Fuzzy Rule-Based Intelligent System for Predicting Hotel Occupancy

Sara Salehi

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Applied Mathematics and Computer Science

Eastern Mediterranean University
January 2020
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Ali Hakan Ulusoy
Director

I certify that this thesis satisfies the requirements as a thesis for the degree of Doctor of Philosophy in Applied Mathematics and Computer Science.

Prof. Dr. Nazim Mahmudov
Chair, Department of Mathematics

We certify that we have read this thesis and that in our opinion, it is fully adequate, in scope and quality, as a thesis of the degree of Doctor of Philosophy in Applied Mathematics and Computer Science.

Prof. Dr. Rashad Aliyev
Supervisor

Examining Committee

1. Prof. Dr. Rashad Aliyev

2. Prof. Dr. Hamza Erol

3. Prof. Dr. Benedek Nagy

4. Prof. Dr. Efendi Nasiboğlu

5. Assist. Prof. Dr. Ersin Kuset Bodur

ABSTRACT

Accuracy and interpretability have always been significant issues in forecasting methods, and it is very important to have a balance between these issues when developing a system in tourism demand forecasting.

There are various clustering algorithms used in many branches. The efficiency of the clustering technique is stipulated by the performance of the clustering results. Fuzzy c-means algorithm is highly efficient for unbiased clustering. In this thesis, fuzzy c-means algorithm is applied on monthly number of guest arrivals in one of the hotels of North Cyprus over 40 months to find the optimal number of clusters in the analysis problem. Also, the fuzzy rule-based system model for hotel occupancy forecasting is developed, and in order to enhance the comprehensibility and accuracy of this model, Mamdani fuzzy rule-based system is used.

Based on the values of root mean square error and mean absolute percentage error which are metrics for measuring forecast accuracy, it is defined that the forecasting model with 7 clusters and 4 inputs provides an optimal solution of the problem.

Keywords: Forecasting, Time series, Fuzzy c-means clustering, Fuzzy rule-based system, Mamdani model

ÖZ

Doğruluk ve yorumlanabilirlik, öngörülme yöntemlerinde her zaman önemli konular olmuştur ve bu konular arasında dengeyi sağlamak, turizm talebinin öngörülmesi için sistemin geliştirilmesinde çok önemlidir.

Birçok alanda kullanılan çeşitli kümeleme algoritmaları mevcuttur. Kümeleme tekniğinin verimliliği kümeleme sonuçlarının performansı ile belirlenir. Bulanık c-ortalama algoritması tarafsız kümeleme için yüksek verimliliğe sahiptir. Bu tezde, bulanık c-ortalama algoritması, 40 ay boyunca Kuzey Kıbrıs'taki otellerden birine gelen misafir sayısı için uygulanır ve bu algoritmanın amacı, analiz probleminde kümelerin en uygun sayısını bulmaktır. Ayrıca, bulanık kural tabanlı sistem modeli, otel doluluk oranını tahmin etmek için geliştirilir ve bu modelin anlaşılırlığını ve doğruluğunu arttırmak için Mamdani kural tabanlı sistem kullanılır.

Tahmin doğruluk ölçümü için kullanılan ölçevler olan hataların ortalama kare kökü ve ortalama mutlak yüzde hatalarının değerlerine dayanarak, 7 kümeli ve 4 girişli tahmin modeli problemin en uygun çözümünün sağlanmasını tanımlar.

Anahtar Kelimeler: Tahmin, Zaman serisi, Bulanık c-ortalama kümeleme, Bulanık kural tabanlı sistem, Mamdani modeli

DEDICATION

To My Family

ACKNOWLEDGMENTS

I wish I could be able to express all of my appreciation to my supervisor Prof. Dr. Rashad Aliyev for his great support, help, and valuable advice at every time.

I am very glad to have a nice family, and I am always grateful to them for their continuous support at every stage of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

In general, tourism has been one of the fastest growing sectors of the economy. In 2017, tourism contributed approximately $7.9 trillion to the global economy which was around 10.2% of global gross domestic product. In the same year, based on statistics, one out of five new jobs was somehow related and linked to tourism [1].

It is broadly acceptable that promotion involving a considerable amount of investment needs predictions of market penetration and future demand. The promotion of tourism sector can be achieved through forecasting the number of tourists, and therefore, current and past tourism demands should be carefully analyzed.

Due to the competitive and complicated environment in tourism sector, it would be required to observe and enhance the previous standard of performances. To illustrate the importance of accuracy and validity in predicting, many empirical research papers have been compared which highlight the econometric modeling and forecasting methods and models [2].

One of the major components in the study of tourism sector is reaching reasonable forecast accuracy of tourism demand. So having an accurate plan would be crucial in tourism demand forecasting. As a result, not only some policy makers, managers and planners, but also researchers are concerned with tourism demand. Due to remarkable increase in international tourist numbers, both private and public sectors need access to an accurate forecasting method. Using forecasting results, managers and planners are able to make strategic decisions. Under inconsistent and unstable conditions, making a wrong decision could turn away tourists in peak periods. It is important that there should be a reasonable balance between computational overhead and accuracy of the system.

Various methods are available for estimation of tourism demand, and selecting the proper method needs to be based on time horizons and availability of data. Obviously, each method has strengths and drawbacks. These methods are mostly categorized into two main groups: quantitative and qualitative. In studying tourism demand, the use of quantitative methods has been increasing rapidly. It could be as a result of some factors such as the fact that resources have been dedicated to the collection of quantitative data and the maintenance of tourism data sets. Simultaneously, some destination managers and governmental employees in different sectors of the tourism industry prefer to have more reliable information and data, and are interested in the result of quantitative methods. As a result, the usage and application of quantitative methods have become more common in non-academic and academic fields [2,3].

By using mathematical rules, quantitative methods arrange the past information of tourism demand. Quantitative methods have some major subgroups such as time series methods, artificial intelligence methods, and econometric approaches. One of the quantitative approaches is time series analysis which deals with past data, and is applicable when there is no enough knowledge and information available on data generating procedure [4].

By comparing and testing the accuracy of different methods, researchers have found that time series methods give admissible forecasts with reasonable benefits and low cost. The main focus of time series models is based on using previous patterns and trends (such as monthly, seasonality, stationary, etc.) to have predictions based on the identification of the patterns and trends. Therefore, the time series method illustrates variables with respect to its random disturbance term as well as its own past. So the data collection and model estimation processes are reasonably cost efective [5].

There are various applications of classical time series models; however they have some limitations such as the need for assuming a formal model with a probability distribution [6]; the need for examining carefully whether the time series is nonstationary, because in this case, the data would have a stochastic trend which makes an imprecise forecasting result [7]; the reliance on piecewise linear functions of most classical models; the increase in time complexity due to testing and choosing the suitable functional form; the large amount of data required for accurate results [8]. Furthermore, the process of constructing the model is not based on any economic theory; so it can't be helpful for decision making, planning, and policymaking [9].

Fuzzy time series (FTS) was introduced by Song and Chissom to overcome the limitations of classical time series model. The major difference between FTS and traditional time series is that the values of FTS are fuzzy sets whereas the traditional method uses crisp numbers. Fuzzy models have the following advantages over conventional models: applying fuzzy models in complicated and optimization problems successfully [10]; having more capability to deal with nonlinear relationships [11]; being applicable when only a small amount of data is available;

using linguistic values rather than crisp values; dealing with incomplete data; being effectively applicable under unclear and uncertain circumstances [12].

FTS has strong potential to deal with linguistic values and historical data. Song and Chissom designed the forecasting model based on fuzzy logic theory [13]. Later on Song and Chissom modified the model in which the average forecasting error was reduced [14]. Using the same data from the proposed model of Song and Chissom, Chen modified the model by using simplified arithmetic operations instead of complex max-min composition operations to reduce overload computation remarkably [15]. These models are domain independent; however Huarng suggested the heuristic models to consider fluctuations in FTS and to improve forecasting results [16]. After that, Chen proposed the model by using the high order fuzzy time series [17] which was later extended and developed [18]. Fuzzy clustering algorithm for FTS was developed in [19], and this algorithm shows better performance compare to conventional algorithms. FTS method has been applied in tourism demand [20], stock markets [21], temperature prediction [22], etc.

After Zadeh introduced the theory of fuzzy sets [23], the general idea of fuzzy clustering was proposed in [24, 25]. In general, the problem of fuzzy clustering can be classified into three different groups: fuzzy rule learning, fuzzy relation, and optimization of an objective function. The idea of defining fuzzy clustering by optimizing the objective function was mentioned by Bezdek [26], and it was named fuzzy c-means clustering.

Fuzzy c-means as well as its extensions called possibilistic c-means and possibilistic fuzzy c-means algorithms are analyzed in [27].

Mamdani introduced fuzzy control systems by using fuzzy logic to clarify linguistic rules that indicate strategies in the control system qualitatively [28–30]. Since fuzzy rule-based systems work with fuzzy logic rather than classical logic rules, these systems are mentioned as an extension of the classical rule-based system. Because the Mamdani model is based on linguistic variables, this kind of fuzzy rule-based system is called a descriptive or linguistic system. It should be mentioned that formulating and modifying knowledge in this system is much easier than in other fuzzy rule-based systems.

The major aim of the Mamdani fuzzy rule-based system is to design a fuzzy system having the potential and capability to indicate the performance of the real system in a way which is comprehensible for human beings.

Unfortunately there is a contradiction between accuracy and interpretability while designing linguistic fuzzy systems. Various measurements for interpretability and different methods to have more understandable and interpretable fuzzy rule-based system are considered in [31]. Due to the important point of having a suitable balance between accuracy and interpretability, several genetic fuzzy systems are designed and tested based on Mamdani fuzzy rule-based system to obtain both accuracy and interpretability [32].

Chapter 2

LITERATURE REVIEW AND PRELIMINARIES

2.1 General Review

As a result of rapid increasing of scale of international tourism in the past few decades, this industry has attracted attention to research areas related to tourism. The most common techniques for tourism demand forecasting are the ones based on time series models, for instance, seasonal auto-regressive integrated moving average (SARIMA) and multivariate auto-regressive integrated moving average (MARIMA) [33], and those which are based on econometric models, for instance, vector autoregressive (VAR) [34], error correction model (ECM) and autoregressive distributed lag model (ADLM) [35].

In order to identify and distinguish four dimensions of performance in hotel occupancy in England over 279 hotels for two years from 1992 to 1994, time series factor analysis (TSFA) is used [36]. These four dimensions are overall occupancy level, seasonality occupancy, length of season and long term trend. It is noted that in the marketing policy, hotel occupancy could be the starting point but not an endpoint with the aim of having more accurate targeted forecasting and effective hotel marketing advantage.

In [37], three different models are built based on artificial neural network (ANN), ARIMA, and multivariate adaptive regression splines (MARS) for predicting the tourism demand in Taiwan. The result of this study demonstrates that the performance of ARIMA exceeds the performances of other two methods by computing the root mean square error (RMSE), mean absolute percentage error (MAPE), and mean absolute deviation (MAD).

In order to improve the accuracy in time series forecasting, the Artificial Neural Network (ANN) and ARIMA models are combined in [38]. The results show that the combined model has higher forecasting accuracy compare to the accuracy of each model applied separately.

In [39], it is noted that the combination of linear and nonlinear models can increase the forecasting accuracy. To test the accuracy, number of tourists from Taiwan is used. The performance and the accuracy of three individual and six combined models are compared. The result reveals that the combined models have the lowest error.

With the aim of increasing the accuracy of the forecasting system to predict the number of tourists in Hong Kong and Taiwan, a novel forecasting system is developed [40]. The system is designed by combining fuzzy c-means clustering with logarithm least-squares (LLS) support vector regression (SVR) technologies (LLS-SVR). For selecting the proper parameters of LLS-SVR in this system, genetic algorithm (GA) is used simultaneously.

The design process of forecasting system based on time series data has two major properties: there are large fluctuations and there is no enough historical data available. Therefore, a forecasting hybrid model is proposed in [41] for predicting the number of tourists from Taiwan to the U.S. by using adaptive fuzzy time series and particle swarm optimization (PSO). The reason for using PSO is to set the length of each interval in the domain of discourse to increase the accuracy of forecasting. The combination of ARIMA and PSO models with two main steps is considered in [42]; in the first step, the ARIMA model is used to extract useful information and pattern from the time series data; in the second step, by applying the PSO model, the parameters of ARIMA will be estimated for predicting the future values. The capability of the new model is tested on various time series data such as gold price variables and exchange rate. The results show that the new model can effectively improve the accuracy of the forecasting.

The following fuzzy time series based models are used to forecast the number of tourists in Taiwan: the first one is based on neural networks for dealing with nonlinear data [43], and the second one is combined with GA having the proper fuzzy intervals and minimum error [44].

The method for tourism demand forecasting by using fuzzy Takagi–Sugeno system is proposed in [45], and the rules for this system are extracted from trained support vector machine (SVM) techniques. Due to having symbolic structure in fuzzy rules and the ability of generalization in SVMs, the obtained fuzzy rules show high accuracy in forecasting and high level of comprehensibility for practitioners. The prediction is made for tourists' number travelling from nine different countries to China, and the proposed approach is compared with some other forecasting techniques.

In [46], a fuzzy system is designed based on the Mamdani-type fuzzy rule-based system, and the genetic algorithm technique is applied to obtain the rules. By using the GA technique, useful information patterns can be extracted. The model is capable to handle complexity, uncertainty, and non-linearity which can be in the set of actual data related to tourist arrivals. As a case study, the system is tested to forecast the arrival number of tourists to Taiwan.

2.2 Statement of the Problem

Although there is no guarantee to have an exact and precise forecasting method, the need for it is undeniable. The forecasting methods which are based on time series can use some or all available data to make a prediction of future values. However, using all available data is very complicated and expensive.

In competitive and complex operating environment, it is necessary to check out and to enhance some of the performances and fulfillments in hotel sector. Evaluating and comparing the performances of different hotels to raise their standards can be possible by effective hotel management. Occupancy rate is considered as the need for accommodation in the hotel during a specific point of time. Therefore, having information about the hotel occupancy rate would provide a suitable performance measure in addition to guide and help in making a good decision related to hotel management. So the success in enhancing hotel occupancy rate is a consequence of practicing good management. The data available for hotel occupancy can provide the great temporally and non-temporally dis-aggregated means of analyzing the performance of a hotel [47].

There is a general concurrence that in designing forecasting model, the combination of at least two methods or algorithms is preferred with the aim of increasing the accuracy, decreasing both complexity and calculation time.

Fuzzy clustering is used for the following purposes:

- to find center points of all clusters;
- to determine data structure;
- to define the fuzzy partition of data that is one of the main characteristics in fuzzy inference.

This thesis investigates the analysis of the monthly hotel occupancy with time series data taken from one of the hotels in North Cyprus over 40 months. By designing the system based on combination of fuzzy c-means clustering and Mamdani fuzzy rule-based system, we have an accurate system for forecasting the number of guest arrivals. As a major advantage of this combination, we gain some accuracy, interpretability, and the possibility of controlling and managing the total number of linguistic rules.

2.3 Preliminaries

Definition 2.3.1 (Fuzzy set)**.** Let U be the domain of discourse or universal set such that $U = \{u_1, u_2, \ldots, u_n\}$. A fuzzy set $\tilde{A}$ in U is a set of ordered pairs characterized by element u and its membership function $\mu_{\tilde{A}}(u)$.

$$\begin{aligned} \tilde{A} = \quad & \{(u, \mu_{\tilde{A}(u)}) | u \in U\} \\ & \mu_{\tilde{A}} : U \longrightarrow [0,1]. \end{aligned} \tag{2.1}$$

Definition 2.3.2 (Membership function)**.** It is a map that associates each element in U to a real number in the interval $[0,1]$. The value of membership function at u shows the ”grade of membership” or ”degree of membership” of u in $\tilde{A}$. Consequently, whenever the value of $\mu_{\tilde{A}}(u)$ is close to unity, then we have the high membership degree for u in $\tilde{A}$.

If $\tilde{A}$ is not a fuzzy set (crisp set), then $\mu_{\tilde{A}}(u)$ has only two values 1 or 0 which means u either belongs or does not belong to the set $\tilde{A}$, respectively.

The various types of membership functions are used: Triangle, Trapezoidal, Gaussian, Sigmoidal, Generalized Bell membership function, etc.

Definition 2.3.3 (Gaussian membership function)**.** The Gaussian membership function is more common due to its concise notation and smoothness compared to other available types of membership functions. The Gaussian membership function differs from the Gaussian probability distribution. The Gaussian membership function is defined as follows:

$$gaussian(u;c,\sigma) = \mu_{\tilde{A}(u)} = e^{\frac{-(c-u)^2}{2\sigma^2}} \tag{2.2}$$

where c is the center of the fuzzy set $\tilde{A}$ and σ is its width.

Definition 2.3.4 (Logical operation: Union, Intersection)**.** Let $\tilde{A}$ and $\tilde{B}$ be two fuzzy sets with their membership functions $\mu_{\tilde{A}}(u)$ and $\mu_{\tilde{B}}(u)$, respectively. The union of $\tilde{A}$ and $\tilde{B}$ is the fuzzy set $\tilde{C}$, and $\tilde{C} = \tilde{A} \cup \tilde{B}$ is defined as follows:

$$\mu_{\tilde{C}}(u) = \max\left(\mu_{\tilde{A}}(u), \mu_{\tilde{B}}(u)\right) \qquad u \in U \tag{2.3}$$

The intersection of $\tilde{A}$ and $\tilde{B}$ is the fuzzy set $\tilde{C}$, and $\tilde{C} = \tilde{A} \cap \tilde{B}$ is defined as follows:

$$\mu_{\tilde{C}}(u) = \min\left(\mu_{\tilde{A}}(u), \mu_{\tilde{B}}(u)\right) \qquad u \in U \tag{2.4}$$

Definition 2.3.5 (Fuzzy time series)**.** Let $U(t)$ be a universal set such that $U(t) \subset \mathbb{R}^1$ where $t = \ldots, 0, 1, 2, \ldots$ and $\mathbb{R}^1$ is representing a set of real numbers on which fuzzy sets $f_k(t)$, $k = 1,2,3,\ldots$ are defined. Fuzzy time series $F(t)$ is defined as a collection of $f_k(t)$ on $U(t)$ such that

$$F(t) = \{f_k(t)\}_{k \in I} = \{f_1(t), f_2(t), f_3(t), \ldots\} \tag{2.5}$$

From the equation (2.5), it can be seen that $F(t)$ depends on t (time), because the universal set may be different at different times.

Definition 2.3.6 (Rule-based system (RBS))**.** RBSs are used to convert a knowledge of a human expert, within a specific area, into a system in an automated way. The representation of knowledge in such systems is based on a set of rules. By setting and defining different number of rules, the system is able to make a suggestion in different situations about what to decide or what to do in the next step.

The rules are in the form of IF-THEN statements which are called production rules. The rules have the following structure:

$$\text{IF } P \text{ THEN } Q \tag{2.6}$$

where IF part is called a premise (or antecedent) and THEN part is called a conclusion (or consequent). The necessity of designing the RBS is to express the knowledge on a domain of a certain problem in the form of production rules. If the problem domain is too large, then the number of rules would increase dramatically and cause the inefficient and complicated system. The primary elements in RBS are: a set of facts, a set of rules, and a termination criteria. The facts are normally referred to any assertions and declaration and can be considered as a collected groups of data and conditions. The rules interact with data indirectly, with either multiple or single conditions. All actions in the system are defined by the set of IF-THEN rules. To increase the performance of the system, irrelevant rules should be avoided [48].

Definition 2.3.7 (Fuzzy rule-based system (FRBS))**.** This system is considered as an evolvement of conventional RBS. Therefore the main distinction between them is that instead of using classical logic statements in classical RBS, the FRBS uses fuzzy logic in both premise and conclusion parts of IF-THEN rules.

The linguistic variables are used in fuzzy rules of FRBS. The universal set of the linguistic variables is represented by the range of possible values for them. A fuzzy IF-THEN rule is represented as follows:

$$\text{IF } a \text{ is } \tilde{A} \text{ THEN } b \text{ is } \tilde{B} \tag{2.7}$$

where a and b are input and output linguistic variables, respectively. $\tilde{A}$ and $\tilde{B}$ are representing linguistic values defined by fuzzy sets on the universal sets Φ and U, respectively.

Each fuzzy rule should have more than one premise, and all of the calculations are done in parallel by using fuzzy set operations to find a single number as an output. It is also possible to have multiple parts in the consequent and all of them are equally affected by the premise. However, it is desired to have an exact solution (not a fuzzy one) as an output of the system.

Chapter 3

FUZZY C-MEANS FOR DATA CLUSTERING

3.1 Fuzzy Clustering

Clustering is a principal technique for extracting some knowledge from a data set. There is a challenging problem when a data set is erroneous, incomplete, and uncertain for the traditional methods of clustering because these methods have been proposed for analyzing and examining complete data sets.

Clustering is normally used as a significant data mining technique to know about how data points are distributed in a data set. The reason of using clustering is a partition of considered data set into some groups (clusters). The data points in each cluster are as similar as possible. Accordingly, data points in different clusters are as dissimilar as possible.

In conventional data clustering techniques, each data point is in exactly one cluster; so the results of clustering will be less informative and there would be a problem with boundary data points. Some information related to clustering structure will be lost if there exist any overlapping clusters in data set. To overcome the mentioned drawbacks, in fuzzy clustering technique, each data point is assigned to each cluster with the membership degree.

The membership value 1 shows the certain allocation of the data point to its belonging cluster while as the membership value 0 shows that the data point can't be the member of that cluster. The data points which are in the overlap of clusters have the membership values that are roughly equal to the overlapping clusters. Therefore, in comparison with a model based on hard clustering, a model based on fuzzy clustering can be more understandable to human perception. Additionally, by examining the results of fuzzy clustering, the information about overlaps between clusters as well as the structure of clustering can be derived.

To sum up, the results from fuzzy clustering are useful for [49]:

i. the interim result that will be used for the processing in the future;

ii. the output of the process in data mining.

In the next section, the algorithm of fuzzy c-means clustering is explained.

3.2 Fuzzy C-Means Clustering

Fuzzy c-means (FCM) clustering is an unsupervised clustering method for classifying a set of objects (elements) based on their similarity. In the concepts of fuzzy clustering and FCM, the membership functions of clusters are defined based on a distance function, and as a result, the membership degrees indicate proximity of each data point to the different cluster centers (multi-cluster centers).

Let $\mathbb{R}^p$ be defined as a set of p tuples of real numbers ($\mathbb{R}$), given a set X, where $X = \{x_1, x_2, \ldots, x_n\} = \{x_j\}_{j=1}^{n}$ which is finite such that $X \subseteq \mathbb{R}^p$, and c is an integer representing the total number of clusters where $c \in \{2, 3, \ldots, n\}$. There exists a matrix A which represents a fuzzy c partition of the set X such that $A \in M_{cn}$. The M_{cn} is defined as the set of real matrices with real number entries and with the size $c \times n$. The matrix $A = [a_{ij}]$ is named a partition matrix and the fuzzy clusters are characterized by it. The entries of matrix A have the following conditions [26]:

i. A_i and A^j represent the ith row and jth column of A such that $A_i = (a_{i1}, \ldots, a_{in})$ and $A^j = (a_{1j}, \ldots, a_{cj})$. A_i and A^j are ith membership function of X and the value of c membership function of jth observation (datum) in the set X, respectively. In other words, A_i is determined as ith fuzzy subset of X;

ii. a_{ij} is defined as $a_i(x_j)$ which is the membership degree of the ith fuzzy subset for the jth datum, i.e., it is the membership degree for x_j in the ith cluster;

iii. $a_{ij} \in [0,1] \qquad \forall i, j$;

iv. $\sum_i a_{ij} = 1 \qquad \forall j$;

v. $\sum_j a_{ij} > 0 \qquad \forall i$;

vi. $\sum_j a_{ij} < n \qquad \forall i$.

FCM clustering uses an optimization algorithm with the aim of minimizing an objective function J_m [26, 50]:

$$J_m(A, v) = \sum_{j=1}^{n} \sum_{i=1}^{c} a_{ij}^m \|x_j - v_i\|^2 \rightarrow min \tag{3.1}$$

where A is a fuzzy c partition of X, $v = \{v_i\}$, $i = 1, 2, \ldots, c$; the center of each cluster is represented by v_i where $v_i \in \mathbb{R}^p$; $\| * \|$ is any kind of inner product norm metric, and m is defined to control the degree of fuzzy overlap $m \in (1, \infty)$.

From the definition of FCM it would be clear that the object may be in more than one cluster based on the value of membership degree which is the major difference between fuzzy clustering and hard clustering. This is due to the fact that in fuzzy clustering we have a range from 0 to 1 for membership degree rather than selecting 0 or 1.

The main steps of FCM algorithm are given below:

1. Select the value for parameters such as c, n, and m, $c \in [2, n-1)$ where n is the total number of objects or data items, $m \in (1, \infty)$, select also any inner product $\| * \|$, e.g. the Euclidean distance between cluster center and data item.

2. Initialize a_{ij}.

3. Compute cluster centers v_i as follows:

$$v_i = \frac{\sum_{j=1}^{n} (a_{ij})^m x_j}{\sum_{j=1}^{n} (a_{ij})^m}, \quad 1 \leq i \leq c. \tag{3.2}$$

4. Update a_{ij} as follows:

$$a_{ij} = \left[\sum_{k=1}^{c} \left(\frac{\|x_j - v_i\|}{\|x_j - v_k\|} \right)^{\frac{2}{m-1}} \right]^{-1}. \tag{3.3}$$

5. Calculate J_m by equation (3.1).

6. Repeat steps 3-5 until the special number of iterations is reached or the difference between the values of J_m in the last two steps is less than δ, where δ is the minimum threshold.

As a result of applying the fuzzy clustering method, there would be c number of local behaviors of the available data around the clusters' centers, i.e. fuzzy relations captured by behaviors. It should be noted that in FCM clustering technique it is possible to use any other kind of methods to measure the distances since the main point of clustering here is to define the position of centroids.

Some implementations of FCM clustering are described below:

– reformulating and changing some criteria in FCM algorithm to create a specialized algorithm and then comparing it with the original algorithm [51];

- modifying the original algorithm to take turns estimating the fuzzy membership degrees and clusters' centers; so the new modification causes the decrease in time complexity [52];
- editing the original algorithm for matching the desired level of parallelism, decreasing complexity of time as well as making the hardware implementation become more simpler [53].

3.3 Application of Fuzzy C-Means for Data Clustering

The methodology is based on fuzzy data mining and fuzzy approximate reasoning. Knowledge mining is based on data-driven approach. For this purpose, we use fuzzy clustering. Fuzzy clustering is performed by using FCM technique. This technique is used to find the center points and the interval of each partition in the universal set.

To forecast the number of guest arrivals, the system is designed based on combination of FCM clustering and Mamdani FRBS. To prove the efficiency and accuracy of this system, the data on monthly number of guest arrivals over 40 months are collected from one of the hotels in North Cyprus (Figure 3.1).

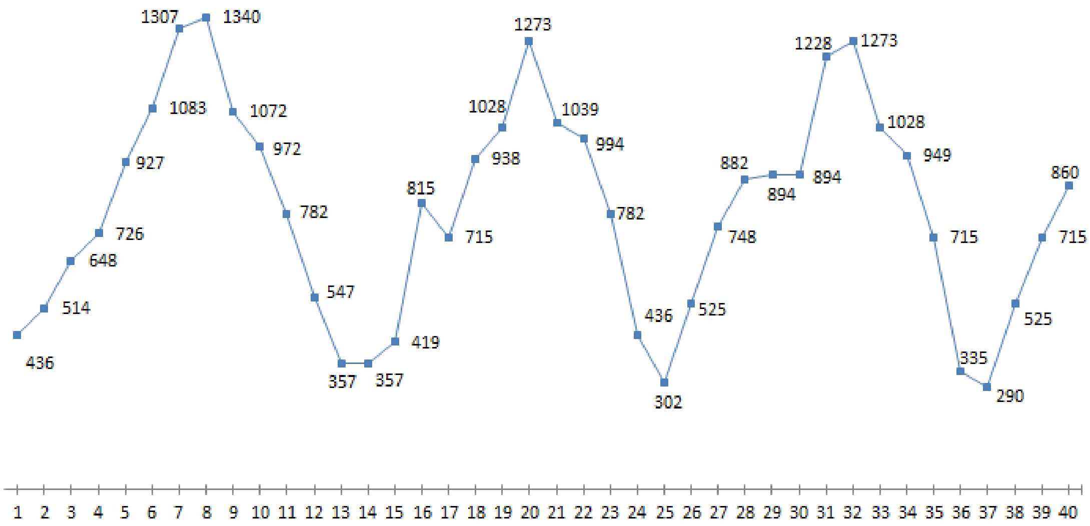


Figure 3.1: Number of guest arrivals in one of the hotels of North Cyprus over 40 months

The first step is clustering the data, and FCM clustering is selected with parameters shown in Table 3.1. The data are categorized (clustered) into different number of clusters $c = 5, 6, 7$. The results of clustering are shown in Tables 3.2 – 3.4. The data in each table show the coordinates of the cluster centers [50, 54].

Table 3.1: Parameters of fuzzy c-means clustering

Parameters	**Value**
m	2
Maximum number of iterations	25
δ	0.001
n	40
c	5, 6, 7

Table 3.2: The coordinates of the cluster centers with 5 clusters.

Cluster Centers	
1.0e+03 * v_i	
v_1	(0.0193 , 1.2849)
v_2	(0.0208 , 0.7155)
v_3	(0.0186 , 1.0334)
v_4	(0.0255 , 0.8945)
v_5	(0.0201 , 0.4005)

Table 3.3: The coordinates of the cluster centers with 6 clusters.

Cluster Centers	
1.0e+03 * v_i	
v_1	(0.0185 , 1.0371)
v_2	(0.0214 , 0.7348)
v_3	(0.0193 , 1.2854)
v_4	(0.0245 , 0.3333)
v_5	(0.0251 , 0.9033)
v_6	(0.0176 , 0.5095)

Table 3.4: The coordinates of the cluster centers with 7 clusters.

Cluster Centers	
1.0e+03 * v_i	
v_1	(0.0177 , 0.5081)
v_2	(0.0235 , 0.9150)
v_3	(0.0246 , 0.3327)
v_4	(0.0180 , 0.7923)
v_5	(0.0227 , 0.7139)
v_6	(0.0193 , 1.2857)
v_7	(0.0185 , 1.0407)

The data are divided into two groups of sets: testing set and training set. The testing set includes 30% of data and the training set includes the remaining 70% of data.

Clearly, the partition matrix that is emerged from considered monthly number of guest arrivals over 40 months, for the clustering with number of clusters $c = 7$ has the dimension 7×40. In Table 3.5, the partition matrix related to the testing data (30% of data) is shown.

Table 3.5: Partition matrix related to the testing data.

	894	**894**	**1228**	**1273**	**1028**	**949**	**715**	**335**	**290**	**525**	**715**	**860**
c_1	0.0028	0.0029	0.8504	0.9883	0.0053	0.0090	0.0004	0.0001	0.0019	0.0011	0.0008	0.0093
c_2	0.0029	0.0030	0.0057	0.0005	0.0013	0.0052	0.0034	0.0044	0.0385	0.9467	0.0058	0.0136
c_3	0.0201	0.0206	0.0836	0.0059	0.9549	0.1176	0.0014	0.0003	0.0033	0.0025	0.0024	0.0509
c_4	0.0014	0.0014	0.0037	0.0004	0.0007	0.0027	0.0010	0.9932	0.9341	0.0178	0.0017	0.0061
c_5	0.9177	0.9159	0.0301	0.0025	0.0277	0.8062	0.0037	0.0004	0.0047	0.0043	0.0062	0.5096
c_6	0.0134	0.0137	0.0111	0.0010	0.0036	0.0184	0.9667	0.0009	0.0103	0.0184	0.9441	0.0778
c_7	0.0416	0.0425	0.0155	0.0014	0.0064	0.0410	0.0234	0.0006	0.0073	0.0092	0.0391	0.3328

The rows are labeled by clusters 1 to 7 represented as from c_1 to c_7, and the columns are labeled by the number of guest arrivals in each month. For example, in the first column of the matrix the number of guest arrivals is 894; this number has the highest probability (approximately 91%) to be in the cluster c_5, and lower probabilities to be in other clusters: 0.28% in cluster c_1, 0.29% in the cluster c_2, and so on.

Chapter 4

FUZZY RULE-BASED SYSTEMS

4.1 Background and Motivation for Fuzzy Rule-Based Systems

One of the significant applications of fuzzy set theory is fuzzy rule-based systems (FRBSs). Conventional rule-based systems use bivalent logic for representing knowledge, and have such crucial disadvantage as being improper and inappropriate in uncertain situations. As a result, classical methods are not able to provide a suitable and reasonable framework which is familiar to human being. By using fuzzy logic and fuzzy statements in FRBSs, it would be possible to handle and capture the uncertainty that make inference methods more flexible and robust. There is a common point in all fuzzy inference applications that different factors and parameters must be available before finding a final conclusion. In such problems, the input and output relations as well as relations between inputs are tricky and complex. Therefore, there exists a difficulty for formulating indirect relations or interactive relations.

In general, fuzzy inference systems, and in particular, FRBSs are useful in the following cases: they simplify the process; they make the computing process faster; they give a result that is accurate enough; IF-THEN rule-based fuzzy inference systems should be designed by practical experience. Since fuzzy inference system would be suitable and helpful in the field of decision making, its applications are found in a wide area of different problems related to uncertainty and vagueness such as industry, robotics, business, economy, etc [55, 56].

One of the significant features of FRBSs is having two levels for knowledge representation. The lower level defines the semantics of FRBSs which includes defining the fuzzy sets in terms of their membership functions, and defining the aggregation function to have the final result (conclusion), i.e. this level is responsible for input/output mapping. On the higher level, rules are representing knowledge. Formal structure is defined by rules, and in this level, linguistic variables are defined and connected by different types of formal operators such as AND, OR, THEN, etc. The inputs/outputs of the system correspond to linguistic variables. The assumption values which are mapped to linguistic terms, are mapped to fuzzy sets which are defined in the first stage of knowledge representation. In the same way, the operators are mapped to the aggregation functions.

For comprehensibility, Michalski mentioned that it is better to describe computer induction in a symbolic way. Symbols are required for communicating knowledge and information since pure numerical models are not suitable to meet the interpretability [57].

A key factor of FRBS is interpretability for which some of the main components are represented below:

1. **Integration.** In an interpretable model of FRBS, the acquired knowledge is verified and related to the domain knowledge of an expert easily. In particular, verifying whether the obtained knowledge can express new relations about data, and controlling the possibility that the obtained knowledge can be integrated and improved with expert knowledge, are important.
2. **Interaction.** By using natural languages, it is possible to have interaction

between a model and a user. Practically, it should be done by editing the existing rules or adding new ones (at the symbolic level) by modifying and improving fuzzy sets represented by linguistic terms, or in the worst case, adding new linguistic terms to represent new fuzzy sets (at the numerical level).

3. **Validation.** The validation of obtained knowledge against domain-specific and common-sense knowledge is possible in a simple way. Due to this characteristic of FRBS, noticing semantic inconsistencies would be enabled. The inconsistencies cause different problems such as misleading data, overfitting the data in the inductive procedure. Detecting this kind of inconsistency is essential to have an improvement of the obtained knowledge in a qualitative way to derive the inductive procedure.

4. **Trust.** This is a capability of convincing the end user about the accuracy and reliability of the model. The advantage of FRBS is the capability to explain its inference process; so the user should be assured about the way of having the final output. This is especially important in the field of medical diagnosis [58].

4.2 Mamdani Fuzzy Rule-Based Systems

A procedure of mapping given input variables to an output space by using fuzzy logic is defined as fuzzy inference. In Mamdani FRBS, both premise and conclusion parts include linguistic variables.

Therefore, considering the system with multiple inputs and single output, fuzzy IF-THEN rules have the form:

$$\text{IF } x_1 \text{ is } A_1 \text{ and} \ldots \text{and } x_n \text{ is } A_n \text{ THEN } y \text{ is } B \tag{4.1}$$

where each x_i represents an input linguistic variable, y is the output, and both B and each A_i represent the linguistic values associated with the corresponding linguistic variables [54].

Mamdani FRBS is the fuzzy inference system which includes five main phases described below.

4.2.1 Fuzzification of Input Variables

The first phase of Mamdani FRBS starts with the transformation of the crisp input variables (numerical values) into linguistic values via a membership function in order to determine the membership degree of each input related to the proper fuzzy sets. Therefore, through the first phase, the output would be the membership degree within the interval $[0,1]$ in the corresponding linguistic fuzzy sets. The crisp inputs are limited by the domain of discourses which are determined by an expert. In this method, by using the fuzzy rules, each of the input variables is fuzzified over all of the membership functions.

4.2.2 Rule Evaluation

In the second phase of Mamdani FRBS, the fuzzified inputs are taken from the first phase and then applied to the fuzzy rules. In the system with multiple input variables, the premise of IF-THEN rules should be described by multiple fuzzy linguistic sets. Therefore, the fuzzy operator is required to combine the membership values to have only one single number which is representing the premise evaluation's result.

4.2.3 Performing Implication Method

The conclusion part in fuzzy IF-THEN rules represented in equation (4.1) is also a fuzzy linguistic set which is defined by appropriate membership function. This phase is used to calculate the output or final result of each fuzzy rule's conclusion/consequent.

By applying this phase on each rule, the following should be summarized:

- Input: the single number provided with the premise.
- Output: fuzzy set.

For performing implication method, min operator is used which is defined in equation (2.4).

Before explaining the next phase, the following clipping and scaling methods should be considered.

Clipping/correlation minimum: In order to associate the consequent with the truth value (logical value) of the premise in IF-THEN rules, membership function of consequent is clipped (cut) at the level of premise's truth value. Since the membership function is clipped from the top, some information will be lost by a clipped fuzzy set. However, this method is used, because it is simple, mathematically fast, and easy for defuzzification process represented in fifth phase of Mamdani FRBS.

Scaling/correlation product: The rule consequent's membership function is originally adjusted by product of all its membership values by premise's truth value.

4.2.4 Performing Aggregation Method

To be able to make a decision, in this phase the outputs are unified in all rules.

- Input: the list of either clipped or scaled conclusion membership functions.
- Output: fuzzy set for each of the output variables.

For performing aggregation method, max operator is used which is defined in equation

(2.3).

4.2.5 Defuzzification Process

In this phase, the output is a single crisp number which is defined by reverse process of fuzzification.

- Input: the combination of fuzzy set outputs from the aggregation phase.
- Output: single crisp number.

As mentioned in [59], there are various methods to be used for the defuzzification process; however the most common among them is the centroid method. The formula for the center of area (COA) is represented as follows:

$$\text{COA} = \frac{\int_{\alpha}^{\beta} \mu_{\tilde{A}}(u) u \, \mathrm{d}u}{\int_{\alpha}^{\beta} \mu_{\tilde{A}}(u) \, \mathrm{d}u} \tag{4.2}$$

where $\tilde{A}$ represents the aggregated fuzzy set and $\mu_{\tilde{A}}(u)$ is its membership function with respect to u.

To describe all the steps and structure of Mamdani FRBS, let's consider an example with two inputs and one output. Temperature and humidity are input variables which are taken through fuzzy reasoning process, and the rules are given below:

- Rule 1: IF x is A_1 and v is B_1 THEN z is C_1.
- Rule 2: IF x is not A_1 THEN z is C_2.
- Rule 3: IF x is A_1 and v is B_2 THEN z is C_3.

where x, v, and z (*temperature, humidity, climate*) are linguistic variables; A_1 *(mild)* is a linguistic value which is determined by a fuzzy set on the universal set X (*temprature*);

B_1 and B_2 (*high, low*) are linguistic values which are determined by fuzzy sets on the universal set V (*humidity*); C_1, C_2, and C_3 (*harsh, livable, comfortable*) are linguistic values which are determined by fuzzy sets on the universal set Z (*climate*).

The result from these three IF-THEN rules are combined and transformed to find a crisp number about the level of climate comfortability.

For above example, Figures 4.1, 4.2, and 4.3 illustrate fuzzification phase of Mamdani FRBS, working principle of fuzzy min operator, and implication method in Mamdani FRBS, respectively. Figure 4.4 illustrates all phases of Mamdani FRBS.

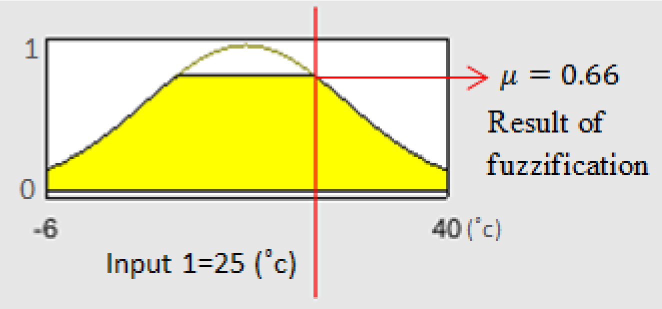


Figure 4.1: Fuzzification phase of Mamdani FRBS

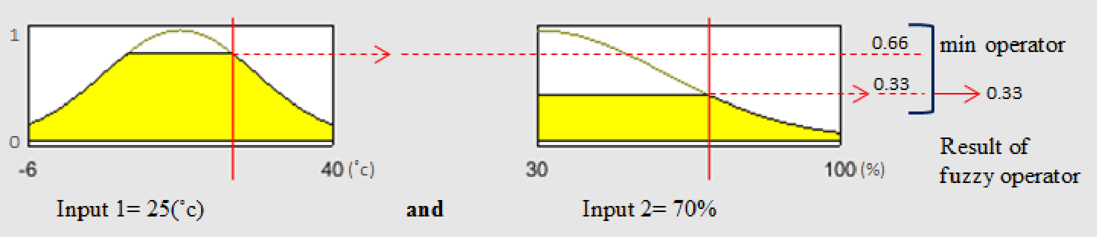


Figure 4.2: Working principle of fuzzy min operator

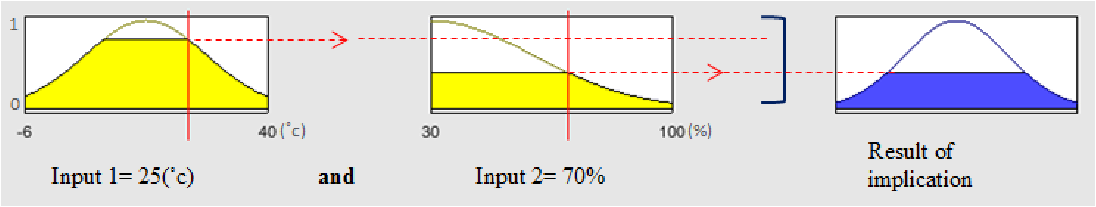


Figure 4.3: Implication method in Mamdani FRBS

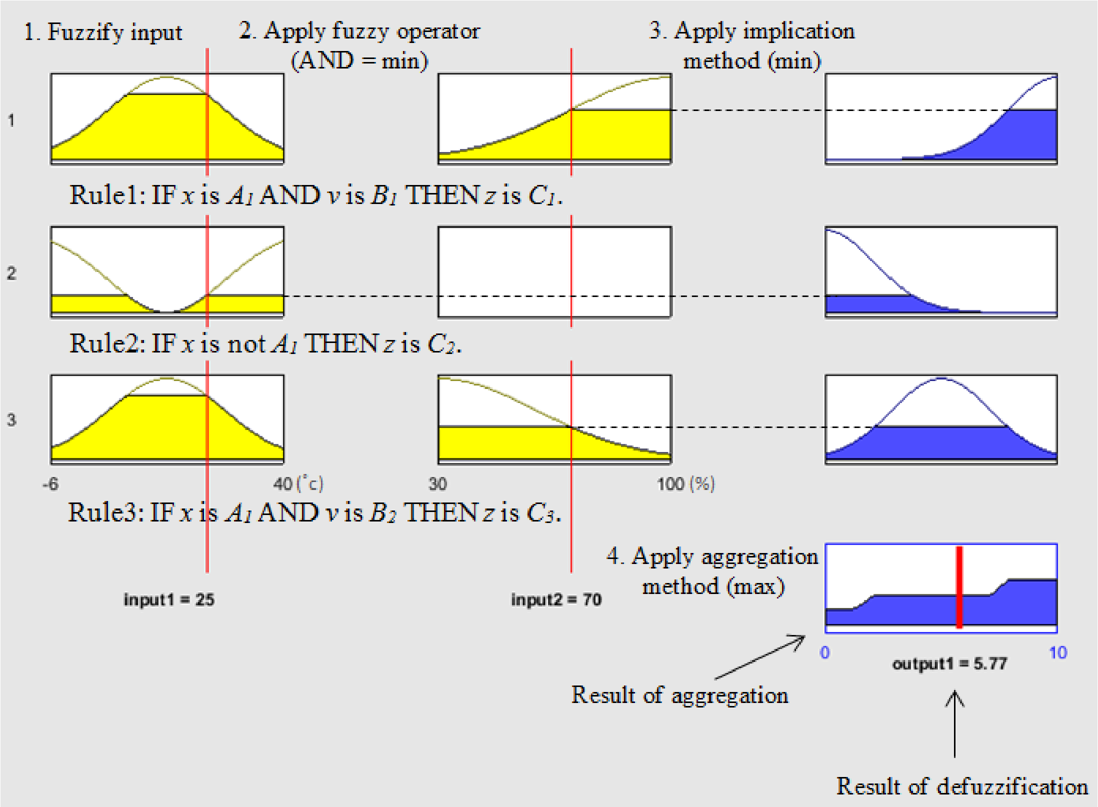


Figure 4.4: All phases of Mamdani FRBS

4.3 Application of Mamdani Fuzzy Rule-Based System for Hotel Occupancy Forecasting

We obtain Mamdani fuzzy rule-based model which approximates a dynamical relationship between forecasted value and previous value of a time series. Mamdani method is used to perform the approximate reasoning that enables to calculate the forecasting value. The describtion of the methodology is given below [54]:

1. Create IF-THEN rules by using FCM approach which is discussed in Section

3.2;

2. Choose the optimal number of clusters and inputs;

3. This procedure is realized by using the Fuzzy Logic Toolbox in Matlab software (R2015a, MathWorks, Natick, Massachusetts, USA);

4. Use Mamdani reasoning approach to obtain the forecasted value on base of given current inputs. For this purpose, we use Fuzzy Inference System Editor (FIS Editor) in Matlab software;

5. The value of fuzzy output for each rule is calculated. As implication operator the following min operator is used:

$$\mu_{B'_i}(X_t) = \min\left(\mu_{A_{i_1}}(X_{t-1}), \ldots, \mu_{A_{im}}(X_{t-m}), \mu_{B_i}(X_t)\right) \tag{4.3}$$

6. The calculated fuzzy outputs of all rules are aggregated by using max operator:

$$\mu_B(X_t) = \max_{i=1,\ldots,n} \mu_{B'_i}(X_t) \tag{4.4}$$

To find a system with high accuracy, different values for the parameters of the system are tested and examined. There are nine different Mamdani FRBSs which are designed with different parameters: different numbers of clusters $c = 5,6,7$ and different number of inputs $x = 2,3,4$. Clearly, the systems have neither same results nor same accuracies.

In Figures 4.5, 4.6, and 4.7, the forecasting results of Mamdani FRBSs with $c = 5,6,7$ clusters using testing data are demonstrated with $x = 2$, $x = 3$, and $x = 4$ inputs, respectively. In order to find the accuracy of the systems, the data in testing sets are used. As an accuracy measure, the mean absolute percentage error and the root mean square error are used.

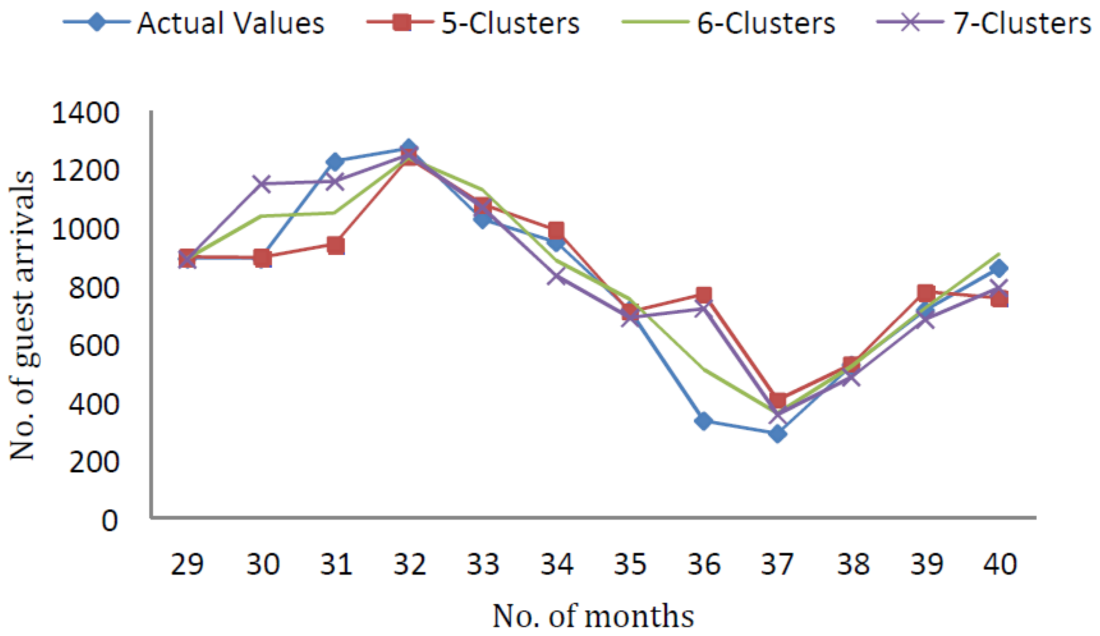


Figure 4.5: Forecasting results of Mamdani FRBSs using testing data with $c = 5, 6, 7$ clusters and $x = 2$ inputs

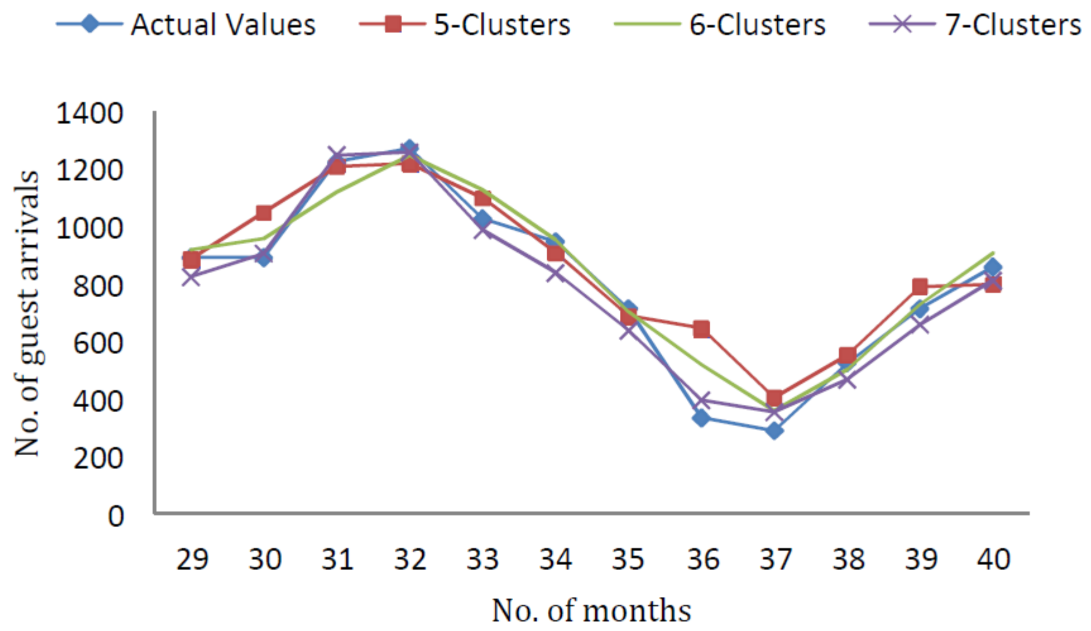


Figure 4.6: Forecasting results of Mamdani FRBSs using testing data with $c = 5, 6, 7$ clusters and $x = 3$ inputs

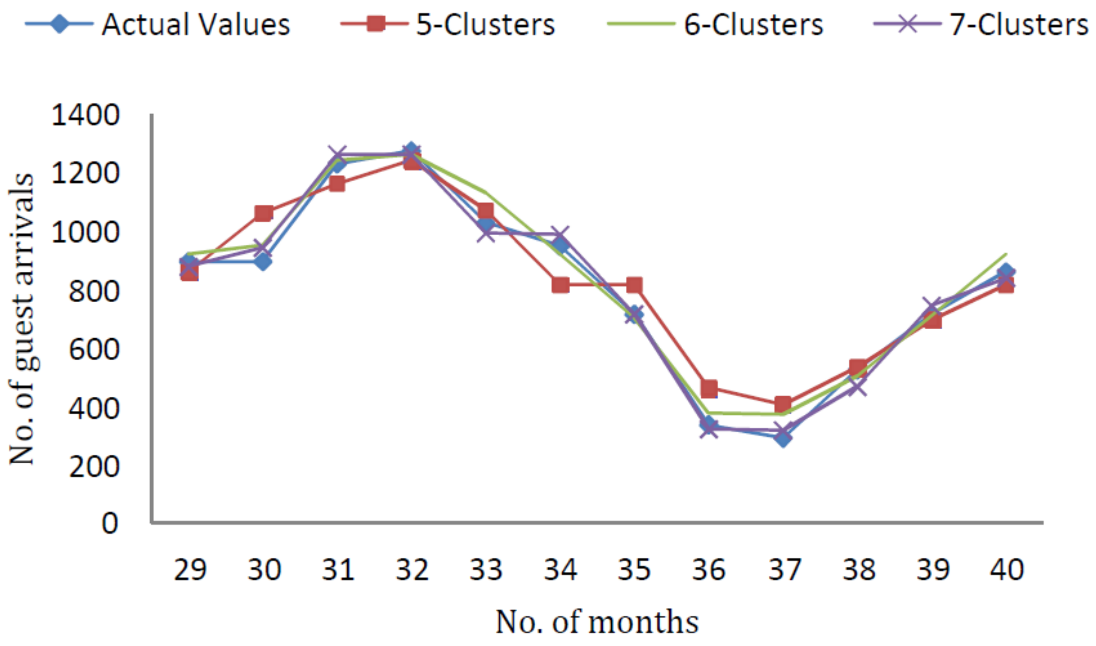
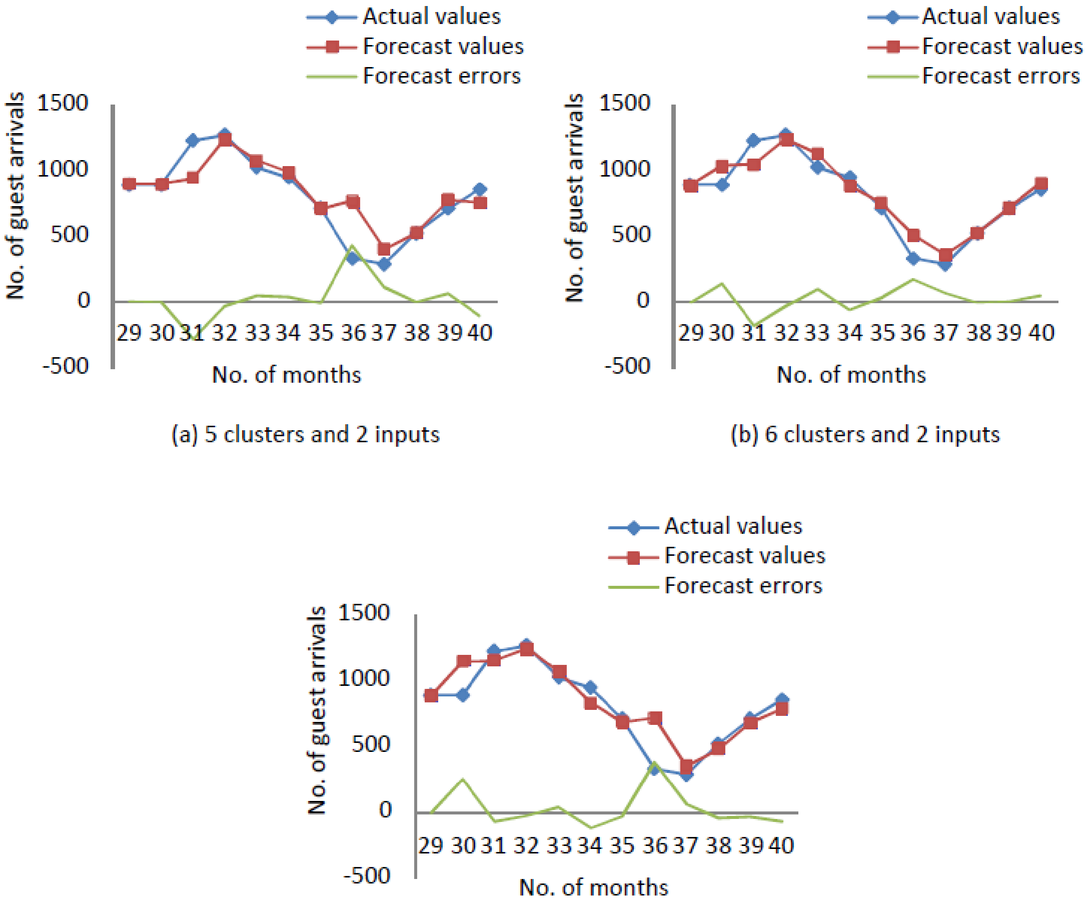


Figure 4.7: Forecasting results of Mamdani FRBSs using testing data with $c = 5,6,7$ clusters and $x = 4$ inputs

To compare each system with the original data, the point to point errors are shown in Figures 4.8, 4.9, and 4.10.

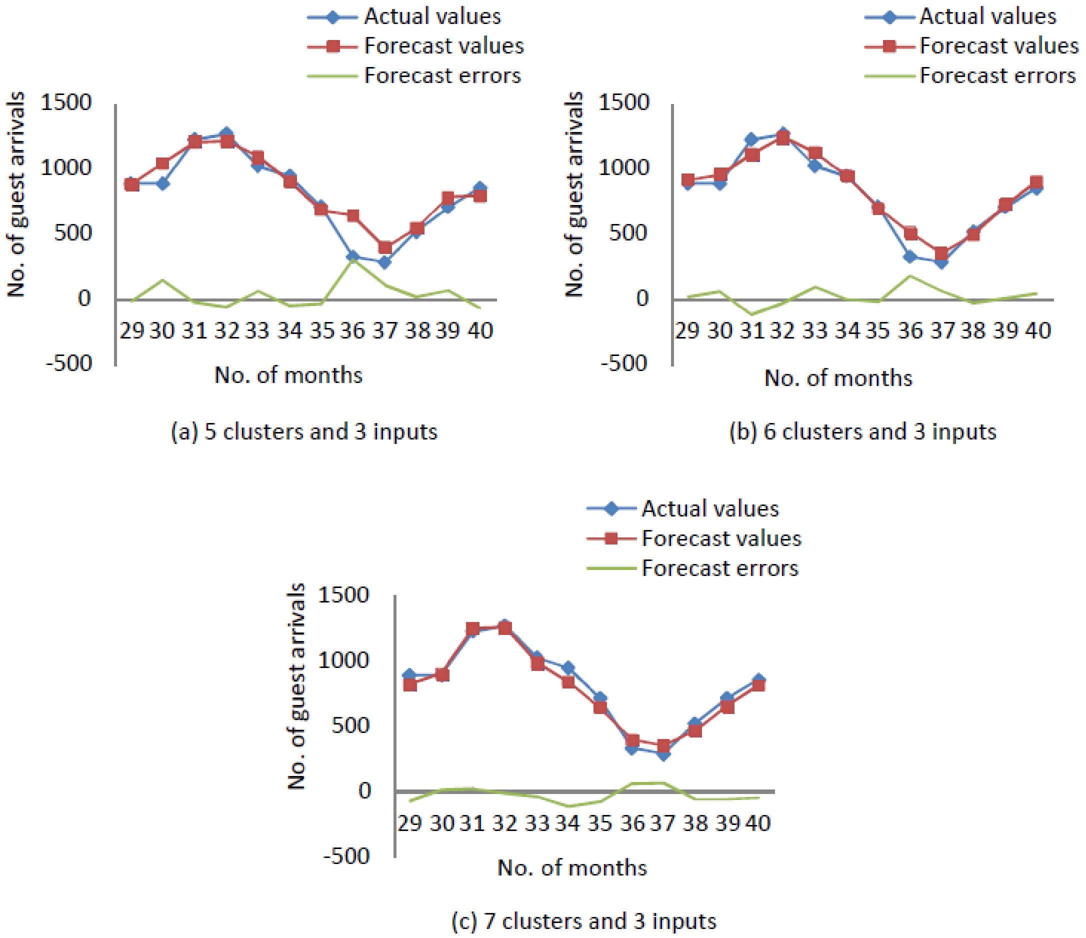


(a) 5 clusters and 2 inputs

(b) 6 clusters and 2 inputs

(c) 7 clusters and 2 inputs

Figure 4.8: Actual values, forecast values and forecast errors of a time series with $c = 5, 6, 7$ clusters and $x = 2$ inputs

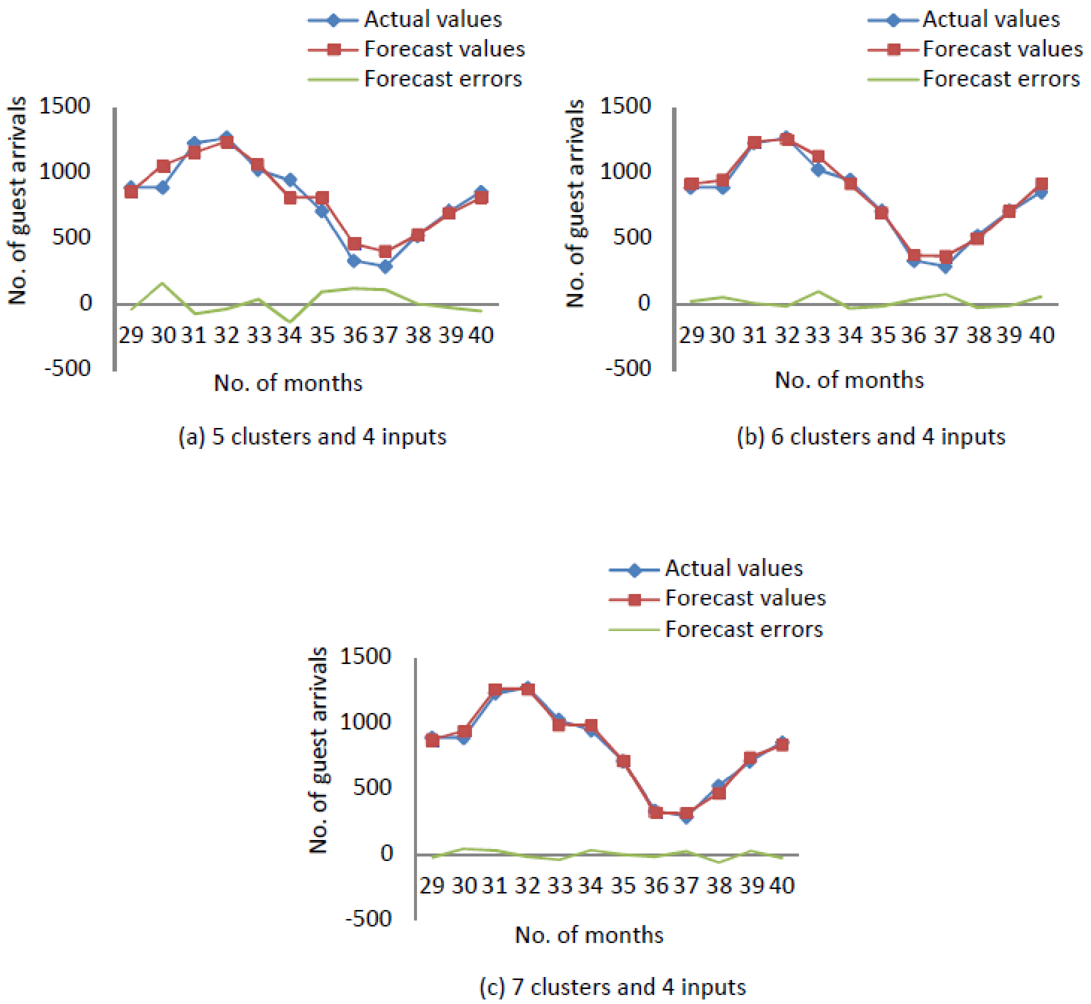


(a) 5 clusters and 3 inputs

(b) 6 clusters and 3 inputs

(c) 7 clusters and 3 inputs

Figure 4.9: Actual values, forecast values and forecast errors of a time series with $c = 5, 6, 7$ clusters and $x = 3$ inputs



(a) 5 clusters and 4 inputs

(b) 6 clusters and 4 inputs

(c) 7 clusters and 4 inputs

Figure 4.10: Actual values, forecast values and forecast errors of a time series with $c = 5, 6, 7$ clusters and $x = 4$ inputs

By looking through the Figures 4.8, 4.9, and 4.10 carefully which show the comparison of point to point errors between actual values and forecasting values in nine systems, it becomes clear that the system with $c = 7$ clusters and $x = 4$ inputs has the lowest error compare to all other systems.

The input and output variables in Mamdani FRBS are usually represented by some linguistic terms instead of quantity terms. Figure 4.11 illustrates the Gaussian membership function plot and homogeneous fuzzy partitions with 7 clusters and 4 inputs.

An advantage of using linguistic variables rather than crisp information is that they are consistent with the uncertain and imprecise nature. Traditional methods can't address such issues to deal with imprecision. Therefore, decision makers are able to deal with complexity, non linearity and uncertainty that might be in the set of actual values with respect to number of guest arrivals because of measurement errors and unstable responses.

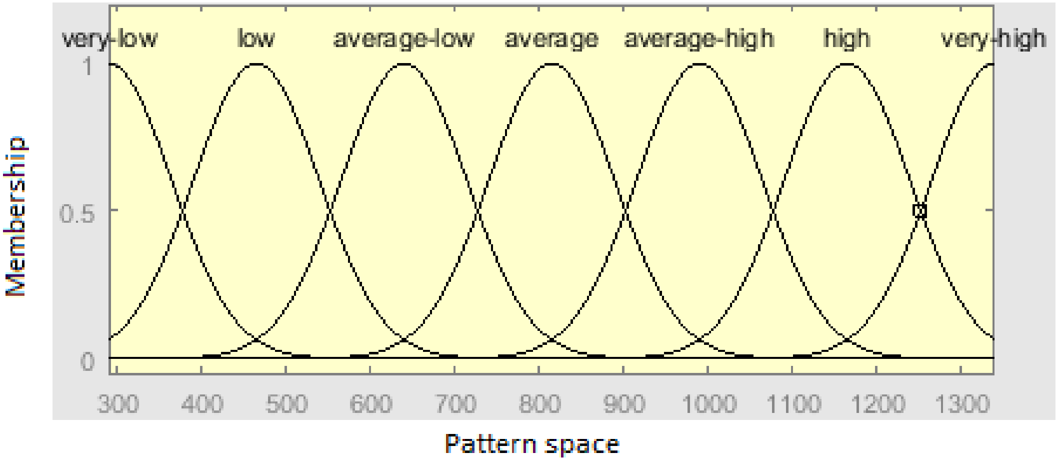


Figure 4.11: Gaussian membership function plot and homogeneous fuzzy partitions with 7 clusters and 4 inputs

Fuzzy sets and their corresponding membership functions for inputs and output of fuzzy rules are shown in Figure 4.12. Here, we use the method of center of area for the defuzzification step and a minimum operator for performing the implication method.

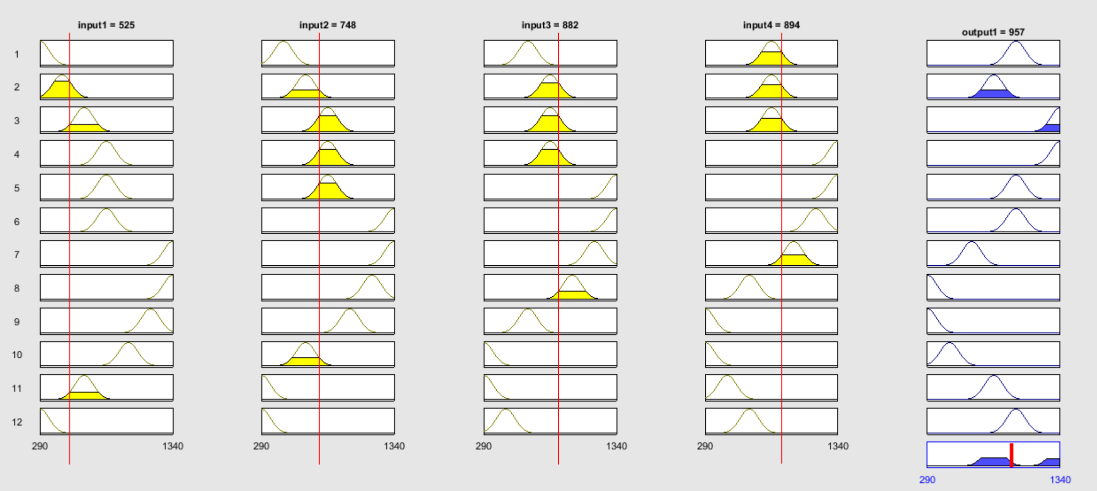
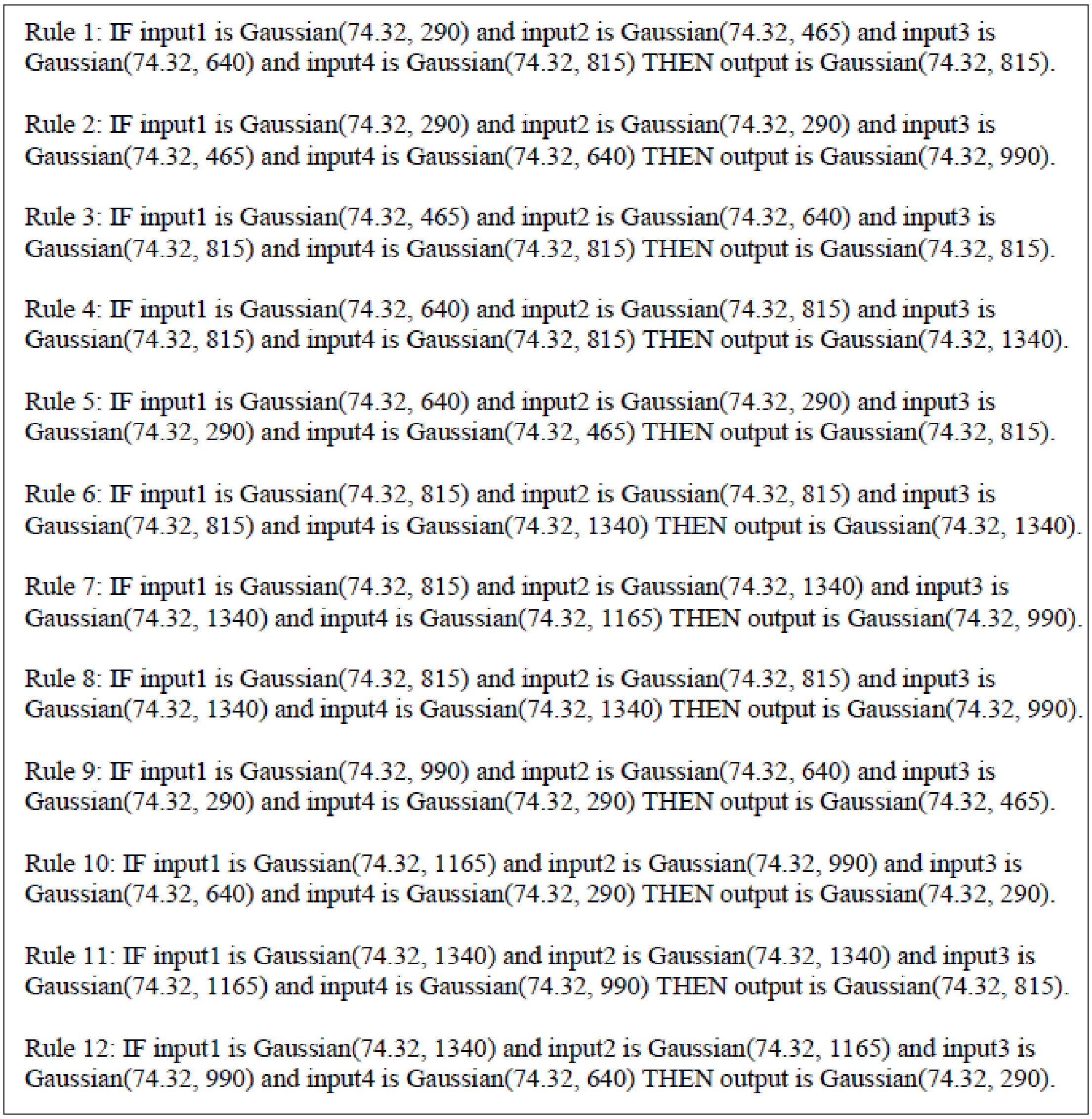


Figure 4.12: Fuzzy sets and their corresponding membership functions for inputs and output of fuzzy rules

As a result, the obtained rules of the Mamdani FRBS with 7 clusters and 4 inputs from testing data are illustrated in Figure 4.13.



Rule 1: IF input1 is Gaussian(74.32, 290) and input2 is Gaussian(74.32, 465) and input3 is Gaussian(74.32, 640) and input4 is Gaussian(74.32, 815) THEN output is Gaussian(74.32, 815).

Rule 2: IF input1 is Gaussian(74.32, 290) and input2 is Gaussian(74.32, 290) and input3 is Gaussian(74.32, 465) and input4 is Gaussian(74.32, 640) THEN output is Gaussian(74.32, 990).

Rule 3: IF input1 is Gaussian(74.32, 465) and input2 is Gaussian(74.32, 640) and input3 is Gaussian(74.32, 815) and input4 is Gaussian(74.32, 815) THEN output is Gaussian(74.32, 815).

Rule 4: IF input1 is Gaussian(74.32, 640) and input2 is Gaussian(74.32, 815) and input3 is Gaussian(74.32, 815) and input4 is Gaussian(74.32, 815) THEN output is Gaussian(74.32, 1340).

Rule 5: IF input1 is Gaussian(74.32, 640) and input2 is Gaussian(74.32, 290) and input3 is Gaussian(74.32, 290) and input4 is Gaussian(74.32, 465) THEN output is Gaussian(74.32, 815).

Rule 6: IF input1 is Gaussian(74.32, 815) and input2 is Gaussian(74.32, 815) and input3 is Gaussian(74.32, 815) and input4 is Gaussian(74.32, 1340) THEN output is Gaussian(74.32, 1340).

Rule 7: IF input1 is Gaussian(74.32, 815) and input2 is Gaussian(74.32, 1340) and input3 is Gaussian(74.32, 1340) and input4 is Gaussian(74.32, 1165) THEN output is Gaussian(74.32, 990).

Rule 8: IF input1 is Gaussian(74.32, 815) and input2 is Gaussian(74.32, 815) and input3 is Gaussian(74.32, 1340) and input4 is Gaussian(74.32, 1340) THEN output is Gaussian(74.32, 990).

Rule 9: IF input1 is Gaussian(74.32, 990) and input2 is Gaussian(74.32, 640) and input3 is Gaussian(74.32, 290) and input4 is Gaussian(74.32, 290) THEN output is Gaussian(74.32, 465).

Rule 10: IF input1 is Gaussian(74.32, 1165) and input2 is Gaussian(74.32, 990) and input3 is Gaussian(74.32, 640) and input4 is Gaussian(74.32, 290) THEN output is Gaussian(74.32, 290).

Rule 11: IF input1 is Gaussian(74.32, 1340) and input2 is Gaussian(74.32, 1340) and input3 is Gaussian(74.32, 1165) and input4 is Gaussian(74.32, 990) THEN output is Gaussian(74.32, 815).

Rule 12: IF input1 is Gaussian(74.32, 1340) and input2 is Gaussian(74.32, 1165) and input3 is Gaussian(74.32, 990) and input4 is Gaussian(74.32, 640) THEN output is Gaussian(74.32, 290).

Figure 4.13: Rules of the Mamdani FRBS with 7 clusters and 4 inputs from testing data

4.4 Analysis of Results

As mentioned above, fuzzy IF-THEN rules with linguistic variables are used in Mamdani FRBS. Dealing with linguistic terms is consistent with the vague and uncertain information. The experimental values with respect to the number of guest arrivals in the hotel include uncertainty, complexity, and non linearity. The traditional methods using a quantitative analysis disenable to address the matter of such imprecision and inaccuracy, and therefore, are unsuitable in these situations.

Before producing rules, the number of linguistic variables should be alleviated to a considerable size to prevent creating extra rules. With n linguistic variables, m inputs and one output, there are totally n^{m+1} rules to be produced. However, in this thesis the number of rules is significantly reduced as it is depicted in Table 4.1 by classifying time series data.

The accuracy of forecasting model is measured by applying such metrics as root mean square error (RMSE) and mean absolute percentage error (MAPE). It is obvious that the lowest values of RMSE and MAPE are the desired ones. RMSE and MAPE are calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i^{pr.} - X_i^{exp.}\right)^2} \tag{4.5}$$

$$\mathrm{MAPE} = 100 \times \frac{1}{N}\sum_{i=1}^{N}\frac{|X_i^{pr.} - X_i^{exp.}|}{X_i^{exp.}} \tag{4.6}$$

where $X_i^{pr.}$ is the prediction (forecast) value, $X_i^{exp.}$ is the experimental value of ith testing data to be defined from the model, and N is the number of data used for testing.

Table 4.1 describes the performances of forecasting models. It is defined that the optimal forecasting model is the one with 7 clusters and 4 inputs with the value of RMSE equal to 31.8355 and value of MAPE equal to 4.1155%.

Table 4.1: Summary of performances of forecasting models

Clusters	**5**			**6**			**7**		
Inputs	2	3	4	2	3	4	2	3	4
No. of rules	15	26	30	19	26	28	20	27	30
RMSE	159.1043	114.1644	89.5023	94.9759	76.1517	48.2485	143.4837	58.5335	31.8355
MAPE(%)	19.0397	16.2267	12.503	11.6201	10.3544	6.5161	17.8701	8.5701	4.1155
Ranking	9	7	6	5	4	2	8	3	1

Chapter 5

CONCLUSION

In this thesis the time series forecasting model for hotel occupancy is analyzed using fuzzy c-means clustering and fuzzy rule-based system. It is crucial issue to improve the rule-based system in tourism demand forecasting with the aim of performing a high accuracy while keeping the system interpretable. Therefore, to keep the balance between accuracy and interpretability of the system, the combination of fuzzy c-means clustering and Mamdani fuzzy rule-based system is used to design the forecasting model. The fuzzy c-means clustering is used to find the center points and the interval of each partition of the discourse universe. Another advantage of using fuzzy c-means clustering is that the number of required rules is reduced significantly. It reveals a capability of the system for designing a supportive system which is applicable in decision making to achieve the optimal number of rules which is very interpretable for both experts and non-experts.

Mamdani fuzzy rule-based systems with some clusters and inputs are applied to analyze and forecast the number of guest arrivals in one of the hotels of North Cyprus over 40 months. With respect to lowest root mean square error and mean absolute percentage error, it is determined that Mamdani fuzzy rule-based system model with 7 clusters and 4 inputs provides optimal performance, and is considered as most suitable forecasting model for hotel occupancy.

REFERENCES

[1] Travel & Tourism: Economic Impact 2017 (World Travel & Tourism Council, 2017). https://www.wttc.org/-/media/files/reports/economic-impact-research/regions-2017/world2017.pdf.

[2] Witt, S. F., & Witt, C. A. (1995). Forecasting tourism demand: A review of empirical research. *Int. J. Forecast*, 11(3), 447-475.

[3] Martin, C. A., & Witt, S. F. (1989). Forecasting tourism demand: A comparison of the accuracy of several quantitative methods. *Int. J. Forecast*, 5(1), 7-19.

[4] Dwyer, L., Gill, A., & Seetaram, N. (Eds.). (2012). *Handbook of research methods in tourism: Quantitative and qualitative approaches*. Edward Elgar Pub.

[5] Song, H., & Li, G. (2008). Tourism demand modeling and forecasting - A review of recent research. *Tour. Manag.*, 29(2), 203-220.

[6] Hansen, J. V., McDonald, J. B., & Nelson, R. D. (1999). Time series prediction with genetic-algorithm designed neural networks: An empirical comparison with modern statistical models. *Comput. Intell.*, 15(3), 171-184.

[7] Wang, C.-H., & Hsu, L.-C. (2008). Constructing and applying an improved fuzzy time series model: Taking the tourism industry for example. *Expert Syst. Appl.*, 34(4), 2732-2738.

[8] Shahrabi, J., Hadavandi, E., & Asadi, S. (2013). Developing a hybrid intelligent model for forecasting problems: Case study of tourism demand time series. *Knowl.-Based Syst.*, 43, 112-122.

[9] Peng, B., Song, H., & Crouch, G. I. (2014). A meta-analysis of international tourism demand forecasting and implications for practice. *Tour. Manag.*, 45, 181-193.

[10] Konar, A. (2005). An introduction to computational intelligence. *Computational Intelligence: Principles, Techniques and Applications*. Springer.

[11] Hung, J.-C. (2009). A fuzzy GARCH model applied to stock market scenario using a genetic algorithm. *Expert Syst. Appl.*, 36(9), 11710-11717.

[12] Li, S.-T., Cheng, Y.-C., & Lin, S.-Y. (2008). A FCM-based deterministic forecasting model for fuzzy time series. *Comput. Math. Appl.* 56(12), 3052-3063.

[13] Song, Q., & Chissom, B. S. (1993). Forecasting enrollments with fuzzy time series — Part I. *Fuzzy Sets Syst.*, 54(1), 1-9.

[14] Song, Q., & Chissom, B. S. (1994). Forecasting enrollments with fuzzy time series — part II. *Fuzzy Sets Syst.*, 62(1), 1-8.

[15] Chen, S.-M. (1996). Forecasting enrollments based on fuzzy time series. *Fuzzy Sets Syst.*, 81(3), 311-319.

[16] Huarng, K. (2001). Heuristic models of fuzzy time series for forecasting. *Fuzzy Sets Syst.*, 123(3), 369-386.

[17] Chen, S.-M. (2002). Forecasting enrollments based on high order fuzzy time series. *Cybern. Syst.*, 33(1), 1-16.

[18] Own, C.-M., & Yu, P.-T. (2005). Forecasting fuzzy time series on a heuristic high-order model. *Cybern. Syst.*, 36(7), 705-717.

[19] Askari, S., Montazerin, N., & Zarandi, M. H. F. (2015). A clustering based forecasting algorithm for multivariable fuzzy time series using linear combinations of independent variables. *Appl. Soft Comput.*, 35, 151-160.

[20] Wang, C.-H. (2004). Predicting tourism demand using fuzzy time series and hybrid grey theory. *Tour. Manag.*, 25(3), 367-374.

[21] Cheng, C.-H., Chen, T.-L., Teoh, H. J., & Chiang, C.-H. (2008). Fuzzy time-series based on adaptive expectation model for TAIEX forecasting. *Expert Syst. Appl.*, 34(2), 1126-1132.

[22] Chen, S.-M., & Hwang, J.-R. (2000). Temperature prediction using fuzzy time series. *IEEE Trans. Syst. Man Cybern. B Cybern.*, 30(2), 263-275.

[23] Zadeh, L. A. (1965). Fuzzy sets. *Inf. Control*, 8(3), 338-353.

[24] Bellman, R. Kalaba, R. & Zadeh, L. A. (1966). Abstraction and pattern classification. *J. Math. Anal. Appl.*, 13(1), 1-7.

[25] Ruspini, E. H. (1969). A new approach to clustering. *Inf. Control*, 15(1), 22-32.

[26] Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York, Pelenum Press.

[27] Suganya, R., & Shanthi, R. (2012). Fuzzy C-means algorithm - a review. *Int. J. Sci. Res.*, 2(11), 440-442.

[28] Mamdani, E. H. (1974). Applications of fuzzy algorithms for simple dynamic plant. *Proc. IEEE*, 121(12), 1585-1588.

[29] Mamdani, E. H. (1976). Application of fuzzy logic to approximate reasoning using linguistic synthesis. In *Proceedings of the sixth international symposium on Multiple-valued logic* (pp. 196-202). IEEE Computer Society Press.

[30] King, P. J., & Mamdani, E. H. (1977). The application of fuzzy control systems to industrial processes. *Automatica*, 13(3), 235-242.

[31] Gacto, M. J., Alcalá, R., & Herrera, F. (2011). Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Inf. Sci.*, 181(20), 4340-4360.

[32] Cordón, O. (2011). A historical review of evolutionary learning methods for Mamdani-type fuzzy rule-based systems: Designing interpretable genetic fuzzy systems. *Int. J. Approx. Reason.*, 52(6), 894-913.

[33] Goh, C., & Law, R. (2002). Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention. *Tour. Manag.*, 23(5), 499-510.

[34] Song, H., & Witt, S. F. (2006). Forecasting international tourist flows to Macau. *Tour. Manag.*, 27(2), 214-224.

[35] Wong, K. K. F., Song, H., Witt, S. F., & Wu, D. C. (2007). Tourism forecasting: To combine or not to combine? *Tour. Manag.*, 28(4), 1068-1078.

[36] Jeffrey, D., & Barden, R. R. D. (2000). Monitoring hotel performance via occupancy time series analysis: The concept of occupancy performance space. *Int. J. Tourism Res.*, 2(6), 383-402.

[37] Lin, C.-J., Chen, H.-F., & Lee, T.-S. (2011). Forecasting tourism demand using time series, artificial neural networks and multivariate adaptive regression splines: evidence from Taiwan. *Int. J. Bus. Admin.*, 2(2), 14-24.

[38] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50, 159-175.

[39] Chen, K.-Y. (2011). Combining linear and nonlinear model in forecasting tourism demand. *Expert Syst. Appl.*, 38(8), 10368-10376.

[40] Pai, P.-F., Hung, K.-C., & Lin, K.-P. (2014). Tourism demand forecasting using novel hybrid system. *Expert Syst. Appl.*, 41(8), 3691-3702.

[41] Huang, Y.-L., Horng, S.-J., Kao, T.-W., Kuo, I.-H., & Takao, T. (2012). A hybrid forecasting model based on adaptive fuzzy time series and particle swarm optimization. In *2012 International Symposium on Biometrics and Security Technologies* (pp. 66-70). IEEE.

[42] Asadi, S., Tavakoli, A., & Hejazi, S. R. (2012). A new hybrid for improvement of auto-regressive integrated moving average models applying particle swarm optimization. *Expert Syst. Appl.*, 39(5), 5332-5337.

[43] Huarng, K.-H., Moutinho, L., & Yu, T. H.-K. (2007). An advanced approach to forecasting tourism demand in Taiwan. *J. Travel Tour. Mark.*, 21(4), 15-24.

[44] Sakhuja, S., Jain, V., Kumar, S., Chandra, C., & Ghildayal, S. K. (2016). Genetic algorithm based fuzzy time series tourism demand forecast model. *Ind. Manag. Data Syst.*, 116(3), 483-507.

[45] Xu, X., Law, R. Chen, W., & Tang L. (2016). Forecasting tourism demand by extracting fuzzy Takagi–Sugeno rules from trained SVMs. *CAAI Trans. Int. Tech.*, 1(1), 30-42.

[46] Hadavandi, E., Ghanbari, A., Shahanaghi, K., & Abbasian-Naghneh, S. (2011). Tourist arrival forecasting by evolutionary fuzzy systems. *Tour. Manag.*, 32(5), 1196-1203.

[47] Jeffrey, D., Barden, R. R. D., Buckley, P. J., & Hubbard, N. J. (2002). What makes for a successful hotel? Insights on hotel management following 15 years of hotel occupancy analysis in England. *Service Industries Journal*, 22(2), 73-88.

[48] Grosan, C., & Abraham, A. (2011). *Intelligent Systems: a Modern Approach.* Berlin: Springer.

[49] Himmelspach, L. (2016). *Fuzzy Clustering of Incomplete Data.* (Doctoral dissertation). Düsseldorf, Germany.

[50] Aliev, R. R., & Salehi, S. (2016). Implementation of fuzzy c-means clustering technique for the hotel occupancy problem. In Proceedings of the *Ninth World Conference on Intelligent Systems for Industrial Automation, WCIS-2016* (pp. 14-19). Tashkent, Uzbekistan.

[51] Hathaway, R. J., & Bezdek, J. C. (1995). Optimization of clustering criteria by reformulation. *IEEE Trans. Fuzzy Syst.*, 3(2), 241-245.

[52] Kolen, J. F., & Hutcheson, T. (2002). Reducing the time complexity of the fuzzy c-means algorithm. *IEEE Trans. Fuzzy Syst.*, 10(2), 263-267.

[53] Lázaro, J., Arias, J., Martín, J. L., & Cuadrado, C. (2003). Modified fuzzy C-means clustering algorithm for real-time applications. In *International Conference on Field Programmable Logic and Applications* (pp. 1087-1090). Springer, Berlin, Heidelberg.

[54] Aliyev, R., Salehi, S., & Aliyev, R. (2019). Development of fuzzy time series model for hotel occupancy forecasting. *Sustainability*, 11(3), 793.

[55] Aliev, R. A., & Aliev, R. R. (2001). *Soft Computing and its Applications*. World Scientific, Singapore.

[56] Aliev, R. A., Fazlollahi, B., & Aliev, R. R. (2004). *Soft Computing and its Applications in Business and Economics*. Springer, Berlin, Germany.

[57] Michalski, R. S. (1983). A theory and methodology of inductive learning. In R. S. Michalski (Ed.), *Machine learning* (pp. 83-134). Springer-Verlag, Berlin.

[58] Alonso, J. M., Castiello, C., & Mencar, C. (2015). Interpretability of fuzzy systems: Current research trends and prospects. In J. Kacprzyk & W. Pedrycz (Eds.), *Handbook of computational intelligence* (pp. 219-237). Springer, Berlin, Heidelberg.

[59] Cox, E. (1999). *The Fuzzy Systems Handbook: A Practitioner's Guide to Building, Using, and Maintaining Fuzzy Systems*. Academic Press, San Diego, CA.