HOTSPOT IDENTIFICATION FOR TRAIN DELAYS

A PROJECT REPORT

Submitted by

CB.EN.U4CSE16621 MYLAVARAPU UMA HEMA SRI

CB.EN.U4CSE16633 PULI HARIKA REDDY

CB.EN.U4CSE16639 R SAI NAVADEEP REDDY

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

AMRITA SCHOOL OF ENGINEERING, COIMBATORE

AMRITA VISHWA VIDYAPEETHAM

COIMBATORE 641 112

OCTOBER 2019

AMRITA VISHWA VIDYAPEETHAM

AMRITA SCHOOL OF ENGINEERING, COIMBATORE, 641112

BONAFIDE CERTIFICATE

This is to certify that the project report entitled "**HOTSPOT IDENTIFICATION FOR TRAIN DELAYS**" submitted by MYLAVARAPU UMA HEMA SRI (CB.EN.U4CSE16621), PULI HARIKA REDDY (CB.EN.U4CSE16633) and R SAI NAVADEEP REDDY (CB.EN.U4CSE16639) in partial fulfillment of the requirements for the award of the Degree **Bachelor of Technology in Computer Science and Engineering** is a bonafide record of the work carried out under our guidance and supervision at Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore

PROJECT GUIDE	CHAIRPERSON
Dr. R. Karthi	**Dr. (Col) P.N. Kumar**
Associate Professor	Professor
Dept. of Computer Science and Engg.	Dept. of Computer Science and Engg.

This project report was evaluated by us on :. . . . . .

INTERNAL EXAMINER EXTERNAL EXAMINER

Acknowledgment

We express our gratitude to our beloved Satguru Sri Mata Amritanandamayi Devi for providing a bright academic climate at this university, which has made this entire task appreciable. This acknowledgement is intended to be a thanks giving measure to all those people involved directly or indirectly with our project.

We would like to thank our Vice Chancellor Dr. Venkat Rangan. P and Dr. Sasangan Ramanathan Dean Engineering of Amrita Vishwa Vidyapeetham for providing us the necessary infrastructure required for completion of the project.

We express our thanks to Dr.(Col.P.N.Kumar), Chairperson of Department of Computer Science Engineering, Dr.C.Shunmuga Velayutham and Dr. G. Jeyakumar, Vice Chairpersons of the Department of Computer Science and Engineering for their valuable help and support during our study. We express our gratitude to our guide, Dr. R. Karthi, for the guidance, support and supervision. We feel extremely grateful to Dr. G. Jeyakumar, Dr. Senthil Kumar M, A. Baskar and Ms. Dhanya M Dhanlakshmy for their feedback and encouragement which helped us to complete the project. We also thank the staff of the Department of Computer Science Engineering for their support. We would like to extend our sincere thanks to our family and friends for helping and motivating us during the course of the project.

Abstract

In 2017-18, Indian Railways had the worst punctuality performance in three years. 30 percent trains ran late in 2017-18, according to official data. From April, 2017 - March, 2018, the punctuality of mail and express trains was 71.39 percent, down from 76.69 percent in 2016 - 17, which is deterioration of 5.30 percent. Hence it is necessary to identify stations or regions where these issues majorly occur and those stations or regions can be termed as **Hotspots**. With this technique, similar delay patterns can be formed among the stations or regions considered. This model is useful for analysis of the delay patterns across the stations and the reasons for the same can be analyzed for better society.

Table of Contents

List of Figures

List of Abbreviations

EMD Earth Mover's Distance

Chapter 1

Introduction

1.1 Background

Rail transport is an important mode of transport in India. As of March 2017, the rail network comprises 121,407 km of track over a route of 67,368 km and 7,349 stations. It is the fourth-largest railway network in the world and being one of the busiest networks it is obvious to face many issues in day to day life. One of the major issues is the delay of trains, where many stations face this delay throughout the year.

1.2 Problem Statement

Over the past two years, Indian Railways faced a deterioration of 5.30 percent in maintaining the punctuality of trains. Finding similar delay patterns across various stations where frequent delays occur, by trains passing through them and projecting the analysis results i.e., patterns on Maps, is the main objective.

1.3 Specific Objectives

In this project, the region is restricted to Tamil Nadu and Kerala. The data collected is processed and clustering is done to obtain Dendograms to be projected. The latitude and longitude data of stations are gathered and the delay patterns i.e., Dendograms with various colour representations,. are plotted on Maps.

1.4 Findings

The said analysis, can be further increased to the scope of considering all the stations along with all the trains running across the country and the results are presented onto Maps (Figure 3.1) for further analysis.

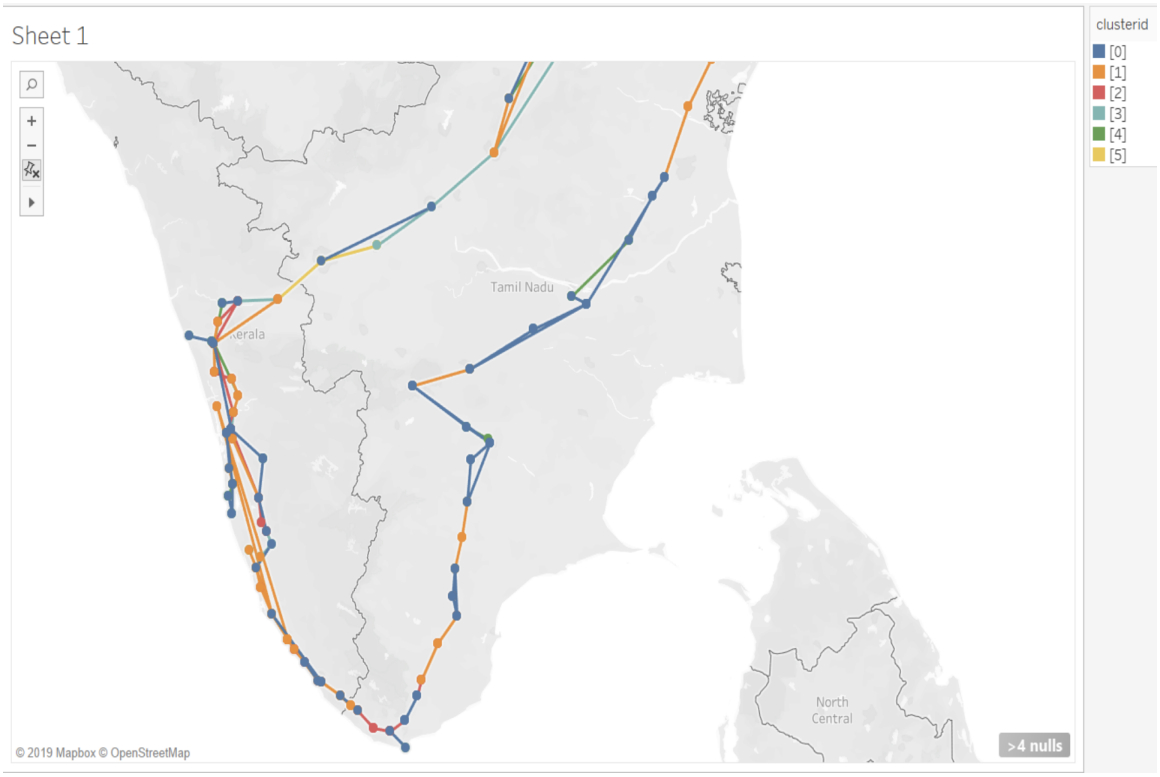

Figure 1.1: Delay Patterns on Maps

Chapter 2

LiteratureSurvey

The following section provides a review of the literature related to the analysis of the railway schedules, that provides the delay regions for the Indian railway trains in the region of Tamil Nadu and Kerala. The project execution is based on Spatio-Temporal Profiling of delays based on the railway running schedules. It is implemented using the preprocessed raw data of train delays across various stations and plotting the results on Maps using the latitude longitude of the selected stations.

The development was made based on the methods and processes discussed in Spatio-Temporal Profiling of Public Transport Delays Based on Large-Scale Vehicle Positioning Data From GPS, a survey published by Piotr Szyman ski , Michał Żołnieruk, Piotr Oleszczyk, Igor Gisterek, and Tomasz Kajdanowicz [1], that was submitted to IEEE in 2018. The authors provide a detailed explanation of data collection and its preprocessing methods on the similar type of data we require in this project. The methodology and implementation has been discussed with much effective in execution and giving wonderful results.The discretized raw data edges have to undergo Agglomerative Clustering [2], which needs the distance cost matrix from the EMD Algorithm [3]. We used a python wrapper pyemd [4] [5] [6] for calculating earth mover's distance. This project is completely useful for the Indian Railway department to analyse the results for the reasons and use it effectively for

a better cause.

The main challenge lies in the stage of data collection and processing it to our convenience for application of further methodology. The large delay data and official schedules data of the trains is crucial and very much limited availability of the same makes it a challenging situation as it is the data from past time period. The manual data collection makes the process a bit costly, which we can overcome using web scraping.

Chapter 3

Proposed System

3.1 System Architecture

The architecture flow for the analysis of the Hotspot identification for Train Delays is explained below. The analysis can be started as follows:

1. Data gathering of delay schedules and official schedules of the trains considered from the official railway websites for analysis.
2. Data filtering based upon the scope of the project i.e., stations selected in the region of Tamil Nadu and Kerala part of the country.
3. Preprocessing raw data to formulate the links for the edges between all the stations along with the delay and actual arrival time at the stations.
4. Discretization of station link data points projected on time-delay matrix for simpler calculations.
5. Formulation of cost of transforming one edge to another using EMD Algorithm and computing distance matrix.
6. Distance matrix is given as input for Agglomerative clustering and six different clusters are obtained consisting of similar delay patterns.

7. The cluster data along with the latitude longitude data of stations is provided to the Tableau Data Visualization Tool and the patterns are generated on the Map.

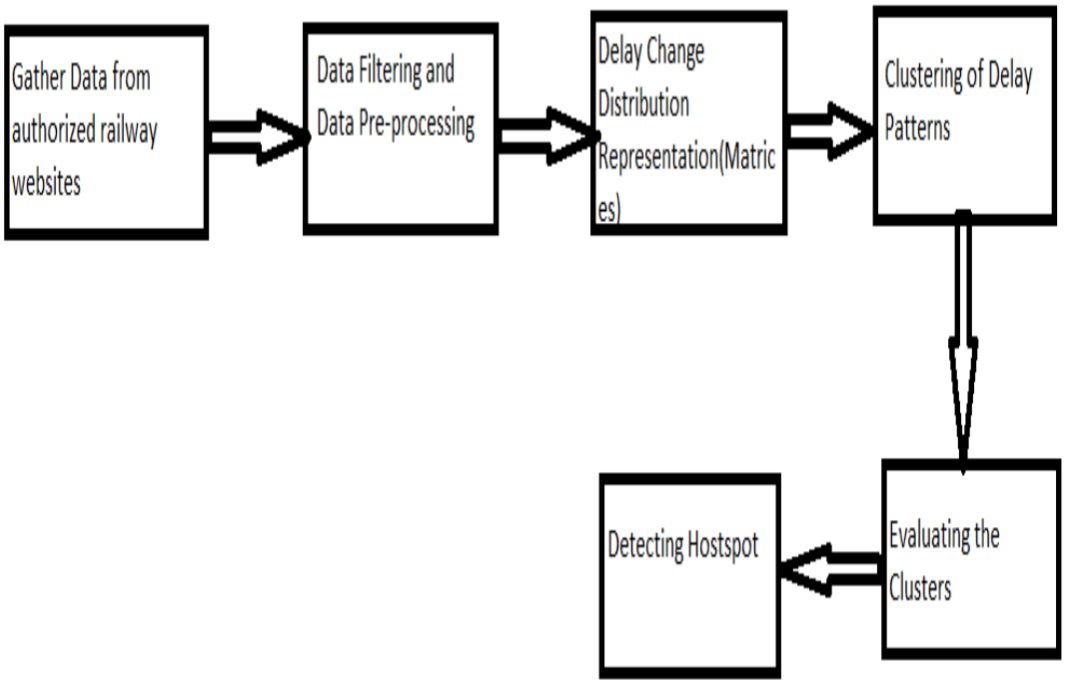

Figure 3.1: Architecture Flow

3.2 System Specification

This project requires following specifications for analysis:

- Raw data from official railway websites
- Data filtering and preprocessing techniques in python
- Discretization of bins knowledge
- Clustering Algorithm
- Tableau Data Visualization Tool

3.3 Methodology

The raw data that has been preprocessed is now obtained into format of links representing edges consisting of two consecutive stations (source and destination) with the actual arrival times of all the trains considered based upon the delay at the destination station.

This data of each edge has to be represented in the form of a two dimensional matrix. A perfect representation would take into account:

- Two dimensions of data: both delay change value and time it occurred should be considered
- Density of points: the chosen representation should recognise the density of delay changes of given value in given time, i.e. be resilient to the fact that for certain times of day we may have more points than in others due to differentiation in traffic on the route

1. Delay Change Distribution Representation

   We propose a representation based on the discretization of all delay changes points in both dimensions. The figures 3.2 and 3.3 represent the structure of data representation. The next step is to decide the bins for time discretization. So, we have taken six bins into consideration with each bin of four hours on the x-axis and variable range bins on the y-axis. The next step is to calculate the total number of delay points in each bin (taking into account both delay change discretisation and time discretisation). Then each bin for time is going to be considered separately and normalized. For each cell in the specific time bin, the value is going to be normalised by dividing the value by the sum of values in the whole column (time bin). This should be done for all rows. This process of normalisation guarantees that the number of delay points in whole time bin is going to be neglected during comparison with different time bin. The last step is to normalise the whole matrix - dividing each cell by the sum

of all cells. Thanks to this process all the values of the edge representation sum to 1 and it is easier to analyse.The edges which have no values in any of the above time bins were deleted.

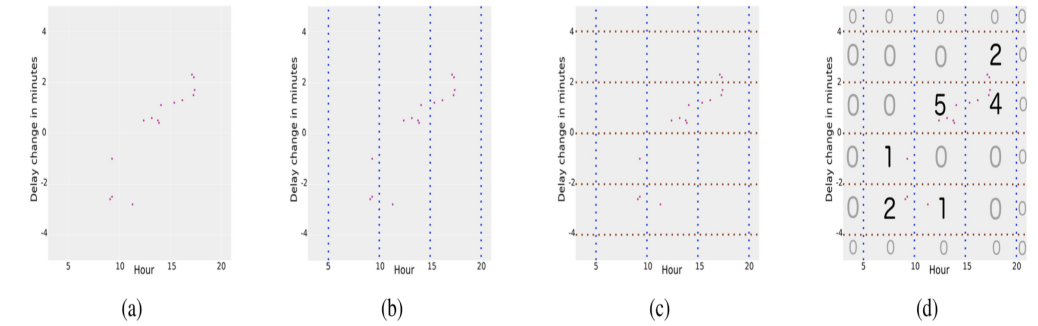

Figure 3.2: Binning Process

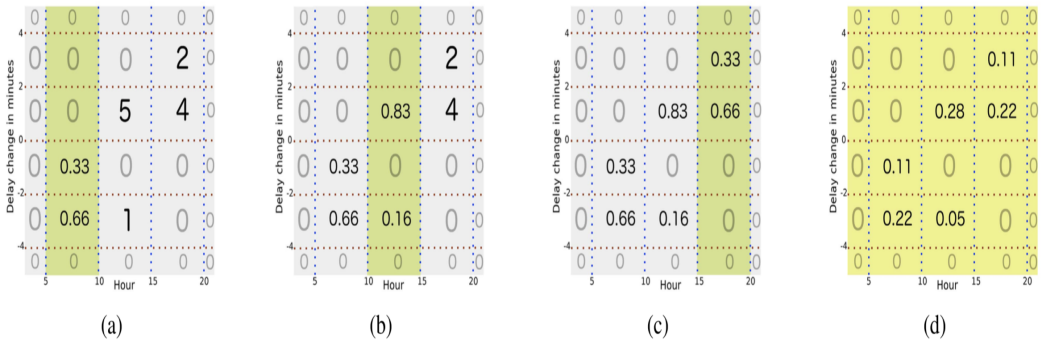

Figure 3.3: Normalization

2. Clustering

We have decided to use Agglomerative Clustering. This approach allows us to follow the dendogram and understand each of the splits, allowing additional analysis and deeper understanding of the clustering process. Agglomerative clustering starts with assigning every edge to be a separate cluster. Then in each step, it performs cluster merge between two nearest clusters. Clusters distances are maintained in a distance matrix, which at the beginning of the process is the same as the distance matrix between all objects (edges). After the merge of two closest clusters, the distance matrix has to be updated - algorithm

removes distances between pairs containing merged clusters and recomputes the distance between new cluster and old clusters using a selected linkage metric.

3. Distance Metric Between Edges

   Now, we have all the edges in the form a uniform matrix structure, which make the computations between the edges simpler.It is crucial for any clustering to choose a correct and domain fitting distance metric for comparing two objects. In our domain the distance metric should satisfy few requirements:

   - It should reflect the difference in cell values
   - It should take into account the distance between cells which values are being compared
   - It should reflect the distance in a way that allows us to consider small variances of data as noise - really similar objects should be close to each other

   Earth Mover's Distance (EMD):

   This distance is proportional to the amount of work that needs to be performed to transform one distribution (one field with piles of earth) into another (different field with different piles of earth). In our case, the total amount of earth is 1, since we normalised all of our bins at the end of processing edges to our representation for clustering. We used a python wrapper pyemd for calculating earth mover's distance was used [4]. Earth mover's distance considers all bins of our edge representation to be different piles of earth. Its goal is to find a minimal (in regard to performed work) way of transforming one edge representation to another. It does it using:

   - Values inside cells (probability of given delay value at given time)
   - Distances between cells

   For each bin, a distance to all other bins should be calculated. It should reflect the real distance between bins in both dimensions: delay and time changes.

4. Plotting

   The analysis requires visualization of data on a Map. The clusters obtained from the clustering process are projected onto Map using the Tableau Data Visualization Tool. Each and every cluster can be individually analysed for the delay incurred in the regions with the above tool.

3.4 Implementation

1. Data Preprocessing

   The analysis of Indian railway train delays is an extensive project which requires collection of large raw data. The raw data includes gathering of several train delays across various stations. As part of an analysis, ten days delays of each train is collected across all the stations. The scope of the analysis is restricted to the south region of the Indian Country (Tamil Nadu and Kerala). As part of analysis, there is a requirement of official running schedules of selected trains, so that the variation with the actual running schedules can be calculated.

   Processing: The raw data gathered is represented in excel workbooks and has to be preprocessed before it is available for the analysis.

   - Filtering based on Official Schedules:

     The trains official schedules data is collected from the official website of the Indian railway and represented in the excel workbook. The data gathered is filtered with the required stations in the considered region.

   - Filtering based on running schedules

     The train delay data is also filtered with the required stations in the considered region representing all the delays at the stations.

   - Obtaining the link edges

     The filtered data is taken into consideration for creating links between each successive station representing an edge between two stations.

   The filtered data is processed to obtain a final data of links and corresponding

delay times with the official schedules, producing the actual arrival times of the trains at the stations considered.

Now, a two dimensional grid matrix has to be generated where, the strongly connected station edges are represented on a two dimensional grid matrix with binary values as connection exists or not, by plotting all the stations on both the x axis and y axis accordingly. So, there has to be at least one connection between all the stations included.

2. After preprocessing, now the data points (delay,time) are represented into matrix and normalization is done using the bins consisting the data points. The outliers and empty bins are eliminated.

3. We have taken six bins on both axes of the matrix (time and delay) with variable range on the delay axis.

4. For Agglomerative Clustering we have chosen, distance matrix needs to be given as input which is obtained by the EMD Algorithm.

5. Each and every edge matrix is transformed into the other edge matrix by getting the edge matrices and the distance matrix of our time-delay matrix consisting cost of moving each and every bin to the other bin.

6. With these three inputs i.e., edge 1 matrix, edge 2 matrix and the distance matrix of bins to the EMD Algorithm, we get the distance matrix which becomes the input to the Agglomerative Clustering process.

7. Agglomerative Clustering gives the results in Dendograms (Figure 3.4), forming six clusters in our analysis, represented using different colours.

8. The analysis has to be made using the clusters being projected on the Map of Indian Country using the Tableau Data Visualization tool which takes the cluster data as input along with the latitude and longitude data of the selected stations.

9. Tableau gives the link representation of each and every cluster as shown in Fig and also the visualization of all the clusters on a single Map.

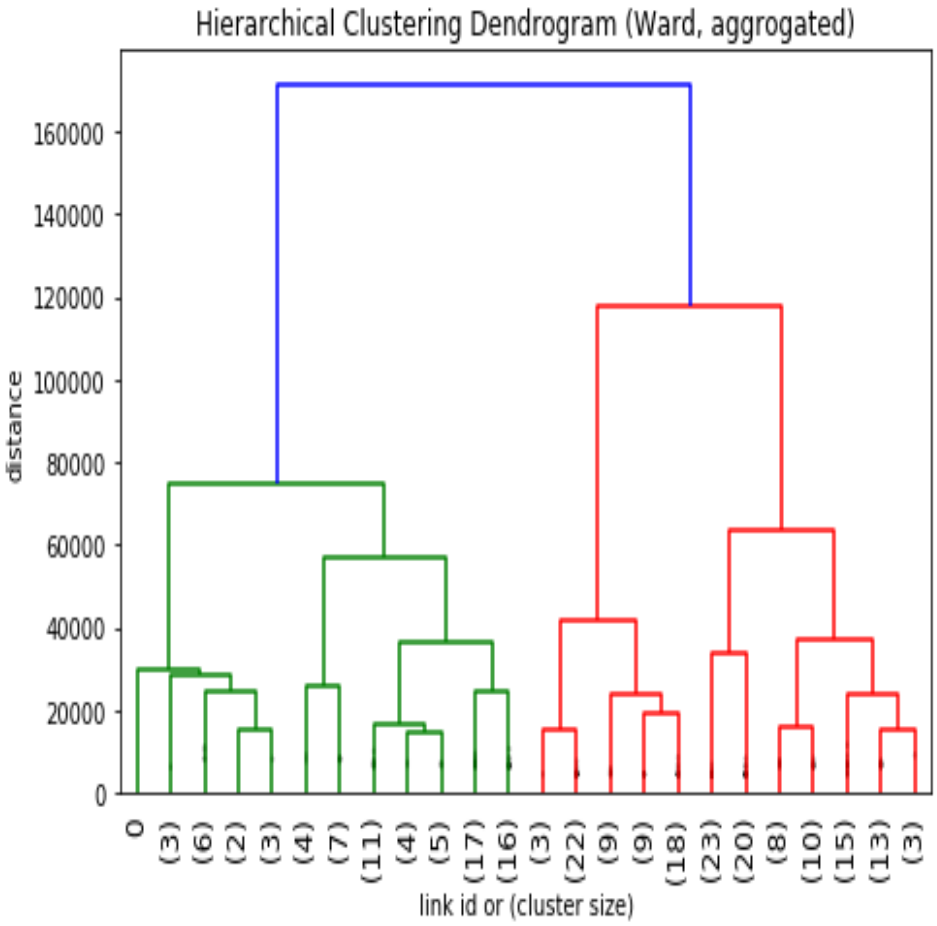

Figure 3.4: Agglomerative Clustering - Dendograms

Chapter 4

Results and Discussion

Analysis results from the clusters sent to the Tableau tool, gives the following visualizations for the states Tamil Nadu and Kerala (Refer Figure 4.1).

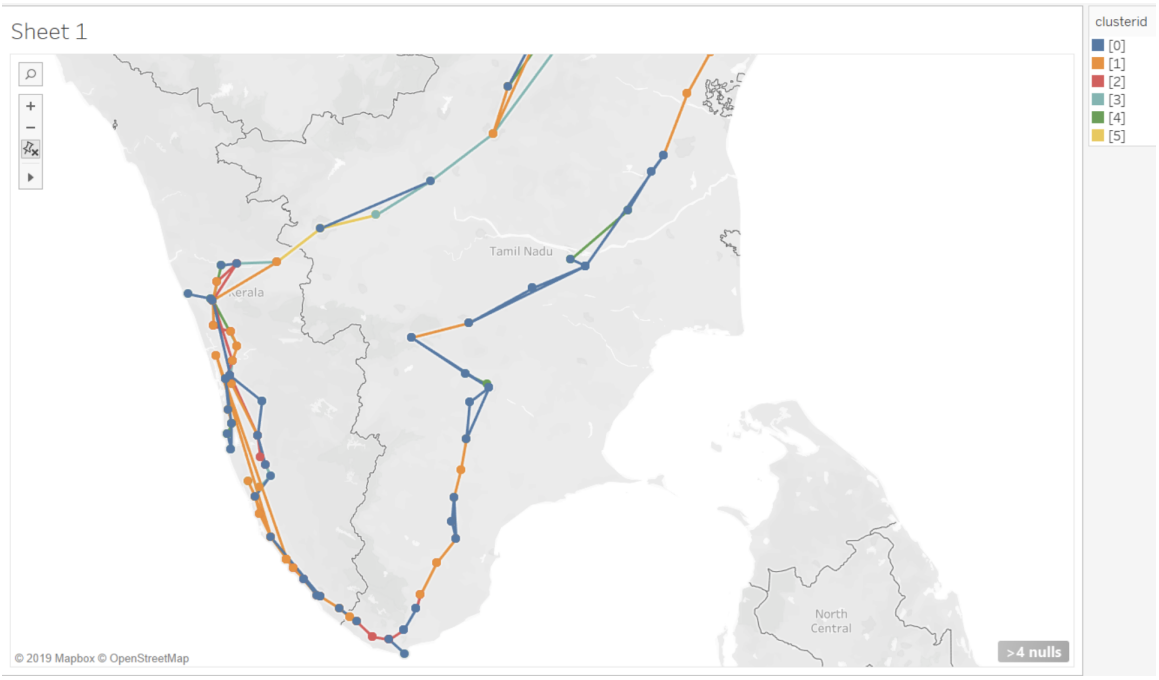

Figure 4.1: Clusters of Delay Patterns

We have dual axis map, which maps the edges and the stations onto a single map,after which all the links with stations are categorized into clusters (Refer Figure 4.2).

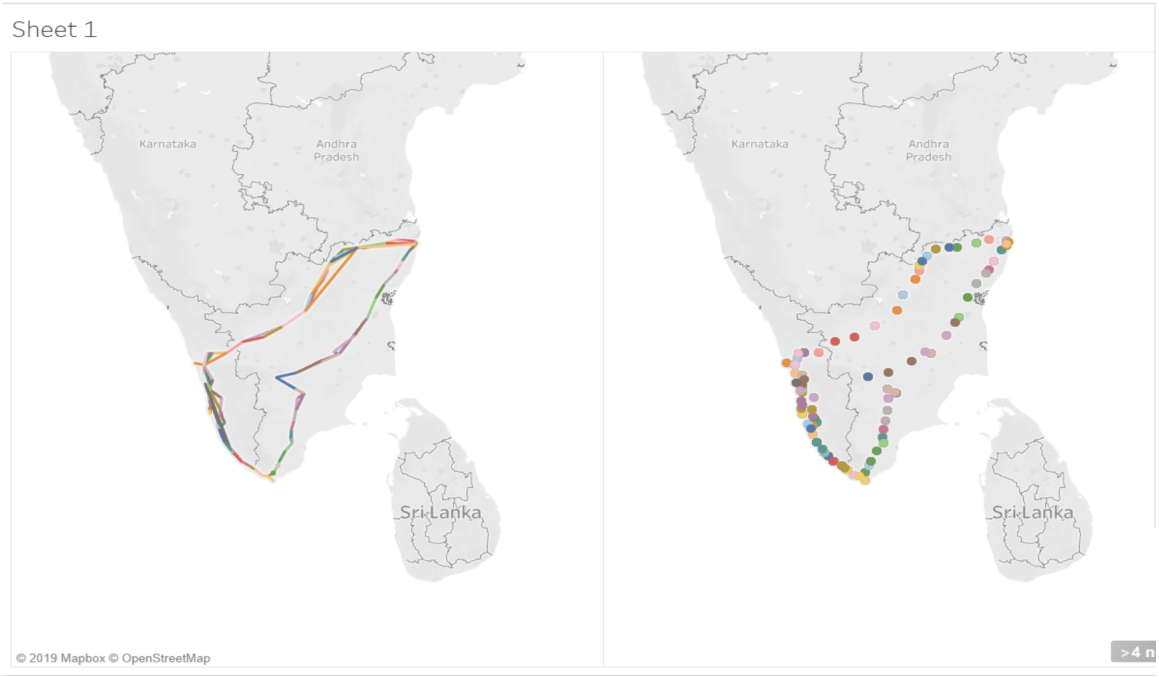

Figure 4.2: Dual Axis Map

Cluster-wise visualization makes it easier for analysis of similar delay patterns and their corresponding stations (Refer Figure 4.3).

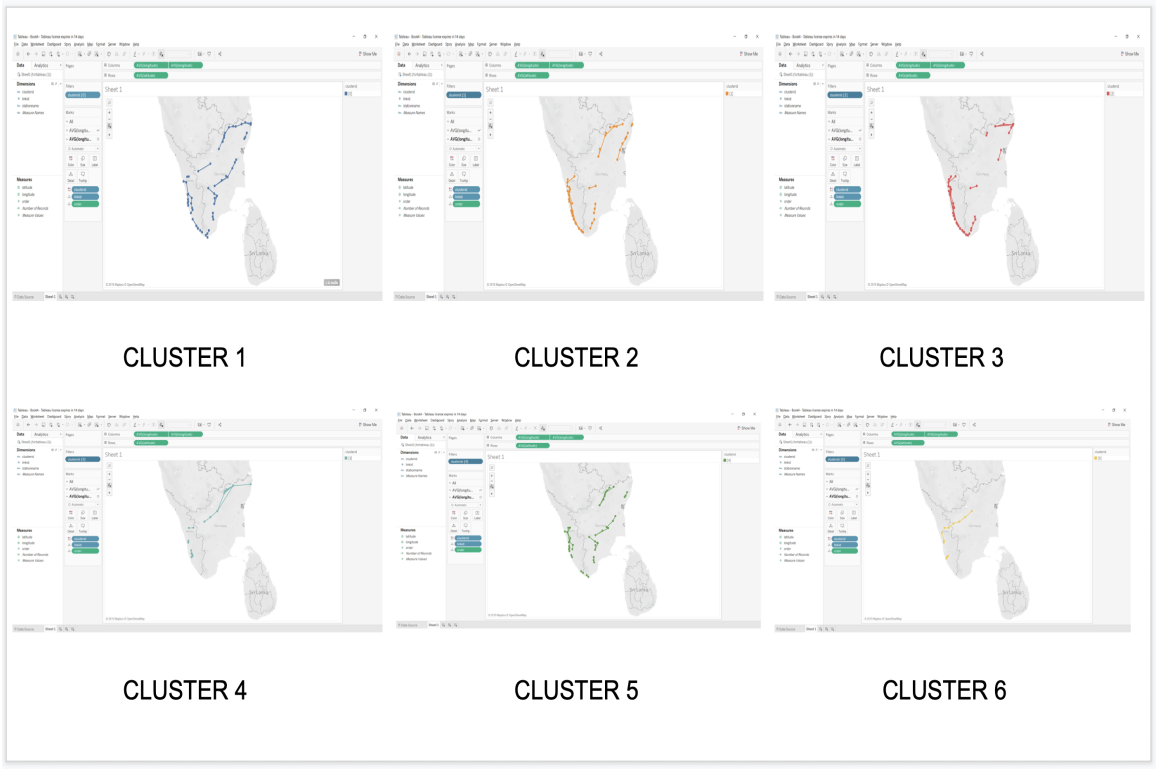

Figure 4.3: Cluster-Wise Representation

Chapter 5

Conclusion and Future Work

The analysis can be further developed by considering various locations in the Indian country. Potential stations where most of the delays occur can be taken into consideration for finding out the patterns of similar delay and projected for analysis. These delay patterns help to find out the regions where trains face issue with the travel time. The same can be taken into consideration by the Indian Railway department for citing various issues related to the delays at stations.

The said analysis would require a large data collection process of the running status of various trains over a period of time for accurate results. This would help in gathering the delays at stations for the considered trains. The extended scope of the project requires a large amount of space and time for the algorithm execution.

Chapter 6

Bibliography

[1] Piotr Szyman ski , Michał Z ołnieruk, Piotr Oleszczyk, Igor Gisterek, and Tomasz Kajdanowicz, "Spatio-Temporal Profiling of Public Transport Delays Based on Large-Scale Vehicle Positioning Data From GPS in Wroclaw", IEEE 2018.

[2] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, Cluster Analysis, 4th ed. Hoboken, NJ, USA: Wiley, 2009.

[3] G. Monge, Mémoire Sur la Théorie Des Déblais et Des Remblais. Paris, France: De l'Imprimerie Royale, 1781.

[4] W. Mayner. PyEMD: A Python Wrapper for Pele and Werman' s Implementation of the Earth Mover' s Distance Metric. Accessed: Jun. 14, 2017. [Online]. Available: https://github.com/wmayner/pyemd.

[5] O. Pele and M. Werman, "A linear time histogram metric for improved SIFT matching," in Computer Vision–ECCV. Berlin, Germany: Springer, 2008, pp. 495–508.

[6] O. Pele and M. Werman, "Fast and robust earth mover's distances," in Proc. IEEE 12th Int. Conf. Comput. Vis., Sep./Oct. 2009, pp. 460–467.