FUZZY ASSOCIATION RULE MINING FROM SPATIO-TEMPORAL DATA: AN ANALYSIS OF METEOROLOGICAL DATA IN TURKEY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SEDA ÜNAL ÇALARGÜN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

JANUARY 2008

Approval of the thesis:

"FUZZY ASSOCIATION RULE MINING FROM SPATIO-TEMPORAL DATA: AN ANALYSIS OF METEOROLOGICAL DATA IN TURKEY"

submitted by **Seda Ünal Çalargün** in partial fullfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen _____
Dean, **Graduate School of Natural and Applied Sciences**

Prof. Dr. Volkan Atalay _____
Head of Department, **Computer Engineering**

Prof. Dr. Adnan Yazıcı _____
Supervisor, **Computer Engineering Dept., METU**

Examining Committee Members:

Prof. Dr. İsmail Hakkı Toroslu _____
Computer Engineering Dept., METU

Prof. Dr. Adnan Yazıcı _____
Computer Engineering Dept., METU

Assoc. Prof. Dr. Şebnem Düzgün _____
Mining Engineering Dept., METU

Assist. Prof. Dr. Pınar Şenkul _____
Computer Engineering Dept., METU

Assist. Prof. Dr. Tansel Özyer _____
Computer Engineering Dept., TOBB

Date: 29.01.2008

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Seda Ünal Çalargün

Signature :

ABSTRACT

FUZZY ASSOCIATION RULE MINING FROM SPATIO-TEMPORAL DATA: AN ANALYSIS OF METEOROLOGICAL DATA IN TURKEY

Çalargün, Seda Ünal

M.S., Department of Computer Engineering

Supervisor: Prof. Dr. Adnan Yazıcı

January 2008, 105 pages

Data mining is the extraction of interesting non-trivial, implicit, previously unknown and potentially useful information or patterns from data in large databases. Association rule mining is a data mining method that seeks to discover associations among transactions encoded within a database. Data mining on spatio-temporal data takes into consideration the dynamics of spatially extended systems for which large amounts of spatial data exist, given that all real world spatial data exists in some temporal context. We need fuzzy sets in mining association rules from spatio-temporal databases since fuzzy sets handle the numerical data better by softening the sharp boundaries of data which models the uncertainty embedded in the meaning of data. In this thesis, fuzzy association rule mining is performed on spatio-temporal data using data cubes and Apriori algorithm. A methodology is developed for fuzzy spatio-temporal data cube construction. Besides the performance criteria interpretability, precision, utility, novelty, direct-to-the-point and visualization are defined to be the metrics for the comparison of association rule mining techniques. Fuzzy association rule mining using spatio-temporal data cubes and Apriori algorithm performed within the scope of this thesis are compared using these metrics. Real meteorological data (precipitation and temperature) for Turkey recorded between 1970 and 2007 are analyzed using data cube and Apriori algorithm in order to generate the fuzzy association rules.

Keywords: Data mining, fuzzy association rules, fuzzy spatio-temporal data cube, association rule mining, association rule mining comparison metrics

ÖZ

UZAYSAL VE ZAMANSAL VERİDEN BULANIK İLİŞKİ KURALLARI BULUNMASI: TÜRKİYEDE ÖLÇÜLMÜŞ OLAN METEOROLOJİ VERİSİNİN ANALİZİ

Çalargün, Seda Ünal

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Yöneticisi: Prof. Dr. Adnan Yazıcı

Ocak 2008, 105 sayfa

Veri madenciliği, büyük veri tabanlarından ilginç, önemli, dolaylı, daha önceden bilinmeyen ve yararlı olma olasılığı yüksek bilgi veya desen çıkarılmasıdır. Veri tabanında saklanan bilgileri birbirleriyle ilişkilendirmeyi hedefleyen veri madenciliği yöntemine ilişki kuralları madenciliği denir. Zamansal ve uzaysal veri üzerindeki madencilik, gerçek dünya uzaysal verilerinin çeşitli zamansal bağlamda var oldukları düşünüldüğünde çok miktarda veriye sahip uzaysal sunulan sistemlerin dinamiklerini dikkate alır. Zamansal ve uzaysal veri tabanlarındaki ilişki kurallarının madenciliğinde bulanık kümelere duyulan ihtiyaç, verinin keskin sınırlarını yumuşatarak verinin anlamındaki belirsizliği modelleyebilen bulanık kümelerin rakamsal veriyi daha iyi ele alabilmesinden dolayıdır. Bu tezde veri küpleri ve Apriori algoritması kullanılarak bulanık ilişki kuralları madenciliği gerçekleştirilmiştir. Bulanık zamansal ve uzaysal veri kübü oluşturulması için bir metodoloji geliştirilmiştir. Performans kriterinin yanısıra anlaşılırlık, netlik, kullanılabilirlik, yenilikçilik, alakalılık ve görüntülenebilirlik metrikleri ilişki kural madenciliği tekniklerini karşılaştırmak için tanımlanmıştır. Bu metrikler kullanılarak veri kübü ve Apriori algoritmasının bulanık ilişki kural madenciliği yetenekleri karşılaştırılmıştır. 1970 ve 2007 yılları arasında kaydedilmiş, Türkiye'ye ait gerçek meteoroloji verisi (sıcaklık ve yağış), bulanık ilişki kuralları çıkarılmak üzere veri küpleri ve Apriori algoritması ile analiz edilmiştir.

Anahtar Kelimeler: Veri madenciliği, bulanık ilişki kuralları, bulanık uzaysal ve zamansal veri kübü, ilişki kuralları madenciliği, ilişki kuralları madenciliği karılaştırma metrikleri

To my family…

For being with me, all the time…

ACKNOWLEDGMENTS

I would like to express my inmost gratitude to my supervisor Prof. Dr. Adnan Yazıcı. His patience, vision, sweet communication and friendly approach is the key reason to vitalize this work. He was the one believing in me more than anybody else. It is an honor for me to share his knowledge, wisdom and humanity.

I am also indebted to Dr. Zuhal Akyürek and Pınar Aslantaş Bostan for Turkey's Meteorological data set and comments. I would be lost in this domain without their help.

I would like to express my heart-felt thanks to Canku, my dear love. Without his unconditional love, joy and support, this thesis could not be completed.

I also would like to thank my parents and parents in law for their complimentary love and support.

Finally, I would like to dedicate this work particularly to our grand grandmother Sadriye Kaya and grandfather Ahmet Sönmez. I will always remember...

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

With the popular use of satellite telemetry systems, remote sensing systems, medical imaging, and other computerized data collection tools, a huge amount of spatial data has been stored in spatial databases, geographic information systems, spatial components of many relational or object-relational databases, and other spatial information repositories. The spatial data are gathered in spatio-temporal databases over time. As a result, it is becoming necessary to develop efficient methods for the analysis and understanding of such huge amount of spatio-temporal data and utilize them effectively. The relations embedded in the spatio-temporal data have to be revealed for the use of decision support systems in order to take the advantage of the collected information. Decision support systems are database systems that run queries efficiently to make decisions. They are very useful in understanding the trends and making predictions, which makes them the focus of the database industry.

Spatio-temporal data mining is one of the techniques for the analysis of geospatial data such as meteorology data. The common mistake of early researches on geospatial data mining is that they only take a static view of the data (e.g. location, dimensionality, geometry, and topology) in order to capture the spatiality [4]. However, the temporal dimension has to be integrated for the analysis of geospatial data since all geographic phenomena evolve in time. Spatio-temporal data mining eases the prediction of spatial processes or events with the knowledge extracted from spatio-temporal data.

The need to handle imperfect data that are either uncertain or imprecise has motivated fuzzy OLAP. Fuzzy logic has been introduced to provide a way of modeling the uncertainty of natural language. With the use of fuzzy logic, fuzzy OLAP enables the extraction of relevant knowledge in a more natural language and give results to the queries with a certain precision about reliability of knowledge. Numerical data dimensions, such as temperature,

precipitation, can be better expressed using fuzzy logic.

Fuzzy spatio-temporal data mining can be performed through constructing a data cube. Data cubes presents the data in a multidimensional view which allows to view the data from a number of different perspectives. Apriori algorithm represents classical data mining methods that can work on fuzzy spatio-temporal data to mine association rules. Mining association rules through data cube construction and Apriori algorithm gives the opportunity to make a comparison between them. Association rules mined using different methodologies can be compared according to interpretability, precision, utility, novelty, direct-to-the-point, performance and visualization metrics.

In this thesis, fuzzy association rule mining is performed on spatio-temporal data using data cubes and Apriori algorithm. A methodology is developed for fuzzy spatio-temporal data cube construction. Interpretability, precision, utility, novelty, direct-to-the-point, performance and visualization are used as the metrics for the comparison of fuzzy association rule mining using data cubes and Apriori algorithm. Mean monthly temperature and precipitation data for Turkey recorded between 1970 and 2007 is analyzed using data cube and Apriori algorithm in order to generate the fuzzy association rules. The use of real data for the realization of theoretical concepts has revealed the applicability of this study. The visualization of association rules based on their significance and support values is performed in order to provide a complete analysis tool for a decision support system in meteorology domain.

1.1 Related Work

Traditionally, analysts have performed the task of extracting useful information from recorded data. But, the increasing volume of data in modern business and science, requires computer based approaches. As data sets have grown in size and complexity, there has been an inevitable shift away from direct data analysis toward indirect, automated data analysis using more complex and sophisticated tools. The studies have been directed to capture spatial and time-varying characteristics of the data collected all around the world.

G. Pestana and M. M. da Silva have studied a multidimensional data model that is specialized to store spatial data and analyzes the requirements to implement efficiently on-line analytical processing of spatial data [2]. The proposed data model uses a visual modeling tool, named Perceptory [5], to support spatiotemporal data modulations. Their approach preserves the traditional star schema while bringing new spatial OLAP capabilities into

the decision process. They analyze the efficiency of exploiting business data and the corresponding evolution of spatial object over time, taking into account the following three perspectives:

- Conceptual multidimensional modeling of spatial data to spatially support the decision-making process.
- Theories and methods of spatiotemporal reasoning, to develop ontology studies for geospatial data interoperability, aggregation and multirepresentation.
- Shortening spatiotemporal heterogeneities and enhance exploratory spatial data analysis through spatial and topological operations.

In this thesis, the star schema is used. The data model approach in [2] is not adapted since the main concern is mining fuzzy association rules from meteorological data using data cubes. The efficiency in mining association rules is achieved through data cubes instead of designing a specialized multidimensional data view.

The second order Hidden Markov models (HMM2) can also be used to discover frequent sequences of events in temporal and spatial data. J.F. Mari and F. Le Ber describe a clustering method on spatial and temporal data based on a second order Hidden Markov models [6]. They use HMM2 to map the observations into a set of states generated by a second order Markov chain. Their classification is performed both in time domain and spatial domain by using a posteriori probability that the stochastic process is in a particular state assuming a sequence of observations. In this thesis, spatio-temporal data mining is performed by finding association rules instead of classification and clustering which is targeted in [6].

Spatial association rule mining is studied by L. Wang et al [7] through partitioning the set of rows with respect to the spatial relations in a relation table. Concepts, introduced by spatial domain, such as a reference object, task-relevant objects and concept hierarchy tree are used as keys for mining multilevel spatial association rules. A reference object is the main subject of the description while a task relevant object is a spatial object that is relevant for the task at hand and spatially related to the reference object. A concept hierarchy tree is a tree structure that forms taxonomy of concepts ranging from the root (the most general concept) to some representation of all special values at the leaves. Higher level concepts are generalization of the lower level concepts while lower level objects are the specialization of the higher level objects. The algorithm in [7] works as follows:

- Spatial relations between reference objects and task relevant objects are extracted through spatial query and spatial computation.
- The spatial predicates are transformed into a relation table.
- Equivalence partition trees for spatial relation, reference object and task relevant object in relation table are created by scanning the relation table.
- Frequent predicate sets are computed at each level of the concept hierarchy tables [8], [9].
- For each concept level, the association rules are found.

N. Stefanovic, J. Han and K. Koperski focus on a method for spatial data cube construction called object-based selective materialization, which is different from traditional cuboid-based selective materialization for non-spatial data cube construction [10]. This approach uses a single cell of a cuboid as the atomic structure during selective materialization instead of a cuboid. Three algorithms namely, spatial greedy, pointer intersection and object connection, have been developed for partial materialization of the spatial objects resulted from spatial OLAP operations. Although the main concern of [10] is the selective materialization, the construction of spatial data warehouses on a spatial data cube has also been studied. This thesis provides spatio-temporal data cube construction for meteorological data of Turkey. The ideas for handling spatial data in [10] have also been utilized. Different than [10] the temporal aspect has been introduced. The selective materialization is not used within the scope of this thesis. The materialization is done at cell level for each possible combination of temperature, precipitation, snowy days and elevation dimensions since the relation between these dimensions is interesting in meteorology domain.

A spatial cube is constructed in [11] for the analysis of spatial movement of RFID data sets. Data cube is divided into RFID cuboids in order to aggregate data at different abstraction levels. The idea is based on the fact that RFID data tend to move together in bulky mode. A model for efficiently answering of wide range of RFID queries is proposed. This model is suitable supply chain management applications. It cannot be applied to meteorology domain which is used as case study in this thesis.

Fuzzy temporal association rule mining has been studied in [12]. They developed a fuzzy calendar algebra to help the construction of desired time intervals in which interesting patterns are discovered and presented in terms of fuzzy temporal association rules. The time dimension has been fuzzified. Temporal descriptions such as 'on the weekends'

and 'at the end of a year' and 'at the very beginning of a month' are defined on the fuzzy calendars with a specific membership function. In this thesis, the time dimension is used in the seasonal information and year intervals which are not fuzzy at all. The fuzziness is introduced in the temperature, precipitation, snowy days and elevation dimensions.

Another paper focusing on the temporal dimension in association rule mining is [13]. This paper proposes a temporal association rule mining algorithm that differs the order of mining execution. First of all candidate frequent item sets are found using Apriori algorithm. Then the frequent item set is restricted using time constraints. Apriori algorithm is also studied within the scope of this thesis to mine association rules. For each record in the database, the time constraint is a simple date instead of a complex time expression. As a result, in this thesis the time restriction is considered before applying Apriori algorithm [14].

I. Narin [15] has contributed to spatial knowledge discovery by generating more understandable and interesting knowledge from spatial data by extending spatial generalization with fuzzy memberships, extending the spatial aggregation in spatial data cube construction by utilizing weighted measures, and generating fuzzy association rules from the constructed fuzzy spatial data cube. This thesis is similar to her work in the following aspects:

- In spatio-temporal data cube constructed, all the dimensions and the measures are fuzzified.
- Extended spatial data generalization for fuzzy spatial data cube to consider the membership values is utilized.
- Each cell of the data cube is constructed so that it has its own membership value for each dimension.

Different than [15], in this thesis, temporal dimension is also considered in the construction of the data cube. Temporal aggregation is developed to generate seasonal generalizations in this respect. I. Narin [15] calculates significance and certainty of fuzzy association rules such that they reflect the reliability of generalization, instead of the frequency of the data. In this thesis, the definition of significance and certainty is extended to reveal the frequency of the data, because the frequency is necessary to compare how a discovered association rule is supported on each spatial location. Apriori algorithm [14] is implemented as a second approach to mine fuzzy association rules from the fuzzy spatio-temporal data

cube. Fuzzy spatio-temporal association rule mining using data cubes and Apriori algorithm are compared according to some metrics, namely interpretability, precision, utility, novelty, direct-to-the-point, performance and visualization. Turkey's real meteorological data set, which includes a huge amount of data, is used to mine association rules for comparison. A method is proposed for the visualization of fuzzy association rules.

The comparison between different association rule mining algorithms is performed with respect to their performance [16], [17]. The quality of the association rules are also evaluated according to the support and confidence values [18]. Support and significance concepts are used to compare the quality of two different association rules instead of comparing the same association rule discovered by different methods. In this thesis, the fuzzy association rules are mined using data cube and Apriori algorithm. Interpretability, precision, utility, novelty, direct-to-the-point, performance and visualization are defined and studied as metrics to compare these methodologies in terms of their ability to find association rules and the quality of the discovered relationships.

1.2 Motivation Behind the Proposed Systems

There are several association rule mining techniques. Data cube approach and Apriori algorithm are two techniques for discovering spatio-temporal relationships. However, it is hard to say which one is a better association rule miner. Their ability to mine fuzzy association rules has to be compared. The commonly known criteria for comparison is the performance [16], [17]. Performance metric specifies which one is a faster or less resource consumer in mining association rules. Besides the performance criteria, the metrics that can express the quality of an association rule discovered shall be specified.

In this thesis, interpretability, precision, utility, novelty, direct-to-the-point, performance and visualization are defined to be the metrics for the comparison of association rule mining techniques [16], [17], [18]. Fuzzy spatio-temporal association rule mining using data cubes and Apriori algorithm are compared using these metrics. The data mining techniques are experimented in the analysis of Turkey's meteorological data collected since 1970.

A data cube for the discovery of the relationships among the meteorology data in Turkey has been constructed. The relationships are presented in terms of association rules which are found with the exploration of the data cube. Apriori algorithm is used as a second approach to mine association rules from the fact table of the fuzzy spatio-temporal data

cube. By this way, the resulting fuzzy association rules discovered by data cube mining and Apriori algorithm are compared in terms of metrics defined for the quality of association rules. An application is developed ready to be used by a domain analyst. The domain analyst does not need to know how to mine and interpret fuzzy association rules, since the relations discovered are mapped within GIS to enable further inspection.

1.3 Thesis Outline

In Chapter 2, the background information on which this thesis has been evolved is given. It details data mining, association rule mining and data warehouse concepts. How a fuzzy spatio-temporal data cube is constructed and association rule mining performed on the constructed data cube within the scope of this thesis is given in Chapter 3. The use of Apriori algorithm for mining fuzzy association rules and visualization technique developed by this thesis is also detailed. Chapter 4 includes the comparison between association rule mining with data cube and Apriori algorithm. It also studies the reliability of the association rules discovered by the techniques studied in this thesis with respect to the meteorology domain. The implementation guide necessary for the realization of the concepts detailed in Chapter 3 is given in Chapter 5. Chapter 6 provides the fuzzy spatio-temporal data mining application for Turkey's meteorological data set. Finally, Chapter 7 makes a conclusion of this thesis and points out future directions in this domain.

CHAPTER 2

BACKGROUND

2.1 Data Mining

Data mining is the extraction of interesting non-trivial, implicit, previously unknown and potentially useful information or patterns from data in large databases. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets.

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include [19]:

- association (patterns where one event is connected to another event, such as purchasing a pen and purchasing paper),
- sequence or path analysis (patterns where one event leads to another event, such as the birth of a child and purchasing diapers),
- classification (identification of new patterns, such as coincidences between duct tape purchases and plastic sheeting purchases),
- clustering (finding and visually documenting groups of previously unknown facts, such as geographic location and brand preferences),
- and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities, such as the prediction that people who join an athletic club may take exercise classes).

The Data Mining process usually consists of three phases [20]: 1) pre-processing or data preparation, 2) modelling and validation and, 3) post-processing or deployment. During

the first phase, the data may need cleaning and transformation according to possible constraints imposed by some tools, algorithms, or users. The second phase consists of choosing or building a model that better reflects the application behaviour. Such a model should be evaluated in terms of its efficiency and accuracy of its predictive results. Finally, the third step consists of using the selected model to effectively study the application behaviour. Usually, the model output requires some post-processing in order to exploit it. This step can benefit from data visualization, since interactivity and user expertise are very important in the final decisionmaking and data interpretation.

2.1.1 Spatio-temporal Data Mining

Spatial data mining is an emerging research area dedicated to the development and application of novel, typically inductive, computational techniques for the analysis of very large, heterogeneous spatial databases [21]. Spatio-temporal data mining takes into consideration the dynamics of spatially extended systems for which large amounts of data exist. Given that all real world spatial data exists in some temporal context, and knowledge of this context is often essential in interpreting it, spatial data mining is inherently spatio-temporal data mining to some degree.

A number of authors have noted that there are problematic issues in applying association rule mining to spatial data, and, analogously, spatio-temporal data. Non-spatial association rule mining seeks to find associations among transactions that are encoded explicitly in a database whereas spatial association rule mining seeks to find patterns in spatial relationships that are typically not encoded in a database but are rather embedded within the spatial framework of the geo-referenced data [22]. These spatial relationships must be extracted from the data prior to the actual association rule mining.

One challenge in spatial data mining is the performance/data storage tradeoff between preprocessing spatial relationships among geographic objects and computing those relationships on the fly [23]. A number of approaches have been developed to address this issue, including the use of R* trees and minimum bounding rectangles for fast computation of spatial relationships [24], the use of spatial relationship indices [25], and the encoding of spatial relationships among certain target sets of geographic objects prior to data mining [26].

Spatial data mining tasks and techniques can be roughly classified into five categories including segmentation, dependency analysis, deviation and outlier analysis, trend discov-

Table 2.1: A possible classification of spatio-temporal data mining tasks and techniques [1]

Spatio-Temporal data mining task	**Descriptions**	**Techniques**	
		Static spatial data	**Spatio-temporal data**
Segmentation	Clustering Classification	Cluster analysis Bayesian classification Decision tree Artificial neural networks	Temporal extensions to clustering Temporal extensions to classification
Dependency analysis	Finding rules to predict the value of some attribute based on the value of other attributes over time	Association rules Bayesian networks	Temporal association rules Temporal extensions to Bayesian networks
Deviation and outlier	Finding data items that exhibit unusual deviations from expectations clustering and other data mining methods	Clustering and other data mining methods Outlier detection	Temporal extension to techniques in the left column
Trend discovery	Prediction of lines and curves Summarizing the database, often over time Discover correlations among the events in sequences	Discovery of common trends Regression	Sequence mining
Generalization and characterization	Compact descriptions of the data	Bayesian networks Attribute oriented induction	Temporal extension to techniques in the left column

ery, and generalization and characterization. Table 2.1 applies this classification to spatio-temporal data mining tasks.

2.2 Association Rule Mining

Association rule mining seeks to discover associations among transactions encoded within a database [14]. An association rule takes the form A → B where A (antecedent) and B (consequent) are sets of predicates. For example, consider a database that encodes transactions made at a supermarket. An association rule may state that "customers that purchase bagels also purchase cream cheese". This statement may be expressed as:

If a bagel is purchased than cream_cheese is purchased [support%, confidence%]

Association rule mining uses the concepts of support and confidence to identify rules that are particularly interesting or unexpected. The support is the probability that a transaction in the database contains the predicates contained in the antecedent and the consequent, for instance the probability that a record in the database contains both the purchase of a bagel and a cream cheese in the example above. Support is formulated in equation 2.1.

$$Support\ of\ A \rightarrow B = \frac{\#\ of\ transactions\ containing\ A\ and\ B}{total\ \#\ of\ transactions} \tag{2.1}$$

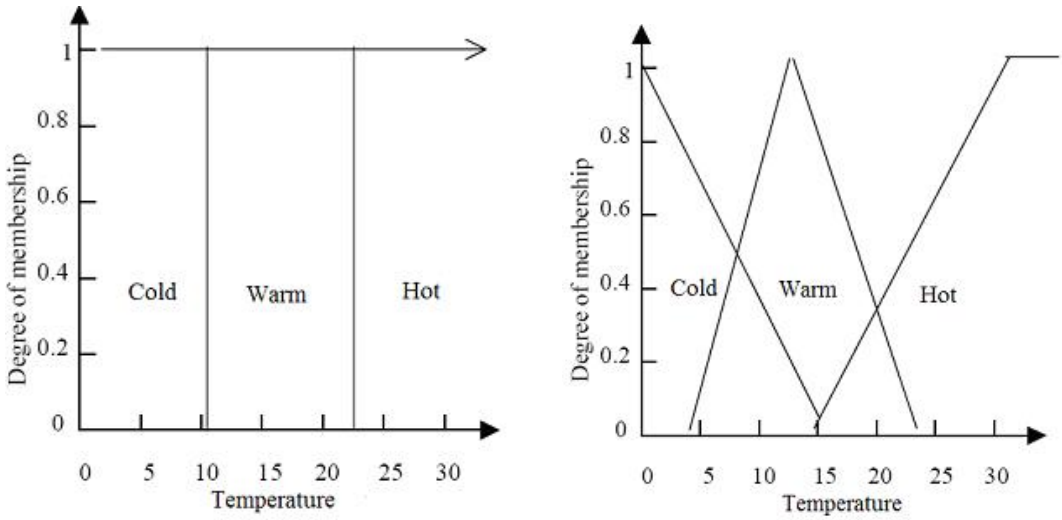
The confidence is the probability that a record that contains the antecedent also contains the consequent, for instance the probability that a record containing the purchase of a bagel in the database contains also the purchase of cream cheese in the example above. Confidence is formulated in equation 2.2.

$$Confidence\ of\ A \rightarrow B = \frac{\#\ of\ transactions\ containing\ both\ A\ and\ B}{\#\ of\ transactions\ containing\ A} \tag{2.2}$$

The support and confidence of a rule are typically reported in support-first order in parentheses following the rule, i.e. "(support%, confidence%)". Thresholds for support and confidence can be set to eliminate rules that are not of interest for a particular data mining application.

2.2.1 Fuzzy Association Rules

Fuzzy sets theory has been shown to be a very useful tool in data mining in order to represent the so-called association rules in a natural and human-understandable way. Fuzzy



(a) Non-fuzzy set

(b) Fuzzy set

Figure 2.1: Non-fuzzy and fuzzy representations of sets for quantitative variables. The x-axis is the value of a quantitative variable. The y-axis is the degree of membership in the sets cold, warm, and hot.

sets handles the numerical data better since they soften the sharp boundaries of data. Another advantage of using fuzzy logic is that it allows one to represent concepts that could be considered to be in more than one category (or from another point of view - it allows representation of overlapping categories). In standard set theory, each element is either completely a member of a category or not a member at all. In contrast, fuzzy set theory allows partial membership in sets or categories. The second technique, data mining, is used to automatically learn patterns from large quantities of data. The integration of fuzzy logic with data mining methods helps to create more abstract and flexible patterns.

As shown in Figure 2.1a, one would typically divide the range of possible values into discrete buckets, each representing a different set. The y-axis shows the degree of membership of each value in each set. The value 10, for example is a member of the set cold to the degree 1 and a member of the other two sets, warm and hot, to the degree 0. In fuzzy logic, a particular value can have a degree of membership between 0 and 1 and can be a member of more than one fuzzy set. In Figure 2.1b, for example, the value 10 is a member of the set cold to the degree 0.4 and a member of the set warm to the degree 0.75. In this example, the membership functions for the fuzzy sets are piecewise linear functions. Using fuzzy logic terminology, the temperature is a fuzzy variable (also called a linguistic variable), while the possible values of the fuzzy variable are the fuzzy sets cold, warm, and hot. In general,

Table 2.2: Record containing membership values

	Temperature		**Precipitation**		**Elevation**	
	label	**membership**	**label**	**membership**	**label**	**membership**
t1	hot	0.9	wet	0.2	low	0.5
t2	hot	0.7	wet	0.4	low	0.8
t3	hot	0.8	wet	0.3	high	0.2

fuzzy variables correspond to nouns and fuzzy sets correspond to adjectives.

Fuzzy association rules have the form "If X is A then Y is B" where X and Y are disjoint sets of attributes and A and B are fuzzy sets that describe X and Y respectively [27]. The semantics of the rule is when "X is A" is satisfied it can be implied that the consequent part "Y is B" is also satisfied. Interesting rules have enough significance and high certainty factors. Significance is for the satisfiability of the item-sets and certainty if for the satisfiability of the rules.

While generating fuzzy association rules, first large item-sets, those with a significance higher than a user specified threshold, are found. The significance is calculated by summing the votes of all records for the specified item-set and by dividing that sum to the count of the records. A vote of the record corresponds to the production of the membership values for the fuzzy sets in A that are described for X.

For example, X={temperature, precipitation}, A={hot, wet}, Y={elevation} and B={low}, and we have the following records T=t1, t2, t3 depicted in Table 2.2. Significance of the rule for is S(XY,AB)=(0.9*0.2*0.5+0.7*0.4*0.8)/3=0.104. After discovering large item-sets, interesting rules are generated according to the certainty factor that is computed as C((X,A),(Y,B)) =S(XY,AB)/S(X,A). For the example above, certainty is C((X,A),(Y,B))=0.104/((0.9*0.2+0.7* 0.4+0.8*0.3)/3)=0.452. The association rule found for this example can be represented as follows:

If temperature is hot and precipitation is wet then elevation is low [10.4%, 45.2%]

In short, if the rule has enough significance, it is determined as one of the large item-sets. Then, if it also has enough certainty it is specified as one of the interesting association rules in the database.

2.2.2 Apriori Algorithm

Apriori is an efficient algorithm that generates all significant association rules between items in a large database of customer transactions where each transaction consists of items purchased by a customer in a visit. In this algorithm, the problem of rule mining can be decomposed into two subproblems [14]:

1. Generate all combinations of items that have fractional transaction support above a certain threshold, called minimum support. Call those combinations large item sets, and all other combinations that do not meet the threshold small item sets.

2. For a given large item set Y = I_1 I_2 ... I_k, k $\geq$ 2, generate all rules (at the most k rules) that use items from the set I_1,I_2,...,I_k. The antecedent of each of these rules will be a subset X of Y such that X has k $-$ 1 items, and the consequent will be the item Y $-$ X. To generate a rule, take the support of Y and divide it by the support of X. If the ratio is greater than minimum confidence then the rule is satisfied with the confidence factor; otherwise it is not. Note that if the item set Y is large, then every subset of Y will also be large, and we must have available their support counts as the result of the solution of the first subproblem. Also, all rules derived from Y must satisfy the support constraint because Y satisfies the support constraint and Y is the union of items in the consequent and antecedent of every such rule.

Having determined the large item sets, the solution to the second subproblem is rather straightforward. As a result Apriori mainly concentrates on the discovery of large itemsets. Algorithm 1 provides the Apriori algorithm. The formulation of support and confidence is given in equation 2.1 and 2.2, respectively.

2.3 Data Warehouses

Data warehousing is a collection of decision support technologies, aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions. A data warehouse is a "subject-oriented, integrated, time varying, non-volatile collection of data that is used primarily in organizational decision making." [28]. Typically, the data warehouse is maintained separately from the organization's operational databases. There are many reasons for doing this. The data warehouse supports on-line analytical processing (OLAP), the functional and performance requirements of which are quite different from

Algorithm 1 Algorithm for Apriori

```
input: transactions, minSupport, item set
output: large k item sets

1. set largeItemSets to null
2. find item sets with 1 element
3. add 1 element large item sets with the frequency greater
   than minimum support to largeItemSets
4. set k to 2
5. while k is smaller than size of item set
6.   get large k-1 item sets
7.   if large k-1 item set is not empty
8.     generate candidate k element item sets by joining
       large k-1 item set with itself
9.     eliminate candidates with frequency smaller than
       minimum support
10.     add k element large items to largeItemSets
       increment k by 1
11.   end if
12. end while
```

Table 2.3: Differences between Data Warehouses and Operational Databases

	Data Warehouses	**Operational Databases**
Processing	OLAP	OLTP
Operations	Rollup, drill-down, etc.	Read, update
Task Types	Complex, ad-hoc	Structured, repetitive
Transactions	-	Short, atomic, isolated
Data	Historical, summarized and consolidated, read-only	Detailed, up-to-date
Data Structure	De-normalized, redundant	Normalized
Data Access	Millions of records	A few (hundreds of) records
Size	GBs to TBs	MBs to GBs
Critical issues	Data integration	Consistency, recoverability
Performance metric	Query throughput and response time	Transaction throughput
Target	Market	Customer
Database design	Star/Snowflake schema	ER, UML
Users	Knowledge worker	IT professional
Function	Decision support	Day-to-day operations

those of the on-line transaction processing (OLTP) applications traditionally supported by the operational databases. The differences between data warehouses and operational databases are presented in Table 2.3.

Typical data warehouse architecture is as displayed in Figure 2.2. It consists of back-end tools, metadata repository, data warehouse and optional data marts and front-end tools. Before running queries on data warehouses, consolidated data is cleaned [3], in other words corrected. Since large volumes of data from multiple are involves, there is high probability of errors and anomalies. After cleaning, data are loaded to data warehouses [3]. This process involves sorting, indexing, summarization, aggregation, integrity constraints checking, building derived tables and indices and materializing views. The front end tools consist of some query tools, report writers, analysis tools and data mining tools [3]. OLAP operations are executed by these front end tools to enable the end user to query in terms of domain specific business data.

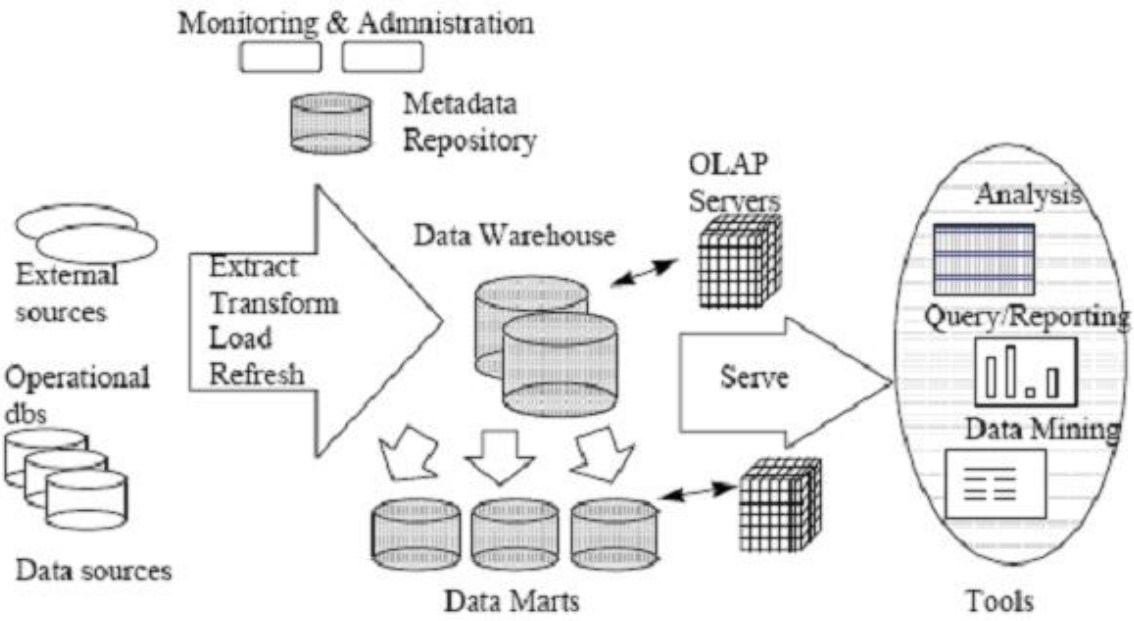


Figure 2.2: The Data Warehouse Architecture [3]

2.3.1 Spatio-temporal Data Warehouses

Spatial as well as non-spatial data can change their values over time. Particularly, in the data warehouses where the data is stored for long periods of time and where the changes to these data cannot overwrite the already existing values, the important consideration is how to represent the time during which these values are valid. Data warehouses consider the temporal aspect in a very limited way by including the time dimension and offering to represent the changes in time referring them only to the measures. Thus, if the changes to dimensions occur and the user wants to preserve them, firstly none of the conceptual multidimensional models represent that, and secondly, the logical level needs special applications and additional storage to manage them.

In the database community, temporality has been studied for more than a decade. Temporal databases include different types of time. The most used types are transaction time and valid time, which allow representing the changes of data during its existence. The transaction time serves to represent the time when the changes are applied to data using insert, delete, or update operations; the valid time indicates the period of time in which data is valid in the real-world regardless of when this information is recorded in the database. In the data warehouse context, the data cannot be deleted and modified, thus it is not necessary to include the transaction time. However, future work can incorporate the valid

time for dimensions that change over time. As a result, the graphic modeling of valid-time periods should be proposed in the conceptual data warehouse model.

[29] addresses spatio-temporal data warehouse schema modeling in order to facilitate those queries and more generally, on-line analytical processing (OLAP) on spatio-temporal databases. Some problems linked to the integration of data in a spatio-temporal data warehouse is presented in [30].

2.3.2 OLAP

OLAP (Online Analytical Processing) is a computing technique for summarizing, consolidating, viewing, applying formulae to and synthesizing data according to multiple dimensions in data warehouses [31]. OLAP has been growing in popularity due to the increase in data volumes and the recognition of the business value of analytics. Until the mid-nineties, performing OLAP analysis was an extremely costly process mainly restricted to larger organizations.

OLAP allows business users to slice and dice data at will. Normally data in an organization is distributed in multiple data sources and are incompatible with each other. For example, point-of-sales data and sales made via call-center or the Web are stored in different location and formats. It would be a time consuming process for an executive to obtain OLAP reports such as - What are the most popular products purchased by customers between the ages 15 to 30? As a result, part of the OLAP implementation process involves extracting data from the various data repositories and making them compatible. Making data compatible involves ensuring that the meaning of the data in one repository matches all other repositories. An example of incompatible data is as follows: Customer ages can be stored as birth date for purchases made over the web and stored as age categories (i.e. between 15 and 30) for in store sales.

OLAP necessitates a multidimensional data model to facilitate analysis and visualization. That model is designed differently from relational data models which were designed by entity relationship diagrams. Furthermore, in a relational data model, data are stored in tables; but in multidimensional data model, data are stored in n-dimensional spreadsheet (i.e. data cubes).

It is not always necessary to create a data warehouse for OLAP analysis. Data stored by operational systems, such as point-of-sales, are in types of databases called OLTPs. OLTP (Online Transaction Process) databases do not have any difference from a structural per-

spective from any other databases. The main difference, and only, difference is the way in which data is stored. OLTPs are designed for optimal transaction speed. Examples of OLTPs can include ERP, CRM, SCM, Point-of-Sale applications, Call Center.

There are different OLAP servers depending on the kind of DBMS the data warehouse is implemented on, access methods and query processing techniques like ROLAP, MOLAP, HOLAP, DOLAP and JOLAP.

ROLAP (Relational OLAP) stores data in relational databases [3], [34]. They are extended relational DBMSs or intermediate servers in front of the relational DBMSs. MOLAP (Multidimensional OLAP) stores the multidimensional data in special structures over which the OLAP operations are directly implemented [3], [34]. Special index structures, which are better than ROLAP, are used. HOLAP (Hybrid OLAP) combines both ROLAP and MOLAP [34]. At the lowlevel, relational tables are used while at the high-level array-based multidimensional storage is preferred. DOLAP (Directory OLAP) extracts data for manipulation from a relational or multidimensional database and access them via an OLAP engine. Small amounts of data are stored in files on a user's desktop computer. JOLAP (J2EE object-oriented interface to OLAP) is being developed by Java community. A standard set of object classes and methods for business intelligence are provided in that interface.

2.3.3 Multidimensional Data Model

The structure of a data warehouse is usually represented using the star/snowflake schema, also called multidimensional schema, made up of a set of fully denormalized dimension tables, one for each dimension of analysis, and a fact table whose primary key is obtained by composing the foreign keys referencing the dimension tables.

A star schema model ([3] [32]) contains a large central table (fact table) and a set of smaller attendant tables (dimensional tables) joined to the central table. The fact table stores the keys of multiple dimensions and the numerical measures and the dimensional tables store the textual description of the dimensions. A template star schema is given in Figure 2.3.

A variant of a star schema model is called a snowflake (schema) model ([3] [32]), where some dimension tables are normalized, further split into more tables, forming the shape similar to a snowflake. Snowflake schema is depicted in Figure 2.4.

With such a star/snowflake schema model, multidimensional databases or data cubes

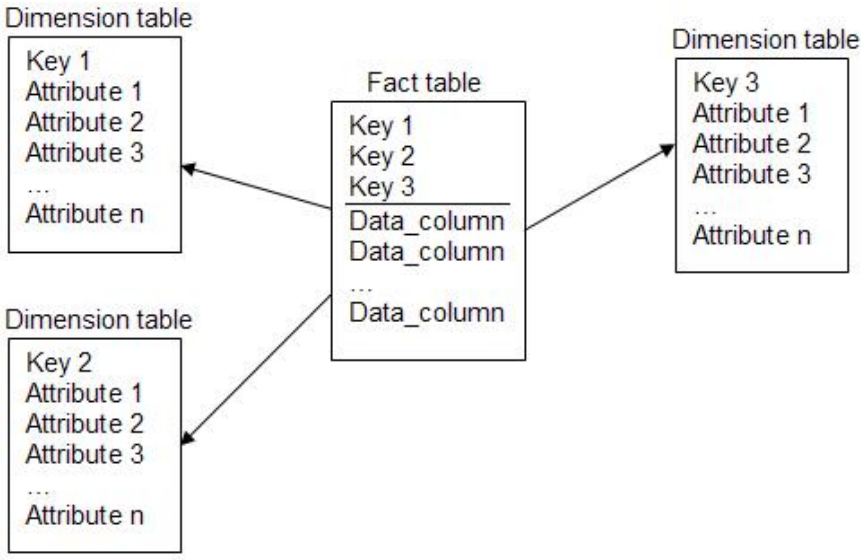


Figure 2.3: The star schema

can be constructed to facilitate typical OLAP operations, such as drill-down, roll-up, dicing, slicing, pivoting which are detailed in Section 2.3.5.

Mining in a data warehouse structured with a multidimensional data model, has been recognized to be an efficient way to find interesting patterns from large amount of data. However traditional multidimensional data model does not support the complex spatial data. It has to be extended within the scope of the requirements specified in Table 2.4.

2.3.4 Data Cube

Data cube is a data abstraction that allows one to view aggregated data from a number of perspectives. Conceptually, the cube consists of a core or base cuboid, surrounded by a collection of sub-cubes/cuboids that represent the aggregation of the base cuboid along one or more dimensions. The dimension to be aggregated is called the measure attribute, while the remaining dimensions are known as the feature attributes.

A d-dimensional base cube is associated with 2^d cuboids. Each cuboid represents a unique view of the data at a given level of granularity. Not all these cuboids need actually be present since any cuboid can be computed by aggregating across one or more dimensions in the base cuboid. However, some or all of these cuboids may be computed so that users may have rapid query responses at run time.

Data cubes are multidimensional extensions of 2-D tables, just as in geometry a cube is a three-dimensional extension of a square. The word cube brings to mind a 3-D object,

Table 2.4: Requirements for a multidimensional data model with spatial data [2]

Data Model Requirements	**Description**
Explicit and multiple hierarchies in dimensions	Explicit hierarchies are useful in data analysis because they aggregate data to the right level of detail for roll-up/drilldown operations efficiency Support for multiple hierarchies means that multiple aggregation paths are possible. These are important because multiple hierarchies exist naturally in much data and to handle imprecision in queries.
Partial containment	A multidimensional data model should provide built-in support for dimensions with partial containment relationships (e.g., districts would, though approximately, roll up to cities).
Non-normalized hierarchies	Situations occur where a hierarchy value has more than one parent, a value has no relationship to any value in the category immediately above it in the dimension hierarchy, or a value has no relationship to any value in any category below it (e.g., land parcel value may be related to several district parent values, or a city value may have no cell child values.
Different levels of granularity	Support for different levels of granularity enables a request to refer to other values than those in the category at the lowest level of a dimension hierarchy.
Many-to-many relationships between facts and dimensions	This requirement implies that a fact may be related to more than one value in a dimension.
Handling of imprecision	When facts are characterized by dimension values from different levels, imprecision in the data occurs because data for a query is missing or the transitive relationships between members of aggregation paths may become imprecise.

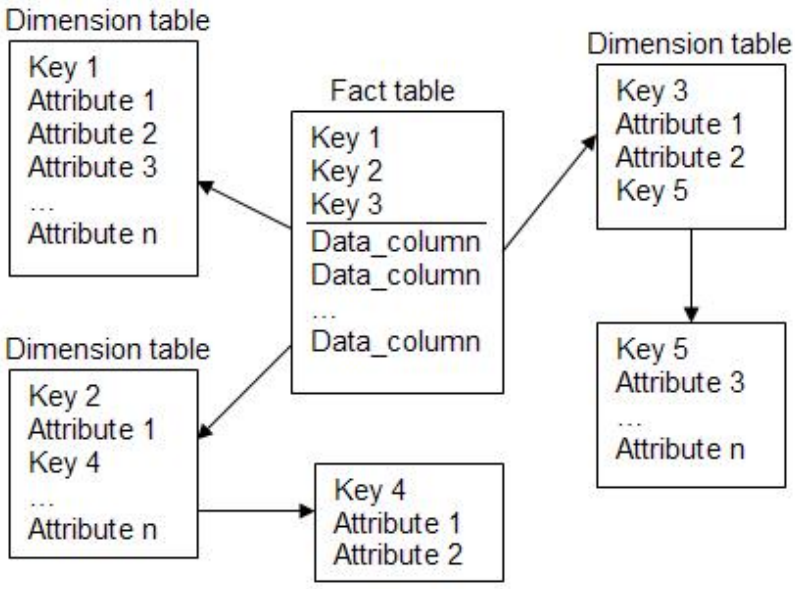


Figure 2.4: The snowflake schema

and it can be thought of a 3-D data cube as being a set of similarly structured 2-D tables stacked on top of one another. But data cubes aren't restricted to just three dimensions. Data cubes are constructed with many dimensions, but people tend to look at just three at a time. What makes data cubes so valuable is that the data cube can be indexed on one or more of its dimensions. A simple data cube is given in Figure 2.5.

In a spatial warehouse, both dimensions and measures may contain spatial components. A spatial data cube can be constructed according to the dimensions and measures modeled in the data warehouse. There are three types of dimensions in a spatial data cube [10]:

1. **Non-spatial dimension** is a dimension containing only non-spatial data. For example, temperature and precipitation dimensions contain non-spatial data whose generalizations are also non-spatial, such as hot and wet.

2. **Spatial-to-non-spatial dimension** is a dimension whose primitive level data is spatial but whose generalization, starting at a certain high level, becomes non-spatial. For example, state in the US map is spatial data. However, each state can be generalized to some non-spatial value, such as pacific_northwest or big_state, and its further generalization is non-spatial, and thus playing a similar role as a non-spatial dimension.

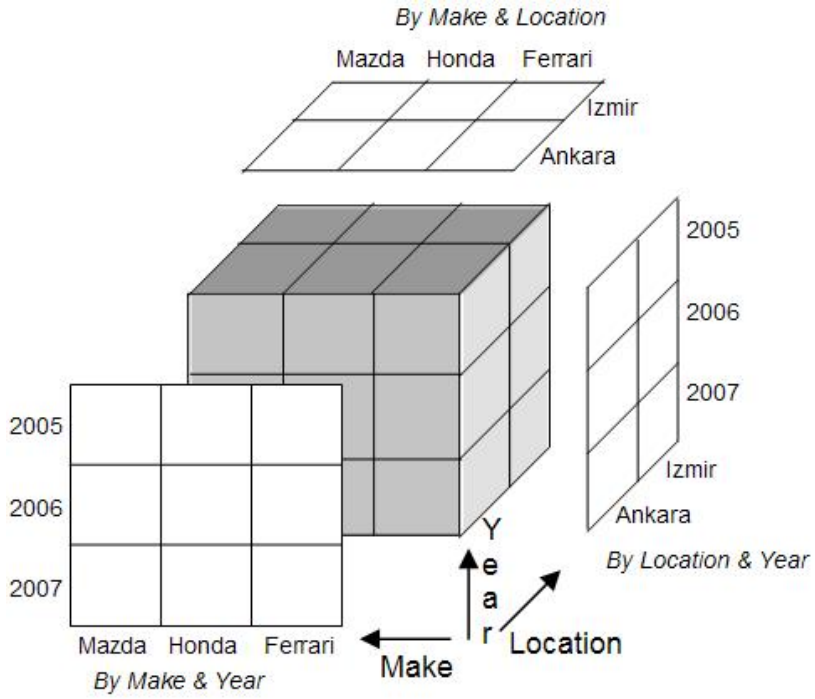


Figure 2.5: A simple data cube

3. **Spatial-to-spatial dimension** is a dimension whose primitive level and all of its high-level generalized data are spatial. For example, equi-temperature-region is spatial data, and all of its generalized data, such as regions covering 0-5_degree, 5-10_degree, and so on, are also spatial.

Measures in a spatial data cube can be grouped as follows [10]:

1. **Numerical measure** is a measure containing only numerical data. For example, one measure in a spatial data warehouse could be monthly revenue of a region, and a roll-up may get the total revenue by year, by county, etc. Numerical measures can be further classified into distributive, algebraic, and holistic. A measure is distributive if it can be computed by cube partition and distributed aggregation, such as count, sum, max; it is algebraic if it can be computed by algebraic manipulation of distributed measures, such as average, standard deviation; otherwise, it is holistic, such as median, most_frequent, rank. The scope of our discussion related to numerical measures is confined to distributive and algebraic measures.

2. **Spatial measure** is a measure which contains a collection of pointers to spatial objects. For example, during the generalization (or roll-up) in a spatial data cube for

weather domain, the regions with the same range of temperature and precipitation will be grouped into the same cell, and the measure so formed contains a collection of pointers to those regions.

2.3.5 Data Cube Operators

Data cube makes use of the OLAP operators to process the multidimensional data. OLAP is an emerging technology that empowers users with complete ease in manipulating their data. An OLAP interface enables users to perform various ad-hoc queries, arbitrarily filter data, rotate a table, drill down, get desired summaries, and rank. Some popular OLAP operations that manipulate data along dimensions are as the following [10]:

- **Roll-up:** Generalizes one or a few dimensions (including the removal of some dimensions when desired) and performs appropriate aggregations in the corresponding measures.
- **Drill-down:** Specializes one or a few dimensions and presents low-level objects, collections or aggregations. It is the reverse operation of roll-up.
- **Slicing:** Selects on one dimension. Summarized data is extracted for a given dimension value. Single item is extracted when selection on all dimensions is done.
- **Dicing:** Projects on dimensions. A sub cube is extracted. Several slices are intersected. Items are compared in a cross-tabulated table.
- **Pivoting:** Presents the dimensions in different cross-tabular layouts. It is a visualization operation that rotates the data axis. Data is examined from different angles.
- **Ranking:** Sorts the data along the selected dimension.
- **Filtering:** Performs selection using some constants.
- **Defining computed attributes.**

CHAPTER 3

FUZZY ASSOCIATION RULE MINING FROM SPATIO-TEMPORAL DATA

There is a huge amount of data accumulated by the use of digital maps, images from satellites, medical equipment and sensor network. Spatio-temporal databases have received considerable attention during the past few years due to the accumulation of such large amounts of multi-dimensional data evolving in time. This voluminous data is useless unless it is processed to discover interesting relationships and summarized for the analysis of the domain experts. Although this information can be obtained from operational databases, its computation is expensive, rendering online processing is inapplicable. A vital solution is the construction of a spatio-temporal data warehouse.

Spatio-temporal data warehouses are very suitable for the systems that need to store large amounts of spatial data evolving in time and analyze them, enable spatial characterization of spatial data, summarize the data in different dimensions at different level of abstraction and facilitate the discovery of knowledge and decision making.

Fuzzy data warehouses in mining association rules from spatio-temporal databases are essential since uncertainty in imperfect data can be modeled by fuzzy logic. Fuzzy logic has been introduced to provide a way of modeling uncertainty of natural language. It enables the extraction of relevant knowledge in a more natural language. It also gives results with a certain precision about the reliability of the mined knowledge. Numerical data dimensions, such as temperature, precipitation, can be better expressed using fuzzy logic.

In this thesis, knowledge discovery from the real meteorological data, which has 108,041 records, gathered between 1970 and 2007 in Turkey is performed for comparing data mining techniques in terms of the quality of association rules they discovered. A fuzzy spatio-

temporal data cube is constructed with this voluminous data and fuzzy association rules are mined from the constructed data cube. The association rules are also mined with the use of Apriori algorithm, which is famous for working on the transactional databases. Two approaches are studied for comparison according to the metrics defined within the scope of this thesis. The association rules are also displayed on a Turkey map. This visual representation enables the domain experts to analyze the results more easily and precisely.

3.1 Constructing a Fuzzy Spatio-Temporal Data Cube

In this section, the characteristic properties of the fuzzy spatio-temporal data cube constructed within the scope of this thesis is detailed. The similarities and the differences between the work performed and the literature are pointed out. The following subsections constitutes the construction of the fuzzy spatio-temporal data cube step by step.

Similar to [15], in fuzzy spatial generalization, besides the spatial generalization, the membership value of the generalization is also computed according to the defined fuzzy labels and membership functions. Membership values can have the value between [0,1] according to the fuzzy set theory and take as input numeric values from the crisp data such as 20 for temperature value. For example, a $20°C$ temperature region will be generalized as hot with membership value [0,82], which can be read as 82% hot. As a result, in the fuzzy spatio-temporal data cube the values are directly generalized to the descriptive fuzzy labels with the corresponding membership values. The use of ranges such as 15 to $25°C$ temperature regions [10] is surpassed. In fuzzy spatio-temporal data cubes, not only the dimensions but also the measures are fuzzified in the same manner.

Spatial data generalization is extended for fuzzy spatio-temporal data cube to consider the membership values. The membership value of the generalized tuple is calculated according to the weight of each tuple taking place in the generalization. In [15], the regions are generalized according to their fuzzy labels without considering the spatial relationships (such as neighborhood) among them. In this thesis, the spatial generalization is performed with the stations in the same basin. For example, the temperature value of the Gediz basin is generalized using its stations with the same fuzzy description. The membership value is calculated by taking the average of the membership values of the stations in the generalization. It is also necessary to keep the number of stations generalized since the count is used to calculate significance and the certainty of the mined association rules. Table 3.1 and Table 3.2 illustrates this fuzzy spatial generalization approach.

Table 3.1: Spatial Generalization

Basin	**Station**	**Temperature**
Gediz	17792	hot
Gediz	17184	hot
Gediz	17750	hot
Gediz	17186	cold

Table 3.2: Fuzzy Spatial Generalization

Basin	**Station**	**Temperature**	**Temperature Membership**	**Count**
Gediz	17792	hot	0.62	1
Gediz	17184	hot	0.84	1
Gediz	17750	hot	0.96	1
Gediz	17186	cold	0.54	1
Gediz	-	hot	0.80	3
Gediz	-	cold	0.54	1

In the fuzzy spatio-temporal data cube each cell has its own membership value for each dimension as it is in [15]. This is different than the work in [31] which proposes that each slice in the cube has the same membership value as given in Figure 3.1. Differentiation of membership values for each cell models the spatial data better, because spatial objects might have common properties, but each of them might posses that property with a different degree. Figure 3.2 includes an example fuzzy spatio-temporal data cube constructed with this approach. Please note that, there exists a spatial dimension, but it is eliminated in the figure in order not to make it complicated.

In this thesis, temporal aggregation is performed to generate seasonal generalizations. In [10], the roll-up on the time dimension is performed to roll-up values from day to month as follows: The roll-up of the temperature dimension is performed by first computing the average temperature grouped by month and by spatial region and then generalizing the values to ranges such as -10 to 0 or to descriptive names such as cold. This leads to loss of information. Using the approach in [10] the generalization of the values in Table 3.3 from month to season results in Table 3.5. Table 3.4 is used as a middl stage. In this thesis

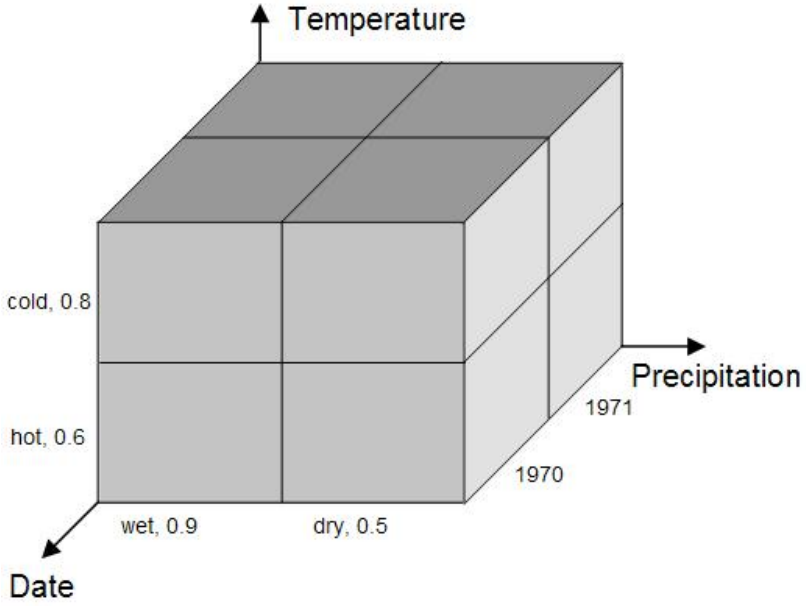


Figure 3.1: Fuzzy Data Cube

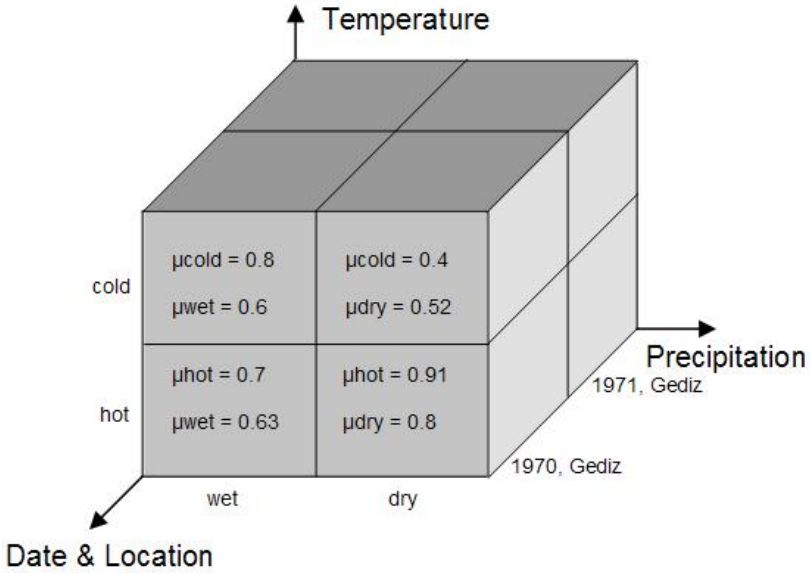


Figure 3.2: Fuzzy Spatio-Temporal Data Cube

Table 3.3: Crisp data per month

Month	**Station**	**Temperature**
March	1781	10
April	1781	14
May	1781	20
June	1781	15
July	1781	30
August	1781	29

Table 3.4: Average values per season

Season	**Station**	**Temperature**
Spring	1781	14.6
Summer	1781	24.6

Table 3.3 is generalized to Table 3.7 where first the crisp data is fuzzified to descriptive labels (Table 3.6) and then temporal aggregation is applied.

In [15], the interesting association rules are determined according to their significance and certainty factors, where these values reflect the reliability to generalization, instead of support and confidence factors which reflect the frequency of the data. In this thesis, the definition of significance and certainty is extended to reveal the frequency of the data because the frequency is necessary to compare how a discovered association rule is supported on each spatial location. Visualization of the association rules is based on this concept. For an association rule, the significance and certainty of each weather station is found using the fuzzy spatio-temporal data cube and mapped within GIS to display on Turkey.

Table 3.5: Fuzzy values per season

Season	**Station**	**Temperature**	**Temperature Membership**
Spring	1781	mild	0.6
Summer	1781	hot	0.8

Table 3.6: Fuzzy values per month

Month	**Station**	**Temperature**	**Temperature Membership**
March	1781	cold	0.5
April	1781	mild	0.6
May	1781	hot	0.72
June	1781	mild	0.62
July	1781	hot	0.8
August	1781	hot	0.78

Table 3.7: Fuzzy generalization per season

Season	**Station**	**Temperature**	**Temperature Membership**
Spring	1781	cold	0.5
Spring	1781	mild	0.6
Spring	1781	hot	0.72
Summer	1781	mild	0.8
Summer	1781	hot	0.79

Table 3.8: An example from Turkey's meteorological data

Date mm/dd/yyyy	Station	Min Temp	Max Temp	Avg Temp	Precipitation	Snowy Days	Elevation
2/1/2005	17050	-10.80	16.80	3.40	144.70	13	51.00
2/1/2005	17052	-10.00	15.00	2.50	94.10	5	232.00
2/1/2005	17054	-12.00	16.80	3.50	61.70	11	183.00
2/1/2005	17056	-9.50	17.00	4.20	74.90	10	4.00
2/1/2005	17059	-6.80	18.00	6.10	100.90	9	30.00
2/1/2005	17061	-4.00	17.00	5.80	151.00	10	58.00
2/1/2005	17062	-4.80	17.40	6.10	134.50	9	33.00
2/1/2005	17066	-2.50	19.30	7.40	96.70	10	76.00
2/1/2005	17069	-3.00	19.80	7.40	83.20	6	31.00
2/1/2005	17070	-9.60	14.60	2.40	41.50	12	742.00
2/1/2005	17074	-10.00	15.10	1.40	22.50	4	800.00
2/1/2005	17080	-12.70	15.00	1.20	27.40	5	751.00
2/1/2005	17083	-11.00	15.50	3.10	21.60	1	759.00

3.1.1 Meteorology Data Set

Climate summarizes the average, range and variability of weather elements, e.g. rain, wind, temperature, fog, thunder, and sunshine, observed over many years at a location or across an area. In order to understand the climate better, a wide range of data from the atmosphere, oceans and land surface have been collected. In this thesis, the data collected by the Turkish State Meteorological Service [33], which is the only legal organization providing all meteorological information in Turkey have been used.

The data set covers the measurements taken from 263 big climate stations in Turkey. The monthly averages for temperature, precipitation and the number of snowy days per station from 1970 to 2007 are included in the data set. The monthly minimum and maximum values for temperature are also recorded in the stations.

The data provided by the Turkish State Meteorological Service are in text format. They are processed and gathered in a database table shown in Table 3.8. There are a total of 108,041 records in the crisp data table.

3.1.2 Dimensions and Measures

The dimensions of the Turkey's meteorological data and their properties are as follows:

- Date: The date of measurement.
- Station: The weather station where the measurement is taken.
- MinTemperature: The minimum temperature (in °C) recorded in a month.
- MaxTemperature: The maximum temperature (in °C) recorded in a month.
- AvgTemperature: The average temperature (in °C) recorded in a month.
- Precipitation: Total of precipitation (in mm) in a month.
- SnowyDays: The number of days with snow in a month.
- Elevation: The altitude of the weather station.

The dimensions interested can be selected for the construction of the data cube. Within the scope of this thesis, the measures need not to be specified. Each dimension other than date and station can be selected as the measure at the time of association rule mining. In meteorology domain, the relation among the above dimensions are crucial. As a result, before association rules are mined, the dimension on which the effects of other dimensions are interested can be specified as the measure.

3.1.3 Fuzzifying Dimensions

For the dimensions other than date and station fuzzy sets and membership functions can be defined. The membership functions are applied to the values in the crisp data table. Then the values are described by the fuzzy label whose membership is bigger than the memberships of the other labels in the fuzzy set. Algorithm 2 states the algorithm for this process. An example fuzzy set for temperature dimension is given in Figure 3.3.

The algorithm for fuzzifying dimensions is given in Algorithm 2. Let us fuzzify a 18°C temperature region according to the fuzzy set {cold, mild, hot} and the membership function given in Figure 3.3. According to the membership function, 18°C temperature can be generalized as 'cold' with membership value [0,1], as 'mild' with membership value [0,8] and as 'hot' with membership value [0]. The fuzzy label with the greatest membership value labels the crisp data. As a result, 18°C temperature is labeled as 80% 'mild'.

Algorithm 2 Algorithm for fuzzyfying dimensions

```
input: a crisp value, fuzzy set, membership function
output: fuzzy label, membership value

1. set membershipValue and tmpMemValue to zero
2. set fuzzyLabel to null
3. for each fuzzy label in the fuzzy set
4.    set tmpMemValue to the result of the membership function
      applied to the crisp value
5.    if tmpMemValue is greater than the membershipValue
6.      then set tmpMemValue to membershipValue and set fuzzy
        label to fuzzyLabel
7.    end if
8. end for
```

Many popular distribution functions can be used as membership functions. Some common membership functions are listed in Figure 3.4. It is a good practice to decide on the fuzzy sets and their distributions with a domain expert.

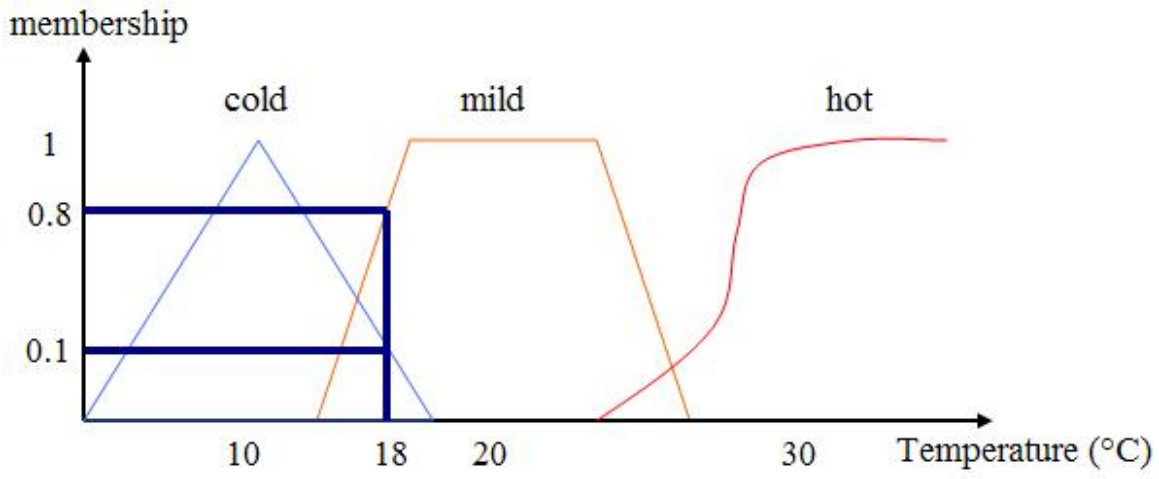


Figure 3.3: An example fuzzy set for temperature

3.1.4 Aggregating Data

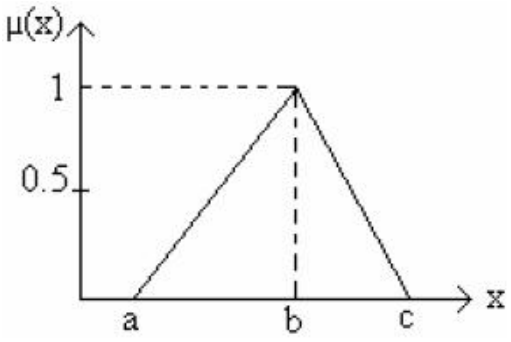
Each tuple in the meteorology data set includes a measurement value of a weather station for a specific date. After fuzzifying the dimensions, the data cube can be constructed once the aggregation of the data is completed.

The seasonal relationships among the dimensions are mined from the meteorological data set. Hence, temporal aggregation has to be performed. For a station, there are 12 recordings, one for each month, per year. In the fact table of the constructed data cube temporal dimension is represented with the following dimensions:

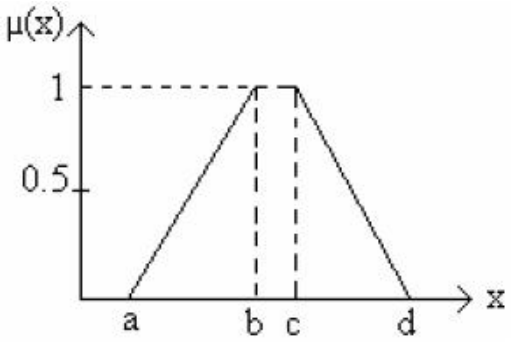
- Year
- Fall temperature & Fall temperature membership
- Winter temperature & Winter temperature membership
- Spring temperature & Spring temperature membership
- Summer temperature & Summer temperature membership
- Fall precipitation & Fall precipitation membership
- Winter precipitation & Winter precipitation membership
- Spring precipitation & Spring precipitation membership
- Summer precipitation & Summer precipitation membership
- Fall snowy days & Fall snowy days membership
- Winter snowy days & Winter snowy days membership
- Spring snowy days & Spring snowy days membership
- Summer snowy days & Summer snowy days membership

Three methods can be evaluated to project the monthly values on seasonal basis:

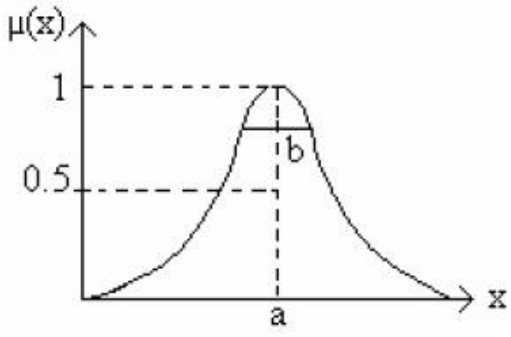
1. **Seasonal averages:** The seasonal average of the crisp values is calculated. Then the average values are fuzzified and a new tuple is generated which includes seasonal averages. This leads to loss of information, because monthly climate behavior is lost by averaging. The tuple generated using the Table 3.5 with this approach is given in Table 3.9.



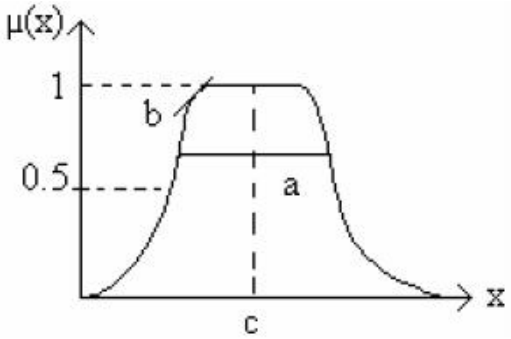
(a) Triangular



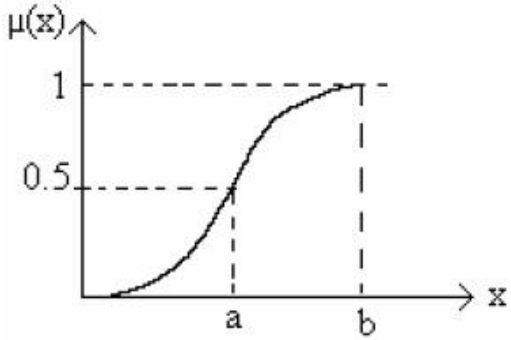
(b) Trapezoidal



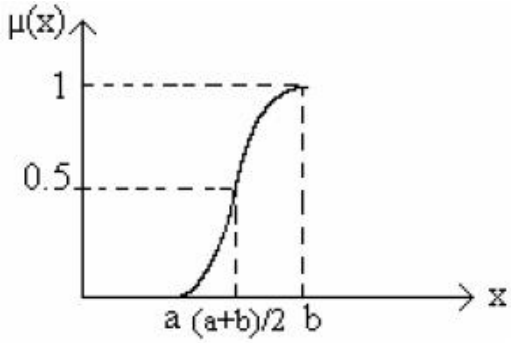
(c) Gaussian



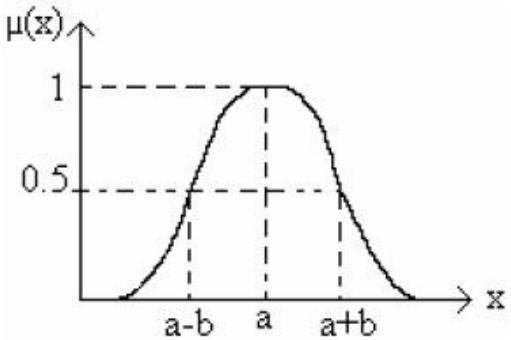
(d) Bell Shaped



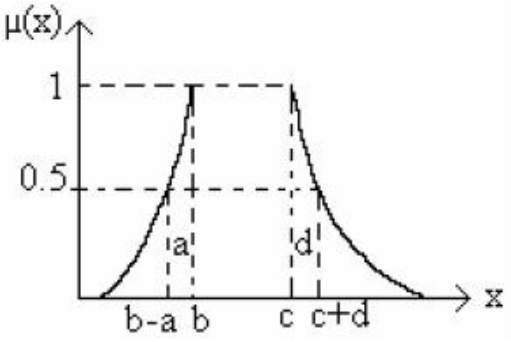
(e) Sigmoidal



(f) S Shaped



(g) $\prod$ Shaped 1



(h) $\prod$ Shaped 2

Figure 3.4: Membership function types

2. **Aggregating by sampling:** After fuzzifying the dimensions, one value from each season is combined to create a new tuple. As a result, 3 new tuples are stored. This is better than the first method, but still do not represent the correct yearly climate picture.

3. **Fuzzy aggregation:** Each possible combination of seasonal fuzzy values are combined in a new tuple. The tuple generated using the Table 3.7 with this approach is given in Table 3.10. As it is seen, although Table 3.9 and Table 3.10 are the results of aggregation of the same crisp data presented in Table 3.3, they are very different from each other. This approach is time and space consuming, but since the data cube will be constructed once and utilized for several time, it is acceptable. The correctness of the data included in the data cube is much more important then how long it takes to construct the data cube.

The third approach is utilized within the scope of this thesis for temporal aggregation. The algorithm for the fuzzy temporal aggregation is provided in Algorithm 3.

Table 3.9: Temporal aggregation by averaging

Station	**Spring Temperature**	**Spring TemperatureMem**	**Summer Temperature**	**Summer TemperatureMem**
1781	mild	0.6	hot	0.8

Table 3.10: Fuzzy temporal aggregation

Station	**Spring Temperature**	**Spring TemperatureMem**	**Summer Temperature**	**Summer TemperatureMem**
1781	cold	0.5	mild	0.8
1781	cold	0.5	hot	0.79
1781	mild	0.6	mild	0.8
1781	mild	0.6	hot	0.79
1781	hot	0.72	mild	0.8
1781	hot	0.72	hot	0.79

Algorithm 3 Algorithm for fuzzy temporal aggregation

```
input: fuzzy labels and membership values of
       fuzzified dimensions
output: temporal aggregated fact table

1. for each year (from 1970 to 2007)
2.   for each station in the fact table
3.     get records for the station in that year
4.     put fuzzy labels and memberships in fall, winter,
       spring and summer lists
5.     take the cartesian product of season lists
6.     for each element in the cartesian product
7.       store new tuple in the fact table
8.     end for
9.   end for
10. end for
```

As an example, the result of temporal aggregation performed on Table 3.7 are given in Table 3.10. In Table 3.7, we have a total of 5 tuples where 3 of them are recorded as {cold[0,5], mild[0,6], hot[0,72]} in 'Spring' and 2 of them are recorded as {mild[0,8], hot[0,79]} in 'Summer' seasons. In order to perform temporal aggregation we have to take the cartesian product of the sets {cold[0,5], mild[0,6], hot[0,72]} and {mild[0,8], hot[0,79]}. Then, we get {cold[0,5] mild[0,8], cold[0,5] hot[0,79], mild[0,6] mild[0,8], mild[0,6] hot[0,79], hot[0,72] mild[0,8], hot[0,72] hot[0,79]}. These new 6 tuples now represent both the behavior in 'Spring' and 'Summer' seasons.

Without the hierarchy concept, data warehouse cannot be pre-aggregated on spatial dimension [34]. It is logical that grouping data heavily contributes to the global query cost and such a cost can be reduced by pre-computing the aggregated data that are useful to answer a given workload. The fact table generated by the temporal aggregation of the fuzzified crisp data includes 718164 tuples. It is a large number that results in long response time for answering queries which is mining association rules in this thesis. It is necessary

to take the advantage of data cubes through aggregation on spatial dimensions. This is achieved by creating sub-cuboids as a result of spatial aggregation.

Each station belongs to a Basin. Regions possessing differentiated weather characteristics from their neighbors are specified as Basins by the Turkish State Meteorological Service. There are 26 basins in Turkey. The spatial hierarchy in the meteorological data set is as Station → Basin → Country.

The weather characteristics of the basins are derived by analyzing the data gathered in their stations. Following this approach, the values per stations are spatially aggregated by basins and stored in a sub-cuboid for each possible group by expression of the dimensions. The group by expressions for the construction of the fuzzy spatio-temporal data cube are as follows:

- {Temperature}
- {Precipitation}
- {SnowyDays}
- {Temperature, Precipitation}
- {Temperature, SnowyDays}
- {Temperature, Elevation}
- {Precipitation, SnowyDays}
- {Precipitation, Elevation}
- {SnowyDays, Elevation}
- {Temperature, Precipitation, SnowyDays}
- {Temperature, Precipitation, Elevation}
- {Temperature, SnowyDays, Elveation}
- {Precipitation, SnowyDays, Elevation}
- {Temperature, Precipitation, SnowyDays, Elevation}

The group by expressions correspond to the set of all possible unique elements of the dimensions. Please note that group by expression for Elevation is not needed to be found

Algorithm 4 Algorithm for fuzzy spatial aggregation based on group by expression

```
input: fact table
output: group by table

1. for each basin
2.   create a set of tuples whose fuzzy labels are unique
3.   for each element in the set
4.     find the tuples in the fact table whose fuzzy labels
       are same as the element
5.     count the tuples
6.     sum up the membership values for each dimension
7.     calculate new membership for each dimension by dividing
       the sum to the count
8.     insert one new tuple to the group by table with the
       fuzzy values in element, new membership values and
       count value
9.   end for
10. end for
```

Table 3.11: Fact table before spatial aggregation

Basin	Station	Spring Temperature	Spring TemperatureMem	Summer Temperature	Summer TemperatureMem
Meric	17050	mild	0.5	hot	0.8
Meric	17050	hot	0.6	hot	0.79
Meric	17052	hot	0.62	mild	0.54
Meric	17052	hot	0.54	hot	0.79
Meric	17054	mild	0.72	hot	0.8
Meric	17054	hot	0.58	hot	0.75

Table 3.12: Spatial aggregation based on group by expression on station

Basin	Spring Temperature	Spring TemperatureMem	Summer Temperature	Summer TemperatureMem	Count
Meric	mild	0.61	hot	0.8	2
Meric	hot	0.57	hot	0.77	3
Meric	hot	0.62	mild	0.54	1

for Turkey's meteorological data set, because the elevation has a fixed value for each dimension and does not vary in time.

The Algorithm for spatial aggregation based on a group by expression is provided in Algorithm 4. Table 3.12 presents the result of an example aggregation according to group by temperature of the sample fact table given in Table 3.11. The tuples {mild[0,5], hot[0,8]} and {mild[0,72], hot[0,8]} in Table 3.11 can be grouped into one tuple {mild[0,61], hot[0,8]} in Table 3.12 by averaging the membership values. It is important to note that, methods for incremental update of the data cube to make it consistent and up-to-date with the introduction of new data is not addressed in this thesis.

3.1.5 Finding Fuzzy Association Rules From Data Cube

Association rules are mined from the fuzzy spatio-temporal data cube, constructed with Turkey's meteorological data set, with some restrictions specified by the user:

- Basin: The relationships in the specified basin is searched. Generally, the basin is the smallest unit for the analysis in meteorology domain.

- From-To Year: The analysis can focus on a specific year interval. The data belonging to this year interval is used in mining the association rules.
- Measure: One dimension is specified as the measure. The effects of the other dimensions on the measure is searched through association rules.
- Dimensions: The dimensions of interest are specified by the user.
- Minimum significance: The rules above the minimum significance are considered as frequent. Other rules are eliminated.
- Minimum certainty: The rules above the minimum certainty are considered as interesting. Other rules are eliminated.

Calculation of the support and confidence is the heart of association rule mining. The support and confidence concepts are extended to reflect the effects of the fuzzy memberships values. Significance and certainty are defined corresponding to support and confidence, respectively.

Significance of a rule in a data cube can be formulated as follows:

$$Significance\ of\ A \rightarrow B = \frac{\sum(\prod(\mu_i)) * Count}{the\ size\ of\ the\ cube}\ where, \tag{3.1}$$

i is the dimension and μ_i is the membership value of dimension i for the tuples containing fuzzy values A and B.

Certainty of a rule in a data cube can be formulated as follows:

$$Certainty\ of\ A \rightarrow B = \frac{Significance\ A\ and\ B}{Significance\ A} \tag{3.2}$$

The algorithm for mining association rules in fuzzy spatio-temporal data cube is presented in Algorithm 5.

Table 3.13 can be used as a group by table in order to give an example that clarifies Algorithm 5. The example rule definition is to find relations between 'Spring precipitation' and 'Summer temperature' in Meric basin with minimum significance and minimum certainty greater than 10%. Fuzzy set for precipitation is given as {dry, fair, wet} and fuzzy set for temperature is given as {cold, mild, hot}. The cartesian product of these fuzzy sets is {dry cold, dry mild, dry hot, fair cold, fair mild, fair hot, wet cold, wet mild, wet hot}. For each element in the cartesian product, the significance is found. Significance of 'dry cold' is zero which is smaller than minimum significance. Then, we will continue with the next

Algorithm 5 Algorithm for mining fuzzy association rules from data cube

```
input: fact table, sub-cuboids, basin, dimensions, measure,
       minSignificance, minCertainty
output: association rules

1. find the sub-cuboid suitable for the dimensions and measure
2. compute the cartesian product of the fuzzy labels in the
   dimensions and measure
3. for each element in the cartesian product
4.   find significance of A->B
5.   if significance is greater than or equal to minSignificance
6.     exclude measure and find significance of A
7.     calculate certainty by dividing significance A->B by
       significance A
8.     if certainty is greater than or equal to minCertainty
9.       find new membership values
10.      print association rule
11.    end if
12.  end if
13. end for
```

Table 3.13: Example cuboid for association rule mining

Basin	**Spring Precipitation**	**Spring PrecipitationMem**	**Summer Temperature**	**Summer TemperatureMem**	**Count**
Meric	dry	0.61	hot	0.8	2
Meric	fair	0.57	hot	0.77	3
Meric	wet	0.62	mild	0.54	1
Firat	dry	0.7	hot	0.8	6

element. Suppose it is time for 'dry hot'. The significance for 'dry hot' is ((0.61*0.8)*2)/4 = 0.244. Since the significance is greater than minimum significance, certainty has to be found. The certainty of 'dry hot' is (0.244/(0.61*2/4)) = 0.8 which is greater than minimum certainty. As a result, the first association rule is found:

From 1970 to 2007 in Meric if spring is 61% dry then summer is 80% hot [24.4%, 80%]

An example of an association rule mined from the data cube constructed with the real meteorological data set of Turkey is:

From 1970 to 2007 in Gediz if summer is 87% hot then summer is 54% dry [36%, 43%]

This rule means that after analyzing the data gathered in Gediz basin from 1970 to 2007 the followings can be inferred:

- Summer temperature affects the precipitation in Summer.
- When summer is really hot (i.e 87% hot), summer is considerably dry (i.e. 54% dry).
- One third (36%) of all measurements at the stations in Gediz basin can be said to record summer as hot and dry.
- Half of (43%) all measurements stating summer is hot, it can be said that it is also recorded as dry.

The data cube is aggregated according to all possible combinations of the temperature, precipitation, snowy days and elevation dimensions. When the query is restricted to a specific time period other than 1970-2007, it is required to access the fact table although the entire spatial data cube has been pre-aggregated since the aggregation is not performed for time dimension.

3.1.6 Finding Association Rules Using Apriori

Apriori algorithm is famous for mining association rules from transactional data sets. Fuzzy databases are very similar to transactional databases. Transactional databases include binary features and items (i.e. items in market basket). Fuzzy data sets present quantitative features with fuzzy labels which can be thought as items [35], [27]. As a result, Apriori

Table 3.14: Example cuboid for Apriori Algorithm

Basin	**Spring Precipitation**	**Summer Temperature**
Meric	dry	mild
Meric	dry	hot
Meric	wet	mild
Meric	dry	hot

algorithm can also be used to mine association rules from the data cube constructed within the scope of this thesis. The fuzzy labels of the dimensions selected by the user in the fact table is provided to the Apriori algorithm as the item set input.

An example execution of Algorithm 1 can be demonstrated using Table 3.14. Let minimum support and minimum significance be 0.4 and fuzzy elements for precipitation and temperature be {dry, fair, wet} and {cold, mild, hot}, respectively. In this table large item sets with one element is {dry, hot}. 'mild' and 'wet' is eliminated since their frequency is 1/4=0.25 which is smaller than 0.4. The other fuzzy elements are also eliminated since their frequency is 0. Now it is time to find large item sets with two element by joining 1 element large item sets with itself and checking support values of each candidate. {dry hot, hot dry} are two element large item sets. As a result, the following rules are found by Apriori:

From 1970 to 2007 in Meric if spring is dry then summer is hot [50%, 66.6%]

From 1970 to 2007 in Meric if summer is hot then spring is dry [50%, 100%]

The following association rule mined from Turkey's meteorological data set using Apriori algorithm is:

From 1970 to 2007 in Gediz if summer is hot then summer is dry [76%, 78%]

This rules means that after analyzing the data gathered in Gediz basin from 1970 to 2007 the followings can be inferred:

- Summer temperature affects the precipitation in Summer.
- 76% of all measurements at the stations in Gediz basin recorded summer hot and dry.

- When summer is recorded as hot, it is also recorded as dry 78% of the time.

The membership values are not processed by the Apriori algorithm. As a result, Apriori treats each fuzzy label with a membership value of 1. This degrades the quality of discovered association rules.

3.2 Visualization of Association Rules

An association rule of the following form is informative for a meteorological domain expert. But it is not complete. In the GIS applications, the visualization of the outputs reveals the power of the analysis. As a result, the visualization of the mined association rules is necessary.

From 1970 to 2007 in Gediz if summer is 87% hot then summer is 54% dry [36%, 43%]

The strength of the relationship represented by an association rule is stated with its significance and certainty. Therefore, these two metrics can be used for visualization. The association rules mined within the scope of this thesis shows seasonal meteorological relationships in a specific basin. It is very natural for a domain expert desire to see the global picture, that is the map of this rule.

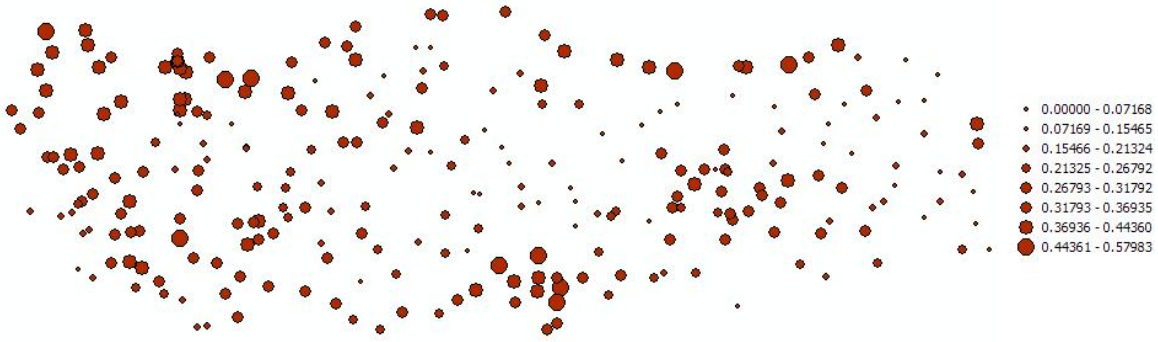


Figure 3.5: Significance map with symbols - From 2000 to 2007 in Gediz if summer is 87% hot then summer is 54% dry [36%, 43%]

For a specific association rule found for a basin, the significance and certainty of this rule for each station in Turkey is calculated and stored in a .dbf file. This .dbf file is processed with ArcGIS engine [36] to visualize the association rule in two different maps,

significance map and certainty map, respectively. Throughout this thesis, interpolation is performed on the discrete data in .dbf file using IDW (Inverse Distance Weighting) technique. Any other visualization technique and symbology can be applied by the domain experts according their needs. Figure 3.5 presents an example significance map which is visualized using discrete symbols. The .dbf file for the above rule is provided in Appendix C.

Significance map presents where an association rule is valid on the map. The colors show the strength of validity, that is the association rule is supported more frequently in the dark colored areas and it is rare for lighter colored areas. Certainty map presents the expectancy of an association rule among the measurements in the stations satisfying the first part of the rule. It is a good indicator for the prediction of an association rule when the antecedent is satisfied.

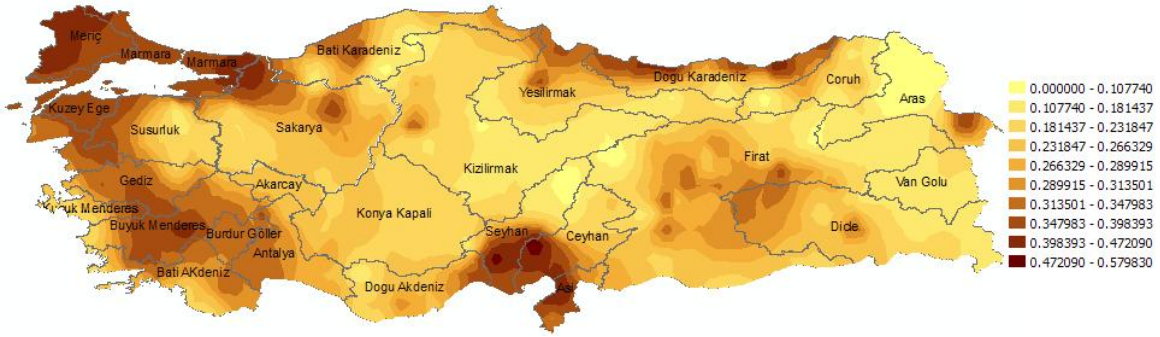


Figure 3.6: An example significance map - From 1970 to 2007 in Gediz if summer is 87% hot then summer is 54% dry [36%, 43%]

The significance and certainty maps for the association rule given above are presented in Figure 3.6 and Figure 3.7, respectively. In Figure 3.6, it is seen that Meric, Marmara, Buyuk Menderes and Asi basins have darker colors than the other regions. This means that in these basins, years with hot summer and dry summer have been observed more frequently than the other parts since 1970. In Figure 3.7, it is seen that Meric, Marmara, Bati Karadeniz and Dogu Karadeniz basins are colored with dark brown. This can be interpreted as follows: in these basins, summer can be expected to be dry if it is hot. This information is extracted with the analysis of meteorological data collected since 1970. It is seen that although Bati Karadeniz and Dogu Karadeniz are colored with dark in Figure 3.7, they are colored with light colors in Figure 3.6. This means that dry and hot summers

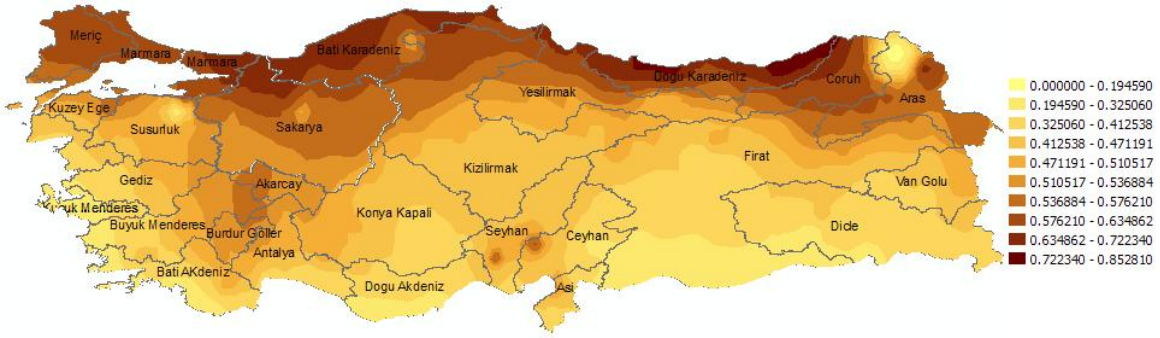


Figure 3.7: An example certainty map - From 1970 to 2007 in Gediz if summer is 87% hot then summer is 54% dry [36%, 43%]

are not frequently observed in these basins, but the chance of having dry summers when summer is hot, is higher.

Support and confidence maps are also created for the association rules mined using Apriori algorithm. These maps enable the comparison between fuzzy spatio-temproal association rules mined using data cube and Apriori approaches.

Association maps extend the meaning of a fuzzy association rule to spatial and temporal level. An association rule is discovered from meteorological data for a specific basin. In order to create the map of an association rule, the significance and certainty of the rule in each station are calculated. Then, these values are interpolated to visualize the map. By this way, the lowest level of the spatial hierarchy is used. When the map is created, the rule can be thought as a spatial rule. For example using Figure Figure 3.6 the following spatial rule can be extracted:

From West to East of Turkey, the effect of hot summers on dry summers decreases

The association maps can also enable the extraction of temporal knowledge. For this purpose, the same rule can be searched for successive time periods from 1970 to 2007. By analyzing each association map for each period, we can infer the temporal variation of the effects of hot summers on dry summers. The significance map of the following rules are presented on Figure 3.8 and Figure 3.9, respectively.

From 1970 to 1980 in Gediz if summer is 81% hot then summer is 57% dry [38%, 47%]

From 2000 to 2007 in Gediz if summer is 89% hot then summer is 55% dry [36%, 41%]

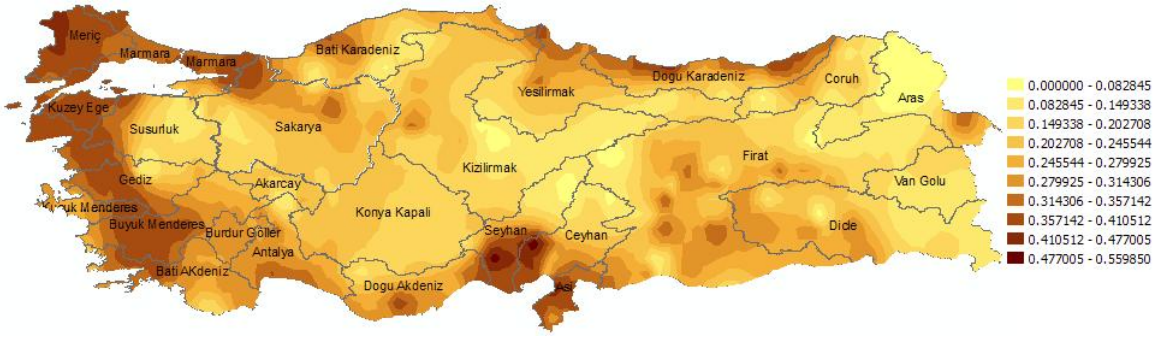


Figure 3.8: Significance map between 1970 and 1980 - From 1970 to 1980 in Gediz if summer is 81% hot then summer is 57% dry [38%, 47%]

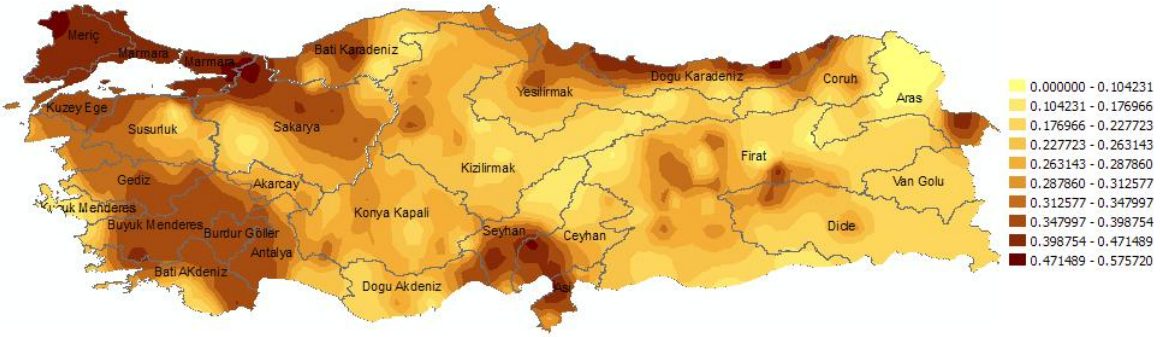


Figure 3.9: Significance map between 2000 and 2007 - From 2000 to 2007 in Gediz if summer is 89% hot then summer is 55% dry [36%, 41%]

The difference map can be visualized to analyze the temporal variation of the association rule between these time periods. The difference map where the significance values per station calculated for 1970-1980 are extracted from the significance values per station calculated in 2000-2007 is given in Figure 3.10. According to the difference map, the following temporal rules can be inferred between 2000-2007 and 1970-1980 time periods:

In recent years in Asi, the effect of hot summers on dry summers has not been changed

In recent years in West Turkey, the effect of hot summers on dry summers has been extended to a larger region

In recent years in Southeast Turkey, the effect of hot summers on dry summers has been decreased

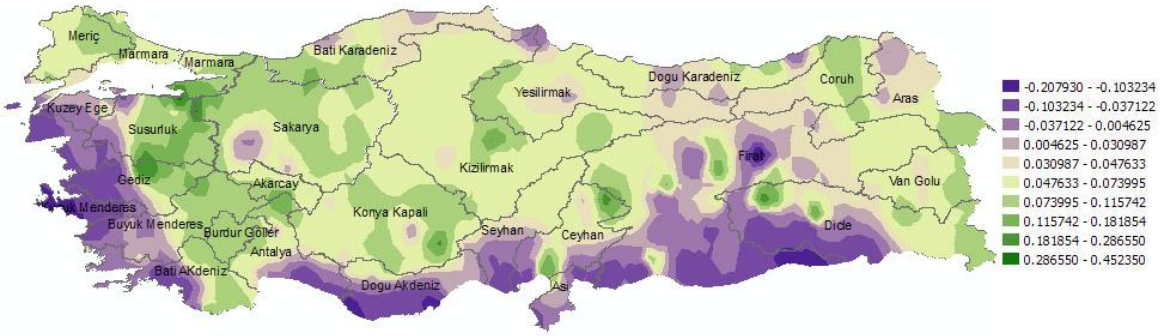


Figure 3.10: Difference significance map for the effects of hot summers on dry summers

CHAPTER 4

COMPARISON BETWEEN DATA CUBE AND APRIORI APPROACH

In this chapter, for the comparison of data cube and Apriori algorithms, metrics that express the quality of the discovered association rules are defined. Example association rules discovered using data mining application for Meteorological domain developed within the scope of this thesis are also presented and analyzed according to the Meteorology domain aspects. A questionnaire (provided in Appendix A) is prepared for the evaluation of association rules by the domain experts from GGIT department in METU. The questionnaire contains a set of fuzzy association rules which are mined by Data cube and Apriori algorithms using the same relation specifications. The rule set includes thirty rules for each technique. A part of the questionnaire is presented in Figure 4.1. Two domain experts have provided their comments. The results of the questionnaire are used to classify the association rule mining techniques in this chapter.

4.1 Metrics for Comparing Association Rule Mining Algorithms

In this thesis, data mining is performed using two different methods: association rule mining through data cubes and association rule mining through Apriori algorithm, which represents classical data mining methods. Using different methods provides the opportunity to make a comparison between them. The following metrics are used to compare different association rule mining approaches:

- Interpretability: This is the complexity of the association rules discovered and the size of the decision tree constructed.

	SIMPLICITY clear 10-8 fully understandable 8-5 not very understandable 5-2 meaningless 2-0	CERTAINTY& UTILITY exact 10-8 precise 8-5 rough 5-2 imprecise 2-0	NOVELTY surprise 10-7 novel 7-4 ordinary 4-0	DIRECT-TO-THE-POINT Very connected 10-8 related 8-5 less-related 5-2 out-of-scope 2-0
Apriori				
A1. From 1970 to 2007 in Buyuk_Menderes if fall is dry and winter is dry then summer is mild [2%, 2%]	9	7	3	8
A2. From 1970 to 2007 in Buyuk_Menderes if fall is dry and winter is dry then summer is hot [88%, 97%]	9	8	3	8
A3. From 1970 to 2007 in Buyuk_Menderes if fall is dry and winter is fair then summer is hot [0%, 100%]	8	7	3	8
A4. From 1970 to 2007 in Buyuk_Menderes if fall is fair and winter is dry then summer is hot [0%, 100%]	8	7	3	8
Data cube				
D1. From 1970 to 2007 in Buyuk_Menderes if fall is 63% dry and winter is 68% dry then summer is 54% mild [0%, 1%]	6	5	4	5
D2. From 1970 to 2007 in Buyuk_Menderes if fall is 66% dry and winter is 83% dry then summer is 86% hot [41%, 81%]	8	7	4	8
D3. From 1970 to 2007 in Buyuk_Menderes if fall is 66% dry and winter is 54% fair then summer is 90% hot [0%, 89%]	7	6	7	6
D4. From 1970 to 2007 in Buyuk_Menderes if fall is 69% fair and winter is 75% dry then summer is 94% hot [0%, 94%]	8	6	7	7
Apriori				
A5. From 1970 to 2007 in Buyuk_Menderes if winter is dry then summer is mild [2%, 2%]	8	6	6	7
A6. From 1970 to 2007 in Buyuk_Menderes if winter is dry then summer is hot [96%, 97%]	9	8	3	8
A7. From 1970 to 2007 in Buyuk_Menderes if winter is fair then summer is hot [0%, 100%]	5	6	4	7
Data cube				
D5. From 1970 to 2007 in Buyuk_Menderes if winter is 66% dry then summer is 54% mild [0%, 1%]	8	5	4	7
D6. From 1970 to 2007 in Buyuk_Menderes if winter is 81% dry then summer is 84% hot [65%, 82%]	8	7	3	8

Figure 4.1: Metric comparison sheet

- Precision: This metric studies the quality of the association rules which is represented in terms of confidence.
- Utility: Support error extend is another measure to compare association rule mining approaches.
- Novelty: This is the ability to discover non-trivial, implicit, previously unknown and potentially useful association rules.
- Direct-to-the-point: This is the ability to discover related association rules. Trivial rules and the rules out of the scope of the user specifications degrade this property.
- Performance: Working on the analysis of voluminous data brings the necessity for the computational power.
- Visualization: This metric studies which association rule mining technique discovers rules that can be visualized better. More detailed maps enhance this property.

In the following paragraphs each of the items above are discussed in detail.

'Interpretability' is a good measure to compare the understandability of association rules. Since the association rules are consumed by human, they have to be user friendly. Both data cube and Apriori algorithm creates fuzzy association rules. Fuzziness handles the numerical data better by softening the sharp boundaries of data which makes it more understandable for human. Data cube approach utilizes membership values in order to discover more precise rules, whereas Apriori algorithm considers only the fuzzy labels.

With the processing of membership values, the association rules mined through data cube mechanism includes many percentage values (%) which makes it harder to understand.

The following rule is discovered using data cube. As it is seen percentage symbols (%) makes it hard to read and analyze the meaning of the rule. In this thesis, this is overcome with the visualization of the association rules on Turkey's map according to their significance and certainty. As a result, the rule is represented in a more convenient way for a meteorologist.

From 1970 to 2007 in Gediz if summer is 87% hot then summer is 54% dry [36%, 43%]

The same rule found through Apriori algorithm is represented as follows. Although the membership values are needed to precisely model the crisp data in fuzzy domain, their presence degrades the readability of the association rule. In this sense, Apriori algorithm produces more understandable rules.

From 1970 to 2007 in Gediz if summer is hot then summer is dry [76%, 78%]

For 'interpretability' metric, the following classifications are proposed within the scope of these this: clear, fully understandable, not very understandable, meaningless. Data cube and Apriori algorithms can be classified as presented in Table 4.1 according to 'interpretability' metric when the above association rules are considered.

Table 4.1: Comparison according to Interpretability

Class/Technique	**Data cube**	**Apriori**
clear		
fully understandable		X
not very understandable	X	
meaningless		

For the 'precision' of association rules, the support and confidence calculation methods can be inspected for data cube approach and Apriori algorithm. Apriori approach treats fuzzy labels as if they are transactional items. It does not take into account the membership values which represents to what extend a fuzzy label is valid for a record. As a result,

the discovered association rules do not reflect the correct nature of the data. Data cube approach evaluates fuzzy labels with their membership values. The resulting association rules mined through data cube approach are more precise. The quality of the association rules can be enhanced with the design of the fuzzy sets and membership functions together with a domain expert.

Precise, rough, imprecise are specified as the classification terms for the comparison according to 'precision' metrics. The votes for data cube and Apriori techniques are given in Table 4.2. Data cube approach is rated as 'precise', but its success depends on the definition of fuzzy set and membership functions. The better the fuzzy sets and membership values are defined, the more precise association rules are discovered.

Table 4.2: Comparison according to Precision

Class/Technique	**Data cube**	**Apriori**
precise	X	
rough		X
imprecise		

'Utility' metric defines the usefulness of the discovered association rules. This metric is domain dependent and can be composed of other metrics. For example, for a real-time system the application with better performance is more useful. For the case of fuzzy association rule mining, the 'precise' and 'direct-to-the-point' metrics make up of the 'utility' metric.

Useful, less-useful, impractical are specified as the classification terms for the 'utility' metric. Data cube is rated as more useful than the Apriori algorithm. Because, data can be modeled according to the user needs while constructing the data cube. Additionally, data cube approach enables the user to create more specific queries. It presents only the related relationships. Table 4.3 presents the rating of data cube approach and Apriori algorithm according to 'utility' metric.

The association rules reveals the hidden relationships in a large database. The 'novelty' of an association rule mining method is evaluated by its ability to discover non-trivial, implicit, previously unknown and potentially useful information. In data cube approach, the relationships between a measure and selected dimensions are searched. The discovered

Table 4.3: Comparison according to Utility

Class/Technique	**Data cube**	**Apriori**
useful	X	
less-useful		X
impractical		

rules summarizes only the effects of the dimensions on the measure. In Apriori approach, the measure is not used. All possible interesting and frequent relationships among the selected dimensions are discovered. Therefore, the possibility of Apriori algorithm to find more novel association rules is higher.

Although all data mining techniques seek for interesting, non-trivial, implicit, previously unknown and potentially useful relationships, some techniques have more chance to do so. Surprise, novel, ordinary are specified as the classifications for the 'novelty' metric. Table 4.4 presents the comparison of data cube and Apriori algorithm according to 'novelty' criteria.

Table 4.4: Comparison according to Novelty

Class/Technique	**Data cube**	**Apriori**
surprise		X
novel	X	
ordinary		

'Direct-to-the-point' metric seems to be the reverse of 'novelty' metric. But it is not. It also means the ability to discover association rules within the scope of the user needs. In order to query for associations using data cube approach, user has to specify the dimensions and a measure. The relations discovered represent the effects of the dimensions on the measure attribute. In Apriori algorithm, it is not possible to specify a measure. The discovered rules include all possible relations between the selected dimensions. As a result, the user is provided with irrelevant association rules.

Very connected, related, less-related and out-of-scope are the classifications defined for 'direct-to-the-point' metric. Data cube and Apriori algorithms are classified according to

this metric as presented in Table 4.5.

Table 4.5: Comparison according to Direct-to-the-point

Class/Technique	**Data cube**	**Apriori**
very connected	X	
related		
less-related		X
out-of-the-scope		

In order to compare data cube approach and Apriori algorithm in terms of performance issues, how long it takes each method to find association rules of different complexities can be inspected. Table 4.6 presents such a comparison. The time to find association rules for different number of dimensions is calculated. When the number of dimensions is increased, the complexity of the relationships increases since the relationships among more parameters are searched. When Table 4.6 is analyzed, it is obvious that the performance of data cube approach is better than the Apriori algorithm. The performance test is performed with a data cube whose fact table includes 718,164 records. Actually, the Apriori algorithm is used to discover interesting relationships among large databases for once that is all the relationships are mined and then the results are utilized for further analysis. On the other hand, data cubes provide an efficient query tool for the analysts. The construction of the cube is performed once where several queries for inspecting different aspects of the data are created later on. The power of data cubes comes with the group by expressions. The time to discover association rules decreases with the increasing number of aggregated group by tables prepared in the data cube. The performance of the data cube construction is not compared since it is only performed by the data cube approach.

The complexity of the algorithms for association rule mining using data cube and Apriori algorithm also supports the performance table in Table 4.6. The complexity of Algorithm 5 which presents association rule mining using data cubes is O(n ($\Pi\ f_i$)), where n is the number of records in the cuboid and ($\Pi\ f_i$) is the product of the number of elements in each fuzzy set. O(nm2^m) is the complexity for Apriori algorithm presented in Algorithm 1. n is the number of records in the fact table and m is the sum of the number of elements in the fuzzy sets. The relation between m and f_i is as follows: $\Sigma\ f_i = \text{m}$.

Table 4.6: Time to mine association rules

# of dimensions	Data cube(msec)	Apriori(msec)
2	47	344
3	94	3234
4	235	76430
5	1297	3212686
6	3562	...
7	10312	...
8	30592	...
9	242269	...

'Visualization' metric studies which association rule mining technique discovers rules that can be visualized as a map better. This metric is related with 'precision' metric. Because, the more precise support and confidence calculation is performed the better the map is visualized. Figure 4.2 and Figure 4.3 present example maps for support values of an association rule. As it is seen the transition regions in significance map generated by data cube approach is more detailed than the support map generated by the Apriori algorithm for the same rule. When an association rule is visualized on Turkey's map for meteorological analysis, it is seen that the pattern of the discovered rule on the map are the same for both data cube and Apriori. However, the created maps are more detailed for data cube approach as a result of the use of membership values. The colored regions are smoothly evolving for data cube approach where they are roughener for the Apriori algorithm.

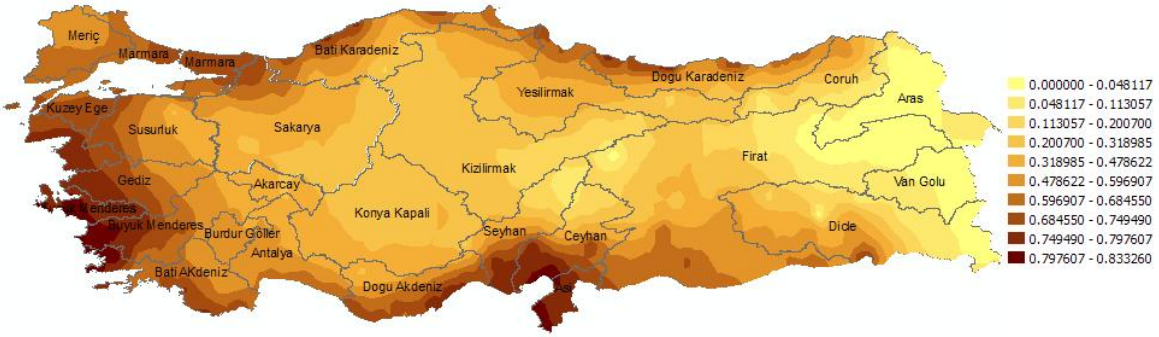


Figure 4.2: Significance map by Data Cube 'detailed' - From 1970 to 2007 in Firat if winter is 75% mild then winter is 78% dry [30%, 78%]

Figure 4.3: Support map by Apriori 'less-detailed' - From 1970 to 2007 in Firat if winter is 75% mild then winter is 78% dry [30%, 78%]

For 'visualization' metric, the following classifications are proposed within the scope of these this: Detailed, less-detailed, rough. Data cube and Apriori algorithms are classified as presented in Table 4.7.

Table 4.7: Comparison according to Visualization

Class/Technique	**Data cube**	**Apriori**
detailed	X	
less-detailed		X
rough		

4.2 Analyzing Example of Meteorological Association Rules

A data mining application for the analysis of Turkey's meteorological data set accumulated since 1970 is developed in this thesis. This data mining application can be extended for other domains. In order to present the reliability of this work, a comparison between the previously known results and the association rules discovered using data mining application shall be performed. If the known results can be proved, the confidence for the previously unknown relationships will be high.

In the comparison the significance maps are used. Because, significance map presents where an association rule is valid on the map. The colors shows the strength of validity,

that is the association rule is supported more frequently in the dark colored areas and it is rare for lighter colored areas. On the other hand, certainty map presents the expectancy of an association rule among the measurements in the stations satisfying the first part of the rule. Since the meteorological findings presented in this section are the known results instead of being predictions, significance map is a good method to make a comparison.

Rule1: In Meteorology domain, it is expected that in Gediz, Kucuk Menderes and Buyuk Menderes basins, the increasing summer temperature results in decreasing summer precipitation.

The following rule is mined using the data mining application:

From 1970 to 2007 in Buyuk_Menderes if summer is 85% hot then summer is 54% dry [34%, 42%]

The rule tells that when the data since 1970 is analyzed, it can be said that summers have been hot and it has resulted in dry summers. In order to prove the meteorological finding, the general picture for the Turkey shall be investigated. The significance of each station to that rule is found and they are displayed on a Turkey map. The significance map of Rule1 is presented in Figure A.2. It is seen that, on the significance map, Gediz, Kucuk Menderes and Buyuk Menderes basins support this rule to a great extend since they are darker than the other regions. In addition to that in Asi, Meric and Marmara basins the increasing summer temperature also results in decreasing summer precipitation.

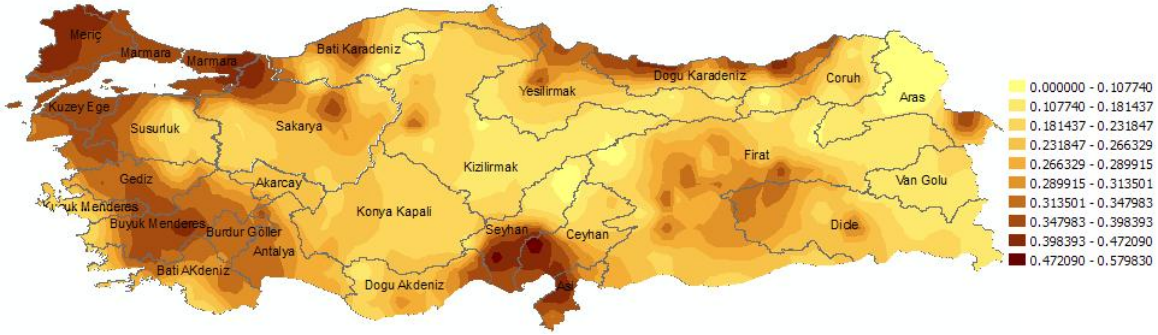


Figure 4.4: Significance map of Rule1 - From 1970 to 2007 in Buyuk_Menderes if summer is 85% hot then summer is 54% dry [34%, 42%]

Rule2: In Meteorology domain, it is expected that in the basins to the west of Kizilirmak

basin, the increasing fall temperature has no effect on the decrease in fall precipitation.

The following rule is mined using the data mining application:

From 1970 to 2007 in Buyuk_Menderes if fall is 83% mild then fall is 75% dry [42%, 70%]

The significance map of Rule2 is presented in Figure 4.5. It is seen that, on the significance map, in the basins to the east of Kizilirmak, increasing fall temperature causes decrease in fall precipitation. In line with the meteorological rule, in the basins to the west of Kizilirmak, the increasing fall temperature is not very much related with the decreasing fall precipitation even if Akarcay and Susurluk can be thought as an exception.

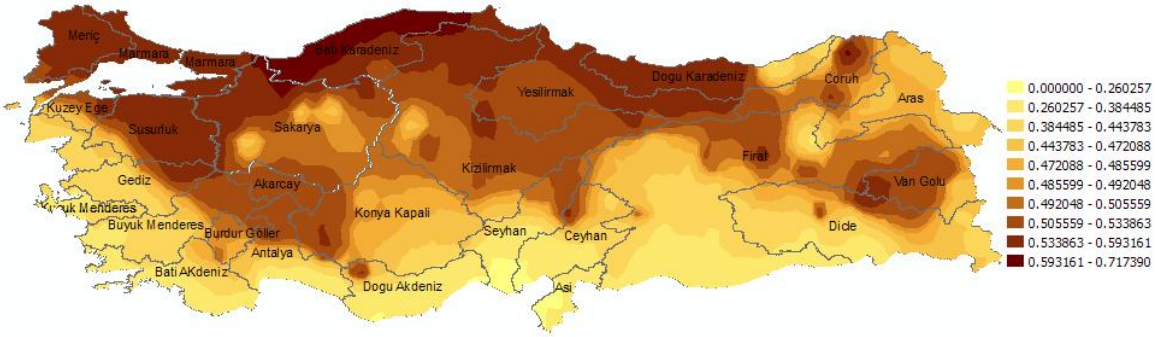


Figure 4.5: Significance map of Rule2 - From 1970 to 2007 in Buyuk_Menderes if fall is 83% mild then fall is 75% dry [42%, 70%]

Rule3: In Meteorology domain, it is expected that in the water dam regions of Firat basin, the increasing winter temperature results in increasing winter precipitation.

The following rule is mined using the data mining application:

From 1970 to 2007 in Firat if winter is 75% mild then winter is 78% dry [30%, 78%]

The significance map of Rule3 is presented in Figure 4.6. It is seen that, on the significance map, in Firat basin the relation of mild winters and dry winters is not supported. However, it cannot be said that in Firat basin mild winters causes fair or dry winters. This meteorological rule cannot be discovered precisely through data mining application.

The same rules are also discovered using Apriori algorithm through data mining application. The support maps corresponding to the Apriori algorithm are presented in Figure A.1, Figure 4.8 and Figure 4.9. As it is seen support maps are also inline with the

Figure 4.6: Significance map of Rule3 - From 1970 to 2007 in Firat if winter is 75% mild then winter is 78% dry [30%, 78%]

significance maps. However, they are not as precise as the significance maps since Apriori algorithm discards the membership values in the discovery of fuzzy association rules. As a result, the transition regions are rougher.

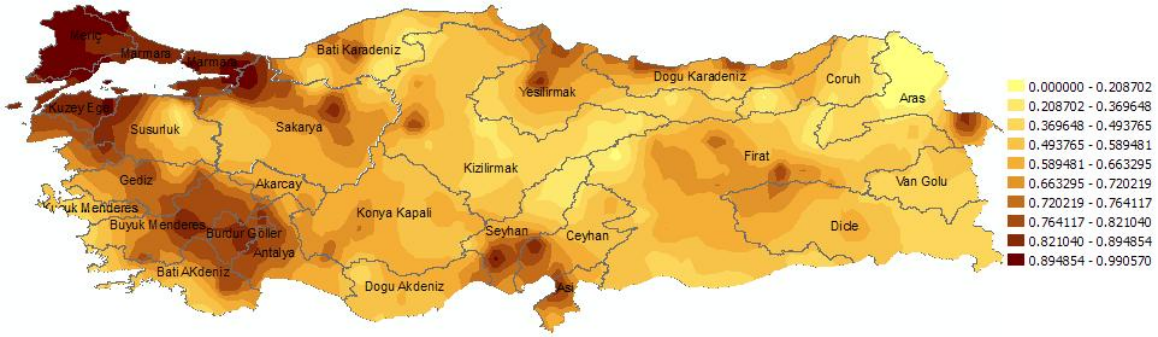


Figure 4.7: Support map of Rule1 - Significance map of Rule1 - From 1970 to 2007 in Buyuk_Menderes if summer is 85% hot then summer is 54% dry [34%, 42%]

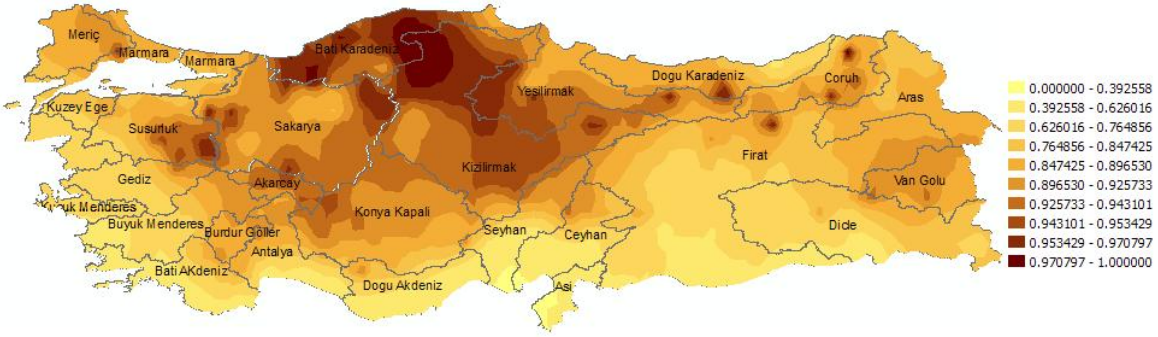


Figure 4.8: Support map of Rule2 - From 1970 to 2007 in Buyuk_Menderes if fall is 83% mild then fall is 75% dry [42%, 70%]

Figure 4.9: Support map of Rule3 - From 1970 to 2007 in Firat if winter is 75% mild then winter is 78% dry [30%, 78%]

CHAPTER 5

IMPLEMENTATION

Chapter 3 details the methodology and algorithms for the construction of a fuzzy spatio-temporal data cube and mining association rules on the constructed data cube within the scope of this thesis. The use of Apriori algorithm for mining fuzzy association rules and visualization technique developed is also detailed. This Chapter presents the implementational choices for the realization of the concepts given Chapter 3.

5.1 Implementation Details

In this thesis, a data miner application based on data cubes is developed and it is capabable of:

- Constructing a fuzzy spatio-temporal data cube
- Mining association rules using data cube
- Mining association rules using Apriori algorithm
- Visualizing association rule mapping within GIS

The system architecture for the whole application is given in Figure 5.1.

Java 1.5.0_02−b09 is used through Eclipse 3.2.1 development environment. The data cube is stored in Microsoft Access 2002 database. Database connection is established through MicroSoft Access Driver for JDBC access. The following external libraries are utilized:

- jdbf.jar
- jxl.jar
- arcGIS.jar [36]

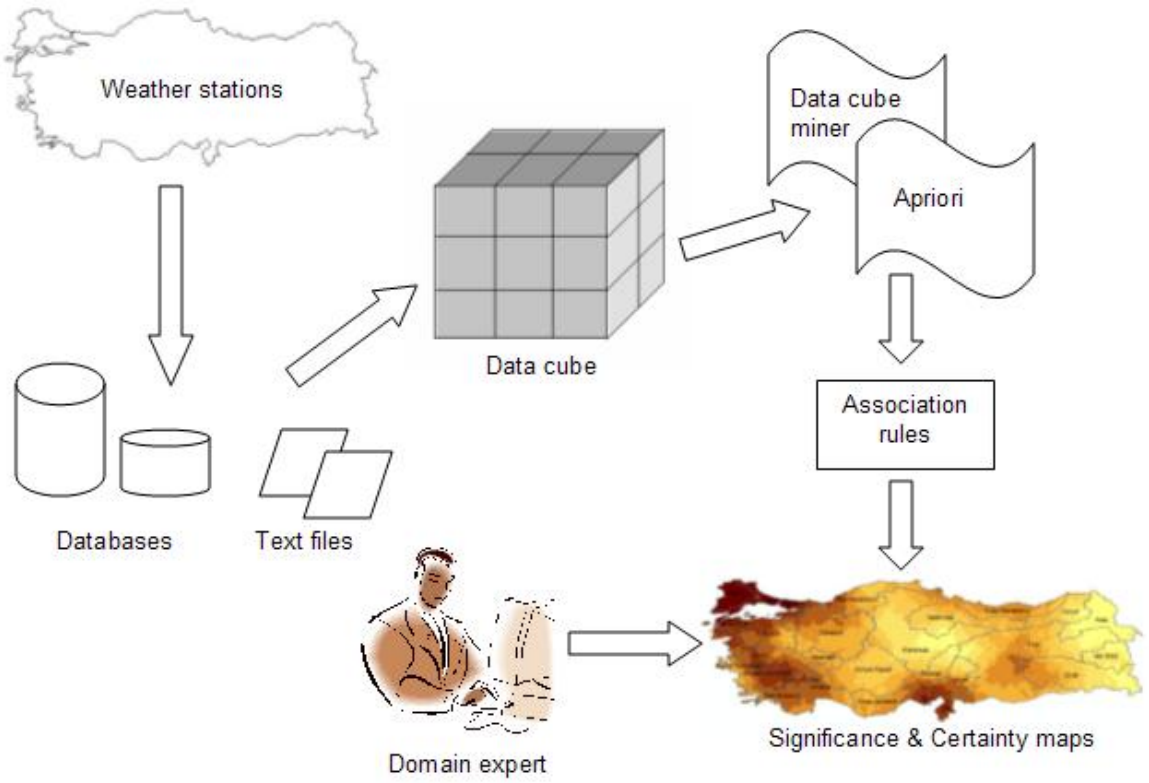


Figure 5.1: System architecture

The main components of the data miner application is presented in Figure 5.2. GUI component is composed of Java Swing panels. It is the bridge between the user and the data processor component. It also accesses the database to display the data necessary to construct a data cube. It guides the user for the construction of a fuzzy spatio-temporal data cube and can be used as a query screen for mining association rules from the constructed data cube. Chapter 6 presents the visual display of the data miner application.

Data processor component is the heart of the application. It is composed of the following units:

- **DataGeneralizer**: This unit is responsible for fuzzifying the crisp data. The fuzzy sets are defined and stored by the help of this component. The fuzzy logic is applied to all specified crisp data fields and the generalized data is stored in the database.
- **DataCubeGenerator**: This unit is responsible for temporal and spatial aggregation. It constructs the fact table and sub-cuboids. The fact table and the group by cuboids are created by the DataCubeGenerator. This component works for once and performs a time consuming operation. After it has constructed the fuzzy spatio-temporal data

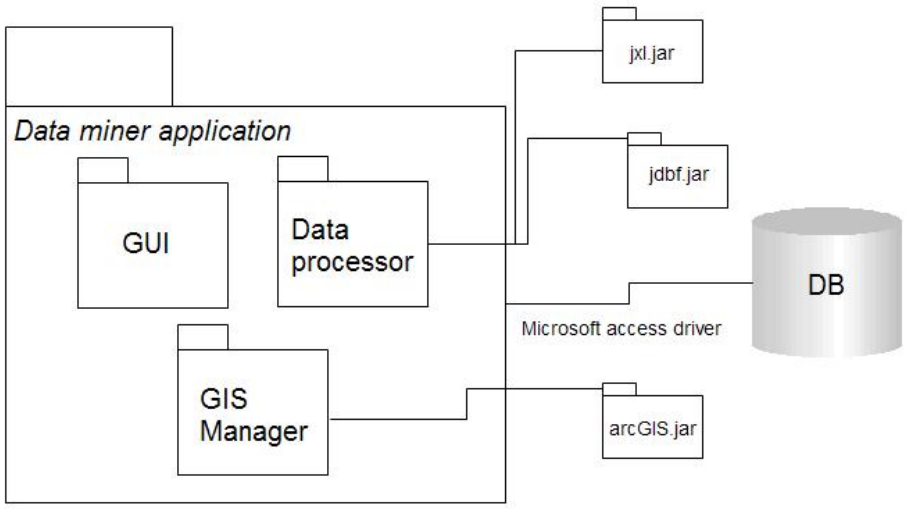


Figure 5.2: Architecture of Data Miner

cube, DataMiner, MyAppriori and DBFExporter units works on that data cube.

- **DataMiner**: This unit performs the discovery of association rules using the data cube. The association rules discovered by this unit reflect the membership values coming from the fuzzy logic.

- **MyApriori**: Apriori algorithm is implemented in this unit. It finds association rules in Apriori way. The fuzzy elements in the fact table are processed as if they are items in a transactional database. The discovered association rules reflect the frequency of the common relationships.

- **DBFExporter**: This unit generates .dbf files which are used as an input for the GIS operations in order to visualize the association rules. It utilizes jxl.jar and jdbf.jar libraries. The group by cuboids are used to compute the support and significance of an association rule for each station.

GIS manager component can be thought as a separate application. It uses a GIS engine in order to process the .dbf files. .dbf files include the data necessary for the creation of significance and certainty maps. GIS manager maps .dbf files within GIS. The operations managed by GIS manager can be performed by a user through a GIS tool such as ArcMap [36]. The following GIS operations are applied to the .dbf file to create significance and certainty maps:

- The .dbf file is imported as a layer.

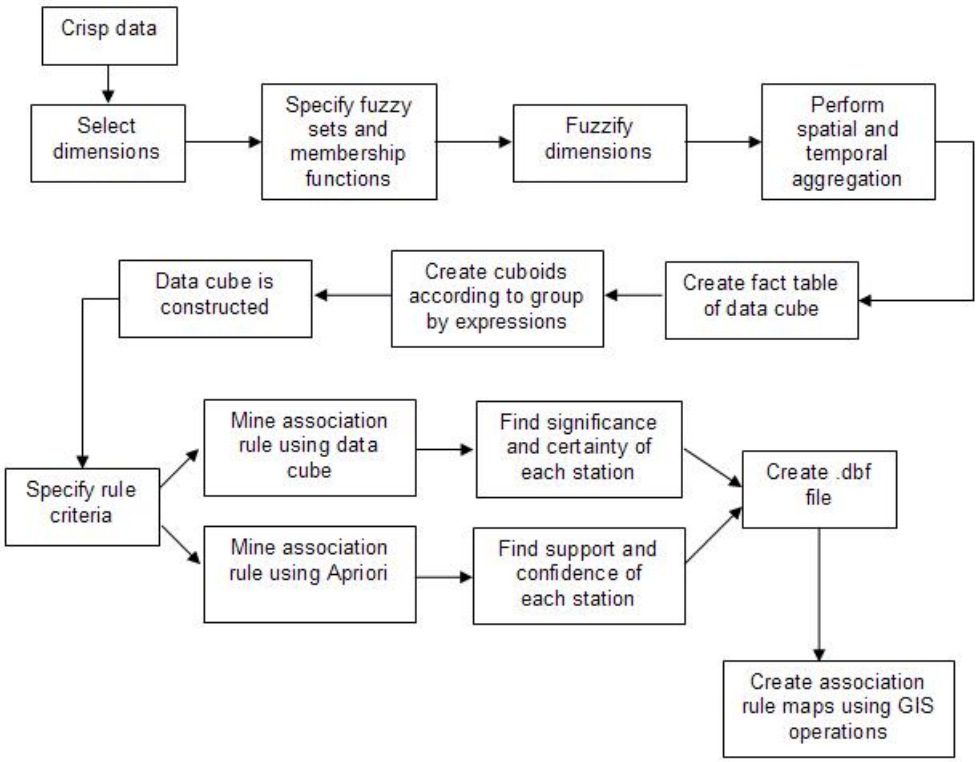


Figure 5.3: Activity diagram for Data Miner

- Interpolation is performed on the significance and certainty data computed for each station using IDW (Inverse Distance Weighting).
- The map is cropped according to Turkey border.

The activity diagram that summarizes the steps of the execution procedures of this implementation is given in Figure 5.3.

5.2 Database Design

Database component stores the data cube. The database design is given in Figure 5.4.

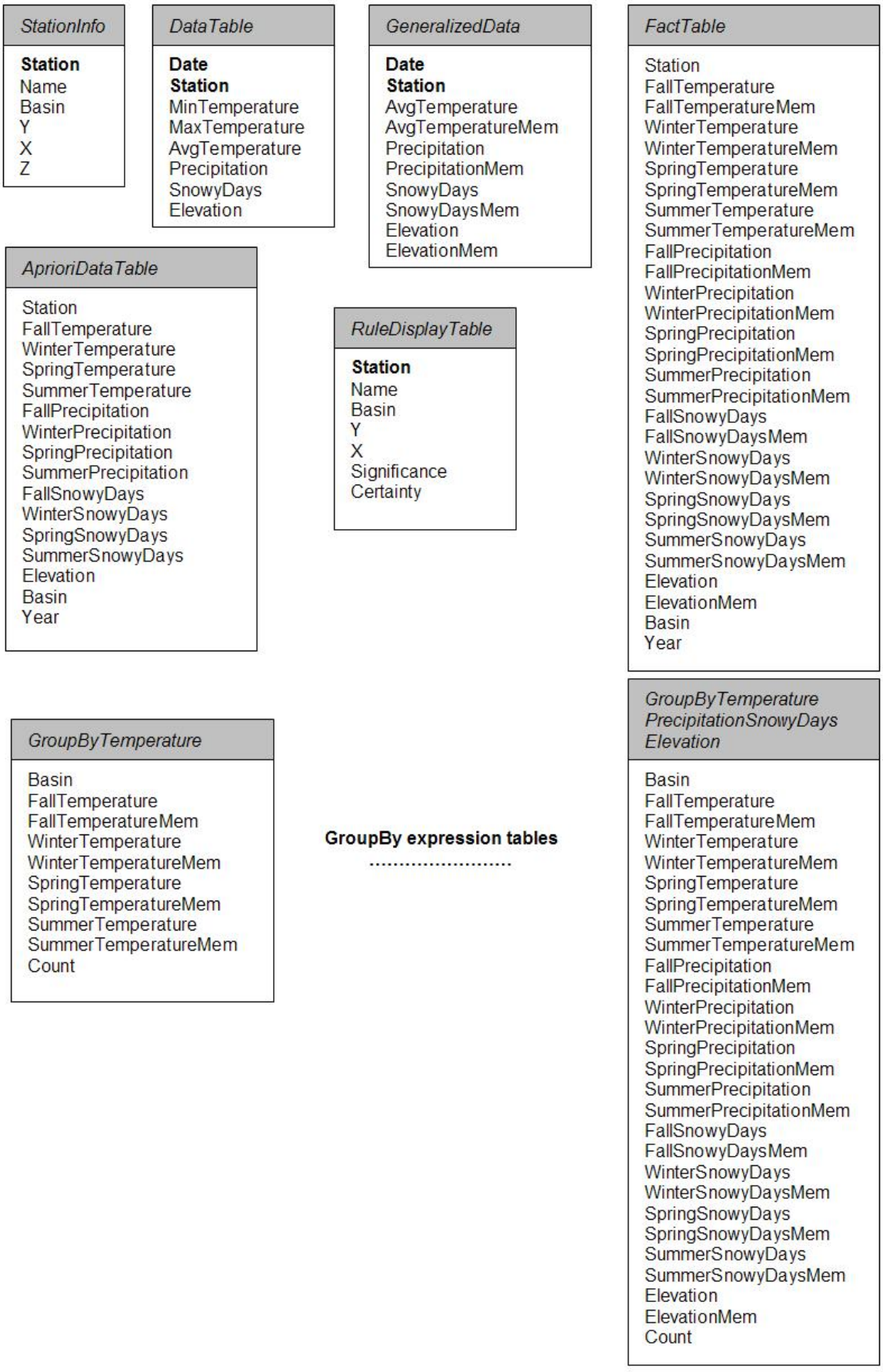


Figure 5.4: Database design

CHAPTER 6

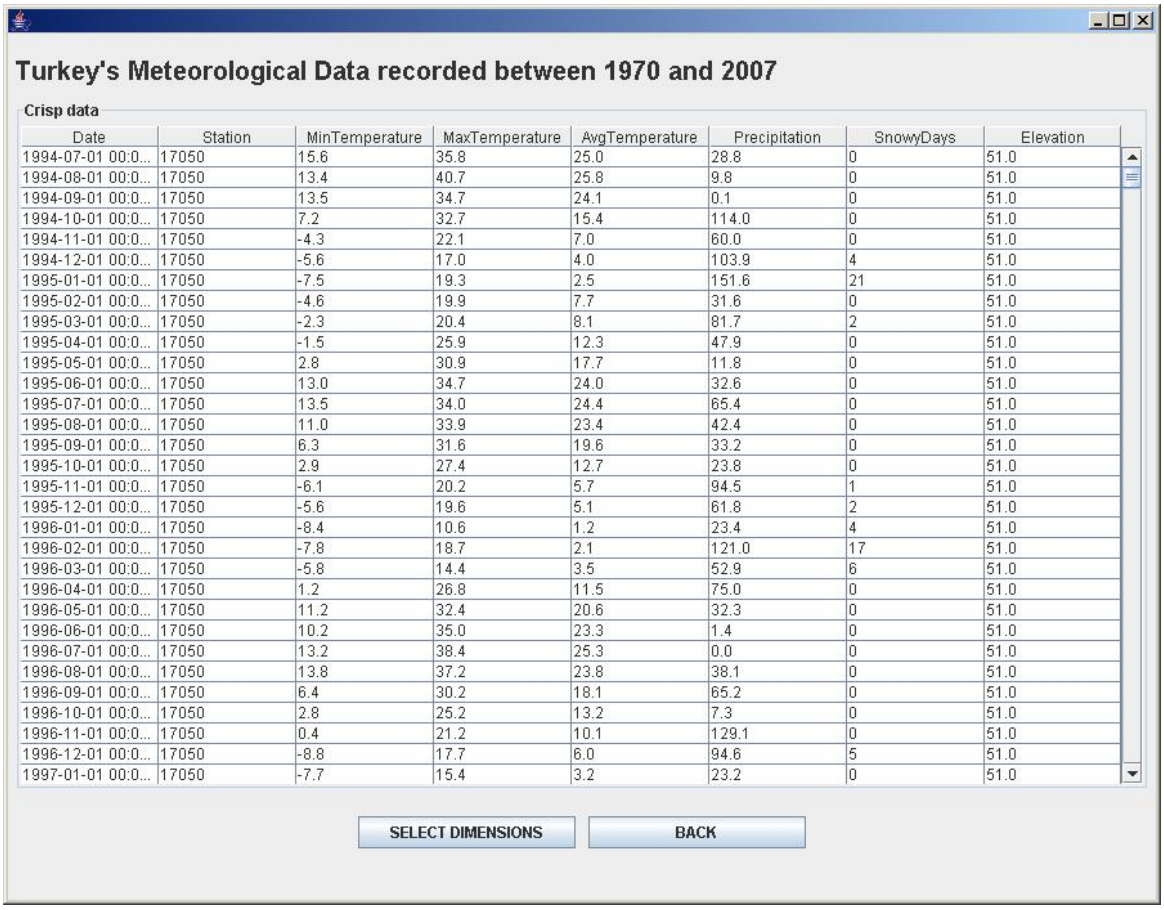
DATA MINING APPLICATION FOR METEOROLOGICAL DATA OF TURKEY

Meteorology is one of the domains which includes increasing data accumulation. There are sensors all around the world collecting weather information. The main purpose is not only to inform people about the daily weather reports but also to figure out the interesting climate relations. Meteorological data measured by the sensors in the weather stations has to be evaluated according to the GIS characteristics. The traditional GIS applications are working with spatial data separately from the business data. This loosely coupled approach has a data integrity drawback since these two data types are managed apart. There are several methods that can be used to mine interesting relationships among spatio-temporal data.

In this thesis, data cube approach and Apriori algorithm are the techniques used for association rule mining. A case study has been implemented to demonstrate the fuzzy association rule mining using spatio-temporal data cubes and Apriori algorithm. Analysis of Turkey's real meteorological data is performed. This chapter illustrates step by step how this analysis is done through the application developed in the light of the concepts of this thesis.

The case study is composed of two parts: (1) Construction of a new data cube, (2) Mining association rules from the data cube. The construction of the data cube is performed once and association rules are mined from the constructed cube for several times. Figure 6.1 presents the start page of the application.

Figure 6.1: Welcome screen



Date	Station	MinTemperature	MaxTemperature	AvgTemperature	Precipitation	SnowyDays	Elevation
1994-07-01 00:0...	17050	15.6	35.8	25.0	28.8	0	51.0
1994-08-01 00:0...	17050	13.4	40.7	25.8	9.8	0	51.0
1994-09-01 00:0...	17050	13.5	34.7	24.1	0.1	0	51.0
1994-10-01 00:0...	17050	7.2	32.7	15.4	114.0	0	51.0
1994-11-01 00:0...	17050	-4.3	22.1	7.0	60.0	0	51.0
1994-12-01 00:0...	17050	-5.6	17.0	4.0	103.9	4	51.0
1995-01-01 00:0...	17050	-7.5	19.3	2.5	151.6	21	51.0
1995-02-01 00:0...	17050	-4.6	19.9	7.7	31.6	0	51.0
1995-03-01 00:0...	17050	-2.3	20.4	8.1	81.7	2	51.0
1995-04-01 00:0...	17050	-1.5	25.9	12.3	47.9	0	51.0
1995-05-01 00:0...	17050	2.8	30.9	17.7	11.8	0	51.0
1995-06-01 00:0...	17050	13.0	34.7	24.0	32.6	0	51.0
1995-07-01 00:0...	17050	13.5	34.0	24.4	65.4	0	51.0
1995-08-01 00:0...	17050	11.0	33.9	23.4	42.4	0	51.0
1995-09-01 00:0...	17050	6.3	31.6	19.6	33.2	0	51.0
1995-10-01 00:0...	17050	2.9	27.4	12.7	23.8	0	51.0
1995-11-01 00:0...	17050	-6.1	20.2	5.7	94.5	1	51.0
1995-12-01 00:0...	17050	-5.6	19.6	5.1	61.8	2	51.0
1996-01-01 00:0...	17050	-8.4	10.6	1.2	23.4	4	51.0
1996-02-01 00:0...	17050	-7.8	18.7	2.1	121.0	17	51.0
1996-03-01 00:0...	17050	-5.8	14.4	3.5	52.9	6	51.0
1996-04-01 00:0...	17050	1.2	26.8	11.5	75.0	0	51.0
1996-05-01 00:0...	17050	11.2	32.4	20.6	32.3	0	51.0
1996-06-01 00:0...	17050	10.2	35.0	23.3	1.4	0	51.0
1996-07-01 00:0...	17050	13.2	38.4	25.3	0.0	0	51.0
1996-08-01 00:0...	17050	13.8	37.2	23.8	38.1	0	51.0
1996-09-01 00:0...	17050	6.4	30.2	18.1	65.2	0	51.0
1996-10-01 00:0...	17050	2.8	25.2	13.2	7.3	0	51.0
1996-11-01 00:0...	17050	0.4	21.2	10.1	129.1	0	51.0
1996-12-01 00:0...	17050	-8.8	17.7	6.0	94.6	5	51.0
1997-01-01 00:0...	17050	-7.7	15.4	3.2	23.2	0	51.0

Figure 6.2: Turkey's meteorological data recorded between 1970 & 2007

6.1 Constructing a new Data Cube

When "Construct new cube" button is selected in the start page, the crisp meteorology data is presented to the user as given in Figure 6.2. The crisp data is stored in the "DataTable" in database. The "DataTable" is constructed from the text files (.dat files) provided by the Turkish State Meteorological Service. The "DataTable" is constructed previously, independent of the application.

The user is expected to select the dimensions that will be used in the construction of the data cube. When "Select dimensions" button is clicked, panel in Figure 6.3 is displayed. The attributes of "DataTable" is presented to the user for selection. In this case study, date, station, average temperature, precipitation, snowy days and elevation fields are specified as the dimensions of the data cube. Each dimension is added to the data cube using "Add to cube" button.

When "Add to cube" is pressed for date field, the dialog in Figure 6.4 is shown. In

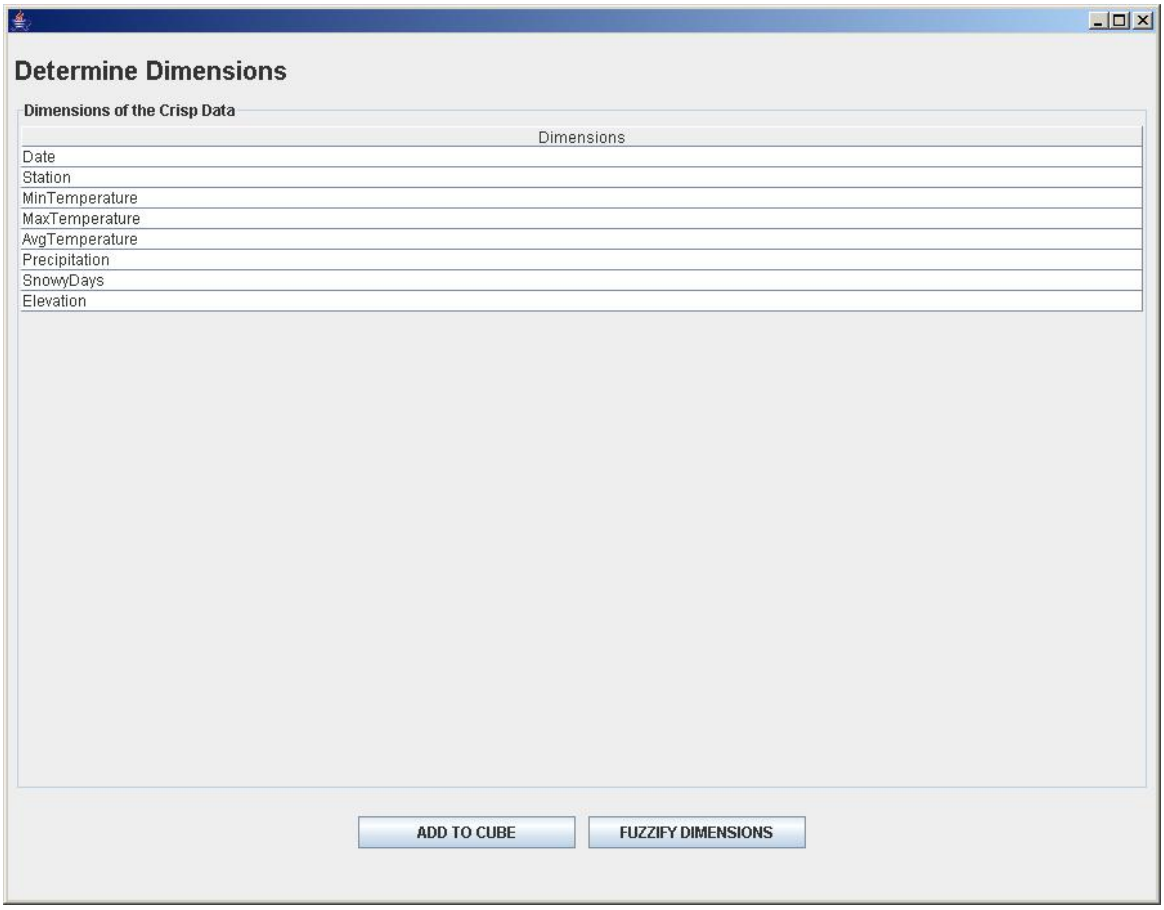


Figure 6.3: Determine dimensions

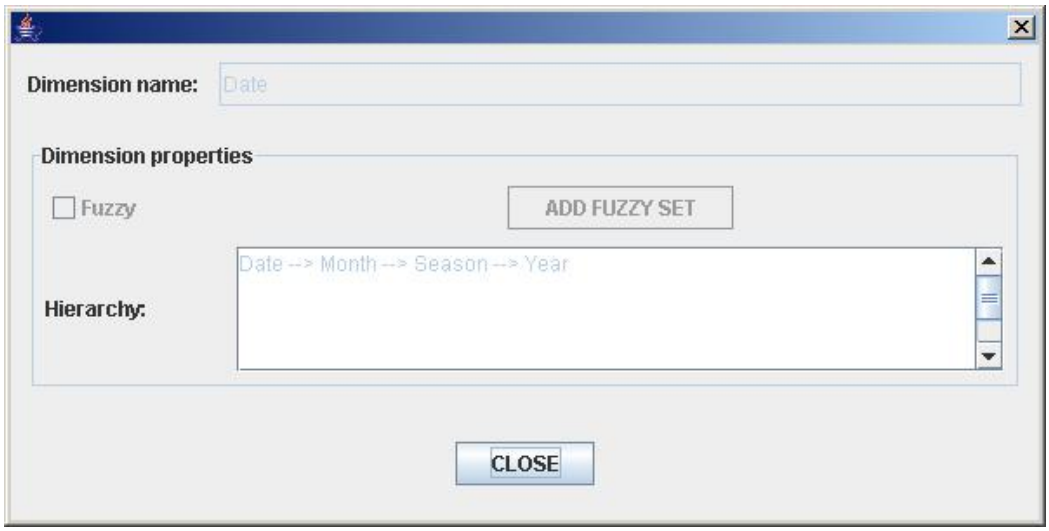


Figure 6.4: Add date dimension

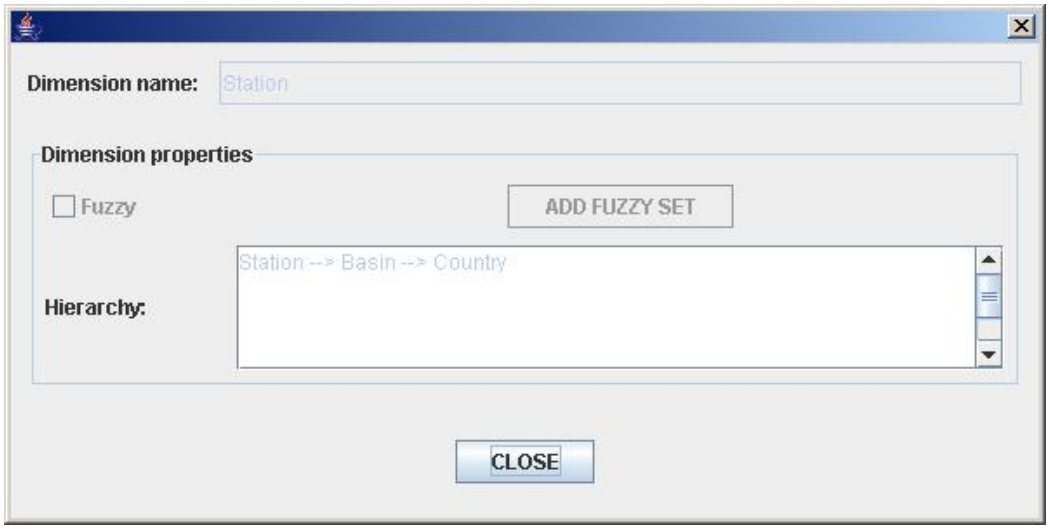


Figure 6.5: Add station dimension

this dialog, the hierarchy for the date field is presented. This field cannot be fuzzified and definition of the fuzzy set is disabled for date field.

The dialog in Figure 6.5 is displayed, when "Add to cube" is clicked for station field. In this dialog, the hierarchy for the station field is presented. This field cannot be fuzzified. So definition of the fuzzy set is also disabled for station field.

Fuzzy sets can be defined for the average temperature field using "Add fuzzy set" button in Figure 6.6, which is displayed when "Add to cube" is clicked for average temperature. "Add fuzzy set" button directs the user to the dialog in Figure 6.7. In this dialog, the fuzzy label and membership function and its parameters are specified. The other elements of the fuzzy set is defined in the same way. When definition of the fuzzy set is completed "Close" button is pressed in Figure 6.8. The status of the fuzzy set defined so far is also presented on this dialog. The fuzzy sets for precipitation, snowy days and elevation are defined similarly. The fuzzy set definitions are provided in Appendix B.

When the determination of the dimensions and the fuzzy sets are completed, "Fuzzify dimensions" button is clicked. The crisp data in Figure 6.2 is fuzzified with this action and generalized values in Figure 6.9 are presented to the user.

When "Aggregate fuzzy spatial data" button in Figure 6.9 is clicked, the temporal and spatial aggregation is performed. The fact table and group by expression tables are created. As a result, the construction of the data cube is completed and the user is directed to the association rule miner part of the application presented in Figure 6.10.

Figure 6.6: Add temperature dimension

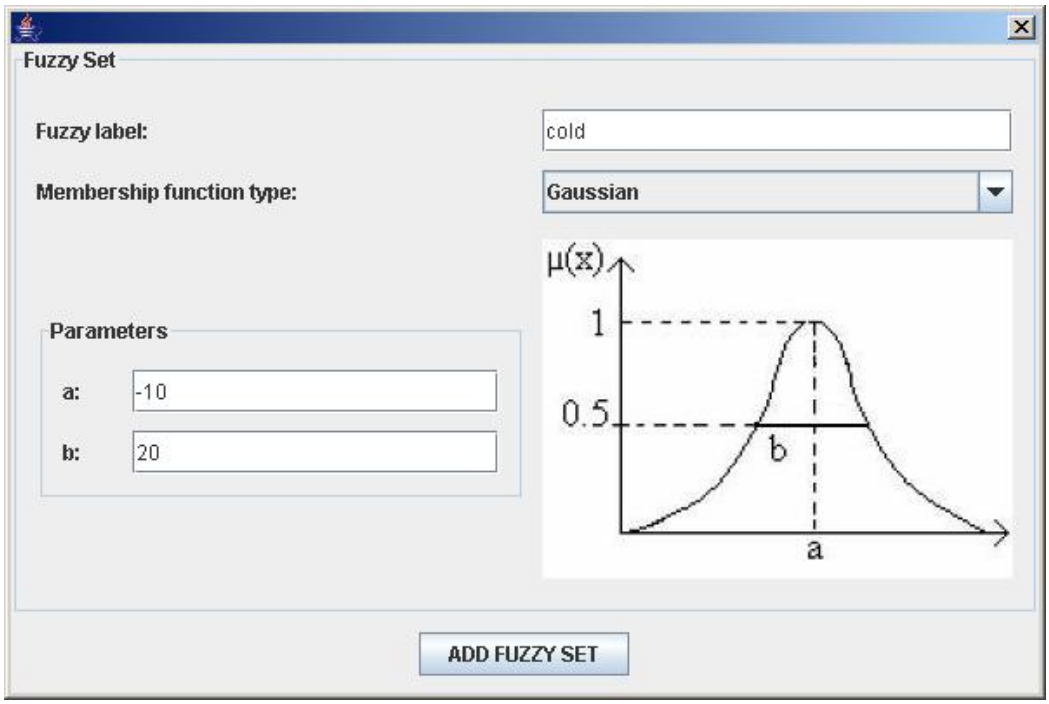


Figure 6.7: Define fuzzy set and membership function

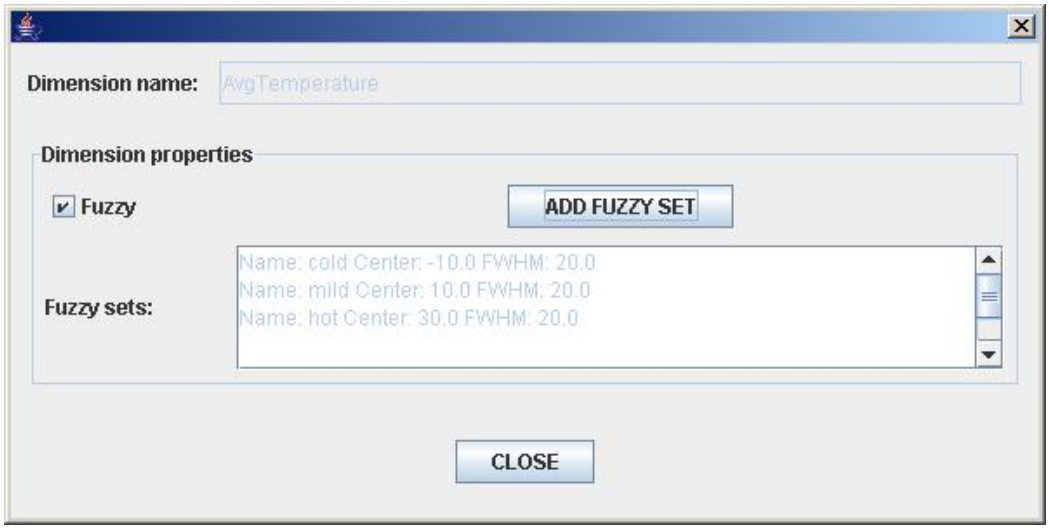
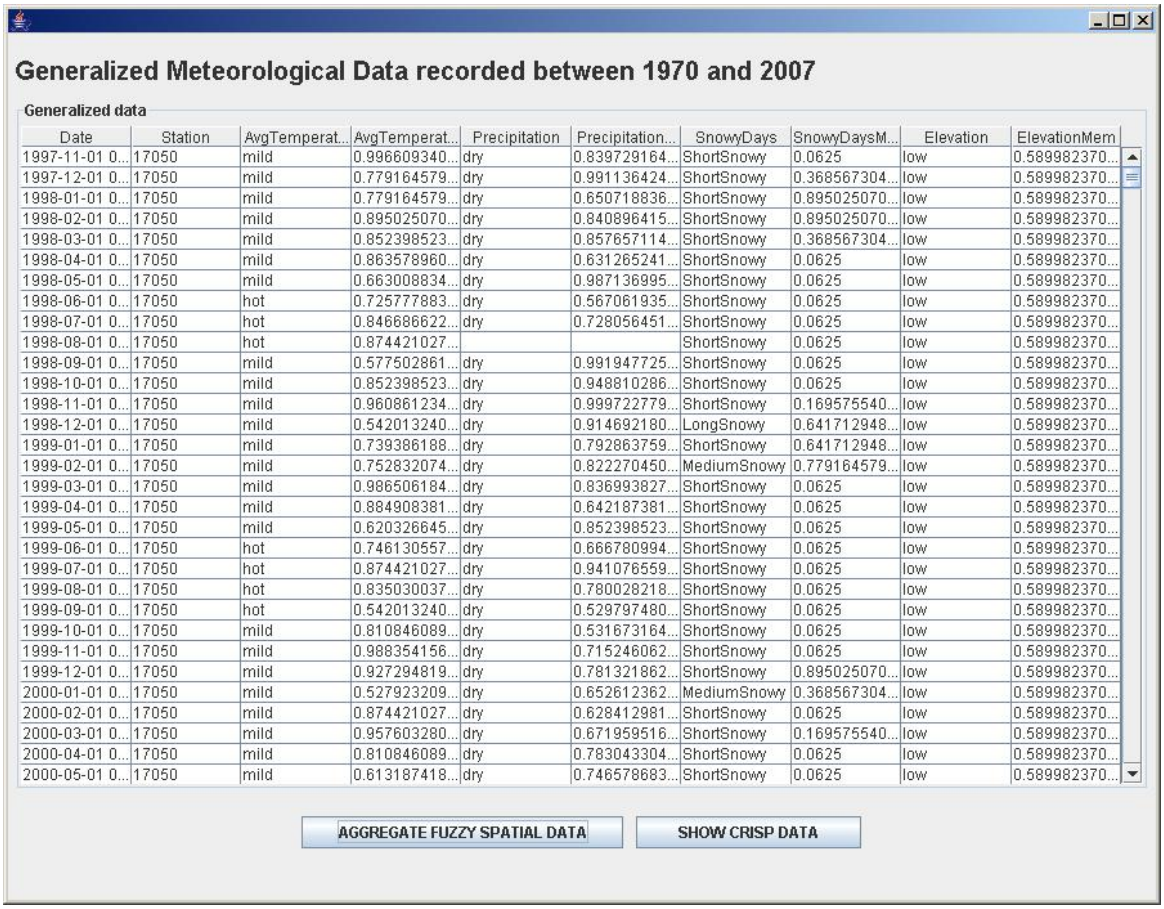


Figure 6.8: Fuzzy set for temperature



Date	Station	AvgTemperat...	AvgTemperat...	Precipitation	Precipitation...	SnowyDays	SnowyDaysM...	Elevation	ElevationMem
1997-11-01 0...	17050	mild	0.996609340...	dry	0.839729164...	ShortSnowy	0.0625	low	0.589982370...
1997-12-01 0...	17050	mild	0.779164579...	dry	0.991136424...	ShortSnowy	0.368567304...	low	0.589982370...
1998-01-01 0...	17050	mild	0.779164579...	dry	0.650718836...	ShortSnowy	0.895025070...	low	0.589982370...
1998-02-01 0...	17050	mild	0.895025070...	dry	0.840896415...	ShortSnowy	0.895025070...	low	0.589982370...
1998-03-01 0...	17050	mild	0.852398523...	dry	0.857657114...	ShortSnowy	0.368567304...	low	0.589982370...
1998-04-01 0...	17050	mild	0.863578960...	dry	0.631265241...	ShortSnowy	0.0625	low	0.589982370...
1998-05-01 0...	17050	mild	0.663008834...	dry	0.987136995...	ShortSnowy	0.0625	low	0.589982370...
1998-06-01 0...	17050	hot	0.725777883...	dry	0.567061935...	ShortSnowy	0.0625	low	0.589982370...
1998-07-01 0...	17050	hot	0.846686622...	dry	0.728056451...	ShortSnowy	0.0625	low	0.589982370...
1998-08-01 0...	17050	hot	0.874421027...			ShortSnowy	0.0625	low	0.589982370...
1998-09-01 0...	17050	mild	0.577502861...	dry	0.991947725...	ShortSnowy	0.0625	low	0.589982370...
1998-10-01 0...	17050	mild	0.852398523...	dry	0.948810286...	ShortSnowy	0.0625	low	0.589982370...
1998-11-01 0...	17050	mild	0.960861234...	dry	0.999722779...	ShortSnowy	0.169575540...	low	0.589982370...
1998-12-01 0...	17050	mild	0.542013240...	dry	0.914692180...	LongSnowy	0.641712948...	low	0.589982370...
1999-01-01 0...	17050	mild	0.739386188...	dry	0.792863759...	ShortSnowy	0.641712948...	low	0.589982370...
1999-02-01 0...	17050	mild	0.752832074...	dry	0.822270450...	MediumSnowy	0.779164579...	low	0.589982370...
1999-03-01 0...	17050	mild	0.986506184...	dry	0.836993827...	ShortSnowy	0.0625	low	0.589982370...
1999-04-01 0...	17050	mild	0.884908381...	dry	0.642187381...	ShortSnowy	0.0625	low	0.589982370...
1999-05-01 0...	17050	mild	0.620326645...	dry	0.852398523...	ShortSnowy	0.0625	low	0.589982370...
1999-06-01 0...	17050	hot	0.746130557...	dry	0.666780994...	ShortSnowy	0.0625	low	0.589982370...
1999-07-01 0...	17050	hot	0.874421027...	dry	0.941076559...	ShortSnowy	0.0625	low	0.589982370...
1999-08-01 0...	17050	hot	0.835030037...	dry	0.780028218...	ShortSnowy	0.0625	low	0.589982370...
1999-09-01 0...	17050	hot	0.542013240...	dry	0.529797480...	ShortSnowy	0.0625	low	0.589982370...
1999-10-01 0...	17050	mild	0.810846089...	dry	0.531673164...	ShortSnowy	0.0625	low	0.589982370...
1999-11-01 0...	17050	mild	0.988354156...	dry	0.715246062...	ShortSnowy	0.0625	low	0.589982370...
1999-12-01 0...	17050	mild	0.927294819...	dry	0.781321862...	ShortSnowy	0.895025070...	low	0.589982370...
2000-01-01 0...	17050	mild	0.527923209...	dry	0.652612362...	MediumSnowy	0.368567304...	low	0.589982370...
2000-02-01 0...	17050	mild	0.874421027...	dry	0.628412981...	ShortSnowy	0.0625	low	0.589982370...
2000-03-01 0...	17050	mild	0.957603280...	dry	0.671959516...	ShortSnowy	0.169575540...	low	0.589982370...
2000-04-01 0...	17050	mild	0.810846089...	dry	0.783043304...	ShortSnowy	0.0625	low	0.589982370...
2000-05-01 0...	17050	mild	0.613187418...	dry	0.746578683...	ShortSnowy	0.0625	low	0.589982370...

Figure 6.9: Generalized data set

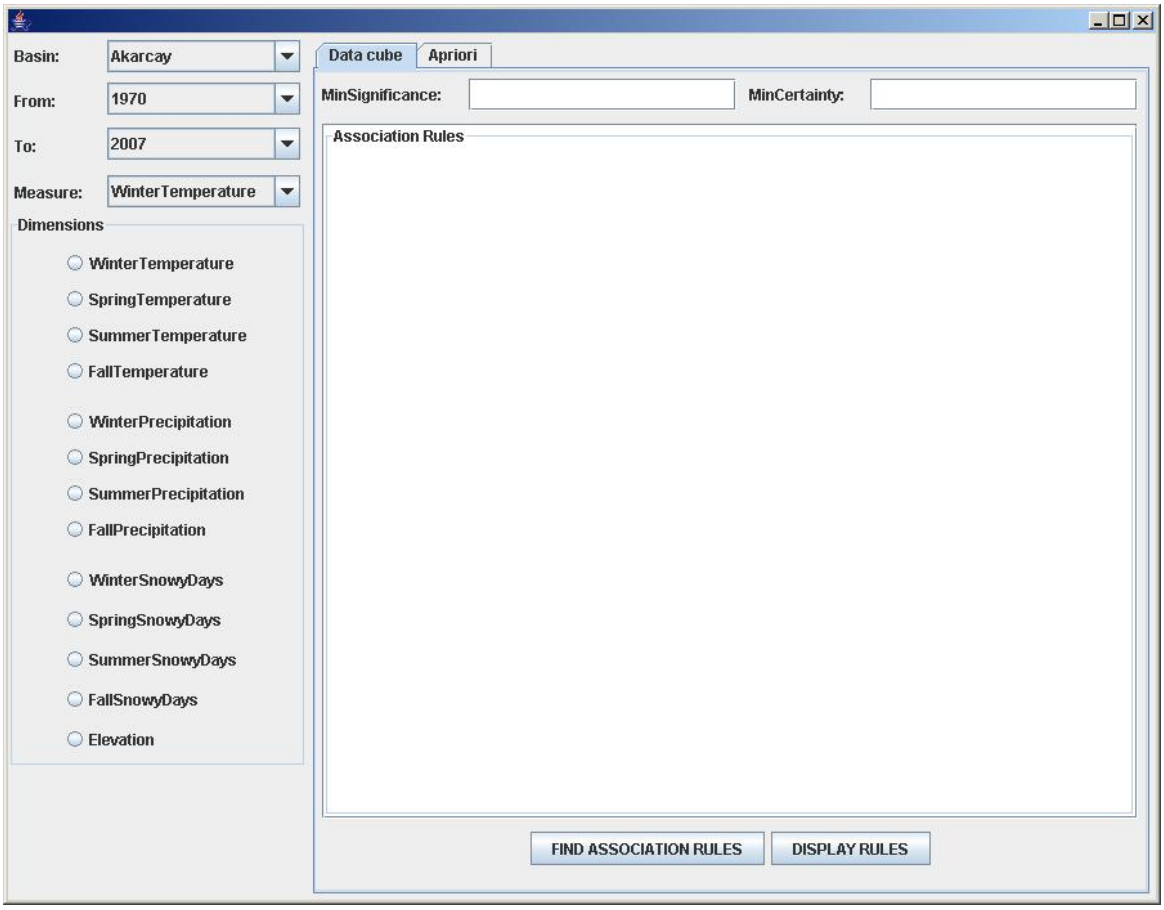


Figure 6.10: Mining association rules

6.2 Mining Association Rules

The association rules can be mined using to different mechanism which are presented in the following subsections. The following fields has to be specified according to the user's needs:

- Basin: The valid association rules among the data of the specified basin is searched.
- From-To Year: The user can restrict the analysis to a specific year interval. The data belonging to this year interval is used in mining the association rules. If not specified, all data recorded between 1970 and 2007 is included to the analysis.
- Measure: One dimension is specified as the measure. The association rules show the effects of the other dimensions on the measure. Measure dimension is not allowed to be specified for the analysis with Apriori algorithm.
- Dimensions: The dimensions of interest are specified by the user.

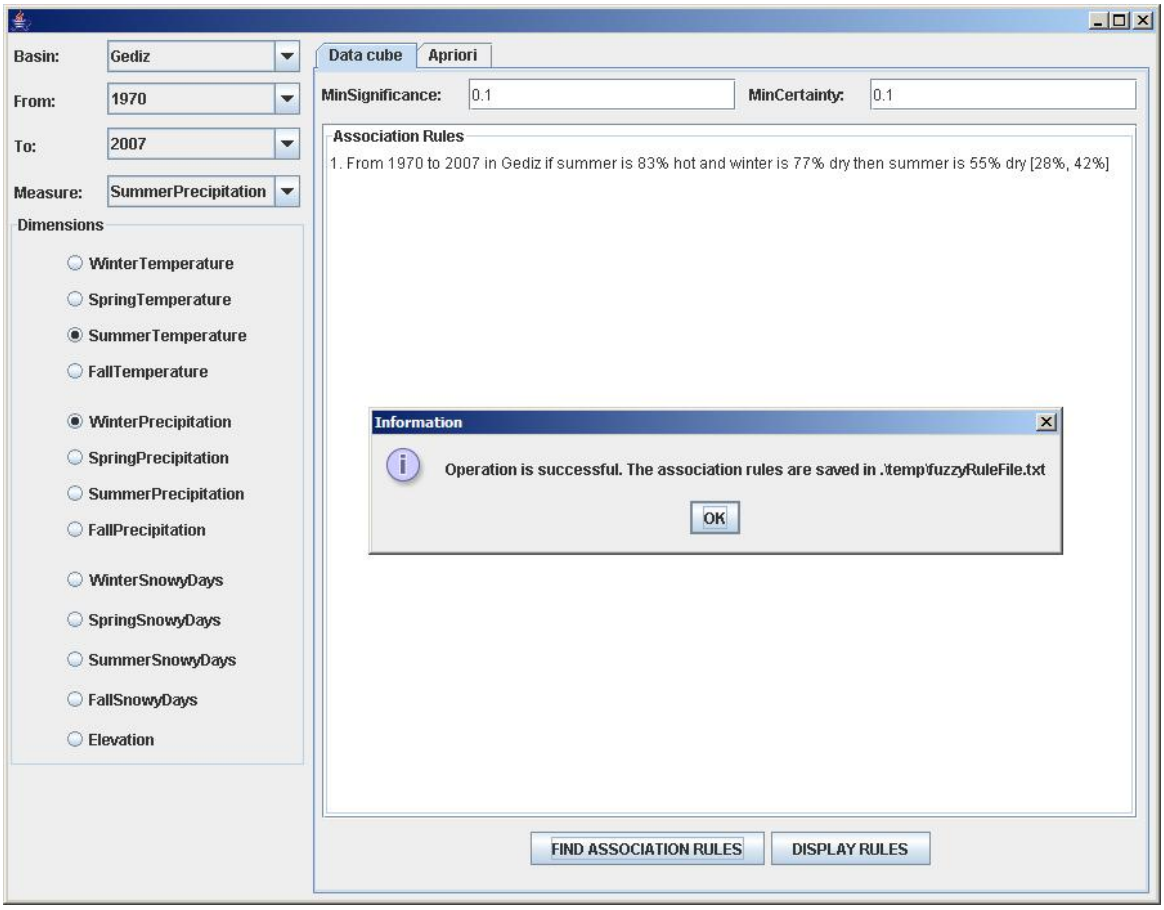


Figure 6.11: Mining association rules using data cube

- Minimum significance/support: The rules above the minimum significance/support are considered as frequent. This field can have values in the range [0, 1] and is used to discard unnecessary rules.
- Minimum certainty/confidence: The rules above the minimum certainty/confidence are considered as interesting. This field can have values in the range [0, 1] and is used to discard unnecessary rules.

6.2.1 Mining Association Rules Using Data Cube

When "Data cube" tab is selected, the fuzzy spatio-temporal association rules are mined using data cube techniques. In Figure 6.11, the association rules, relating winter precipitation and summer temperature with summer precipitation, in Gediz basin is searched using all the information gathered from 1970 to 2007. When "Find association rules" button is clicked, discovered rules are printed on screen The rules are also saved to a text file. The

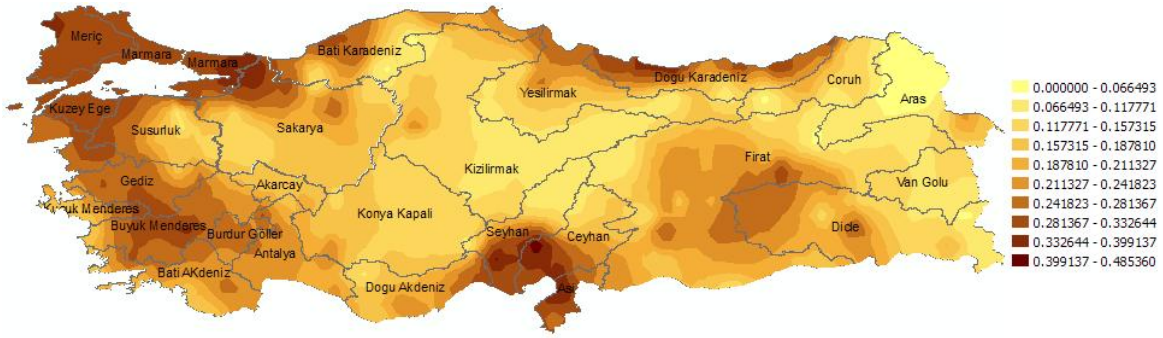


Figure 6.12: Significance map of analysis - From 1970 to 2007 in Gediz if summer is 83% hot and winter is 77% dry then winter is 55% dry [28%, 42%]

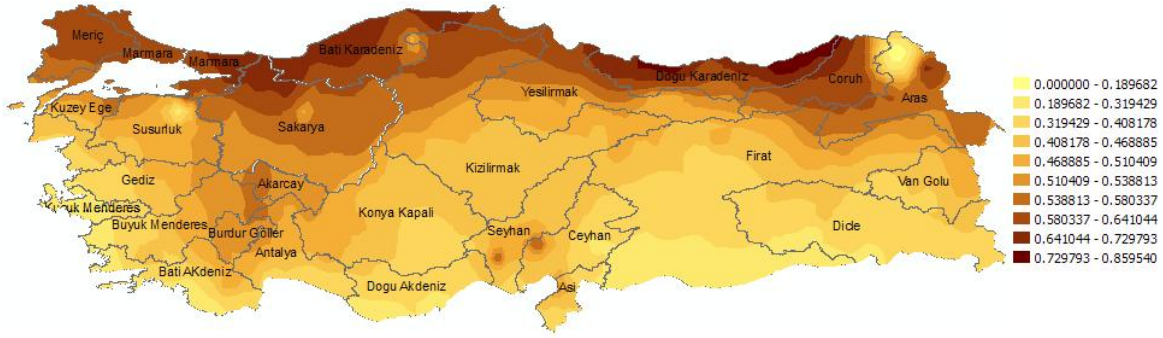


Figure 6.13: Certainty map of analysis - From 1970 to 2007 in Gediz if summer is 83% hot and winter is 77% dry then winter is 55% dry [28%, 42%]

rules are displayed unordered.

The association rules in textual format is hard to digest by the user, because there are a lot of % values. The analysis is completed when an association rule is visualized on Turkey's map within GIS. When "Display rules" button is clicked, the user is asked for the association rule number she is interested in. Then, the supports and confidences of each station for the selected association rule is calculated. These values are displayed to the user in significance and certainty maps presented in Figure 6.12 and Figure 6.13, respectively.

6.2.2 Mining Association Rules Using Apriori Algorithm

When "Apriori" tab is selected, the fuzzy spatio-temporal association rules are mined using Apriori algorithm. In Figure 6.14, the association rules among winter precipitation, summer temperature and summer precipitation in Gediz basin is searched using all the information gathered from 1970 to 2007. When "Find association rules" button is clicked, discovered rules are printed on screen The rules are also saved to a text file. The rules are

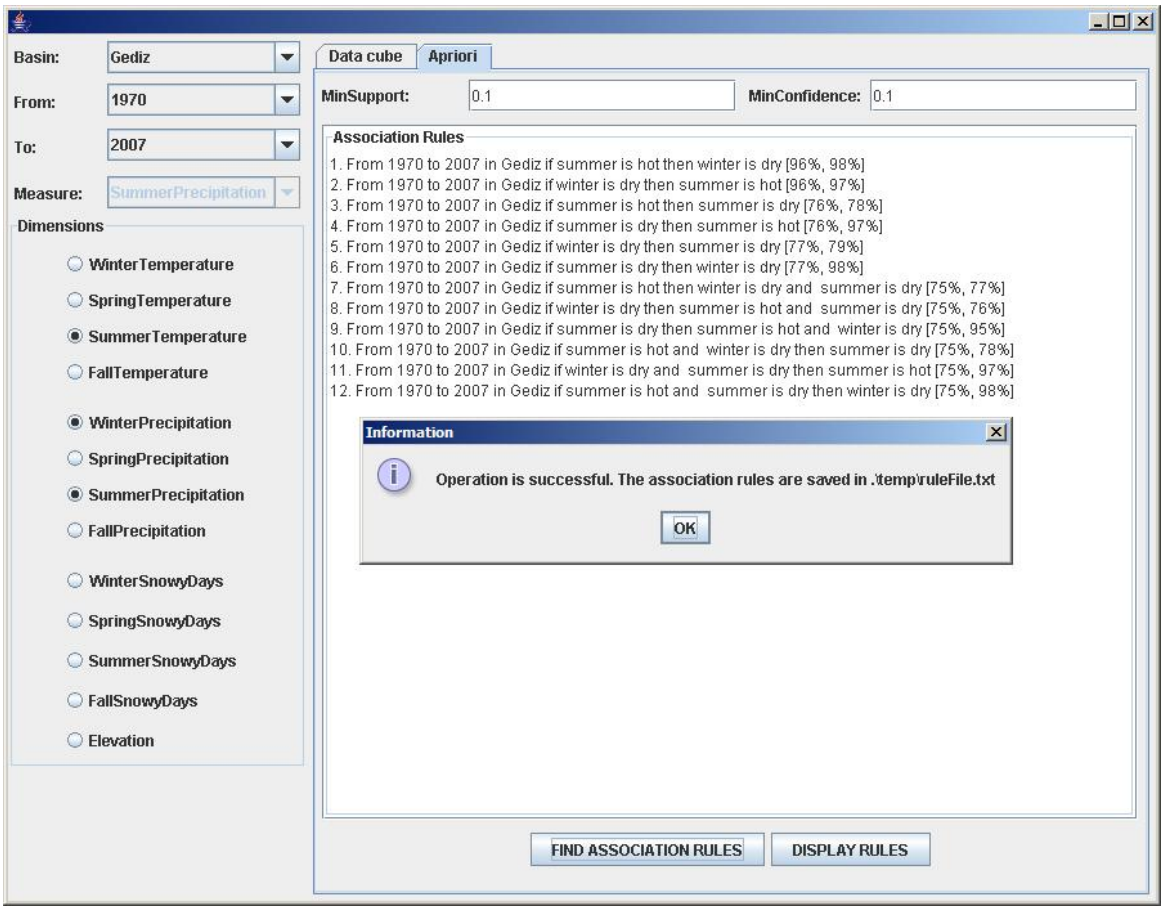


Figure 6.14: Mining association rules using Apriori algorithm

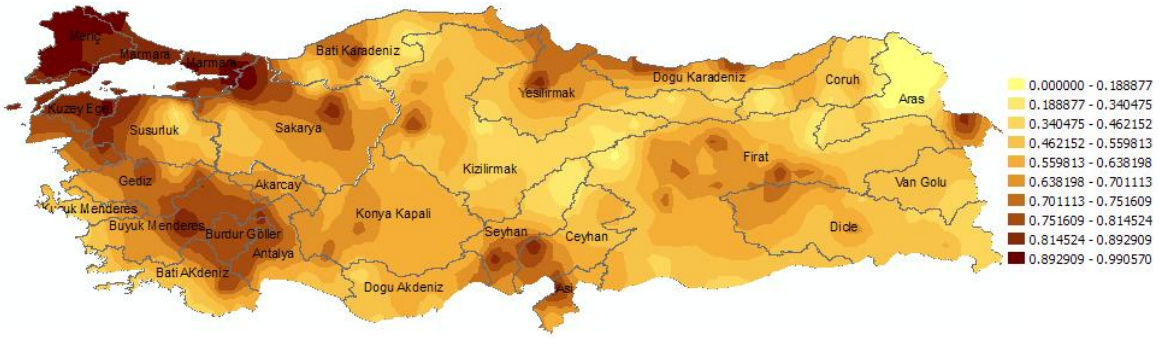


Figure 6.15: Support map of analysis - From 1970 to 2007 in Gediz if summer is hot and winter is dry then winter is dry [75%, 78%]

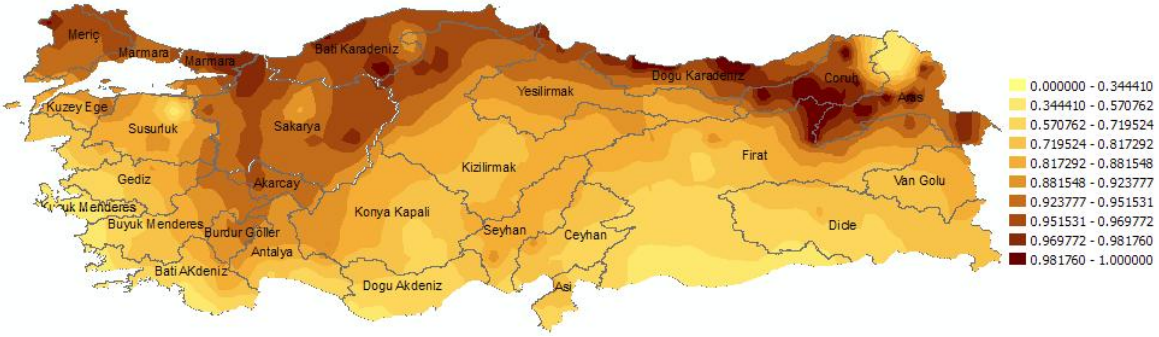


Figure 6.16: Confidence map of analysis - From 1970 to 2007 in Gediz if summer is hot and winter is dry then winter is dry [75%, 78%]

displayed unordered.

The visualization of the association rules found through Apriori algorithm is also possible. Support and confidence maps presented in Figure 6.15 and Figure 6.16 are displayed to the user for the selected association rule.

CHAPTER 7

CONCLUSIONS AND FUTURE DIRECTIONS

This thesis focuses on fuzzy association rule mining from spatio-temporal data by constructing data cubes and using Apriori algorithm. The real life performance of data mining studies is very important since the hypothetical works may fail on real data sets. Turkey's meteorological data set collected since 1970 is analyzed using the data mining approaches introduced within the scope of this thesis. Meteorology data set posses spatial characteristics which are used to visualize discovered association rules to enable domain experts analyze the results better.

In this thesis, fuzzy association rule mining is performed on this data cube using two different methods, namely association rule mining through data cubes and association rule mining through Apriori algorithm. Data cube approach is extended to mine fuzzy association rules for real spatio-temporal data set. Apriori algorithm that represents classical data mining methods is studied to mine fuzzy association rules from the fact table of a fuzzy spatio-temporal data cube. The association rules discovered using different algorithms are analyzed based on predefined metrics for association rules to guide data mining researches in this respect.

Data cube and Apriori algorithms work on the fuzzy spatio-temporal data cube. The data cube construction contains fuzzyfying the dimensions and aggregation on the temporal and spatial dimensions. Crisp values are converted to the fuzzy labels and the membership values according to the fuzzy sets and the fuzzy membership values defined by the user while the data cube construction step. Temporal aggregation transforms monthly collected data into the seasonal aggregated values Additionally, filtering based on a specified time period is provided for the analyzers. Spatial aggregation is done based on the

spatial hierarchy defined on the spatial dimension. According to the spatial hierarchy, the association rules for a specific basin are discovered using the data recorded in the stations which are located within the boundaries of this basin. This regional relationship is resolved to a finer granularity for the visualization of the association rules. The significance and the certainty of the association rules are calculated at the station level to form the data that will be interpolated to create association rule maps. Data cube can be constructed for different fuzzy sets and fuzzy membership values within the scope of this thesis. Once it is constructed, the association rules can be discovered and visualized for several times.

Future studies on this work can be done in the following aspects:

- Although the data cube is constructed once, the performance of this process is also important, which is not considered within the scope of this thesis. This process can be enhanced to minimize the effort for the construction of a fuzzy spatio-temporal data cube.

- Since the construction of a fuzzy spatio-temporal data cube is a costly job, the ways to update the data cube with new data shall be studied.

- The ways to predict missing values for the construction of fuzzy spatio-temporal data cube and mining association rules can be studied. In this thesis, the missing values are kept as "null" in the data cube and discarded while discovering association rules.

- In meteorological domain it is also necessary to define different fuzzy distributions for different regions. For example, the value of 'hot' for Gediz basin is much more different than the understanding of hotness in Firat basin. The temperature value of 20°C is considered 'hot' in Gediz where it is considered as 'mild' in Firat. The fuzzy set definition and usage can be extended in this manner while construction data cube.

- In this thesis, fuzzy association rules are mined from spatio-temporal data. The significance and certainty maps reveal the spatial distribution of the fuzzy relationship. The temporal relationships in the data can be also visualized by creating maps of the association rules mined for successive year periods. The direct discovery of spatio-temporal association rules from spatio-temporal data can be studied.

- The spatial aggregation utilized within the scope of this thesis takes into consideration regional relationship.The association rules in a specific basin are calculated using the data recorded in the stations located in this basin. The neighborhood relationship

may be modeled in the construction of the spatio-temporal data cube. As a result, the effects of close stations on the association rule discovered at a particular station may be investigated.

Appendix A

METRICS QUESTIONNAIRE

Some metrics have been specified for the comparison of the quality of the discovered rules of association rule mining techniques. In this page, the metrics and its classifications are provided. A rule set which includes same rules discovered by Apriori and data cube techniques is provided. I kindly request you to give points between 0-10 to each association rule within the scope of each metric. Visualization metric will be evaluated in the next page based on two maps.

You can also comment on the metrics and metric classification and propose new metrics. THANK YOU.

Interpretability: This is the complexity of the association rules discovered and the size of the decision tree constructed. Interpretability is a good measure to compare the understandability of association rules. Since the association rules are consumed by human, they have to be user friendly. For 'interpretability' metric, the following classifications are proposed within the scope of these this: clear, fully understandable, not very understandable, meaningless.

Precision: This metric studies the quality of the association rules which is represented in terms of confidence. For the precision of association rules, the support and confidence calculation methods can be inspected for data cube approach and Apriori algorithm. Apriori approach treats fuzzy labels as if they are transactional items. It does not take into account the membership values which represents to what extend a fuzzy label is valid for a record. For 'precision' metric, the following classifications are proposed within the scope of these this: precise, rough, imprecise.

Utility: This metric studies the usefulness of the association rules. Useful, less-useful and impractical are specified as the classification terms for the comparison according to 'utility' metric.

Novelty: This is the ability to discover non-trivial, implicit, previously unknown and potentially useful association rules. The association rules reveal the hidden relationships in a large database. The novelty of an association rule mining method is evaluated by its ability to discover non-trivial, implicit, previously unknown and potentially useful information. Surprise, novel, ordinary are specified as the classifications for the 'novelty' metric.

Direct-to-the-Point: This is the ability to discover related association rules. Trivial rules and the rules out of the scope of the user specifications degrade this property. Very connected, related, less-related and out-of-scope are the classifications defined for this metric.

Visualization: This metric study which association rule mining technique discovers rules that can be visualized better. More detailed maps enhance this property. For 'visualization' metric, the following classifications are proposed within the scope of these this: Detailed, less-detailed, rough.

Performance: Working on the analysis of voluminous data brings the necessity for the computational power. This is evaluated using the source code. Hence it is out of the scope of the questionnaire.

EVALUATION CRITERIA

Evaluate Interpretability

clear **10-8**
fully understandable **8-5**
not very understandable **5-2**
meaningless **2-0**

Evaluate Precision

precise **10-7**
rough **7-4**
imprecise **4-0**

Evaluate Utility

useful **8-5**
less-useful **5-2**
impractical **2-0**

Evaluate Novelty

surprise **10-7**

novel **7-4**

ordinary **4-0**

Evaluate Direct-To-The-Point

very connected **10-8**

related **8-5**

less-related **5-2**

out-of-scope **2-0**

RULE SET

Apriori

A1. From 1970 to 2007 in Buyuk_Menderes if fall is dry and winter is dry then summer is mild [2%, 2%]

A2. From 1970 to 2007 in Buyuk_Menderes if fall is dry and winter is dry then summer is hot [88%, 97%]

A3. From 1970 to 2007 in Buyuk_Menderes if fall is dry and winter is fair then summer is hot [0%, 100%]

A4. From 1970 to 2007 in Buyuk_Menderes if fall is fair and winter is dry then summer is hot [0%, 100%]

Data cube

D1. From 1970 to 2007 in Buyuk_Menderes if fall is 63% dry and winter is 68% dry then summer is 54% mild [0%, 1%]

D2. From 1970 to 2007 in Buyuk_Menderes if fall is 66% dry and winter is 83% dry then summer is 86% hot [41%, 81%]

D3. From 1970 to 2007 in Buyuk_Menderes if fall is 66% dry and winter is 54% fair then summer is 90% hot [0%, 89%]

D4. From 1970 to 2007 in Buyuk_Menderes if fall is 69% fair and winter is 75% dry then summer is 94% hot [0%, 94%]

Apriori

A5. From 1970 to 2007 in Buyuk_Menderes if winter is dry then summer is mild [2%, 2%]

A6. From 1970 to 2007 in Buyuk_Menderes if winter is dry then summer is hot [96%, 97%]

A7. From 1970 to 2007 in Buyuk_Menderes if winter is fair then summer is hot [0%, 100%]

Data cube

D5. From 1970 to 2007 in Buyuk_Menderes if winter is 66% dry then summer is 54% mild [0%, 1%]

D6. From 1970 to 2007 in Buyuk_Menderes if winter is 81% dry then summer is 84% hot [65%, 82%]

D7. From 1970 to 2007 in Buyuk_Menderes if winter is 54% fair then summer is 90% hot [0%, 89%]

Apriori

A8. From 1970 to 2007 in Buyuk_Menderes if summer is mild then winter is dry [2%, 100%]

A9. From 1970 to 2007 in Buyuk_Menderes if summer is hot then winter is dry [96%, 99%]

A10. From 1970 to 2007 in Buyuk_Menderes if summer is hot then winter is fair [0%, 0%]

Datacube

D8. From 1970 to 2007 in Buyuk_Menderes if summer is 54% mild then winter is 66% dry [0%, 69%]

D9. From 1970 to 2007 in Buyuk_Menderes if summer is 84% hot then winter is 81% dry [65%, 80%]

D10. From 1970 to 2007 in Buyuk_Menderes if summer is 90% hot then winter is 54% fair [0%, 0%]

Apriori

A11. From 1970 to 2007 in Buyuk_Menderes if fall is dry then summer is mild [2%, 2%]

A12. From 1970 to 2007 in Buyuk_Menderes if fall is dry then summer is hot [89%, 97%]

A13. From 1970 to 2007 in Buyuk_Menderes if fall is fair then summer is hot [0%, 100%]

Data cube D11. From 1970 to 2007 in Buyuk_Menderes if fall is 63% dry then summer is 54% mild [0%, 1%]

D12. From 1970 to 2007 in Buyuk_Menderes if fall is 67% dry then summer is 88% hot [51%, 81%]

D13. From 1970 to 2007 in Buyuk_Menderes if fall is 69% fair then summer is 94% hot [0%, 94%]

Apriori

A14. From 1970 to 2007 in Buyuk_Menderes if summer is mild then fall is dry [2%, 95%]

A15. From 1970 to 2007 in Buyuk_Menderes if summer is hot then fall is dry [89%, 91%]

A16. From 1970 to 2007 in Buyuk_Menderes if summer is hot then fall is fair [0%, 0%]

Datacube

D14. From 1970 to 2007 in Buyuk_Menderes if summer is 54% mild then fall is 63% dry [0%, 64%]

D15. From 1970 to 2007 in Buyuk_Menderes if summer is 88% hot then fall is 67% dry [51%, 62%]

D16. From 1970 to 2007 in Buyuk_Menderes if summer is 94% hot then fall is 69% fair [0%, 0%]

Apriori

A17. From 1970 to 2007 in Buyuk_Menderes if fall is dry then winter is dry [91%, 99%]

A18. From 1970 to 2007 in Buyuk_Menderes if fall is dry then winter is fair [0%, 0%]

A19. From 1970 to 2007 in Buyuk_Menderes if fall is fair then winter is dry [0%, 100%]

Datacube

D17. From 1970 to 2007 in Buyuk_Menderes if fall is 70% dry then winter is 83% dry [50%, 79%]

D18. From 1970 to 2007 in Buyuk_Menderes if fall is 69% dry then winter is 57% fair [0%, 0%]

D19. From 1970 to 2007 in Buyuk_Menderes if fall is 69% fair then winter is 75% dry [0%, 75%]

Apriori

A20. From 1970 to 2007 in Buyuk_Menderes if winter is dry then fall is dry [91%, 91%]

A21. From 1970 to 2007 in Buyuk_Menderes if winter is dry then fall is fair [0%, 0%]

A22. From 1970 to 2007 in Buyuk_Menderes if winter is fair then fall is dry [0%, 100%]

Datacube

D20. From 1970 to 2007 in Buyuk_Menderes if winter is 83% dry then fall is 70% dry [50%, 63%]

D21. From 1970 to 2007 in Buyuk_Menderes if winter is 75% dry then fall is 69% fair [0%, 0%]

D22. From 1970 to 2007 in Buyuk_Menderes if winter is 57% fair then fall is 69% dry [0%, 69%]

Apriori

A23. From 1970 to 2007 in Buyuk_Menderes if summer is mild and winter is dry then fall is dry [2%, 95%]

A24. From 1970 to 2007 in Buyuk_Menderes if summer is hot and winter is dry then fall is dry [88%, 91%]

A25. From 1970 to 2007 in Buyuk_Menderes if summer is hot and winter is dry then fall is

fair [0%, 0%]

A26. From 1970 to 2007 in Buyuk_Menderes if summer is hot and winter is fair then fall is dry [0%, 100%]

Datacube

D23. From 1970 to 2007 in Buyuk_Menderes if summer is 54% mild and winter is 68% dry then fall is 63% dry [0%, 64%]

D24. From 1970 to 2007 in Buyuk_Menderes if summer is 86% hot and winter is 83% dry then fall is 66% dry [41%, 62%]

D25. From 1970 to 2007 in Buyuk_Menderes if summer is 94% hot and winter is 75% dry then fall is 69% fair [0%, 0%]

D26. From 1970 to 2007 in Buyuk_Menderes if summer is 90% hot and winter is 54% fair then fall is 66% dry [0%, 69%]

Apriori

A27. From 1970 to 2007 in Buyuk_Menderes if fall is dry and summer is mild then winter is dry [2%, 100%]

A28. From 1970 to 2007 in Buyuk_Menderes if fall is dry and summer is hot then winter is dry [88%, 99%]

A29. From 1970 to 2007 in Buyuk_Menderes if fall is dry and summer is hot then winter is fair [0%, 0%]

A30. From 1970 to 2007 in Buyuk_Menderes if fall is fair and summer is hot then winter is dry [0%, 100%]

Datacube

D27. From 1970 to 2007 in Buyuk_Menderes if fall is 63% dry and summer is 54% mild then winter is 68% dry [0%, 70%]

D28. From 1970 to 2007 in Buyuk_Menderes if fall is 66% dry and summer is 86% hot then winter is 83% dry [41%, 80%]

D29. From 1970 to 2007 in Buyuk_Menderes if fall is 66% dry and summer is 90% hot then winter is 54% fair [0%, 0%]

D30. From 1970 to 2007 in Buyuk_Menderes if fall is 69% fair and summer is 94% hot then winter is 75% dry [0%, 75%]

Apriori

A31. From 1970 to 2007 in Buyuk_Menderes if fall is dry then summer is hot and winter is fair [0%, 0%]

A32. From 1970 to 2007 in Buyuk_Menderes if summer is hot then fall is dry and winter is

fair [0%, 0%]

A33. From 1970 to 2007 in Buyuk_Menderes if winter is fair then fall is dry and summer is hot [0%, 100%]

A34. From 1970 to 2007 in Buyuk_Menderes if summer is mild then fall is dry and winter is dry [2%, 95%]

A35. From 1970 to 2007 in Buyuk_Menderes if fall is dry then summer is mild and winter is dry [2%, 2%]

A36. From 1970 to 2007 in Buyuk_Menderes if winter is dry then fall is dry and summer is mild [2%, 2%]

A37. From 1970 to 2007 in Buyuk_Menderes if summer is hot then fall is fair and winter is dry [0%, 0%]

A38. From 1970 to 2007 in Buyuk_Menderes if fall is fair then summer is hot and winter is dry [0%, 100%]

A39. From 1970 to 2007 in Buyuk_Menderes if winter is dry then summer is hot and fall is fair [0%, 0%]

A40. From 1970 to 2007 in Buyuk_Menderes if fall is dry then summer is hot and winter is dry [88%, 96%]

A41. From 1970 to 2007 in Buyuk_Menderes if summer is hot then fall is dry and winter is dry [88%, 91%]

A42. From 1970 to 2007 in Buyuk_Menderes if winter is dry then fall is dry and summer is hot [88%, 89%]

Evaluate Visualization

detailed **10-8**

less-detailed **7-4**

rough **3-0**

Please give a point to each map.

Association rule:

From 1970 to 2007 in Buyuk_Menderes if summer is 85% hot then summer is 54% dry [34%, 42%]

Meaning of the rule: In Meteorology domain, it is expected that in Gediz, Kucuk

Menderes and Buyuk Menderes basins, the increasing summer temperature results in decreasing summer precipitation.

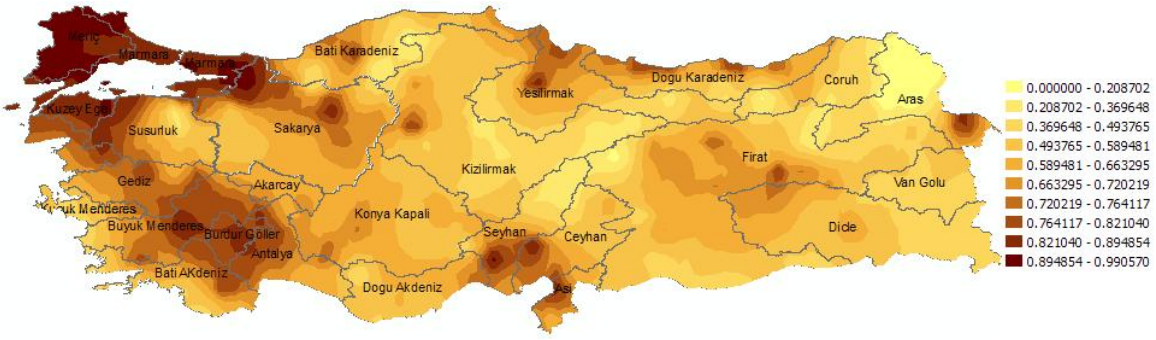


Figure A.1: APRIORI map → Point:

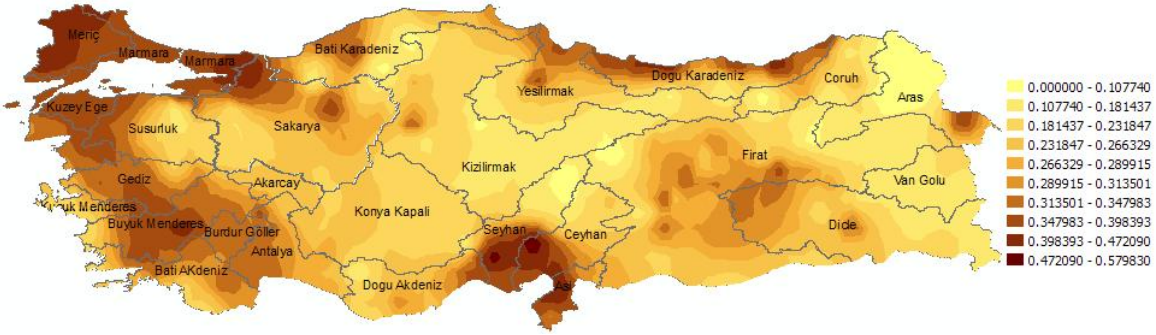


Figure A.2: DATA CUBE map → Point:

Appendix B

FUZZY SETS

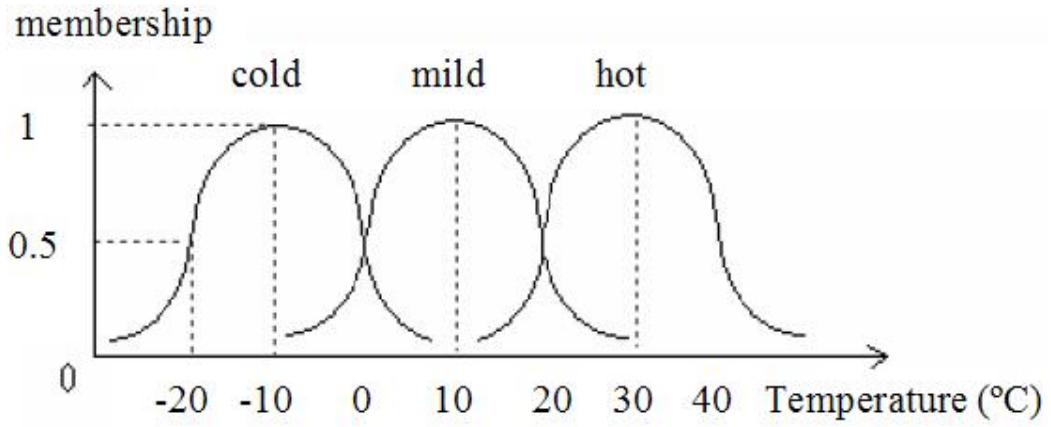


Figure B.1: Fuzzy set for temperature

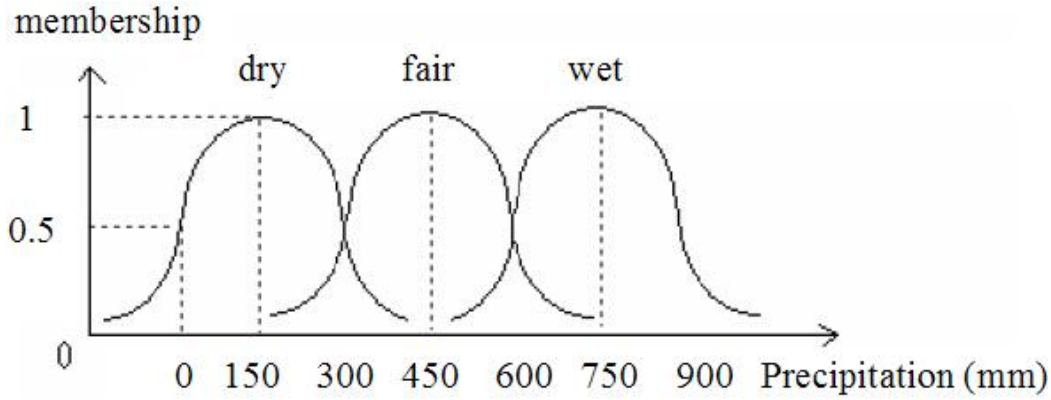


Figure B.2: Fuzzy set for precipitation

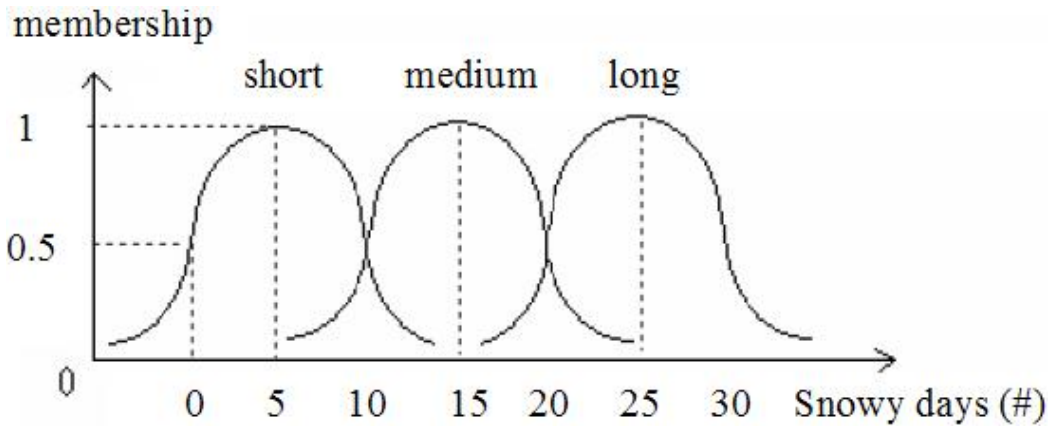


Figure B.3: Fuzzy set for snowy days

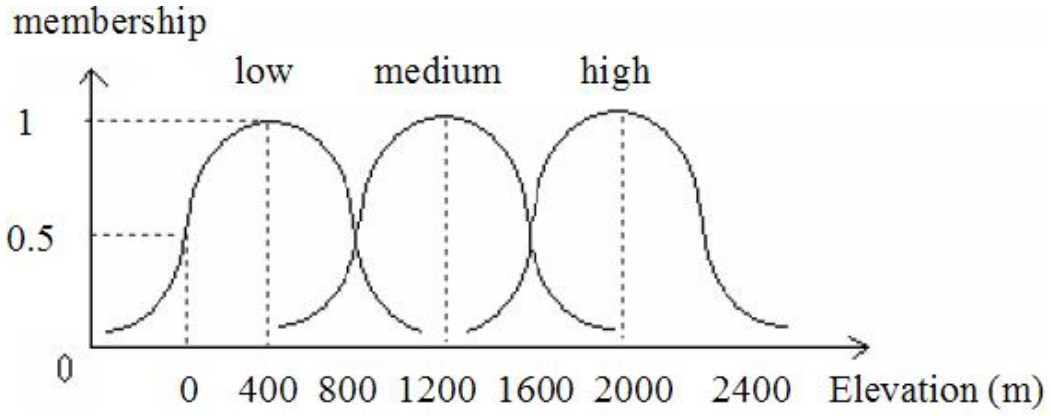


Figure B.4: Fuzzy set for elevation

Appendix C

.DBF FILE FOR VISUAL ANALYSIS

Table C.1: An example .dbf file content

STATION	Y	X	Z	SIGNIFICANCE	CERTAINTY
17718	39.78	40.38	1504.00	0.20	0.49
17928	36.98	32.47	1486.00	0.12	0.43
17180	39.07	26.88	11.00	0.28	0.34
17199	38.35	38.32	965.00	0.35	0.41
17636	40.98	28.80	51.00	0.40	0.59
17784	39.03	43.35	1689.00	0.15	0.43
17820	38.20	26.83	28.00	0.18	0.21
17842	38.28	37.27	1123.00	0.26	0.41
17886	37.35	28.13	406.00	0.39	0.44
17090	39.75	37.02	1288.00	0.10	0.46
17614	41.38	32.20	30.00	0.34	0.72
17614	41.38	32.20	30.00	0.34	0.72
17712	39.49	35.11	1050.00	0.12	0.41
17986	36.08	35.97	8.00	0.35	0.39
17175	39.30	26.70	25.00	0.31	0.35
17696	39.36	27.01	21.00	0.37	0.43
17824	38.15	29.07	688.00	0.35	0.48
17612	41.08	31.17	60.00	0.28	0.76
17940	37.05	36.15	120.00	0.34	0.36
17160	39.15	34.17	1007.00	0.23	0.45

Continued on Next Page...

Table C.1 – Continued

STATION	Y	X	Z	SIGNIFICANCE	CERTAINTY
17880	38.05	44.02	2354.00	0.07	0.47
17962	36.85	36.22	53.00	0.57	0.63
17204	38.73	41.52	1318.00	0.30	0.44
17062	40.97	29.08	45.00	0.43	0.61
17822	38.23	27.97	134.00	0.35	0.38
17981	36.57	35.38	19.00	0.32	0.35
17898	37.42	31.83	1121.00	0.32	0.54
17248	37.50	34.05	1055.00	0.22	0.38
17716	39.90	37.75	1508.00	0.08	0.51
17738	39.19	40.21	1700.00	0.21	0.45
17970	36.15	29.57	25.00	0.20	0.21
17056	40.99	27.55	0.00	0.39	0.58
17220	38.43	27.17	3.00	0.27	0.29
17193	38.62	34.70	1245.00	0.15	0.43
17720	39.55	44.08	1581.00	0.27	0.57
17203	38.87	40.50	1130.00	0.38	0.47
17810	38.77	42.50	1828.00	0.15	0.36
17221	38.29	26.26	19.00	0.21	0.25
17950	37.32	42.18	377.00	0.26	0.28
17748	39.08	28.98	965.00	0.20	0.49
968	36.50	39.55	398.00	0.14	0.14
17201	38.65	39.25	981.00	0.33	0.39
17624	41.13	37.28	21.00	0.41	0.77
17734	39.37	38.12	1094.00	0.30	0.49
17936	37.27	35.07	234.00	0.52	0.58
17630	41.12	42.72	1790.00	0.00	0.00
17152	39.38	27.53	102.00	0.38	0.49
17754	39.10	33.08	995.00	0.19	0.45
17074	41.37	33.78	922.00	0.13	0.62
17906	37.55	34.48	1447.00	0.14	0.39

Continued on Next Page...

Table C.1 – Continued

STATION	Y	X	Z	SIGNIFICANCE	CERTAINTY
17944	37.13	38.17	890.00	0.17	0.18
17632	40.93	26.40	8.00	0.41	0.57
17069	40.78	30.42	29.00	0.47	0.72
17796	38.72	31.05	996.00	0.23	0.53
17162	39.18	36.07	1233.00	0.16	0.45
17600	41.40	27.35	60.00	0.38	0.59
17638	40.91	29.18	59.00	0.44	0.61
17099	39.73	43.05	1638.00	0.18	0.54
17100	39.92	44.05	860.00	0.44	0.58
17626	41.02	39.57	73.00	0.36	0.65
17790	38.47	27.22	54.00	0.29	0.32
17790	38.47	27.22	54.00	0.29	0.32
17282	37.88	41.12	567.00	0.27	0.28
17660	40.39	29.16	4.00	0.43	0.61
620	41.10	29.03	130.00	0.31	0.65
17900	37.58	32.78	1011.00	0.22	0.44
17884	37.32	27.78	56.00	0.29	0.31
17646	40.82	32.90	1152.00	0.03	0.59
17280	37.90	40.23	598.00	0.31	0.32
17066	40.77	29.93	0.00	0.48	0.68
17892	37.32	29.77	1178.00	0.30	0.54
17095	39.55	41.16	1758.00	0.05	0.53
17804	38.80	38.75	717.00	0.38	0.42
17908	37.45	35.82	137.00	0.58	0.61
17690	40.05	42.17	1586.00	0.19	0.58
17684	40.17	38.10	984.00	0.14	0.45
17116	40.18	29.07	231.00	0.44	0.59
17045	41.18	41.82	601.00	0.16	0.63
17706	39.49	30.31	787.00	0.19	0.49
17846	38.24	39.40	1100.00	0.30	0.35

Continued on Next Page...

Table C.1 – Continued

STATION	Y	X	Z	SIGNIFICANCE	CERTAINTY
17966	37.03	37.98	396.00	0.27	0.28
17837	38.45	35.80	1400.00	0.09	0.47
17622	41.58	35.93	22.00	0.34	0.65
17140	39.82	34.80	1284.00	0.07	0.48
17679	40.18	31.35	653.00	0.28	0.48
17924	36.97	28.68	16.00	0.34	0.36
17340	36.81	34.63	16.00	0.38	0.41
17844	38.27	39.19	1240.00	0.23	0.32
17666	40.48	41.00	1185.00	0.32	0.60
17351	36.98	35.35	29.00	0.38	0.41
17070	40.73	31.60	737.00	0.08	0.58
17648	40.92	33.63	878.00	0.19	0.61
17902	37.72	33.55	995.00	0.17	0.33
17806	38.72	39.97	1132.00	0.34	0.39
17702	39.92	30.03	838.00	0.13	0.52
17772	39.07	39.33	980.00	0.35	0.41
17974	36.27	32.32	20.00	0.22	0.25
17265	37.75	38.28	614.00	0.34	0.35
17766	38.57	38.43	900.00	0.29	0.35
17205	38.48	42.30	1661.00	0.16	0.39
17650	41.02	34.03	868.00	0.26	0.61
17725	39.25	29.58	969.00	0.16	0.53
17196	38.75	35.48	1050.00	0.20	0.48
17964	37.03	36.63	526.00	0.28	0.31
17768	39.07	38.92	1031.00	0.30	0.36
17088	40.47	39.47	1226.00	0.09	0.52
17792	38.48	28.13	112.00	0.39	0.43
17920	37.57	44.28	1870.00	0.11	0.38
17122	40.09	29.59	539.00	0.26	0.57
17704	39.55	29.50	869.00	0.20	0.55

Continued on Next Page...

Table C.1 – Continued

STATION	Y	X	Z	SIGNIFICANCE	CERTAINTY
17678	40.15	29.39	220.00	0.35	0.61
17244	37.87	32.48	1027.00	0.28	0.48
17190	38.75	30.53	1197.00	0.24	0.56
17890	37.42	29.33	1039.00	0.33	0.49
17059	41.25	29.03	3.00	0.37	0.62
17261	37.08	37.37	880.00	0.27	0.31
17776	38.97	41.07	1389.00	0.31	0.47
17030	41.28	36.30	102.00	0.37	0.65
17836	38.38	35.50	1281.00	0.20	0.42
17300	36.89	30.73	53.00	0.34	0.36
17096	39.92	41.27	1845.00	0.08	0.57
17664	40.47	32.65	1002.00	0.17	0.57
17375	36.31	30.15	2.00	0.31	0.33
17860	37.92	28.32	93.00	0.37	0.39
17240	37.75	30.55	1116.00	0.30	0.52
17843	38.35	38.48	1290.00	0.24	0.35
17320	36.08	32.83	94.00	0.22	0.24
17774	38.58	40.02	1090.00	0.28	0.40
17808	38.46	40.35	1250.00	0.32	0.42
17752	39.02	31.15	960.00	0.25	0.53
17952	36.75	29.92	1253.00	0.34	0.51
17830	38.35	31.42	1038.00	0.27	0.54
17786	38.98	43.77	1704.00	0.21	0.48
17123	39.46	30.31	801.00	0.17	0.55
17926	37.07	30.20	988.00	0.31	0.50
17092	39.45	39.30	1218.00	0.32	0.53
17750	39.05	29.42	843.00	0.34	0.53
17864	38.08	30.45	1029.00	0.28	0.56
17172	38.47	43.35	1670.00	0.17	0.38
17080	40.60	33.62	730.00	0.29	0.57

Continued on Next Page...

Table C.1 – Continued

STATION	Y	X	Z	SIGNIFICANCE	CERTAINTY
17602	41.75	32.38	13.00	0.33	0.74
17234	37.85	27.85	94.00	0.32	0.34
17740	39.37	41.70	1717.00	0.14	0.53
17870	38.20	37.18	1151.00	0.23	0.39
17726	39.45	31.53	1087.00	0.25	0.56
17292	37.22	28.37	666.00	0.39	0.47
17111	39.83	26.07	27.00	0.28	0.43
17640	40.50	31.10	146.00	0.37	0.70
17979	36.78	35.78	21.00	0.37	0.41
17037	41.00	39.72	39.00	0.40	0.69
17110	40.18	25.90	94.00	0.37	0.48
17802	38.72	36.40	1602.00	0.05	0.43
17233	37.21	27.15	330.00	0.15	0.16
17760	39.20	35.25	1067.00	0.11	0.39
17610	41.18	29.62	0.00	0.34	0.66
17370	36.58	36.17	8.00	0.51	0.55
17290	37.03	27.44	24.00	0.19	0.20
17984	36.20	36.17	100.00	0.29	0.32
17958	36.62	34.30	45.00	0.31	0.34
17686	40.25	40.23	1558.00	0.04	0.54
17728	39.58	32.15	870.00	0.30	0.55
17285	37.57	43.77	1256.00	0.24	0.36
17832	38.28	31.92	1031.00	0.21	0.50
17800	38.39	32.56	969.00	0.24	0.44
17722	39.30	26.59	10.00	0.34	0.39
17700	39.58	28.62	659.00	0.22	0.52
17050	41.67	26.57	38.00	0.48	0.64
17764	39.05	38.50	1142.00	0.31	0.44
17852	38.18	43.06	1696.00	0.14	0.40
17948	37.07	41.22	471.00	0.20	0.20

Continued on Next Page...

Table C.1 – Continued

STATION	Y	X	Z	SIGNIFICANCE	CERTAINTY
17022	41.45	31.80	54.00	0.36	0.76
17186	38.62	27.43	71.00	0.37	0.40
17874	38.13	39.45	662.00	0.32	0.33
17270	37.13	38.77	714.00	0.24	0.25
17668	40.55	41.98	1351.00	0.29	0.63
17188	38.68	29.40	929.00	0.33	0.54
17330	36.38	33.93	19.00	0.26	0.28
17732	39.62	34.42	867.00	0.26	0.48
17778	38.10	41.28	1650.00	0.18	0.46
17608	41.27	26.68	20.00	0.42	0.60
17681	40.30	35.88	744.00	0.23	0.50
17238	37.67	30.33	1255.00	0.39	0.55
9030	38.27	27.05	10.00	0.17	0.19
17052	41.68	27.30	189.00	0.42	0.62
17042	41.40	41.43	139.00	0.43	0.84
17850	37.90	28.15	179.00	0.37	0.40
17606	41.95	34.02	105.00	0.28	0.73
17628	41.18	40.88	0.00	0.35	0.80
17024	41.98	33.78	8.00	0.30	0.70
17688	40.30	41.55	1723.00	0.11	0.61
17730	39.41	33.37	1140.00	0.19	0.47
17862	38.07	30.17	867.00	0.34	0.54
17237	37.78	29.08	392.00	0.46	0.51
17114	40.35	27.97	37.00	0.38	0.54
17872	38.10	37.88	1204.00	0.13	0.30
17135	39.85	33.52	762.00	0.38	0.54
17112	40.13	26.41	8.00	0.36	0.47
17742	39.12	27.18	59.00	0.36	0.41
17954	36.78	31.43	5.00	0.27	0.30
17294	36.75	28.78	7.00	0.24	0.26

Continued on Next Page...

Table C.1 – Continued

STATION	Y	X	Z	SIGNIFICANCE	CERTAINTY
17682	40.30	38.42	1775.00	0.09	0.49
17848	38.37	42.10	1509.00	0.17	0.37
17662	40.52	30.30	92.00	0.41	0.62
17656	40.85	43.32	1715.00	0.02	0.67
17960	37.03	35.82	27.00	0.38	0.42
17040	41.03	40.51	16.00	0.49	0.85
17770	39.07	39.13	1481.00	0.14	0.31
17026	42.02	35.18	90.00	0.34	0.66
17756	39.40	33.78	1041.00	0.14	0.46
17255	37.53	36.95	498.00	0.26	0.28
17868	38.25	36.92	1213.00	0.21	0.41
17835	38.63	34.58	1226.00	0.14	0.41
17296	36.62	29.12	135.00	0.21	0.23
17854	37.95	27.37	12.00	0.21	0.24
17061	41.10	29.03	44.00	0.39	0.68
17680	40.17	31.93	782.00	0.42	0.57
17084	40.55	34.97	830.00	0.16	0.53
17184	38.92	27.85	106.00	0.35	0.39
17866	38.02	36.50	1341.00	0.14	0.43
17083	40.87	35.47	704.00	0.18	0.55
17618	41.36	33.50	1050.00	0.01	0.46
17695	39.92	29.07	613.00	0.04	0.52
17882	37.87	30.83	1157.00	0.33	0.50
17310	36.55	32.00	29.00	0.30	0.33
17847	38.28	39.77	1184.00	0.35	0.37
17762	39.23	37.38	1533.00	0.04	0.49
17826	38.10	30.55	1111.00	0.37	0.55
17098	40.37	43.06	1775.00	0.01	0.57
17896	37.68	31.73	1129.00	0.21	0.49
17780	39.15	42.53	1502.00	0.20	0.47

Continued on Next Page...

Table C.1 – Continued

STATION	Y	X	Z	SIGNIFICANCE	CERTAINTY
17746	39.03	28.39	851.00	0.32	0.50
17634	40.54	26.55	283.00	0.38	0.57
17634	40.54	26.55	283.00	0.38	0.57
17054	41.17	27.78	117.00	0.35	0.61
17912	37.77	39.32	821.00	0.32	0.34
17130	39.95	32.88	921.00	0.32	0.57
17956	36.39	33.26	275.00	0.31	0.32
17978	36.72	37.12	664.00	0.26	0.28
17128	40.12	33.00	957.00	0.20	0.54
17674	40.10	27.65	44.00	0.37	0.55
17834	38.23	34.05	965.00	0.25	0.42
17298	36.85	28.23	163.00	0.24	0.26
17086	40.30	36.57	895.00	0.24	0.50
17798	38.82	31.73	1095.00	0.21	0.52
17232	37.86	27.25	3.00	0.18	0.21
17275	37.30	40.73	729.00	0.23	0.24
17676	40.12	29.02	977.00	0.00	0.00
17736	39.02	39.36	1400.00	0.25	0.38
17812	38.67	43.98	2178.00	0.14	0.51
17828	38.18	31.11	1096.00	0.23	0.50
17380	36.12	29.39	5.00	0.17	0.18
17692	40.33	42.57	2093.00	0.00	0.59
17793	38.35	31.02	1020.00	0.25	0.51
17034	40.92	38.38	23.00	0.45	0.82
17932	37.11	33.13	1025.00	0.25	0.41
17658	40.39	29.07	20.00	0.40	0.64
17250	37.97	34.68	1218.00	0.24	0.48
17129	39.57	32.40	800.00	0.29	0.57
17085	40.65	35.85	620.00	0.39	0.55
17033	40.98	37.90	0.00	0.43	0.76

Continued on Next Page...

Table C.1 – Continued

STATION	Y	X	Z	SIGNIFICANCE	CERTAINTY
17840	38.48	36.50	1566.00	0.04	0.46
17210	37.92	41.95	983.00	0.36	0.38
17616	41.12	32.38	400.00	0.40	0.64

REFERENCES

[1] X. Yao, "Research issues in spatio-temporal data mining," *Geographic Information Science (UCGIS) workshop on Geospatial Visualization and Knowledge Discovery*, November 2003.

[2] G. Pestana and M. M. da Silva, "Multidimensional modeling based on spatial, temporal and spatio-temporal stereotypes," *ESRI International User Conference*, 2005.

[3] S. Chaudri and U. Dayal, "An overview of data warehousing and olap technology," *ACM Sigmod Record*, vol. 26, pp. 65–74, 1997.

[4] D. J. Peuquet, "A conceptual framework and comparison of spatial data models," *Cartographica*, vol. 21, no. 4, pp. 66–113, 1984.

[5] Y. B. et al, "Modeling geospatial databases with plug-ins for visual languages," *ER2004, LNCS 3289*, pp. 17–30, 2004.

[6] J. Mari and F. L. Ber, "Temporal and spatial data mining with second-order hidden markov models," *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, vol. 10, no. 5, pp. 406–414, 2005.

[7] L. Wang, K. Xie, T. Chen, and X. Ma, "Efficient discovery of multilevel spatial association rules using partitions," *Information and software technology*, vol. 47, no. 13, pp. 829–840, 2005.

[8] L. Z. Wang, "A method of the abstract generalization on the bases of the semantic proximity," *Chineese Journal of Computation*, vol. 23, pp. 1114–1121, October 2000.

[9] K. S. Rao, "K-means clusteringfor categorical attributes," Master's thesis, Indian Institute of Technology, December 1998.

[10] N. Stefanovic, J. Han, and K. Koperski, "Object-based selective materialization for efficient implementation of spatial data cubes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 12, no. 6, 2000.

[11] H. Gonzalez, J. Han, X. Li, and D. Klabjan, "Warehousing and analysis of massive rfid data sets," *International Conference on Data Engineering (ICDE'06)*, April 2006.

[12] W.-J. Lee and S.-J. Lee, "Discovery of fuzzy temporal association rules," *IEEE Transactions on Systems, Man and Cybernetics Part B*, vol. 34, pp. 2330–2342, December 2004.

[13] H. Ning, H. Yuan, and S. Chen, "Temporal association rules in mining method," *First International Multi-Symposiums on Computer and Computational Sciences*, vol. 2, pp. 739–742, 2006.

[14] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," *ACM SIGMOD International Conference on Management of Data*, pp. 207–223, 1993.

[15] N. Isik, "Fuzzy spatial data cube construction and its use in association rule mining," Master's thesis, Middle East Technical University, May 2005.

[16] D. Delic, H.-J. Lenz, and M. Neiling, "Improving the quality of association rule mining by means of rough sets," *First International Workshop on Soft Methods in Probability and Statistics SMPS 2002*, September 2002.

[17] J. Hipp, U. Guntzer, and G. Nakhaeizadeh, "Algorithms for association rule mining: A general survey and comparison," *ACM SIGKDD*, vol. 2, July 2000.

[18] C. Zhang and S. Zhang, *Association Rule Mining: Models and Algorithms*. Springer Berlin, Heidelberg, 2002.

[19] J. W. Seifert, "Data mining: An overview," tech. rep., Congressional Research Service of the Library of Congress, 2004.

[20] M.-T. Kechadi and M. Bertolotto, "A visual approach for spatio-temporal data mining," *IEEE International Conference on Information Reuse and Integration*, pp. 504–509, 2006.

[21] H. J. Miller and J. Han, *Geographic data mining and knowledge discovery: An overview*. Taylor and Francis, 2001.

[22] S. Shekhar and S. Chawla, *Spatial Databases: A Tour*. Prentice Hall, 2003.

[23] J. Mennis and J. Liu, *Mining Spatio-Temporal Information Systems*. Kluwer Academic Publishers, 2002.

[24] K. Koperski and J. Han, "Discovery of spatial association rules in geographic information databases," *4th International Symposium Advances in Spatial Databases*, vol. 951, pp. 47–66, 6–9 1995.

[25] M. Ester, A. Frommelt, H. Kriegel, and J. Sander, "Spatial data mining: Database primitives, algorithms and efficient dbms support," *Data Mining and Knowledge Discovery*, vol. 4, pp. 193–216, 7 2000.

[26] D. Malerba, F. Lisi, A. Appice, and F. Sblendorio, "Mining spatial association rules in census data: A relational approach," *Notes of the ECML/PKDD 2002 Workshop on Mining Official Data*, pp. 80–93, 2002.

[27] C. M. Kuok, A. Fu, and M. H. Wong, "Mining fuzzy association rules in databases," *ACM Sigmod Record*, vol. 27, pp. 41–46, 1998.

[28] W. Inmon, *Geographic data mining and knowledge discovery: An overview*. John Wiley, 1992.

[29] L. Savary, T. Wan, and K. Zeitouni, "Spatio-temporal data warehouse design for human activity pattern analysis," *15th International Workshop on Database and Expert Systems Applications*, pp. 814–818, 2004.

[30] M. Miquel, Y. Bedard, A. Brisebois, J. Pouliot, P. Marchand, and J. Brodeur, "Modelling multidimensional spatio-temporal data warehouses in a context of evolving specifications," *Symposium on Geospatial Theory, Processing and Applications*, 2002.

[31] A. Laurent, B. Bouchon-Meunier, and A. Doucet, "Towards fuzzy-olap mining," *Database Support for KDD*, pp. 51–62, 2001.

[32] R. Kimball, *The Data Warehouse Toolkit*. New York: John Wiley & Sons, 1996.

[33] "http://www.meteor.gov.tr," *Turkish State Meteorological Service*.

[34] J. Han and M. Kamber, *Data Mining: concepts and Techniques*. Morgan Kaufmann Publisher, Inc, 2001.

[35] D. W. Xie, "Fuzzy association rules discovered on effective reduced database algorithm," *IEEE Intl Conf on Fuzzy Systems*, pp. 779–784, 2005.

[36] "http://www.esri.com/software/arcgis/," *ArcGIS: The Complete Enterprise GIS*.