2 were choosen based on plotly's interactive tooltip to further examine their time series pattern under 'Comparing Series' plot. Even though all the 3 series are from Cluster 2, docking station at Nashua Street at Red Auerbach Way (green line) has a high morning peak and according to Google map, Nashua Street at Red Auerbach Way station has a multi-storey carpark nearby and users might have parked their cars in the car park before taking a bicycle ride to somewhere nearby. Docking station at 'MIT Stata Center at Vassar St/Main St' (orange line) has a high evening peak and Central Square at Mass Ave/Essex St (blue line) peaked moderately in the morning and the evening. This also shows times series in the same cluster can have slightly different pattern.

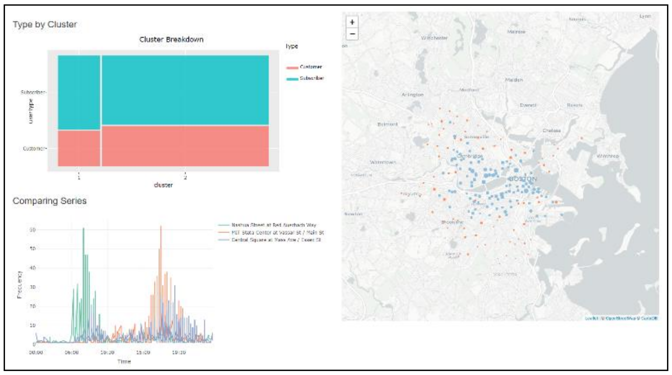


Figure 18: Cluster Features for BLUEbike Data

In conclusion, one week of BLUEbikes data shows that generally stations in Boston can be segregated into cluster which has high demand during both morning peak and evening peak and used more frequently by subscribers and cluster with stations that has lower demand throughout the day with a small peak in the early afternoon.

5.2 Simulated Attacker Data Case Study

200,000 simulated RDP login attempts data based on the attacker data was obtained from github [7] and this data can be used to detect any trend in time series of the attack and to test our application on a dataset with no geographical coordinates.

As there are many categories, filters at the side panel can be used to single out and explore of any of the category of interest. Figure 19 shows the exploratory plots when date range of 1 May to 31 May 2015 was selected and filtered respectively for Asia and America.

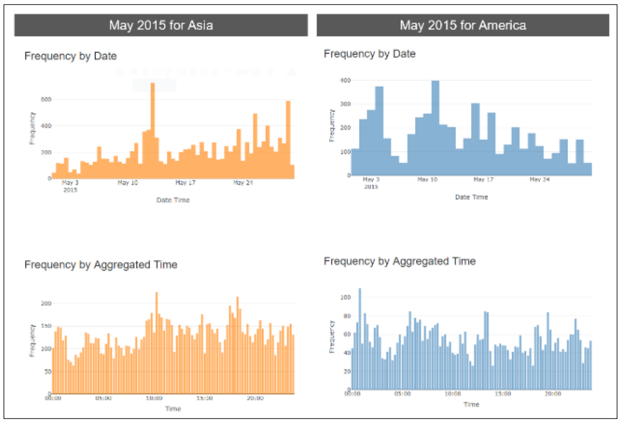
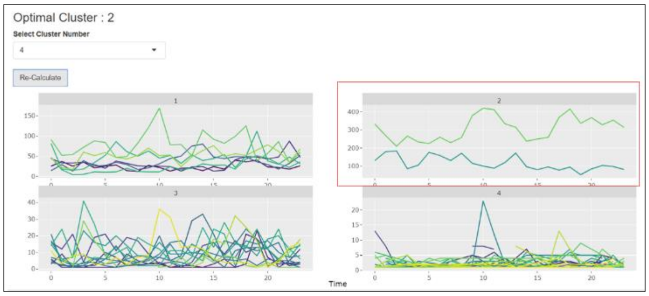


Figure 19: Exploratory Plots for Asia and America continents

It is interesting to note from Figure 19 that there was a weekly peak (seasonality) on weekend for number of attacks from America in May 2015 while there is to obvious seasonally patterns for attacks from Asia.

Next, time series clustering was performed for May 2015 data, aggreagaged by hour, with partitional clustering, DTW distance and DBA centroid. From figure 20, optimal Cluster is 2, however, 4 clusters was selected and the plot for Cluster 2 in Figure 20 further illustrates the methodology of Dynamic Time Warping distance, as the time series which look seemingly different, have the optimal dynamic time warping path based on the constructed warping matrix and thus were classified under the same cluster. The higher frequency series in Cluster 2 is Shanghai and the other series is Los Angles. Both are metropolitan cities.



Optimal Cluster : 2

Select Cluster Number: 4

Figure 20: Illustration of DTW example using Attacker Data

Figure 21 shows the distribution of attacks by clusters in terms of continent. Cluster 1 makes up of cities in Europe, Asia and America with moderately high number of attacks. Cluster 2 consists of Shanghai and Los Angles and highest frequency of attacks. Cluster 3 consists of attacks mainly from European cities with number of attacks being second lowest among the clusters. Cluster 4 is the largest cluster, i.e. most cities, with attacks from all the continents but the frequency of attack from the countries is also the lowest among all the cluster.

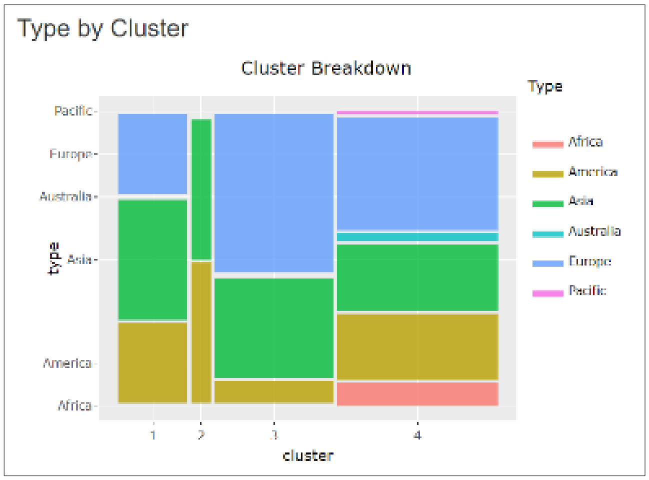


Figure 21: Mosaic Plots of the attacks

This example showcases how dynamic time warping distance comes into play in time series clustering and the application's ability to handle non-geospatial related data. Due to the limitation of attributes in this dataset, evaluation of cluster features was only restricted to the continents and it might be good to incorporate internet penetration of each city.

6. Conclusion

To conclude, with increasing availability of time series data, there is a need to provide users a platform to perform Time Series Clustering analysis for pattern detection and cluster features examination to allow users to have visual understanding of the clustering result.

The VizTS: Clustering Edition Application on RShiny aims to provide such platform for casual users to perform time series clustering on the dataset of choice. In addition, the application provides enhanced features such as time aggregation – a crucial data preparation step, cluster evaluation and cluster recommendation – the capability to select number of clusters to perform time series analysis on, and improved visualization – to further investigate and understand the clustering results' feature and characteristics. The added functionalities, visual and interactivity features bridged the gap of what dtwclust package currently offers.

7. Future Work

The VizTS: Clustering Edition Application however is limited to performing the two more common type of clustering, which are Hierarchical Clustering and Partitional Clustering. Future work to incorporate other type of clustering that are within the dtwclust package such as TADPole and Fuzzy clustering. The application can be further enhanced by including the other distance measures such as LB_Keogh, LB_Improved and Soft-DTW and its by product, the Soft-DTW centroid. However, more work is needed to understand how these distance measure (and centroid) should and can be incorporated.

The application can also be further enhanced to cater for datasets with only date variable instead of date-time variable and to provide users with the option to aggregate time series data by summing up a (value) column. Interconnectivity feature across the charts can also be used to enhance the data exploratory and cluster examination experience.

An extension is to provide Time Series Forecasting module to enable users to use the clustering output to perform time series forecasting to predict the frequency of the occurrences in the future and to determine if aggregated demand forecasting based on clustering output do result in lower mean squared error.

Acknowledgment

The project team would like to thank Prof Kam Tin Seong, Associate Professor of MITB (Practice), from the School of Information Systems (SIS), Singapore Management University (SMU), for his teaching and mentorship.

References

[1] Aghabozorgi, S., Shirkhorshidi, A.S, & Teh, Y.W. (2015). Time-series clustering – *A decade review, Information Systems, Data: Creation, Management and Utilization*, 53, 16-38.

[2] Sardá-Espinosa A. (2017). *Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package*.

[3] Giorgino T. (2009). *Computing and Visualizing Dynamic Time Warping Alignment in R: The dtw Package*. Journal of Statistical Software, 31(7).

[4] Abdullah Mueen, Eamonn J. Keogh: *Extracting Optimal Performance from Dynamic Time Warping*. KDD 2016: 2129-2130

[5] NCSS Statistical Software, (n.d). ©NCSS, LCC, Hierarchical Clustering/Dendrogram, Chapter 445, p.1.

[6] System Data. (2018). In BLUEbikes. Retrieved on 30th November 2018, from, https://www.bluebikes.com/system-data

[7] Simulated RDP login attempts data. (2016). In Github. Retrieved on 30th November 2018, from https://github.com/hrbrmstr/facetedcountryheatmaps/tree/master/data