Thesis no:MSEE-2016:14

Forecasting Trajectory Data

A study by Experimentation

J N S Sri Harsha Vardhan Kamiestty

Faculty of Computing
Blekinge Institute of Technology
SE-371 79 Karlskrona Sweden

This thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science in Electrical Engineering with emphasis on Telecommunication Systems. The thesis is equivalent to 20 weeks of full time studies.

Contact Information:
Author:
J N S Sri Harsha Vardhan Kamisetty
E-mail: harshajns100@gmail.com

University advisor:
Julia Sidorova
Department of computer Science and Engineering

University Co-advisor:
Prof. Lars Lundberg
Department of Computer Science and Engineering

Faculty of Computing
Blekinge Institute of Technology
SE-371 79 Karlskrona, Sweden

Internet: www.bth.se
Phone: +46 4553850 00
Fax : +46 455 38 50 57

ABSTRACT

Context. The advances in location-acquisition and mobile computing techniques have generated massive spatial trajectory data. Such spatial trajectory data accumulated by telecommunication operators is huge, analyzing the data with a right tool or method can uncover patterns and connections which can be used for improving telecom services. Forecasting trajectory data or predicting next location of users is one of such analysis. It can be used for producing synthetic data and also to determine the network capacity needed for a cell tower in future.

Objectives. The objectives of this thesis is, Firstly, to have a new application for CWT (Collapsed Weighted Tensor) method. Secondly, to modify the CWT method to predict the location of a user. Thirdly, to provide a suitable method for the given Telenor dataset to predict the user's location over a period of time.

Methods. The thesis work has been carried out by implementing the modified CWT method. The predicted location obtained by modified CWT cannot be determined to which time stamp it belongs as the given Telenor dataset contains missing time stamps. So, the modified CWT method is implemented in two different methods.

1. Replacing missing values with first value in dataset.
2. Replacing missing values with second value in dataset.

These two methods are implemented and determined which method can predict the location of users with minimal error.

Results. The results are carried by assuming that the given Telenor dataset for one week will be same as that for the next week. Users are selected in a random sample and above mentioned methods are performed. Furthermore, RMSD values and computational time are calculated for each method and selected users.

Conclusion. Based on the analysis of the results, Firstly, it can be concluded that CWT method have been modified and used for predicting the user's location for next time stamp. Secondly, the method can be extended to predict over a period of time. Finally, modified CWT method predicts location of the user with minimal error when missing values are replaced by first value in the dataset.

Keywords: Collapsed weighted tensor method, periodic temporal link prediction, Trajectory data.

ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to my supervisor, Yulia Sidorova and Prof. Lars Lundberg for their valuable guidance and encouragement throughout the period of thesis. Furthermore, I extend my gratitude to Prof. Lars Lundberg for taking time and helping me in technical study for my thesis work.

I like to thank my room mates Srinivas Jayapala, Ranjith Vuppala and Tarun alguri for their support during my thesis in Sweden.

I extend my thanks to all my family and friends for their support at every step of my education.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The advances in location-acquisition and mobile computing techniques have generated massive spatial trajectory data, which represents the mobility of different objects, people and vehicles. Such Spatial Big Data is being accumulated by telecommunication operators, analyzing such massive data will be the key to turn it into business insights.

The data in different analysis applications such as social networks, web analysis and collaborative filtering consists of relationships, which can be considered as link between objects. For example, two people may be linked to each other, if they exchange emails or phone calls. These links can be used to detect the missing links or predict the future links.

The term "link prediction" refers to the following: If data for *T* time steps is given, then we can predict the link at time *T+1*. Temporal link prediction is used to predict temporary links between persons/entities to compensate for missing values, until the data set is complete. Periodical temporal link prediction can be defined as , If *T* time steps is given, predicting the relationship at times *T+1, T+2......, T+L,* where L is length of the periodic pattern [1].

A tensor is a multi-dimensional array with geometric interpretation. Collapsed Weighted Tensor (CWT) is a method used in data analysis to collapse all the given data into a single matrix. CWT is an alternative to Collapsed Tensor method (CT) which gives higher priority to more recent values and proven to be effective technique [1].

This thesis focuses on forecasting trajectory data from given spatial trajectory dataset that is to predict the location of the user at next instances (*T+1, T+2....*) with given data for the user. CWT method is considered and used for forecasting trajectory data.

1.1 Motivation

Spatial trajectory data accumulated by telecommunication operators is huge, analyzing the data with a right tool or method can uncover patterns and connections which can be used for improving telecom services. Forecasting trajectory data or predicting next location of users is one of such analysis. It can be used for producing synthetic data and to determine the network capacity needed for a cell tower. So there is a need to develop a method that can forecast trajectory data.

Data forecasting is handled from different perspectives like Markovian perspective or a graph perspective [1]. This thesis concentrates on a matrix based perspective, because it is more rooted into telecommunication systems research.

1.2 Problem Statement

The CWT method has proven to be effective [1] on link prediction in previous line of work, but it can only be used to predict for one instance and to predict the probability, but not directly a value. The contribution of this thesis is to have a new application for CWT method and to extend the method from t+1 to t+ L.

Telenor database given for research consists of Trajectory data for a period of one week and user size of *27,000*. Computing such huge data is complex and time consuming. So, the CWT method is modified and used for computing a random sample and repeat the experiment to average the outcomes.

1.3 Aim and Objectives

The main aim of this project is to provide a suitable modified version of CWT method for forecasting trajectory data and use the method to predict the location for multiple instances of time.

Objectives:

- Literature review on different methods used for forecasting trajectory data.
- Modifying CWT method to predict the location at next instance of time.
- Using CWT modified methods to predict multiple instance of time.
- Finding a suitable CWT modified method for forecasting trajectory data by analyzing the results.

1.4 Research Questions

Q1. How can CWT method be modified to predict the user's location at next instance of time?

Q2. Given this extension, how can we use the CWT method to predict user's location at multiple instances of time?

Q3. Which method is suitable for the given data among the modified variants of the CWT method for forecasting trajectory data?

The above research questions are answered in this thesis by implementing the CWT method and modifying it to predict location at multiple instances of time.

1.5 Contribution

The contribution of this thesis is:

1. Firstly, to modify CWT method to predict the location of the user at next instance of time as it is used in previous line of work for link prediction to predict the probability of which author may publish in a conference as explained in chapter *3*.
2. Secondly, the thesis adds a temporal aspect to the CWT method, that is to use it for predicting the user's location at next instance of time.
3. Lastly, the thesis finds a suitable method of the modified CWT method for forecasting the trajectory data with the given data set.

1.6 Thesis Outline

Chapter 1, this chapter provides an overview of the research area. It also provides the motivation for this thesis, the problem at hand, research questions and contribution of this work. Chapter 2, contains the related work regarding thesis. Chapter 3, provides the background knowledge and explains the method used in previous line of work. Chapter 4, explains the research method and the variants of the method used to find the suitable method. Chapter 5, contains the result and analysis. Chapter 6, contains discussion on the thesis. Finally, Chapter 7 covers conclusion and future work.

2 BACKGROUND

This chapter describes the concepts required to understand the thesis and the method used in previous line of work.

The data in many domains such as web analysis, social networks, telecommunications, etc. is link based, the link structure can be processed with different data mining procedures. The periodic temporal link prediction problem is, if the given link data is for time T steps predicting the link structure at times T+1, T+2, …… T+L, where L is the length of the periodic pattern.

In [1] for periodic temporal link prediction problem authors have considered DBLP biometric data set. The dataset contains publication data for large number of professional conferences in are related to computer science over a period of ten years from 1991-2000. The authors aim is to know which author is likely to publish in which conference for year 2001. As the data is so large computing is complex, so the data is computed by taking subsets of data. The subsets are divided in such a way that "which author is likely to publish in a particular conference" can be answered. The data is converted into BINARY by a function z which is of size M x N x T where M are the number of authors and N are conferences, and T is the time. The multivariable function z is defined as:

$$Z\,(i,j,t) = \begin{cases} 1 \; if\; object\; i\; links\; to\; object\; j\; at\; time\; t \\ \quad 0 \qquad otherwise \end{cases} \qquad \ldots\ldots\ldots\ldots (1)$$

The authors considered collapse tensor (CT) method for collapsing the data into a single matrix, but the CWT is used as an alternative to CT which is proven to be effective by [2], The matrix is used to calculate the probability of an author to publish in a conference.

CWT method is as follows; it collapses all the data into a single matrix by giving higher priority to the most recent values. The equation for calculating CWT is

$$X_{(i,j)} = \sum_{t=1}^{T} (1-\theta)^{|T-t|} Z_t(i,j) \qquad \text{where } \theta \in (0,1) \;\ldots\ldots (2)$$

This method is modified and used in this thesis by normalizing the equation as shown in equation (3), as to predict the location directly rather than using a binary value to predict probability that a user is most likely to be there at that location.

3 RELATED WORK

This chapter briefly discusses about the previous line of work on this research area which has been a motivation for implementing and completing the thesis.

Hasan et. al [3], has explained a procedure for constructing a dataset to perform link prediction. Authors have identified a short list of features for link prediction in Co-authorship domain. These features provide accuracy and can also be applied for other domains like social networks domain. Authors have evaluated each feature visually and performed a comparative analysis, by comparing the class density distribution through well know ranking algorithms.

In Liu et. al [4], the main task of the author in this thesis is collaborative filtering, that is the objective is to predict interest of users to objects (movies, music, books) based on the interest of the similar users. Authors have considered Netflix data containing 480 thousand users who have given ratings to 18 thousand movie titles. Authors have considered the task as a link prediction problem and solved the task by using selective sub sampling, reviewing scores and with graph topology and by projecting features over time.

Clauset et. al [5], In this the authors have explained different hierarchical structure of the network and predicted the missing links in the network. This helped in thesis to differentiate the temporal link prediction problem from missing link prediction, where the goal is to predict missing links in order to describe a complete picture of overall link structure.

Huang et. al [6], In this author has introduced the time series link prediction problem by considering temporal evolution of link occurrences to predict link occurrence probabilities at a particular time. The author has combined static graph link prediction algorithms and time series models producing significantly better predictions over static graph link prediction methods.

Dunlavy et. al [1], In this the author has considered the problem of temporal link prediction where the goal is to predict links between a time interval in future. Author considered bipartite graphs that evolve over time, matrix and tensor based techniques for predicting future links. The author uses a weighted based method for collapsing all the data into a single matrix. The matrix is used in Katz method which is then extended to bipartite graphs and approximated in a scalable way with truncated singular value decomposition (TSVD). Through several experiments matrix and tensor based methods have proven to be effective for temporal link prediction problem.

In [7] the authors focuses on analysis of spatial data collected by Telenor Sweden. The authors aim at developing method for spatial data analytics and finding an optimal technology to convert research prototype into an industrial application which is scalable.

Authors in [8] have evaluated two cellular network load optimization strategies. Firstly, Tetris optimization to provide most even load in the network. Secondly, cell expansion to selectively expand the capacity of heavily loaded radio cells using cell splitting.

In [9 – 11], authors have used the mobility data to predict home and office locations of the subscribers. This information has been useful for enhancing network services. In [12], [13] authors have developed a subscriber profiling model based on the network load the user generates, and their mobility patters. In [14], authors have characterized subscriber mobility and temporal activity patterns to identify their relation with the traffic volume. Furthermore, the author has investigated the efficient usage of radio resources by different subscribers as well as by different applications.

In [11], [15 – 17], authors have predicted the subscribers future location based on their mobility history. These predictions have been used to develop efficient network paging algorithms and network topologies which led to massive savings in the number of signals made to locate users in the network.

4 METHODOLOGY

This chapter discusses the method followed to achieve the aim, i.e. to provide a suitable method for forecasting trajectory data for given Telenor database.

The CWT method in previous line of work is used to predict the probability whether an author is going to publish in a conference or not. To achieve the aim of the thesis, i.e. to predict the location at next time slot (*T+1*) of the user. The method is modified as shown below. The motivation for modifying CWT method is, by using the method as in previous line of work probability of a user present at a location can be determined, but the location of a user cannot be obtained. So it is modified to predict the location of a user.

Let X_i be the predicted location at time *T+1*, The modified equation used for predicting data is

$$X_i = \sum_{t=1}^{T}(1-\theta)^{|T-t|}x_{(i,t)} \,/\, \sum_{t=1}^{T}(1-\theta)^{|T-t|} \quad \text{where } \theta \in (0,1) \; \ldots (3)$$

In equation (3) the parameter $\theta \in (0,1)$ can be chosen according to experiments on various training data sets. $x_{(i,t)}$ is a *1* x *2* matrix which contains the location of the user at time t, i.e.

$$x_{(i,t)} = [latitude\ at\ t \quad longitude\ at\ t] \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \; (4)$$

The CWT method has been normalized as shown in equation (3), so as to predict the location of the user at next instance of time.

$$f(t) = (1-\theta)^{T-t} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

The equation (5) which is used in modified CWT method in equation (3) gives greater weights to more recent values, which is depicted in Figure 4.1.

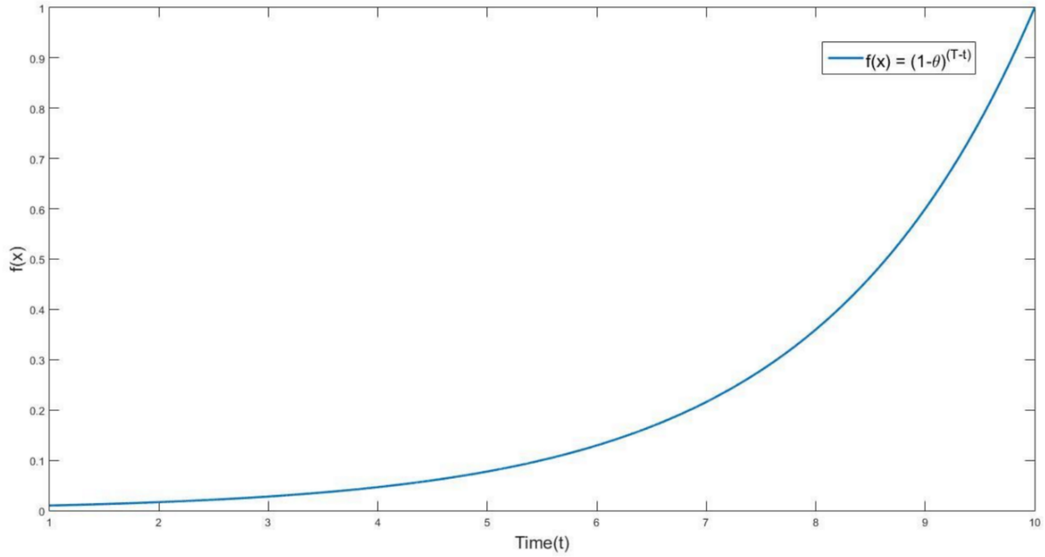


Figure 4-1 Decay function

Figure 4.1 gives us the plot of equation (5) for $\theta = 0.4$ and $T = 10$. $f(t)$ is also known as decay function which gives us the priority for the location at time t.

Modified CWT method is only used to predict the position at next time stamp. To predict position of the user for a periodic temporal aspect, the modified CWT method is extended in a recursive way, where in the predicted value using CWT is added to already existing data to predict the data in next instance of time. For example, the given data consists of T time steps, the location of the user is predicted for the instance of T+1. This predicted value is added to the existing trajectory data and used in prediction of location at T+2 instance and so on.

4.1 Data analysis and preprocessing

The dataset used for forecasting trajectory data was provided by Telenor Sweden. The relevant features in the dataset are presented below, in Table 4.1.

Feature	Description
User Id	Unique identification for each subscriber
Timing	Time stamp of the subscriber connected to the cell tower
Weekday	Time stamp recorded on weekday from Sunday to Saturday
lat	latitude of the subscriber’s location
lon	longitude of the subscriber’s location

Table 4-1Relevant Features in the dataset

The dataset also consists of other features of the user like operating system of the handset used, Site id for the radio cell, store location of the subscription purchase. As the thesis focuses on forecasting the trajectory data, the features mentioned in Table 1 are sufficient.

The dataset contains historical location data of *27010* users in a network of *21801* radio cells during one week in *2015*. The location information of the user in dataset refers to the location of the radio cell to which subscriber is connected. The users or the subscriber’s location has been registered for every 5 minutes, if the subscriber receives or generates a phone call, or a Short Message Service (SMS). For example, if a user receives a phone call in between the time *12:00:00* to *12:04:00* then in the dataset the time f connection will be noted as *12:00:00*. If a subscriber does not receive or send any phone call or SMS in a given *5* minutes’ time slot, then there will be no record of the radio cell to which the subscriber is connected to during this time slot. Hence there will be lot of missing time slots or location in the data for each subscriber.

The dataset was preprocessed to eliminate the duplicate entries in the dataset. Duplicate entries refer to multiple entries of the subscriber with a location of radio cell with in the same *5* minutes’ time slot. For example, duplicate entries of a subscriber refer to the subscriber receiving or sending multiple phone calls or SMS in a given *5* minutes' time slot. The motivation for removing duplicate entries is that the same location at same time and weekday will be registered in the dataset which cannot be used in forecasting trajectory data. The dataset does not contain the day of the week, but not the date when the location of the subscriber has been registered. Hence for this thesis as the week starts from Sunday we consider Sunday as the first day for the subscriber and Saturday as the last.

4.2 Methods

Computing Telenor data consisting of 27000 users at a time is complex and time consuming, so a sample of 100 users at a time is considered by normalizing the length of the sequence in the trajectory over a period of one week. The sample of users are processed through the methods one user at a time to predict the location at next time stamp (*T+1, T+2……* *T+L*) for each user. As explained above there are lot of missing values in the dataset, the data cannot be directly used for predicting the location at next time steps for a user. The motive for replacing the missing values is, the predicted location for next time step of the user with missing values in the data cannot be determined at which instance of time the predicted location is.

As the thesis focuses on predicting location of the user at next instance of time which is different from missing values prediction as explained in chapter1. So two methods for varying the missing values of the dataset are considered

1. By replacing the missing values with the first value in the data set.
2. By replacing the missing values with the previous value in the dataset.

The two variations of the method are then processed through three methods to observe how the prediction varies. These methods are as explained below:

4.2.1 Method 1

In this method by considering the dataset given for a user and replacing the missing values with either first value or previous value in the dataset. Modified CWT method as in equation (3) is performed and the location is predicted at next instance. This method is performed recursively as explained above to predict the locations for next whole week.

For example, for a user *234,* if the first location registered is on Sunday *6:00:00* then all the missing locations at *5* minutes' time interval in the data set are replaced by the location registered at *6:00:00* or by the previous value in the dataset. This data is used in the CWT that is in equation (*1*) to predict the location at next instance of time (*T+1*), by implementing it recursively, i.e. by adding the predicted location to the dataset and by using the equation (*1*), location at next instance of time (*T+2*) is predicted and so on.

4.2.2 Method 2

In this method by removing the last entered value in the dataset and replacing the missing values with either first value or previous value in the dataset. CWT method is performed as in equation (*1*) and the location is predicted at next instance of time. This method is performed recursively and the locations are predicted for next whole week, i.e. from Sunday to Saturday. The motive to remove the last entered value is to see the variation in the prediction.

For example, for a user 234, if the last registered location is on Saturday at time *22:00:00.* Then the registered location is removed and CWT method is performed to predict the user's location for next whole week.

4.2.3 Method 3

In this method as most of the users in the dataset location changes between time *12:30:00* to *13:30:00.* The locations registered at this time interval are considered from Sunday to Saturday. The missing values are replaced either by first value or by previous values and by implementing equation (1) recursively the location prediction for the next whole week is obtained.

For example, for a user 234 the locations in the dataset between time interval *12:30:00* to *13:30:00* are considered. The missing values are replaced by the first value in the dataset. CWT method is performed as in equation (1)

4.3 Experimental procedure

The procedure followed for experimentation is shown below in Figure 4.2.

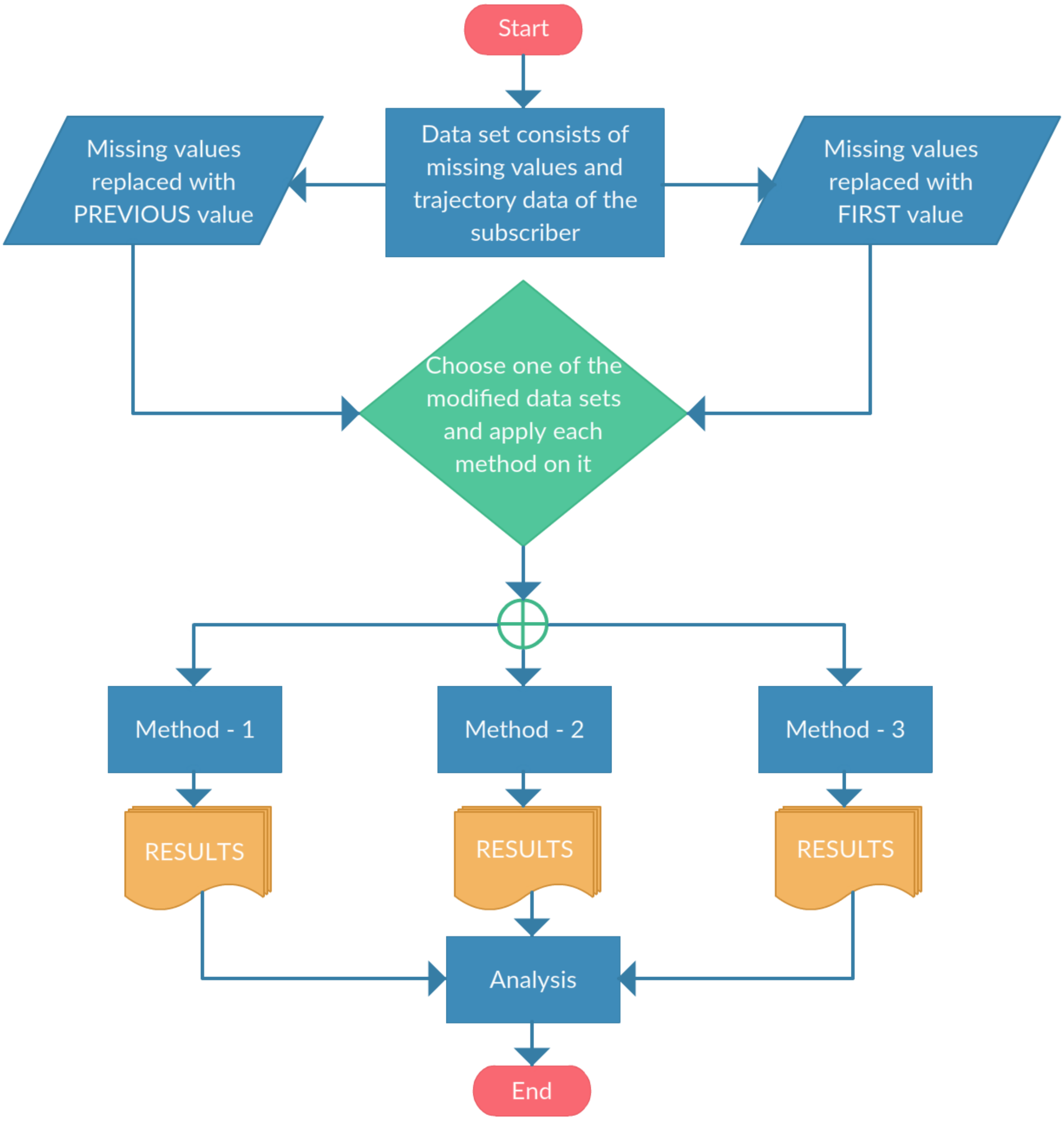

Start
Data set consists of missing values and trajectory data of the subscriber
Missing values replaced with PREVIOUS value
Missing values replaced with FIRST value
Choose one of the modified data sets and apply each method on it
Method - 1
Method - 2
Method - 3
RESULTS
RESULTS
RESULTS
Analysis
End

Figure 4-2 Flow chart representing the experimental procedure

As shown in the above Figure 4.2. There are two methods to replace the missing values in the dataset

1. Replacing missing values with first value in dataset.
2. Replacing missing values with second value in dataset.

Each of this two methods are then processed through 3 methods to observe the variations in prediction

1. Method 1: is to predict the location of the user for next whole week, i.e. the next 2016 location with given data for a period of one week.
2. Method 2: is to predict the location of the user for next whole week by removing the last value in given dataset.
3. Method 3: is to use the data between time interval 12:30:00 to 13:30:00 every day of the week, as most of user's location varies at this time interval. To use this data and predict the location of the user at this time interval for next whole week

The prediction through these methods is observed by varying the data as mentioned in the above methods, and by using the modified CWT method as in equation (3). The parameter θ in equation (3) is varied for the values of 0.2, 0.4, 0.6, 0.8 and the predicted locations are observed for each of the above-mentioned methods.

4.4 Evaluation metrics

Two evaluation metrics for each of the methods have been observed, they are:

1. Computational time
2. Root Mean Square Deviation (RMSD)

4.4.1 Computational time

Computational time refers to the time taken for a method to complete the experiment, i.e. to predict the locations of the user for a method. This metric is used to know which method is faster.

4.4.2 Root Mean Square Deviation (RMSD)

The RMSD represents the sample standard deviation of the differences between predicted values and observed values. The RMSD is given by

$$\text{RMSD} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}} \qquad \text{.... (6)}$$

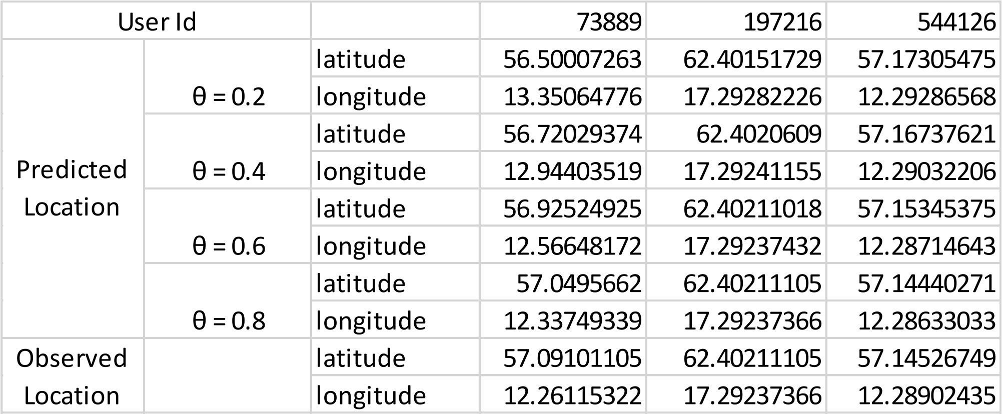
RMSD is calculated by assuming that, the given dataset will be same for the next week. In equation (6), 'n' indicates the number of time stamps for a user in the given dataset. RMSD for a user is calculated using time stamps in the given dataset and the predicted locations at same time stamps without considering missing values, since they are assumed.

5 RESULTS AND ANALYSIS

This chapter analysis the results obtained by implementing the methods explained in chapter 4.

5.1 Prediction with modified CWT method:

CWT method is modified and used for predicting location at next time stamp as explained in chapter 4 equation (3). The results obtained by using equation (3) with the given dataset and without replacing missing values are:



User Id			73889	197216	544126
Predicted Location	θ = 0.2	latitude	56.50007263	62.40151729	57.17305475
		longitude	13.35064776	17.29282226	12.29286568
	θ = 0.4	latitude	56.72029374	62.4020609	57.16737621
		longitude	12.94403519	17.29241155	12.29032206
	θ = 0.6	latitude	56.92524925	62.40211018	57.15345375
		longitude	12.56648172	17.29237432	12.28714643
	θ = 0.8	latitude	57.0495662	62.40211105	57.14440271
		longitude	12.33749339	17.29237366	12.28633033
Observed Location		latitude	57.09101105	62.40211105	57.14526749
		longitude	12.26115322	17.29237366	12.28902435

Table 5-1 Location predicted for next time stamp

Table 5.1 shows the results obtained by the modified CWT method to predict the location at next time stamp. These values are obtained without replacing missing values in the data, and by varying θ at 0.2,0.4,0.6,0.8 as shown in the above Table 5.1. As the predicted locations in the table shows forecasted location is close to the observed location. The error is minimal at θ=0.8, as it gives highest priority to the most recent value than rest of the varying θ values.

The motivation for replacing missing values is, the location predicted by modified CWT method cannot be said to which time stamp it belongs to, as the time interval between the data varies. So, modified CWT method is performed by replacing the missing values in the dataset and the method is extended to periodic temporal aspect as explained in chapter 4.

The two methods to replace missing values in the data set are:

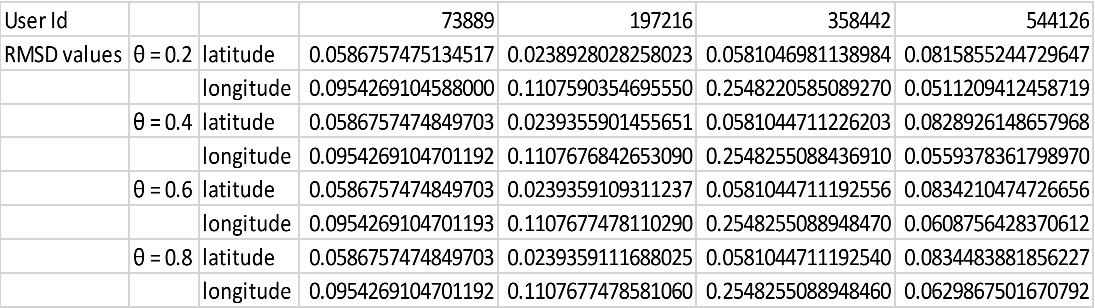
3. Replacing missing values with first value in dataset.
4. Replacing missing values with second value in dataset.

5.2 Replacing missing value with first value:

In this method missing values in the dataset are replaced by the first value in the dataset. As explained in chapter 4.2 and 4.3, to see how the prediction varies this method is performed in three different methods.

5.2.1 Method 1:

As explained in chapter 4, in this method missing values are replaced with first value and the locations of the user for the next whole week are predicted by using equation (3). The RMSD values for this method are calculated by assuming that the locations of the user are same for the next week as the given week data.

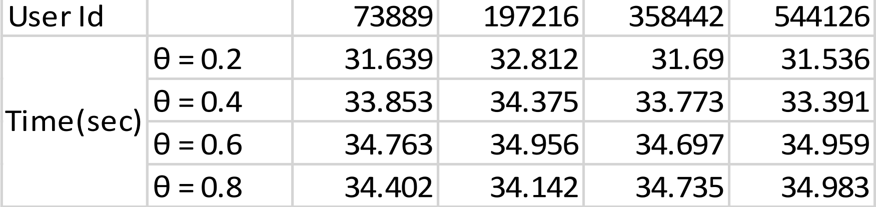


User Id			73889	197216	358442	544126
RMSD values	θ = 0.2	latitude	0.0586757475134517	0.0238928028258023	0.0581046981138984	0.0815855244729647
		longitude	0.0954269104588000	0.1107590354695550	0.2548220585089270	0.0511209412458719
	θ = 0.4	latitude	0.0586757474849703	0.0239355901455651	0.0581044711226203	0.0828926148657968
		longitude	0.0954269104701192	0.1107676842653090	0.2548255088436910	0.0559378361798970
	θ = 0.6	latitude	0.0586757474849703	0.0239359109311237	0.0581044711192556	0.0834210474726656
		longitude	0.0954269104701193	0.1107677478110290	0.2548255088948470	0.0608756428370612
	θ = 0.8	latitude	0.0586757474849703	0.0239359111688025	0.0581044711192540	0.0834483881856227
		longitude	0.0954269104701192	0.1107677478581060	0.2548255088948460	0.0629867501670792

Table 5-2 Method-1RMSD values when replaced by first value

Table 5.2 shows the RMSD values obtained by implementing method 1. As observed in the table for user 73889 for θ = 0.4,0.6, and 0.8 have same RMSD values because more than 80 % of the users given data has same location registered. As observed from the table for the users 197216,358442 and 544126 the RMSD values are small. The change in RMSD values when θ changes is small and only for θ = 0.2 the value differs, because at θ = 0.2 the priority given for the locations is less compared to other θ values.

The computational time is also one of the parameter observed. The computational time is observed by implementing each method and each variation of θ up to 5 times and by averaging the time. The computational time for the users to this method is shown in the below table.



User Id		73889	197216	358442	544126
Time(sec)	θ = 0.2	31.639	32.812	31.69	31.536
	θ = 0.4	33.853	34.375	33.773	33.391
	θ = 0.6	34.763	34.956	34.697	34.959
	θ = 0.8	34.402	34.142	34.735	34.983

Table 5-3 Computational time for Method 1 when replaced by first value

As shown in the above table the average time for a user to perform this method for various values of θ is 33.794sec.

5.2.2 Method 2:

As explained in chapter 4, in this method the last value in the given dataset is removed. The missing values in the dataset are replaced by first value in the dataset. The locations of the user for next whole week are predicted by using equation (3). The RMSD values are calculated by assuming the data of the user is same for the next week as the given week data.

User Id			73889	197216	358442	544126
RMSD values	θ = 0.2	latitude	0.0586757475134517	0.0238948632274360	0.0581039296131267	0.0814845767212889
		longitude	0.0954269104588000	0.1107605310362350	0.2547941764526840	0.0546846315191032
	θ = 0.4	latitude	0.0586757474849703	0.0239233981300990	0.0581037794418144	0.0828055227931507
		longitude	0.0954269104701192	0.1107661948151580	0.2547964552296320	0.0590254466740769
	θ = 0.6	latitude	0.0586757474849703	0.0239235311910140	0.0581037794399972	0.0833819014207970
		longitude	0.0954269104701192	0.1107662212105760	0.2547964552572450	0.0622173905326784
	θ = 0.8	latitude	0.0586757474849703	0.0239235312382769	0.0581037794399956	0.0834417579350964
		longitude	0.0954269104701192	0.1107662212199520	0.2547964552572450	0.0631209244753361

Table 5-4 RMSD values for Method-2 when replaced by first value

Table 5.4 shows the RMSD values by implementing method 2 when missing values are replaced by first value. As observe in Table 5.2 and 5.4 the values of user 73889 does not change because more than 70 % of the locations given in the data set are same and the last value removed in the dataset is same as the first value in the dataset. So, missing values in the dataset when replaced by first value results in the same location prediction as Method 1 for the user 73889. As shown in the Table 5.4 above for other users 197216,358442 and 544126 the change in the RMSD values can be observed after 5 decimal points.

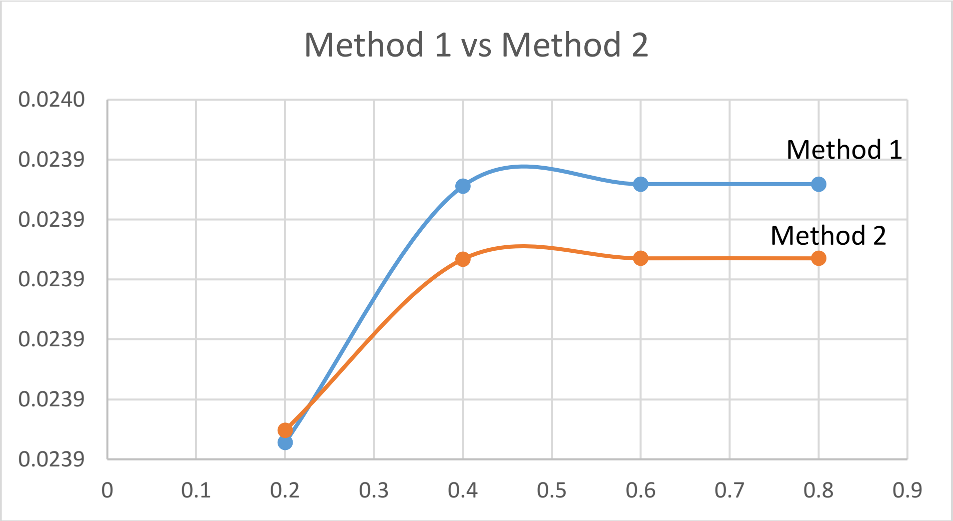
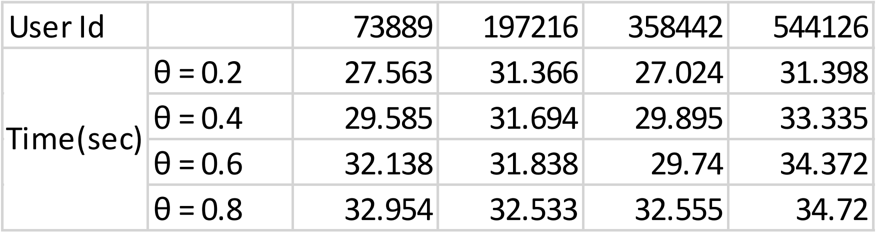


Figure 5-1 Graph between Method 1 and Method 2

For example, Figure 3 shows us the comparison between latitudes of 197216 user for Method 1 and Method 2. As we can observe the RMSD value at θ = 0.2 are almost same, but for θ = 0.4, 0.6, and 0.8 error varies and has less error for Method 2. It shows that even a single value in the dataset effects the prediction.

The computational time observed by implementing Method 2 are given in the below Table 5.5.



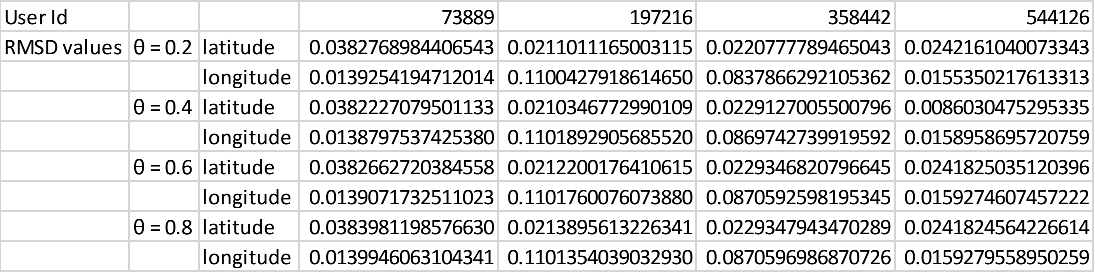
User Id		73889	197216	358442	544126
Time(sec)	θ = 0.2	27.563	31.366	27.024	31.398
	θ = 0.4	29.585	31.694	29.895	33.335
	θ = 0.6	32.138	31.838	29.74	34.372
	θ = 0.8	32.954	32.533	32.555	34.72

Table 5-5 Computational time for Method 2 when replaced by first value

The computational time as observed from table 5.3 and 5.5 show that Method 2 takes less time to compute than Method 1. The average computational time for this method for various values of θ is 31.415 sec.

5.2.3 Method 3

In this method as explained in chapter 4, most of the users in the dataset location changes between time *12:30:00* to *13:30:00.* So, the data between time interval for the week is considered and the data between the same time interval for next week is predicted. RMSD values obtained by this method are:

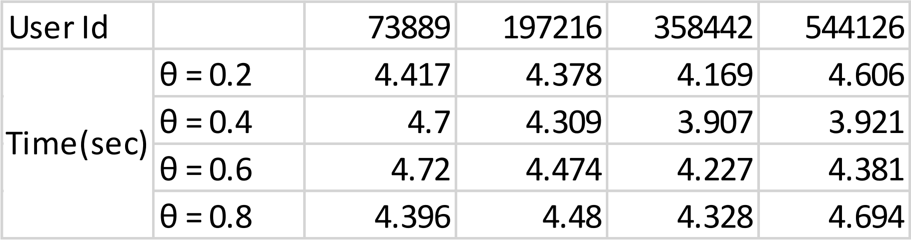


User Id			73889	197216	358442	544126
RMSD values	θ = 0.2	latitude	0.0382768984406543	0.0211011165003115	0.0220777789465043	0.0242161040073343
		longitude	0.0139254194712014	0.1100427918614650	0.0837866292105362	0.0155350217613313
	θ = 0.4	latitude	0.0382227079501133	0.0210346772990109	0.0229127005500796	0.0086030475295335
		longitude	0.0138797537425380	0.1101892905685520	0.0869742739919592	0.0158958695720759
	θ = 0.6	latitude	0.0382662720384558	0.0212200176410615	0.0229346820796645	0.0241825035120396
		longitude	0.0139071732511023	0.1101760076073880	0.0870592598195345	0.0159274607457222
	θ = 0.8	latitude	0.0383981198576630	0.0213895613226341	0.0229347943470289	0.0241824564226614
		longitude	0.0139946063104341	0.1101354039032930	0.0870596986870726	0.0159279558950259

Table 5-6 RMSD values for Method 3 when replaced by first value

As observed in Table 5.6 the RMSD values for Method 3, the error is small compared to other methods. This shows that the modified CWT method can be performed even when location of the user changes more frequently. This also shows that the predicted locations from this method will have minimal error even when the location of user changes frequently.

The computational time observed by implementing Method 3 are given in the below Table 5.7.



User Id		73889	197216	358442	544126
Time(sec)	θ = 0.2	4.417	4.378	4.169	4.606
	θ = 0.4	4.7	4.309	3.907	3.921
	θ = 0.6	4.72	4.474	4.227	4.381
	θ = 0.8	4.396	4.48	4.328	4.694

Table 5-7 Computational Time for Method 3 when replaced by first value

As shown in the Table 5.7, computational time for Method 3 is less compared to other methods, as it predicts only between an interval of time rather than predicting for the whole week. The average computational time for Method 3 for various values of θ is 4.383 sec.

5.3 Replacing missing value with previous value:

In this method missing values in the dataset are replaced by the previous value in the dataset. As explained in chapter 4.2 and 4.3, to see how the prediction varies this method is performed in three different methods.

5.3.1 Method 1:

As explained in chapter 4, in this method missing values are replaced with previous value and the locations of the user for the next whole week are predicted by using equation (3). The RMSD values for this method are calculated by assuming that the locations of the user are same for the next week as the given week data.

User Id			73889	197216	358442	544126
RMSD values	θ = 0.2	latitude	0.0586757484305400	0.0214326084829863	0.0610344248147777	0.0822304305270777
		longitude	0.0954269100939965	0.1101287253942370	0.2119252529629390	0.0327365111213618
	θ = 0.4	latitude	0.0586757474849703	0.0214325625539205	0.0610344248147744	0.0837043239693932
		longitude	0.0954269104701192	0.1101286130572450	0.2119252529629370	0.0325643149145255
	θ = 0.6	latitude	0.0586757474849703	0.0214325625321522	0.0610344248147761	0.0843820418322524
		longitude	0.0954269104701193	0.1101286130037360	0.2119252529629380	0.0325241490306780
	θ = 0.8	latitude	0.0586757474849703	0.0214325625321520	0.0610344248147744	0.0844672842911134
		longitude	0.0954269104701192	0.1101286130037350	0.2119252529629370	0.0325340151987219

Table 5-8 RMSD values for Method 1 when replaced by previous value

Table 5.8 shows the RMSD values obtained by implementing method 1. As observed from Table 5.8 and 5.2 the RMSD values for user 73889 at θ = 0.4,0.6, and 0.8 are same, because more than 80 % of the users given data has same location registered. As observed from the Table 5.8 and 5.2 for the users 197216,358442 and 544126 difference in the RMSD values is small.

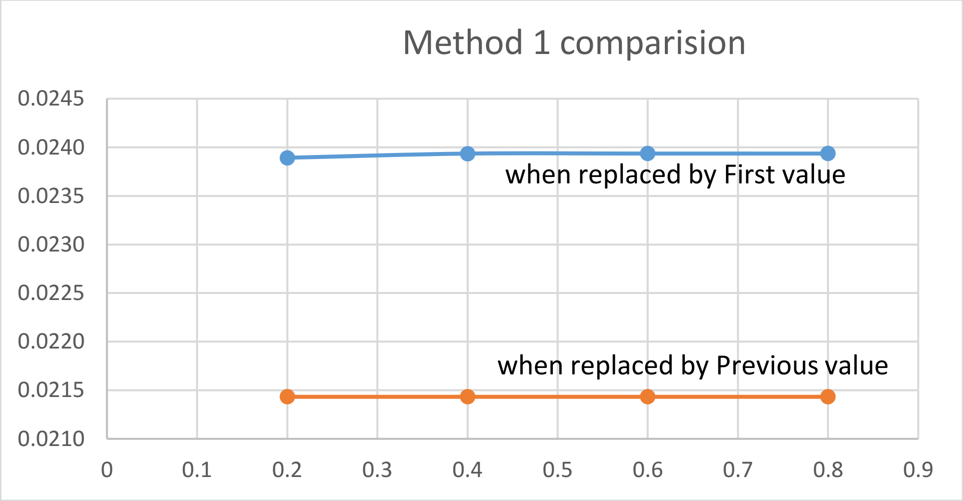


Figure 5-2 Graph between Method 1 when replaced by missing value and previous value

For example, Figure 4 shows us the comparison between latitudes of 197216 user for Method 1, when replaced by missing value and previous value. As we can observe that RMSD values for missing values when replaced by previous value produces minimal error, compared to the missing values when replaced by first value.

The computational time for the users to this Method 1 where missing values are replaced by previous value is shown below in Table 5.9.

User Id		73889	197216	358442	544126
Time(sec)	θ = 0.2	33.726	33.71	31.645	33.107
	θ = 0.4	34.92	35.933	33.68	32.583
	θ = 0.6	35.19	36.103	34.58	34.894
	θ = 0.8	35.841	36.268	34.876	34.851

Table 5-9 Computational Time for Method 1 when replaced by previous value

As observed from Table 5.9 and 5.3, Method 1 when missing values are replaced by first value in the dataset have less computational time. The average computational time to perform this method for various values of θ is 34.494 sec.

5.3.2 Method 2:

As explained in chapter 4, in this method the last value in the given dataset is removed. The missing values in the dataset are replaced by previous value in the dataset. The locations of the user for next whole week are predicted by using equation (3). The RMSD values are calculated by assuming the data of the user is same for the next week as the given week data.

User Id			73889	197216	358442	544126
RMSD values	θ = 0.2	latitude	0.0586757486669315	0.0214326200455634	0.0610344248147777	0.0851699734218633
		longitude	0.0954269099999659	0.1101287535001630	0.2119252529629390	0.0315043911632978
	θ = 0.4	latitude	0.0586757474849703	0.0214325625684331	0.0610344248147744	0.0882182335253034
		longitude	0.0954269104701192	0.1101286130929180	0.2119252529629370	0.0309289508328182
	θ = 0.6	latitude	0.0586757474849703	0.0214325625321526	0.0610344248147761	0.0895159736819741
		longitude	0.0954269104701193	0.1101286130037370	0.2119252529629380	0.0307612897377600
	θ = 0.8	latitude	0.0586757474849703	0.0214325625321520	0.0610344248147744	0.0897401058471485
		longitude	0.0954269104701192	0.1101286130037350	0.2119252529629370	0.0307356368768773

Table 5-10 RMSD values for Method 2 when replaced by previous value

Table 5.10 shows the RMSD values by implementing Method 2 when missing values are replaced by previous value. As observed from Table 5.10,5.8.5.4 and 5.2 for user 73889 at θ = 0.4,0.6, and 0.8 have same RMSD values, because more than 70 % of the given locations in the dataset are same. In Table 5.10 and 5.8 it can be observed that user 358442 has same RMSD values, because last value and the second last value are the same in the given dataset for the user 358442. So, when the last value is removed and missing value replaced with the previous value it results the same.

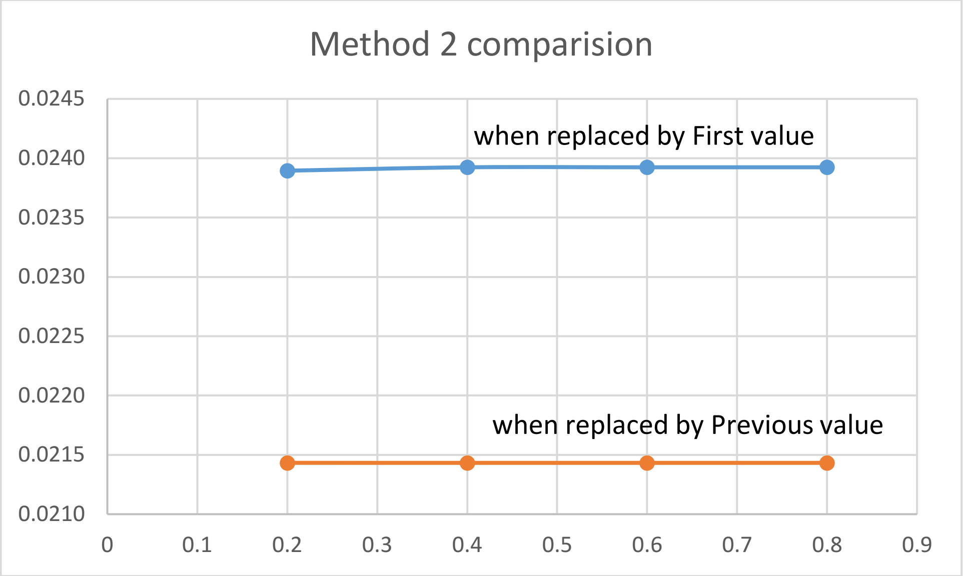


Figure 5-3 Graph between Method 1 when replaced by missing value and previous value

For example, Figure 5 shows us the comparison between latitudes of 197216 user for Method 2, when replaced by missing value and previous value. As we can observe that RMSD values for missing values when replaced by previous value produces minimal error, compared to the missing values when replaced by first value.

The computational time observed by implementing Method 2 are given in the below Table 5.11.

User Id		73889	197216	358442	544126
Time(sec)	θ = 0.2	34.354	32.343	27.135	32.213
	θ = 0.4	34.536	35.315	30.783	34.187
	θ = 0.6	34.717	35.704	32.855	35.038
	θ = 0.8	35.451	35.648	32.608	35.931

Table 5-11 Computational Time for Method 2 when replaced by previous value

As observed from Table 5.11 and 5.5, Method 2 when missing values are replaced by first value in the dataset have less computational time. The average computational time to perform this method for various values of θ is 33.676 sec.

5.3.3 Method 3:

In this method as explained in chapter 4, most of the users in the dataset location changes between time *12:30:00* to *13:30:00.* So, the data between time interval for the week is considered and the data between the same time interval for next week is predicted. RMSD values obtained by this method are:

User Id			73889	197216	358442	544126
RMSD values	θ = 0.2	latitude	0.0453533935898685	0.0208873583346143	0.0491603009293702	0.0971034632288835
		longitude	0.0174025697314695	0.1098694438512440	0.1544092459881680	0.0285649148632129
	θ = 0.4	latitude	0.0452443867961946	0.0208522556706313	0.0520046078854778	0.1023834486017060
		longitude	0.0172339361496294	0.1101616566693930	0.1626797755401450	0.0287274077700723
	θ = 0.6	latitude	0.0450809210817352	0.0209245578079375	0.0520735088471078	0.1025129053555860
		longitude	0.0171184681744491	0.1101462848887390	0.1628822022475110	0.0287324056809622
	θ = 0.8	latitude	0.0450189420805115	0.0210191054049179	0.0520738608515046	0.1025135796400870
		longitude	0.0170769797987330	0.1100973232509950	0.1628832439076510	0.0287324282997757

Table 5-12 RMSD values for Method 3 when replaced by previous value

As observed in Table 5.12 and 5.6 the RMSD values of Method 3 when missing values replaced by first value in the dataset has minimal error. This also proves that modified CWT method can be applied even when location of user changes frequently.

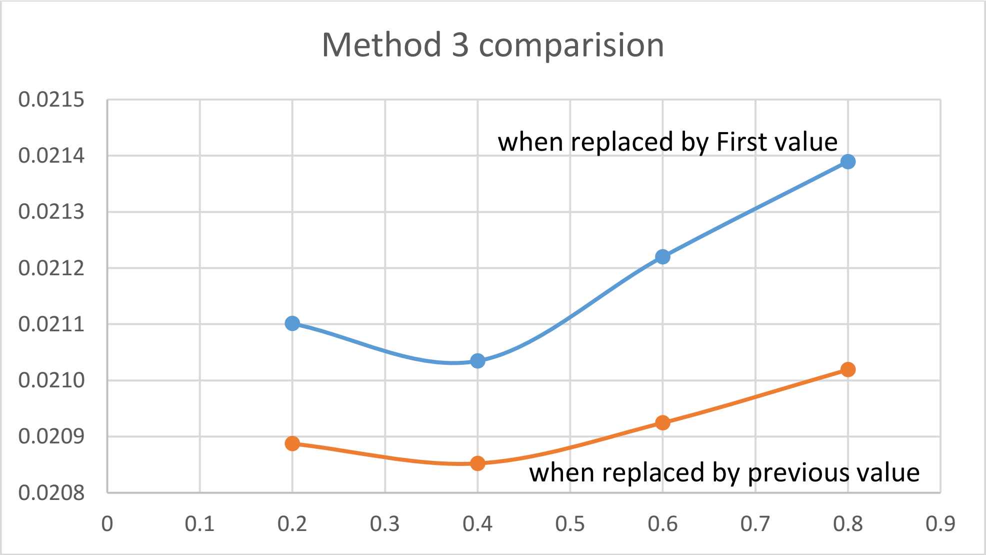
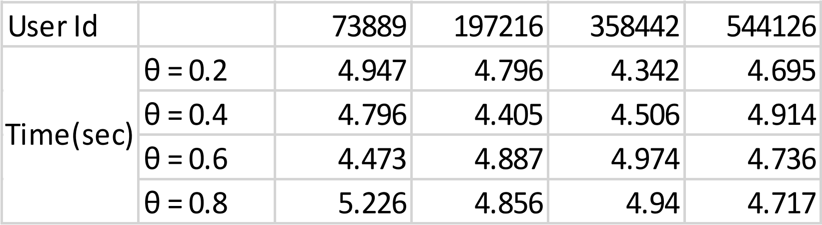


Figure 5-4 Graph between Method 3 when replaced by missing value and previous value

For example, Figure 6 shows us the comparison between latitudes of 197216 user at θ = 0.2 ,0.4, 0.6 and 0.8 for Method 3, when replaced by missing value and previous value. As we can observe that RMSD values for missing values when replaced by previous value produces minimal error, compared to the missing values when replaced by first value in the dataset.

The computational time observed by implementing Method 3 are given in the below Table 5.13.



User Id		73889	197216	358442	544126
Time(sec)	θ = 0.2	4.947	4.796	4.342	4.695
	θ = 0.4	4.796	4.405	4.506	4.914
	θ = 0.6	4.473	4.887	4.974	4.736
	θ = 0.8	5.226	4.856	4.94	4.717

Table 5-13 Computational time for Method 3 when replaced by previous value

As observed from Table 5.13 and 5.7, Method 3 when missing values replaced by first value has less computational time. The average computational time to perform this method for various values of θ is 4.763 sec.

6 CONCLUSION AND FUTURE WORK

From result and analysis presented in Chapter 5, it can be observed that CWT method have been modified and applied for predict location of user at next time step as shown in Section 5.1. It can also be observed that predicted location of the user is close to the observed location of the user with a small difference in values. This method can also be extended to periodical temporal aspect as explained in chapter 4 and as observed from section 5.2 and 5.3. The modified CWT method when applied for the given Telenor dataset has minimal error and less computational time when missing values are replaced by first value as observed from Section 5.2 and 5.3.

Finally, from the results and analysis it was evident that, CWT method have been modified and applied for forecasting trajectory data. It can also be concluded that modified CWT method for the given Telenor dataset will produce better results when missing values are replaced by first value in the dataset.

6.1 Research questions and answers:

RQ1: How can CWT method be modified to predict the user's location at next instance of time?

Answer: The CWT has been modified by normalizing the equation as explained in chapter 4 equation 3. This modification has been done to predict the location of user at next instance of time rather than using the method to predict probability of whether an author is going to publish in a conference as used in previous line of work. The results and analysis of this modification in CWT method are shown in section 5.1

RQ2: Given this extension, how can we use the CWT method to predict user's location at multiple instances of time?

Answer: Modified CWT method is only used to predict the position at next time stamp. To predict position of the user for a periodic temporal aspect, the modified CWT method is extended in a recursive way, wherein the predicted value using CWT is added to already existing data to predict the data in next instance of time. For example, the given data consists of T time steps, the location of the user is predicted for the instance of T+1. This predicted value is added to the existing trajectory data and used in prediction of location at T+2 instance and so on.

RQ3: Which method is suitable for the given data among the modified variants of the CWT method for forecasting trajectory data?

Answer: To the given Telenor dataset there are a lot of missing values in the dataset. There are two methods to replace the missing values.

1. Replacing missing values with first value in dataset.
2. Replacing missing values with second value in dataset.

These two methods are performed in three different methods to see how the prediction varies. The three methods are:

1. Method 1: is to predict the location of the user for next whole week, i.e. the next 2016 location with given data for a period of one week.
2. Method 2: is to predict the location of the user for next whole week by removing the last value in given dataset.
3. Method 3: is to use the data between time interval 12:30:00 to 13:30:00 every day of the week, as most of user's location varies at this time interval. To use this data and predict the location of the user at this time interval for next whole week

The above mentioned 2 methods are performed in these three methods. The results and analysis in chapter 5 determines that when missing values are replaced by the first value in the dataset the prediction has minimal error and has less computational time.

6.2 Future Work

This study is to provide a method for forecasting trajectory data with the given Telenor dataset. As a future work to this study,

1. Modified CWT method can be tested by considering at least with 2 weeks of Telenor data and the prediction can be observed.
2. Telecom operators analyze the data and independent of the conclusions reached data cannot be disclosed due to proprietary reasons. The predicted locations by modified CWT method can be used to produce a synthetic database, i.e. a public database on which different researches can run their algorithm and compare to one another in a fair way.

REFERENCES

[1] D. M. Dunlavy, T. G. Kolda, and E. Acar, “Temporal Link Prediction Using Matrix and Tensor Factorizations,” *ACM Trans Knowl Discov Data*, vol. 5, no. 2, p. 10:1–10:27, Feb. 2011.

[2] U. Sharan and J. Neville, “Temporal-Relational Classifiers for Prediction in Evolving Domains,” in *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 540–549.

[3] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki, “Link prediction using supervised learning,” in *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*, 2006.

[4] Y. Liu and Z. Kou, “Predicting Who Rated What in Large-scale Datasets,” *SIGKDD Explor Newsl*, vol. 9, no. 2, pp. 62–65, Dec. 2007.

[5] A. Clauset, C. Moore, and M. E. J. Newman, “Hierarchical structure and the prediction of missing links in networks,” *Nature*, vol. 453, no. 7191, pp. 98–101, May 2008.

[6] Z. Huang, X. Li, and H. Chen, “Link Prediction Approach to Collaborative Filtering,” in *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries*, New York, NY, USA, 2005, pp. 141–142.

[7] C. Niyizamwiyitira, L. Lundberg, L. Skold, and Sidorova, “ANALYTIC QUERIES ON TELENOR MOBILTY DATA,” *ResearchGate*, Apr. 2016.

[8] J. Sidorova, L. Lundberg, and L. Sköld, “Optimizing the Utilization in Cellular Networks using Telenor Mobility Data and HPI Future SoC Lab Hardware Resources,” *ResearchGate*, Oct. 2016.

[9] B. C. Csáji *et al.*, “Exploring the Mobility of Mobile Phone Users,” *ResearchGate*, vol. 392, no. 6, Nov. 2012.

[10] M. Saravanan, S. V. Pravinth, and P. Holla, “Route detection and mobility based clustering,” in *2011 IEEE 5th International Conference on Internet Multimedia Systems Architecture and Application*, 2011, pp. 1–7.

[11] S. Gatmir-Motahari, H. Zang, and P. Reuther, “Time-Clustering-Based Place Prediction for Wireless Subscribers,” *IEEEACM Trans. Netw.*, vol. 21, no. 5, pp. 1436–1446, Oct. 2013.

[12] A. Hess, I. Marsh, and D. Gillblad, “Exploring communication and mobility behavior of 3G network users and its temporal consistency,” in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 5916–5921.

[13] H. Shi *et al.*, “Segmentation of Mobile User Groups Based on Traffic Usage and Mobility Patterns,” in *2014 IEEE 17th International Conference on Computational Science and Engineering*, 2014, pp. 224–230.

[14] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, “Understanding traffic dynamics in cellular data networks,” in *2011 Proceedings IEEE INFOCOM*, 2011, pp. 882–890.

[15] K. Laasonen, “Clustering and Prediction of Mobile User Routes from Cellular Data,” in *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Berlin, Heidelberg, 2005, pp. 569–576.

[16] S. U and R.S.Bhuvaneswaran, “Mobility Prediction of Mobile Users in Mobile Environment Using Knowledge Grid.” International Journal of Computer Science and Network Security, 2009.

[17] J. Taheri and A. Y. Zomaya, “Clustering Techniques for Dynamic Location Management in Mobile Computing,” *J Parallel Distrib Comput*, vol. 67, no. 4, pp. 430–447, Apr. 2007.