www.dergipark.gov.tr
ISSN:2148-3736
El-Cezerî Fen ve Mühendislik Dergisi
Cilt: 4, No: 3, 2017 (411-428)

El-Cezerî Journal of Science and Engineering
Vol: 4, No: 3, 2017 (411-428)

ECJSE

Research Paper / Makale

The Use of Spatio-Temporal Data Mining for Detection and Interpretation of Trajectory Outliers in Health Care Services

Abdulsamet HASILOGLU[1,*], Seyma Yucel ALTAY[1], Umit ERTAS[2]

[1]Department of Computer Engineering, Faculty of Engineering, Atatürk University, Erzurum, 25240, Turkey
[2]Department of Oral and Maxillofacial Surgery, School of Dentist, Atatürk University, Erzurum, 25240, Turkey
Email: asamet@atauni.edu.tr[1], seyma.yucel@atauni.edu.tr[1], uertas@atauni.edu.tr

Received/Geliş: 19.05.2017 **Revised/Düzeltme:** - **Accepted/Kabul:** 29.05.2017

Abstract: In the last decade, useful information extraction from moving objects has become widespread in the spatial-temporal data mining field with the increasing use of devices such as RFID and GPS. For this purpose, the outlier detection method, which is a subfield of data mining, was applied to the trajectory of patients and diseases in the dental health service. In this article, TRAOD and TOD-SS algorithms combining distance and density-based methods were preferred. These algorithms do not handle the moving object trajectory as a whole unlike other outlier detection techniques. They investigate whether each piece exhibits different behavior according to its neighbors by separating trajectories into pieces. So, they detect outlying trajectory pieces that other algorithms cannot locate. Algorithms preferred in this study were used in a COMB-O model we developed and their performances were compared. In addition, according to the region and clinic, the classification of patients was made. Also, clustering, which is another branch of spatial-temporal data mining, was performed for trajectory. When the COMB-O model was executed, results showed sub-trajectories that deviated from the trajectory data were successfully detected with the help of the trajectory outlier detection algorithms. Inconsistent trajectories perceived provided significant data. In addition to this, successful classification was performed by making use of non-linear classification features of DVM. Moreover, stops and moves in the Faculty of Dentistry were detected by using CB-SMoT and DB-SMoT which are clustering algorithms.

Keywords: Spatio-temporal data mining, moving objects, trajectory data, trajectory outlier detection, support vector machine, stop and moves of trajectory.

Sağlık Hizmetlerinde Aykırı Dataların Kestirimi İçin Mekansal Zamansal Veri Madenciliğinin Kullanımı

Özet: Son on yılda, hareketli nesnelerden yararlı bilgi çıkarma, mekansal-zamansal veri madenciliği alanında RFID ve GPS gibi cihazların kullanımının yaygınlaşması ile yaygınlaşmıştır. Bu amaçla, veri madenciliğinin bir alt alanı olan belirsizlik tespit yöntemi, diş hekimliği servisindeki hastaların ve hastalıkların gidişatına uygulanmıştır. Bu makalede, mesafe ve yoğunluk tabanlı yöntemleri birleştiren TRAOD ve TOD-SS algoritmaları tercih edilmiştir. Bu algoritmalar, diğer aykırı algılama tekniklerinden farklı olarak, hareketli nesne yörüngesini bir bütün olarak ele almaz. Her bir parçanın yörüngeleri parçalara ayırarak komşularına göre farklı davranış sergileyip sergilemediklerini araştırıyorlar. Dolayısıyla, diğer algoritmaların bulamayan yörünge parçalarını algılarlar. Bu çalışmada tercih edilen algoritmalar geliştirdiğimiz COMB-O modelinde kullanılmış ve performansları karşılaştırılmıştır. Buna ek olarak, bölgeye ve klinikte, hastaların sınıflandırması yapılmıştır. Ayrıca, mekansal-zamansal veri madenciliğinin bir başka dalı olan kümeleme, yörünge için gerçekleştirildi. COMB-O modeli yürütüldüğünde, sonuçlar yörünge verilerinden sapmış olan alt yörüngeleri yörünge aykırı değer algılama algoritmaları yardımıyla başarıyla tespit edildiğini gösterdi. Algılanan tutarsız yörüngeler önemli veriler sağlamıştır. Buna ek olarak, DVM'nin doğrusal olmayan sınıflandırma

How to cite this article
Haşıloğlu, A., Altay, S.Y., Ertaş, Ü., "*The Use of Spatio-Temporal Data Mining for Detection and Interpretation of Trajectory Outliers in Health Care Services*" El-Cezerî Journal of Science and Engineering, 2017, 4(3); 411-428.

Bu makaleye atıf yapmak için
Haşıloğlu, A., Altay, S.Y., Ertaş, Ü., *"Sağlık hizmetlerinde aykırı dataların kestirimi için mekansal zamansal veri madenciliğinin kullanımı"* El-Cezerî Fen ve Mühendislik Dergisi 2017, 4(3); 411-428.

özelliklerinden faydalanarak başarılı bir sınıflandırma gerçekleştirildi. Ayrıca, kümeleme algoritmaları olan CB-SMoT ve DB-SMoT kullanılarak Diş Hekimliği Fakültesindeki durmalar ve hareketler tespit edildi.

Anahtar kelimeler: Zamansal-Mekansal veri madenciliği, hareketli cisimler, yörünge verileri, yörünge sapması algılama, destek vektör makinesi, yörünge hareketleri ve durması.

1. Introduction

Spatial-temporal database systems manage data whose geometry varies over time. A spatial-temporal object is an object which has at least one space and time feature. The spatial features of objects are location and geometry of the object and the time property can be defined as the time frame or the time interval when the object is valid [1]. There is also an intense need for knowledge because spatial-temporal databases increase in terms of both number and size rapidly. Spatial-temporal data mining which investigates this knowledge represents machine learning, statistics, geographical visualization and includes information theory by combining various fields [2].

Useful information extraction from moving objects has become increasingly widespread in spatial-temporal data mining over the last decade. The trajectory motion data are obtained in large quantities to determine the position of an object in a certain period of time with location sensitive devices. Trajectories of moving objects are captured with devices such as GPS (Global Positioning System) and RFID (Radio Frequency Identification) and other technologies and spatial-temporal sequences of points are generated [3]. Consequently, trajectory data in large quantities are represented as the raw sample dots. Trajectory data play an increasingly important role in lots of applications such as tourism, traffic management, transportation, fisheries, urban planning, entertainment activities, internet networks and animal migration [4], [5], [6].

Even if most of the techniques of spatial-temporal data mining are oriented to extracting common patterns, finding inconsistent data may be more useful. Outlier detection is expressed as the problem of extracting data patterns which are different to the rest of the data depending on certain criteria. Such a pattern usually contains useful information about the abnormal behavior of the system identified by the data. These abnormal patterns are expressed mostly with the terms such as inconsistent data, noise, anomaly, or error in different application areas. The perception of inconsistent data is used in industrial applications, in the detection of fraud, in military surveillance of enemy activity, in the detection of violations in cyber security, in the abnormal flow problems in pipelines, in the insurance services and in many more areas [7], [8].

There are many methods to detect inconsistent data in data mining. These can be classified as clustering-based, depth-based, distance-based and density-based methods.

In clustering-based methods, although the main objective of the clustering methods such as CLARANS [9], DBSC [10], BIRCH [11] and CURE [12] is to detect clusters, they also capture the outliers. Cateni et al. [13] identified potential outlier data as data that does not settle into any clusters according to a clustering algorithm's perspective. Ben-Gal [14] considers small sized clusters (containing significantly fewer points than other clusters) as clustered outlier data. Namely, if a cluster is different significantly from other clusters, objects in this cluster may be contrary data. Depth-based approaches are based on digital geometry. Outliers are objects with smaller depth. Namely, inconsistent data are usually at the limit of the data surface and normal data is within it. However, in practice this technique is not effective for large data sets [8], [13], [15].

In the distance-based approaches, if at least p part of the objects in T is in a more distant position than distance D from O, object O in a data set T is a distance-based inconsistent object (DB-outlier). The distance-based outlier data concept is well defined for many dimensional data sets. p parameter

data is the minimum number of parts in which the objects in a data space should be out of the *D* neighborhood of the outlier data [16], [17]. Ramasmawy et al. [18] developed a method that does not need to define the *D* distance parameter to detect outlier data. They use *k.* to calculate the outlier data by depending on the distance of the nearest neighborhood. If distance-based methods are used for data sets containing both dense and sparse regions, it may cause certain problems.

Density-based outlier detection techniques estimate the density of the neighborhood of each sample. While a sample in a low-density neighborhood is defined as outlier, the samples in the high density region are defined as normal [19], [20]. The generalized definition of density-based methods is as follows: Local outlier data factor (Local outlier Factor - LOF) is calculated for each object in the data set. This measures how a deviant object deviates. High local outlier data factor represents a low-density neighborhood and accordingly, the potential to provide high inconsistent data [15], [19].

Birant and Kut [21] developed a new technique based on DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm in their study in order to detect ST-Outlier (spatial-temporal outlier data) because clustering is the basic method to detect the outliers. The algorithm suggested here firstly detects spatial outliers and then temporal outliers. However, the fact that firstly the existence of temporal outlier data and then the spatial outlier data is found also suggests the same result. In this study, there is a three-step approach mentioned: clustering, checking and controlling the temporal and spatial neighbors. After applying the clustering algorithm, the objects which do not settle into any clusters are the potential outliers. First of all, whether these objects are really S-outliers (spatial outliers) or not is checked. After spatial outliers are found, whether they are temporal outliers or not is investigated. What exactly is meant by neighborhood in time is time frames following each other such as the same days in consecutive years or consecutive days of the same year. S-Outliers in the same area are compared at different times to support the time dimension. If an S-Outlier's characteristic value differs significantly from that of the temporal characteristics, it is an ST-Outlier neighbor.

Wu et al. [22] also suggested the Outstretch algorithm to detect inconsistent spatial-temporal data. This method uses the spatial-temporal scan statistics; the goal here is to determine the outliers in the spatial-temporal region. Here, for each time period, all outlier sequences over time are found by using spatial top-k outliers which are obtained with the help of Exact-Grid Top-k or Approx-Grid Top-k algorithms and these are stored in a tree structure with the algorithm Outstretch.

To detect and to interpret the trajectory outliers in the health sector, a spatial-temporal data mining model named HONEYCOMB-Outlier (COMB-O) was used in this study for the first time. The COMB-O model program was developed for the detection and interpretation of inconsistent trajectories in health centers. Additionally, an attempt was made to determine clusters showing similar movement patterns temporally and spatially by using CB-SMoT and DB-SMoT algorithms. Moreover, some diseases are unique to specific regions. So, diseases can be classified according to region and clinic.

This article is organized as follows: in section 2, basic concepts about trajectory are mentioned. In section 3, the material method is presented. In section 4, the details of model COMB-O that we developed are mentioned. In section 5, experimental results are presented and the study concludes in section 6.

2. Trajectory

We can talk about moving objects in every application in which there are stops and movement. For instance; if a person is moving from one place to another place, this involves moving object data.

Apart from this, migration of animals, satellite (spacecraft) trajectories, vehicle traffic (trajectory of taxis), aircraft, ship, rocket, etc., may be considered as moving object data. Movement of patients in a health center (clinics, laboratories, etc.) is movement for those patients and spread of disease is also a moving object. For example, where and when it is seen and the onset and spread of cancer create moving object data.

The acquisition of useful information about moving object behavior can only be inferred from the trajectories or patterns (if geographic information with trajectory position is taken into account) in most application areas such as transportation management, animal migration and tourism. Trajectories can be accepted as the way that objects move according to space and time.

Each point in the trajectory symbolizes its position in space at a particular time [23]. From the users' point of view, the concept of trajectory lies in positions of the developing journey of the object in space during a certain time interval. Therefore, trajectory, by definition, is a spatial-temporal concept.

Trajectory is described as a user-defined record containing the position of movements of an object in space at specific time intervals to achieve a goal [24].

trajectory: $[t_{begin}, t_{end}]$ -> space

A trajectory sample consists of a space and time points list, $\{p_0 = (x_0, y_0,t_0), p_1=(x_1,y_1,t_1),\ldots, p_N=(x_N,y_N,t_N)\}$, here $x_i, y_i \in \square, t_i \in \square^+$ object location for i=0,1,…,N; t_0 is start time of travel and t_N is termination time of travel in $t_0 < t_1 < t_2 < \cdots < t_N$ [23].

The trajectory set is represented by $T = \{TR_1, TR_2, \ldots, TR_{number\_of\_trajectories}\}$ and the outlier set is represented by $O = \{O_1, O_2, \ldots, O_{number\_of\_outliers}\}$. A trajectory is multi-dimensional series of points and is described as $TR_i = p_1p_2p_3 \ldots p_j \ldots p_{len_i}$ $(1 \leq i \leq number\_of\_trajectries)$. P_j $(1 \leq j \leq len_i)$ is a point in the trajectory. Length of a trajectory len_i can be different from other ones. $p_{c1}, p_{c2}, \ldots, p_{ck}$ $(1 \leq c1 < c2 < \cdots < ck \leq len_i)$ is a sub-trajectory of TR_i [25].

3. Materials-Methods

3.1 Traod Algorithm

To operate strong outlier data detection algorithms, trajectories are required. There could be cases in moving object data where certain components of trajectory exhibit abnormal behavior but not whole trajectory is not abnormal. In this case, the methods cannot regard this trajectory as outlier data.

However, TRAOD and TOD-SS algorithms find these components with abnormal behavior and at the same time, they research whether these components have sufficient abundance to be accepted as a trajectory outlier.

The TRAOD algorithm firstly divides the trajectory data into pieces to identify whether there are parts that exhibit abnormal behavior. After that, it determines sub trajectories that are inconsistent among these pieces. Regarding detection of the sub trajectories, it is preferred because it is an effective method whatever the size of the distance-based method data. However, some problems may result if the data set contains both dense and sparse regions and to eliminate this, the intensity-based and distance-based approaches are used together.

3.1.1 Close Trajectory

An outlier part of a trajectory is determined depending on the number of trajectories that are close to it. These close trajectories have a certain distance from neighbor trajectories [25], [26].

If a YR_i trajectory ensures $\sum_{L_i \in CP(YR_i, L_j, D)} len(L_i) \geq len(L_j)$ condition, it is close to $L_j \in P(YR_j)(YR_i \neq YR_j)$. Here *D* is determined by the user [25].

3.1.2 Outlying Sub-Trajectory

An L_i part deviates if at least p part of the trajectory in YR is not close to L_i. When formulated, if $L_i \in P(YR_i)$ part Equation 1 is true, it deviates. Here, Y represents the set of trajectories and p is determined by the user [25].

$$|CTR(L_i, D)| \leq \lceil (1-p)|Y| \rceil \tag{1}$$

3.1.2 Trajectory Outlier

If a trajectory contains deviant parts that cannot be ignored, it is an inconsistent trajectory. In other words, if Equation 2 is true, the data is inconsistent. Here, F is a parameter determined by us.

$$Ofrac(YR_i) = \frac{\sum_{L_i \in OP(YR_i, D, p)} len(L_i)}{\sum_{M_i \in P(YR_i)} len(M_i)} \geq F \tag{2}$$

While the number of nearby trajectories is high in dense areas, it is less in sparse areas. As a result, inconsistent sub trajectories may not be detected in dense regions and in the sparse regions, most of the trajectories may be considered inconsistent. In Table 1, the parameters of the TRAOD algorithm are presented.

Table 1- Traod algorithm parameters

Parameter	Explanation
D	Distance
P	To be an outlying L_i partition, minimum partition numbers whose whole trajectories are not close to this partition
F	Required minimum outlying partition numbers to be an outlier trajectory
t-partition	Trajectory partition

The algorithm's time complexity is $O(N_T^2)$. Here, N_T is the total number of t-partitions.

3.1.3 Tod-ss Algorithm

Some trajectory mining algorithms interpret trajectory as a series of static data. They do not handle valuable properties of the trajectory. For example, the initial time of the trajectory may affect the shape, the location and its other properties. In this study, most properties of trajectory are considered. The importance of the properties is adjusted by weighting. In addition, the degree of match is measured by calculating structural features. Structural features not only reflect both internal and external features, but also strengthen the impact of the analysis. The presence of outliers is revealed by comparison of the structural similarity of the segment pairs. Trajectory structure is a set of internal and external properties of a trajectory and all these properties represent the whole trajectory from a wide angle. Trajectory structure contains four features: direction, speed,

angle, and location. The weight of these features is expressed by the vector $W = \{W_D, W_S, W_A, W_L\}$ [26].

SSIM (Structural Similarity) measure is expressed by SDIST (Structural Distance). SDIST consists of 4 comparisons; when L_i and L_j are two segments, $DirDist(L_i, L_j)$ is direction comparison, $SpeedDist(L_i, L_j)$ is speed comparison, $AngleDist(L_i, L_j)$ is angle comparison, and $LocDist(L_i, L_j)$ is location comparison. *N is* distance normalization function and SSIM and SDIST are calculated as:

$$SDIST(L_i, L_j) = DirDist \times W_D + SpeedDist \times W_S + AngleDist \times W_A + LocDist \times W_L \quad (3)$$

$$SSIM(L_i, L_j) = 1 - N(SDIST(L_i, L_j)) \quad (4)$$

Structural similarity of the segments is firstly calculated and the whole trajectory can be matched with structural comparison. The size of the similarity depends on how big SSIM is [26].

3.1.4 ε −Neighborhood

If L_i segment providing $SSIM(L_i, L_j) \geq \boldsymbol{\varepsilon}$ for a SSIM threshold value and L_i exists, it is called L_j in the $\boldsymbol{\varepsilon}$ -neighborhood of L_i [27].

3.1.5 Segment Outlier

For the neighborhood threshold value set by the user σ, if the number of the neighborhood of L_i segment is smaller than σ, it is called an inconsistent segment [26].

3.1.6 Trajectory Outlier

If a trajectory abides by these two conditions, it is inconsistent: 1) if the number of the inconsistent segments in YR_i is greater than the ξ threshold value and 2) if total similarity to MSSIM, namely to other trajectories, is less than the threshold value F [26].

In Table 2, the parameters of TOD-SS algorithm are presented.

Table 2- Tod-ss algorithm parameters

Parameter	Explanation
ε	SSIM threshold
W	Corner threshold
Σ	Neighborhood threshold
F	Structural similarity threshold
Ξ	Ratio threshold

Time complexity of algorithm is $O(N \operatorname{Log} N)$. Here, n is the number of trajectory segments.

3.2 Classification

As the compatibility problem of support vector machines for a model with so many parameters is less than other algorithms and it provides high accuracy and can operate with a large number of independent variables, it is a useful technique for data classification. This method is frequently used in many fields such as handwriting recognition, voice recognition, and prediction of breast cancer, bioinformatics and spatial data analysis [27], [28], [29].

A classification task usually involves the separation of the data into training and test sets. Each sample in the training set contains a target value (i.e. class label) and many properties (i.e. attributes and observed values). The target of DVM (based on the training data) is to produce a model predicting the target values of the test data when only test data properties are given [27]. The simplest model of support vector machines is the model applied to the input that is separated linearly. In non-linear classification, input data are reduced to the property space by being passed through functions named Kernel functions and the classification is made here. The most common kernel functions are given in equations (5), (6) and (7) [27], [28], [29].

$$K(x, y) = (x.y)^d \quad (5)$$

$$K(x, y)=\exp\left(-\|x - y\|^2/(2\sigma^2)\right) \quad (6)$$

$$K(x, y) = \tanh(\gamma(x.y) + \phi) \quad (7)$$

d value in equation (5) is the degree of polynomial, σ value in equation (6) is the radial based function parameter and ϕ value in equation (7) is the sigmoid kernel parameter [27], [28].

3.3 **Clustering**

Spaccapietra discussed a new model handling trajectories called stops and movements [29]. This model has an important place in applications by adding semantic information to process the trajectories. Generally, while the way that the object moves between pauses is defined as movement, pauses are the most important parts of a trajectory in terms of the application. SMoT (Stops and Moves of Trajectories) algorithm is based on the idea of finding the pauses of the candidate points [30]. SMoT targets finding the relevant places where the trajectory points intersect at a certain time to determine the stops as a criterion [31]. The algorithm constantly searches for lower trajectories intersecting with the same candidate point during a minimum time for an application. SMoT algorithm works by testing whether every point in the application is a loop about which the number of trajectories intersects with the candidate point cluster. If there is a constant point of intersection for the same candidate point, the intersection point is recalculated and this value is compared with the current minimum period of time to be candidate. If this intersection continues for at least the minimum period of time, this part of the trajectory is assumed to be a stop. This process is repeated for each point and a list of stops is obtained at the end of this process.

CB-SMoT (Clustering-based Stops and Moves of Trajectories - Clustering-based trajectory stop and motion) is the only trajectory clustering based on the exchange of speed. It is based on low-speed data being important. It is based on the method of determining the places where there are low speed values during an orbital motion. Trajectory data, speed change and the smallest time (minTime) parameters are taken as input and the areas where low speeds exist are marked with time variation. As a result, stop and movement points are determined. This method is preferred for applications where speed is important (such as traffic management) [32].

DB-SMoT is an algorithm that finds direction-based stops and movements of the trajectory. Sets are determined by locating stop and movement points according to direction changes. Trajectory samples, minimal direction variation, the minimum time and maximum tolerance are determined as input. For every trajectory sample, finding the sets whose direction variation is higher than minimal direction variation is performed. Consequently, semantic trajectories are obtained. This method is preferred in applications where variation of direction is important [33].

Trajectory clustering is a significant process for grouping similar points where stops and motions are dense. Trajectory clustering has a very important role in the analysis of data and reveals the basic tendencies of moving objects. In this study, trajectory clusters that show similar movement patterns according to time and space were detected by using CB-SMoT and DB-SMoT algorithms.

4. COMB-O Model: Graphical User Interface

The COMB-O model package program for spatial-temporal data mining was developed by the authors for the detection and statistical interpretation of inconsistent trajectories in a health center. In this program, firstly in order to save the data obtained in the SQL Server database into the relevant tables, an interface named Patient Record was created. Here, Patient Record, Patient Tracking and Inspection tables were updated by relating them with a patient and by choosing appropriate items from the data such as the process previously loaded into the database, clinic, and ICD (International Classification of Diseases) codes.

After the data from the years 2011-2013 were transferred into the database, interfaces were created in order to detect inconsistent data and to complete classification and clustering. In the first step, data selection was made. Here, filtering was performed by choosing items requested from the data such as gender, age range, city, town, illness, disease groups, clinic and history. The required algorithm was selected to detect inconsistent data by presenting the patients obtained from this filtering to the user in a table and the parameters were determined. According to these parameters, after the algorithm was implemented, the results obtained were transferred to the user.

In our other interface, if inconsistent data of the patients derived as the result of filtering are requested, after inconsistent data are obtained, the trajectories are also shown on a map of the Faculty of Dentistry. Moreover, which diseases are seen in patients from which regions is also presented to the user on the map.

In the classification interface, classes are determined according to regions and clinics for the patients in the Dentistry Faculty. Some of the data were used for training and the rest of them were used for the test. Accordingly, the percentage of patients in each class was shown in text and graphically on the screen.

Trajectory clustering was verified depending on speed and direction change in the clustering interface. Accordingly, movements and stops were found and are presented to the user graphically. Figure 1 shows the flow diagram of e COMB-O model.

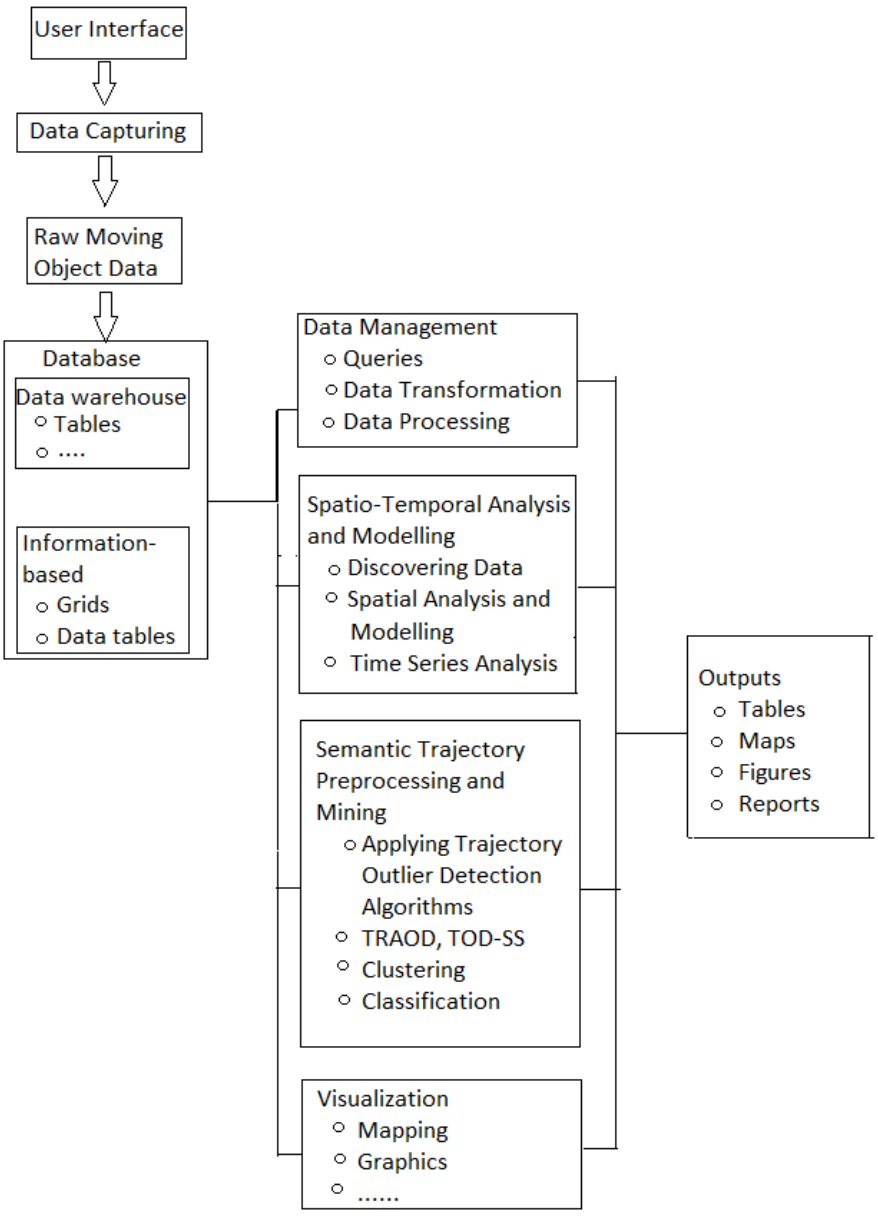

Figure 1- Flow diagram

5. Experimental Results

In this section, the COMB-O model was tested with necessary data obtained from the management of the Faculty of Dentistry in Erzurum Atatürk University which was chosen as the application field. In Figure 2, the data selection interface is shown. By selecting both space and time features in the data selection, spatial-temporal data was retrieved from the database. After data selection was made, the algorithm to be used for this data was indicated and the program was implemented. In the results section, the output of the applied algorithm was obtained.

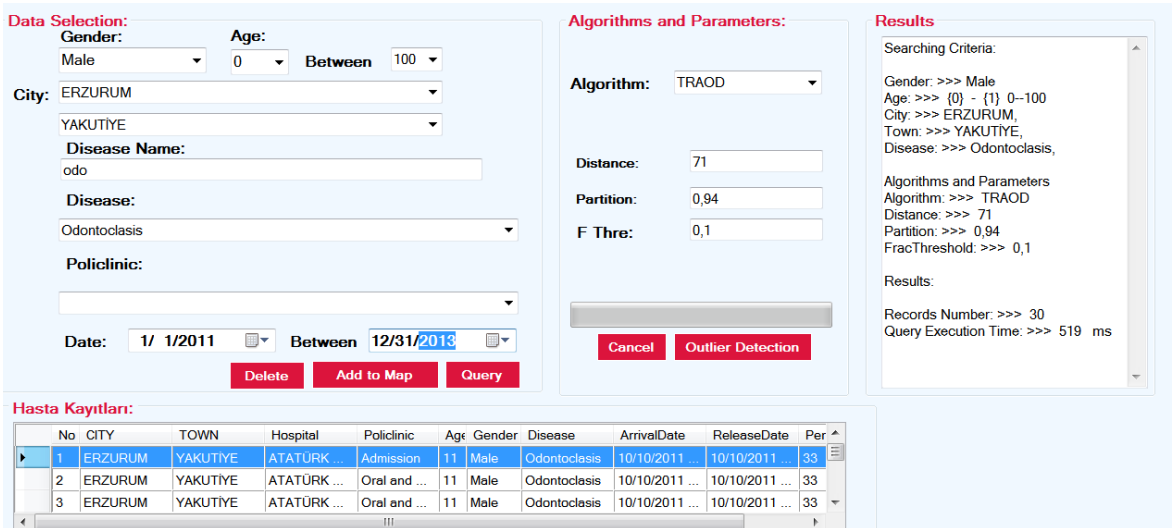

Figure 2- COMB-O Model

As seen in Figure 3, patient tracking could be performed with the COMB-O model. In which clinics the individuals with certain diseases received which treatments could be determined. In addition to this, significant results were obtained for patients following the same trajectory. Patients going to any clinic within a certain period of time with the COMB-O model who went to any other clinics or which clinics they went to were also determined. What's more, the treatments administered to a certain group, deemed important by a group of dentists, were also derived: e.g., which treatments are made for children under the age of 14. Again, another advantage of the COMB-O model is that when the patient has a case in court, documents are available in a very accessible way. Furthermore, which trajectories patients coming from the regions outside Erzurum followed could also be detected, or the diseases from these regions could be investigated. Regional intensity of disease (prevalence) and the number of disease cases (incidence) could be obtained.

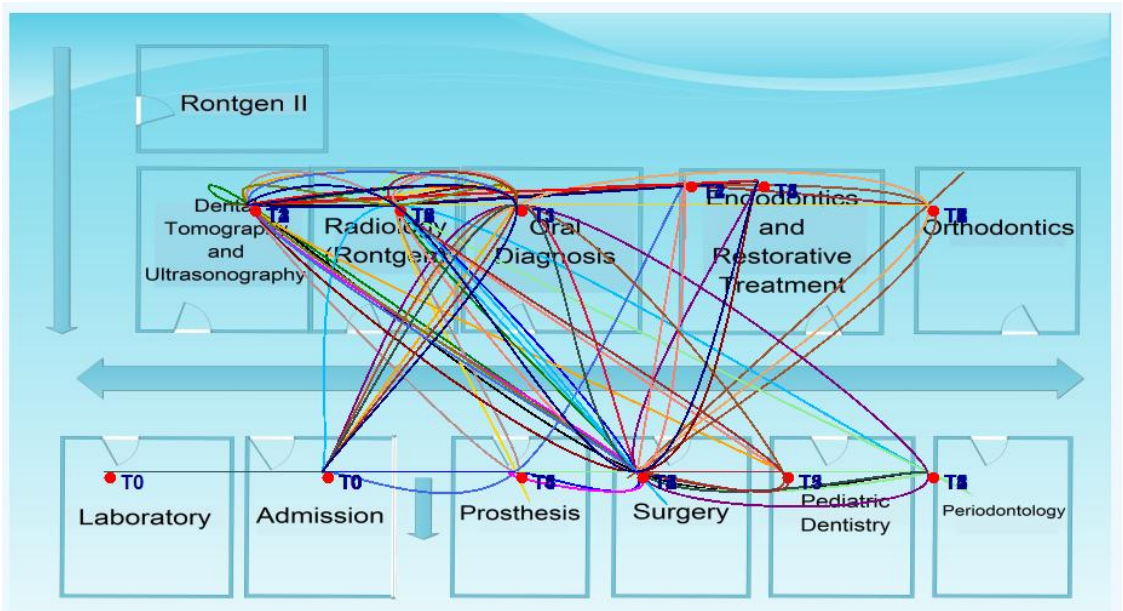

Figure 3- Transferring trajectory data to Faculty of Dentistry map

As seen from Figure 4, by obtaining trajectories of patients from the information in patient records, people who exhibited inconsistent actions were identified. In the inconsistent data determined in accordance with the opinions of members of the Faculty of Dentistry, the disease code might be erroneous. In addition, the idea that the patients applied as emergency cases might be detected as

inconsistent data was noted. Again, detection of inconsistent data allowed identification of patient leakage.

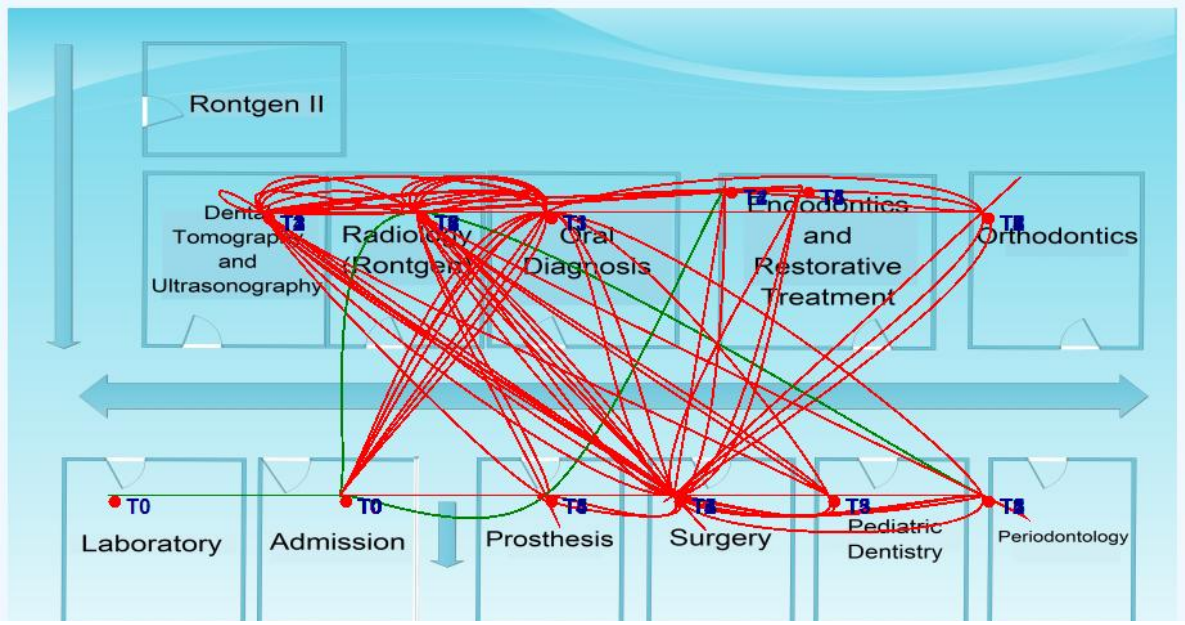

Figure 4- Trajectory outlier detection (green lines are outliers)

Figure 5 shows the regional map. This section illustrates the region a selected portion of patients came from and which diseases they have.

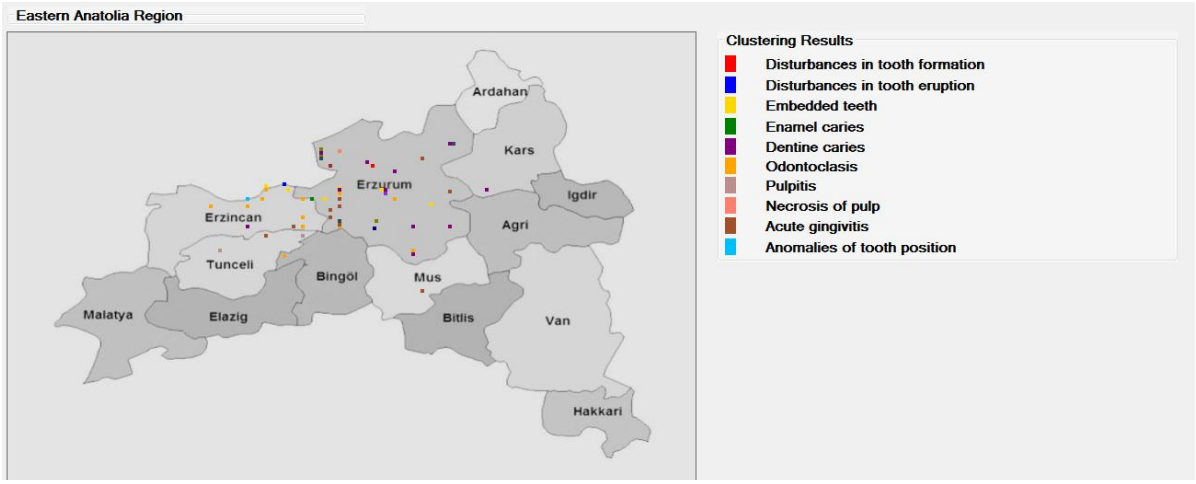

Figure 5- Regional map

Some of the diseases treated in Atatürk University Faculty of Dentistry concentrated in certain areas. For example, patients coming from the Eastern Black Sea Region usually applied to hospital for maxilla-facial deformity problems. The patients going from Eastern Anatolia to the Oral and Maxillofacial Surgery Department often complained about jaw tumors. Apart from this, traumatic cases came from the southern parts of Erzurum. Again, tooth discoloration usually came from East Beyazit. Cleft lip and palate were seen mostly in Eastern Anatolia. Our attributes in the classification are region and clinics. According to these features, classes are shown in the following table. By using the DVM algorithm, the data existing in each of these classes was estimated.

Table 3- Attributes and classes

Attributes		Classes
Region	**Polyclinic**	
Eastern Black Sea Region	Orthodontia	Class 1- Maxilla-Facial Deforı
Eastern Anatolia Region	Oral and Maxillofacial Surgery	Class 2- Jaw Tumors
Southern Parts of Erzurum	Prosthodontics	Class 3- Traumatic Cases
East Beyazit	Conservative Dental Treatment	Class 4- Tooth Discoloration
Eastern Anatolia Region	Pedodontics	Class 5- Cleft Lip and Palate
Eastern Anatolia Region	Periodontics, Endodontics	Class 6- Dental Periphery Disease, Neurovascular Disease

Figure 6 shows results for these six classes with 3012 training data and 3482 test data used. Percentage accuracy is 95.8. Rows have real classes and columns have predicted classes. For example, class 1 has 395 patterns, but 378 of them are classified correctly.

Classes	Class1	Class2	Class3	Class4	Class5	Class6
Class1	378	0	10	0	0	7
Class2	4	714	0	26	0	8
Class3	0	6	508	0	4	0
Class4	12	0	9	274	0	0
Class5	0	0	7	0	290	11
Class6	10	12	4	0	6	1172

Figure 6- Classification results

Belonging to these categories of data for the class above were given as output in the classification. The experimental results of the CB-SMoT algorithm applied to the health services trajectory data are shown in the following chart. According to minTime parameter which the algorithm takes as a different value parameter, algorithm outputs were analyzed as seen in Figure 7. In the data, stops and unknown stops did not change too much after minTime = 600 value.

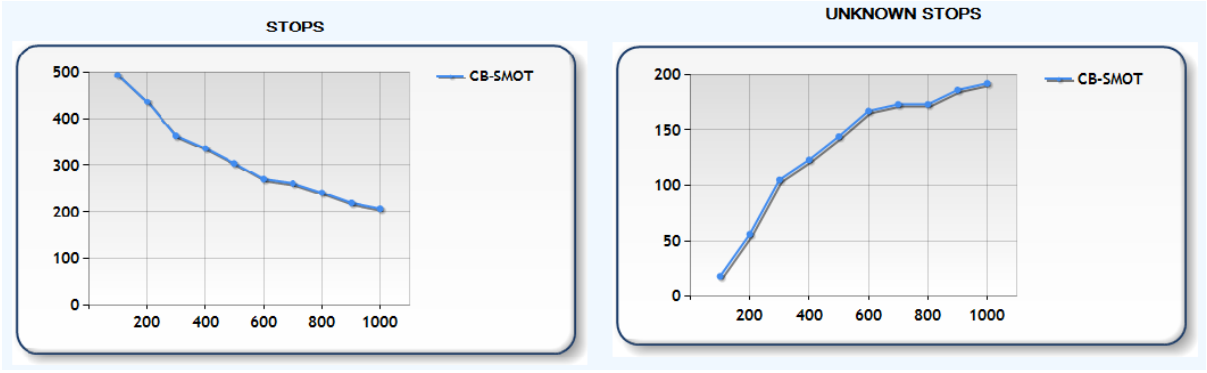

Figure 7- For CB-SMoT algorithm, Stop and unknown stops according to MinTime parameter

Figure 8 shows the number of stops and unknown stops for DB-SMoT algorithm according to minTime parameter. After minTime value exceeded a certain threshold, it was observed that unknown stops did not increase much and stops did not reduce very much.

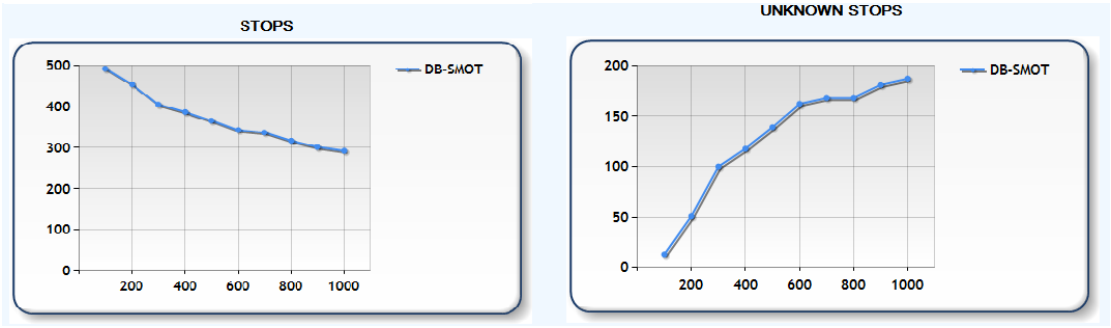

Figure 8- For DB-SMoT algorithm, Stop and unknown stops according to MinTime parameter

As inconsistent trajectories were obtained after TRAOD and TOD-SS algorithms were applied on the trajectories, their performances were compared. Sample algorithm outputs for a certain group are shown in Table 4, Table 5, Table 6, Table 7, Table 8, Table 9 and Table 10.

Table 4- Trajectory outlier detection outputs for Traod algorithm

Algorithm	D	P	F	Running Time (ms)	No. of Outliers
TRAOD	65	0.95	0.1	761	7
TRAOD	75	0.95	0.1	824	5
TRAOD	90	0.95	0.1	362	2
TRAOD	80	0.90	0.1	576	7
TRAOD	80	0.93	0.1	654	5
TRAOD	80	0.98	0.1	114	4

Table 5- TRAOD algorithm outputs for different D values

Algorithm	D	P	F	Running Time (ms)	No. of Outliers
TRAOD	65	0.95	0.1	761	7
TRAOD	68	0.95	0.1	188	7
TRAOD	70	0.95	0.1	327	5
TRAOD	72	0.95	0.1	269	5
TRAOD	75	0.95	0.1	824	5
TRAOD	80	0.95	0.1	79	5
TRAOD	82	0.95	0.1	120	5
TRAOD	85	0.95	0.1	874	2
TRAOD	88	0.95	0.1	387	2
TRAOD	90	0.95	0.1	362	2

Table 6- TRAOD algorithm outputs for different p values

Algorithm	D	p	F	Running Time(ms)	No. of Outliers
TRAOD	80	0.90	0.1	576	7
TRAOD	80	0.91	0.1	425	7
TRAOD	80	0.92	0.1	312	7
TRAOD	80	0.93	0.1	654	5
TRAOD	80	0.94	0.1	741	5
TRAOD	80	0.95	0.1	210	5
TRAOD	80	0.96	0.1	128	5
TRAOD	80	0.97	0.1	451	5
TRAOD	80	0.98	0.1	114	4
TRAOD	80	0.99	0.1	451	4

According to Tables 5 and 6, if D value increases, the number of outliers decreases and in the same way, if p increases, detected outliers decrease, too.

Table 7- Trajectory outlier detection outputs for Tod-ss algorithm

Algorithm	ε	W	σ	F	ξ	Running Time (ms)	Number of Outliers
TOD-SS	5	30	4	15.5	0.1	9702	2
TOD-SS	7	30	4	15.5	0.1	7887	5
TOD-SS	8	30	4	15.5	0.1	5643	8
TOD-SS	9	30	4	15.5	0.1	5115	9
TOD-SS	3	30	3	15.5	0.1	9559	2
TOD-SS	3	36	3	15.5	0.1	2365	4
TOD-SS	3	40	3	15.5	0.1	9625	4
TOD-SS	3	35	7	15.5	0.1	7282	4
TOD-SS	3	35	8	15.5	0.1	2794	6

Table 8- TOD-SS algorithm outputs for different ε values

Algorithm	ε	w	σ	F	ξ	Time (ms)	No. of Outliers
TOD-SS	1	30	4	15.5	0.1	4150	2
TOD-SS	2	30	4	15.5	0.1	8980	2
TOD-SS	3	30	4	15.5	0.1	2370	2
TOD-SS	4	30	4	15.5	0.1	1661	2
TOD-SS	5	30	4	15.5	0.1	9702	2
TOD-SS	6	30	4	15.5	0.1	6512	5
TOD-SS	7	30	4	15.5	0.1	7887	5
TOD-SS	8	30	4	15.5	0.1	5643	8
TOD-SS	9	30	4	15.5	0.1	5115	9
TOD-SS	10	30	4	15.5	0.1	8888	9

Table 9- TOD-SS algorithm outputs for different w values

Algorithm	ε	w	σ	F	ξ	Running Time (ms)	No. of Outliers
TOD-SS	3	30	3	15.5	0.1	9559	2
TOD-SS	3	32	3	15.5	0.1	5775	2
TOD-SS	3	35	3	15.5	0.1	9867	2
TOD-SS	3	36	3	15.5	0.1	2365	4
TOD-SS	3	37	3	15.5	0.1	264	4
TOD-SS	3	38	3	15.5	0.1	5379	4
TOD-SS	3	39	3	15.5	0.1	2871	4
TOD-SS	3	40	3	15.5	0.1	9625	4
TOD-SS	3	42	3	15.5	0.1	10208	2
TOD-SS	3	43	3	15.5	0.1	5786	2

Table 10- TOD-SS algorithm outputs for different σ values

Algorithm	ε	w	σ	F	ξ	Running Time (ms)	No. of Outliers
TOD-SS	3	35	2	15.5	0.1	1881	2
TOD-SS	3	35	3	15.5	0.1	4389	2
TOD-SS	3	35	4	15.5	0.1	8635	2
TOD-SS	3	35	5	15.5	0.1	1542	4
TOD-SS	3	35	6	15.5	0.1	4235	4
TOD-SS	3	35	7	15.5	0.1	7282	4
TOD-SS	3	35	8	15.5	0.1	2794	6
TOD-SS	3	35	9	15.5	0.1	10087	6
TOD-SS	3	35	10	15.5	0.1	4763	6
TOD-SS	3	35	11	15.5	0.1	4950	6

According to Tables 8, 9 and 10, if ε or σ increases, outliers will increase too. If w increases, outliers will firstly increase and then decrease.

When the tables are generally examined, working speed of TODD-SS algorithm is ten times lower than TRAOD algorithm on average for the same data. When the TRAOD algorithm was used with D = 80, p = 0.90 and F = 0.1, it identified seven trajectory outliers. When the TOD-SS algorithm was used with ε = 9, w = 30, σ = 4, F = 15.5, ξ = 0.1 for the same data, it detected two more adverse data as outliers in addition to the seven trajectory outliers that the TRAOD algorithm found. In most applications, it was determined that TOD-SS determined more outliers than TRAOD.

6. Conclusion

In this study, useful, meaningful data from large dimension data sets were obtained by using spatial-temporal data mining techniques for the health care services. The distance-based method, which is one of the techniques for outlier detection, is effective for large dimension data. It can pose some problems only if data contain both sparse and dense regions. The density-based method eliminates this feature of the distance-based method. It successfully detects outliers among data which contain both sparse and dense regions. Therefore, TRAOD and TOD-SS algorithms, combining these two methods, were preferred in our study. These algorithms divide trajectory into different parts by not taking moving objects as a whole, different from other techniques of trajectory outlier detection. Thus, they investigate whether each part exhibits different behavior compared to their neighbors. On this occasion, they perceive lower deviation orbits that other algorithms cannot detect. What's more, the TOD-SS algorithm attempts to detect inconsistent data by comparing the structure of the trajectory segment pairs. It also oversees the structural differences such as angle, speed, location, and direction.

When we compared these two algorithms, it was observed that the TOD-SS algorithm runs relatively more slowly for the same data compared to the TRAOD algorithm. However, as TOD-SS takes structural differences into account, it marked outliers which TRAOD did not determine as anomalous and this also showed that the TOD-SS algorithm is a more flexible, effective and efficient algorithm.

In this study, outlier detection provided these benefits for the Faculty of Dentistry: the identification of defects in Patient Information Management System (PIMS), the identification of incorrect diagnosis codes, leakage patients, and whether the flow of the trajectories are true or not. Illegal situations can be determined with outliers. For example, if a patient comes to the Faculty of

Dentistry for the first time, he has to go to Oral Diagnosis. If he is sent to a different department, faculty management needs to know this situation.

Additionally, the reason most of the patients come from the Eastern Black Sea Region to Ataturk University Faculty of Dentistry because of maxilla-facial deformities was found to be due to preference of the doctors as a result of the classification. Again, it was concluded that tooth discoloration is observed in patients from Eastern Beyazit because of the drinking water. It was determined that traumatic cases mainly came from the southern part of the province of Erzurum because of incompatibility between people. It was also agreed that jaw tumors were observed more because intermarriage of relatives is common in the Eastern Anatolia region. Again, it was found that cleft lip and palate occurred due to the same reason and environmental factors in the Eastern Anatolia Region. Other patients generally had dental periphery or neurovascular disease.

An attempt was made to determine trajectory clusters showing similar movement patterns temporally and spatially by using CB-SMoT and DB-SMoT algorithms. The main tendencies of moving objects were revealed by grouping similar points where stop and movements were intense.
In the future, outlier detection studies will be carried out to monitor sterilization in the Faculty of Dentistry. Whether there is a conflict between the trajectories determined as inconsistent and the trajectories of inconsistent patients obtained from this study will be identified. The development of a model that processes the trajectories of patients and diseases in real time is also being considered.

References

[1] Rao, K.V., Govardhan, A., Rao, K.V.C., "Spatiotemporal Data Mining: Issues, Tasks and Applications", International Journal of Computer Science and Engineering Survey (IJCSES), vol.3, issue 1, 2012, pp. 39-52.
[2] Geetha, R., Sumathi, N., Sathiabama, D. S., "A Survey of Spatial, Temporal and Spatio-Temporal Data Mining", Journal of Computer Applications, vol.1 issue 4, 2008, pp. 31- 33.
[3] Alvares, L.O., Palma, A.T., Oliveira, G., Bogorny, V., "WEKA-STPM: From Trajectory Samples to Semantic Trajectories", in: Proceedings of the Workshop on Open Source Code, Porto Alegre, Brazil 2010, pp. 1–6.
[4] Rocha, J.A., Times, V.C., Oliveira, G., Alvares, L.O., Bogorny, V., "DB-SMoT: A direction-based spatio-temporal clustering method", IEEE Conf. of Intelligent Systems, 2010, pp. 114-119.
[5] Sharma, L.K., Vyas, O.P., Scheider, S., Akasapu, A., "Nearest Neighbor Classification for Trajectory Data ITC", Springer LNCS CCIS 101, 2010, pp. 180–185.
[6] Almeida, V.T., and Güting, R.H., "Indexing the Trajectories of Moving Objects in Networks", Geoinformatica, vol. 9, issue 1, 2005, pp. 33-60.
[7] Gogoi, P., Bhattacharyya, D., Borah, B., and Kalita, J.K., "A survey of outlier detection methods in network anomaly identification", The Computer Journal, vol.54, issue 4, 2011, pp. 570–588.
[8] Kriegel, H.-P., Kroger, P., and Zimek, A., "Outlier Detection Techniques", in: Tutorial at the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2009.
[9] Ng, R.T., and Han, J., "Efficient and effective clustering methods for spatial data mining", Proceeding of *94* VLDB, 1994, pp. 144–155.
[10] Ester, M., Kriegel, H-P., and Xu, X., "A database interface for clustering in large spatial databases", in: Proceedings of 1st International Conference on Knowledge Discovery and Data Mining (KDD-95), 1995.
[11] Zhang, T., Ramakrishnan R., and Livny, M., "BIRCH: An efficient data clustering method for very large databases", Proceedings of the 96 ACM SIGMOD International Conference on Management of Data, New York, NY, USA 1996, pp. 103-114.

[12] Guha, S., Rastogi, R., and Shim, K., "CURE: An efficient clustering algorithm for large databases". SIGMOD Rec., vol. 27 issue 2, 1998, pp. 73–84.
[13] Cateni, S., Colla, V., and Vannucci, M., "Outlier Detection Methods for Industrial Application", in: J. Aramburo, A.R. Trevino (Eds.), Advances in Robotics, Automation and Control, Austria, 2008, pp. 265-281.
[14] Ben-Gal. I., "Outlier Detection", in: O. Maimon, L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, Kluwer Academic Publisher, 2005.
[15] Mansur, M.O., and Noor Md. Sap, M., "Outlier Detection Technique in Data Mining: A Research Perspective", in: Proceedings of the Postgraduate Annual Research Seminar, 2005.
[16] Knorr, E.M., and Ng, R., "Algorithms for Mining Distance-Based Outliers in Large Datasets", Proceedings of VLDB, 1988, pp 392-403.
[17] Knorr, E.M., Ng, R.T., and Tueakov, V., "Distance-Based Outliers: Algorithms and Applications", in: Proc. VLDB Journal, vol.8 issue 3, 2000, pp. 237-253.
[18] Ramaswamy, S., Rastogi R., and Shim, K., "Efficient algorithms for mining outliers from large data sets", Proceedings of the International Conference on Management of Data, Dallas, Texas, USA, 2000.
[19] Breunig, M. M., Kriegel, H-P., Ng, R.T., and Sander, J., "LOF: Identifying Density-Based Local Outliers", Proceedings of the 2000 ACMSIGMOD international conference on management of data, Dallas, Texas, United States, ACM Press New York, NY, USA, 2000, pp. 93–104.
[20] Papadimitriou, S., Kitawaga, H., Gibbons, P., and Faloutsos, V., "LOCI: Fast Outlier detection using the local correlation integral", Proceedings of the International Conference on Data Engineering, 2003, pp. 315-326.
[21] Birant, D., and Kut, A., "Spatio-temporal Outlier Detection in Large Databases", in: 28th International Conference on Information Technology Interfaces, 2006, pp. 179–184.
[22] Wu, E., Liu, W., and Chawla, S., "Spatio-Temporal Outlier Detection In Precipitation Data", Proceedings of the Second international conference on Knowledge Discovery from Sensor Data, Las Vegas, NV,2008, pp. 115-133.
[23] Bogorny, V., "Spatial and Spatio-Temporal Data Mining", Trajectory Knowledge Discovery, IEEE ICDM, Universidade Federal de Santa Catarina, 2010.
[24] Spaccapietra, S., Parent, C., Damiani, M.L., Macedo, J.A.F., Porto, F., and Vangenot, C., "A conceptual view on trajectories", Data Knowl. Eng., vol. 65, issue 1, 2008, pp. 126-146.
[25] Lee, J., Han, J., and Li, X., "Trajectory Outlier Detection: A Partition-and-Detect Framework", 24. Int'l Conf. on Data Engineering, 2008, pp. 140-149.
[26] Yuan, G., Xia, S., Zhang, L., Zhou, Y., and Ji, C., "Trajectory Outlier Detection Algorithm Based on Structural Features", Journal of Computational Information Systems, vol.7, issue 11, 2011, pp. 4137–4144.
[27] . Hsu, C.-W., Chang, C.-C., and Lin, C.-J., "A practical guide to support vector classification", Tech. rep., Department of Computer Science, National Taiwan University, 2003.
[28] Weston, J., "Support Vector Machine", In: Tutorial, 4 Independence Way, Princeton, USA
[29] Spaccapietra, S., Parent, C., Damiani, M.L., Macedo, A.D., Porto, F.,Vangenot, C., "A conceptual view on trajectories", Data and Knowledge Engineering, vol 65, issue 1, 2008, pp. 126-146.
[30] Alvares, L.O., Bogorny, V., Palma, A., Kuijpers, B., Moelans, B., Macedo, J.A.F., "Towards Semantic Trajectory Knowledge Discovery", Technical Report, Hasselt University, Belgium 2007.
[31] Palma A.T., and Bogorny, V., "A Clustering-Based Approach for Discovering Interesting Places in Trajectories, Master of Computer Science Thesis", Unıversıdade Federal Do Rıo Grande Do Sul Instıtuto De Informátıca Programa De Pós-Graduação Em Computação, Porto Alegre, 2008.

[32] Palma, A.T., Bogorny, V., Kuijpers, B., and Alvares, L.O., "A clustering-based approach for discovering interesting places in trajectories", in: ACMSAC, New York, NY, USA, ACM Press, 2008, pp. 863– 868.

[33] Rocha, J.A.M.R., Times, V.C., Oliveria, G., Alvares, L.O., Bogorny, V., "DB-SMoT: A direction-based spatio-temporal clustering method", IEEE Conf. of Intelligent Systems, 2010, pp. 114-119.