FUZZY REGRESSION MODELING OF DEFECT RATES
IN A METAL CASTING PROCESS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ÇANKAYA UNIVERSITY

BY

TUNA KILIÇ

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

SEPTEMBER 2009

Title of the Thesis: FUZZY REGRESSION MODELING OF DEFECT RATES IN A METAL CASTING PROCESS

Submitted by **Tuna Kılıç**

Approval of the Graduate School of Natural and Applied Sciences, Çankaya University

Prof. Dr. Yahya K. Baykal
Acting Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science

Prof. Dr. Levent Kandiller
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Prof. Dr. Gülser KÖKSAL
Co-Supervisor

Inst. Dr. Özlem TÜRKER BAYRAK
Supervisor

Examination Date : 02.09.2009

Asst. Prof. Dr. Ferda Can ÇETİNKAYA	(Çankaya Univ.)
nst. Dr. Özlem TÜRKER BAYRAK	(Çankaya Univ.)
Prof. Dr. Gülser KÖKSAL	(METU)
Assoc. Prof. Dr. İnci BATMAZ	(METU)
Inst. Dr. Haluk AYGÜNEŞ	(Çankaya Univ.)

STATEMENT OF NON-PLAGIARISM

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	: Tuna Kılıç
Signature	:
Date	: 23.10.2009

ABSTRACT

FUZZY REGRESSION MODELING OF DEFECT RATES
IN A METAL CASTING PROCESS

KILIÇ, Tuna
M. Sc., Department of Industrial Engineering
Supervisor : Inst. Dr. Özlem TÜRKER BAYRAK
Co-Supervisor: Prof. Dr. Gülser KÖKSAL

September 2009, 77 pages

This study has two purposes. One of them is to develop model of a metal casting process quality. This model can be used by the company to reduce the number of defects by identifying the process variables which have the most important effects on a certain defect type. The second purpose is to test and develop certain fuzzy regression approaches for the case problem.

In the case study, 36 process variables are observed from the metal casting process. Initially, stepwise linear regression (LR) method is applied to the data set and 8 independent variables are selected as significant. When the process variables are examined, it is realized that instead of identifying them as crisp, expressing them with fuzzy numbers is more appropriate. In the scope of the study, all fuzzy numbers are assumed to be triangular fuzzy numbers.

First the Hojati-Bector-Smimous (HBS2) method developed by Hojati et al. (2005) is generalized to multi variable modeling and then applied to the data set. In order to make a comparison between HBS2 and linear regression (LR) approach, the latter is also applied. Similarly, in order to make a comparison between HBS2 and other fuzzy methods, Fuzzy Functions (FF) method developed by Türkşen and Çelikyılmaz (2006) is used. Furthermore, Non-Parametric Improved Fuzzy Classification Functions developed by Özer (2009) is adapted to the case of fuzzy linear regression. The newly developed method called as Non-Parametric Improved Fuzzy Functions (NIFF) is applied to the same data set for a comparison with the other solutions.

Keywords: *Fuzzy Regression, HBS2, Fuzzy Functions, Non-Parametric Improved Fuzzy Functions, Metal Casting*

ÖZ

BİR METAL DÖKÜM SÜRECİNDEKİ HATA ORANININ
BULANIK REGRESYON YÖNTEMİ İLE MODELLENMESİ

KILIÇ, Tuna

Yüksek Lisans, Endüstri Mühendisliği Anabilim Dalı

Tez Yöneticisi : Dr. Özlem TÜRKER BAYRAK

Ortak Tez Yöneticisi : Prof. Dr. Gülser KÖKSAL

Eylül 2009, 77 sayfa

Bu çalışmanın iki amacı bulunmaktadır. Amaçlardan birisi, bir döküm sürecinde süreç kalitesinin modellenmesidir. Bu model firma tarafından hatalı ürün sayısının azaltılması amacıyla hata üzerinde en fazla etkisi olan süreç değişkenlerinin belirlenmesi için kullanılabilir. Amaçlardan ikincisi metal döküm verisi için çeşitli bulanık regresyon yaklaşımlarının test edilmesi ve geliştirilmesidir.

Bu çalışmada metal döküm sürecinden 36 adet süreç değişkeninin değerleri gözlemlenmiştir. Öncelikle veri kümesine aşamalı doğrusal regresyon analizi (DR) uygulanmış ve sekiz adet değişken seçilmiştir.

Süreç değişkenleri incelendiğinde, söz konusu değişkenlerin kesin (crisp) sayılar olarak ifade edilmesi yerine bulanık sayılar olarak ifade edilmesinin daha uygun olabileceği sonucuna varılmıştır. Çalışma kapsamında kullanılan tüm bulanık sayıların üçgensel bulanık sayılar olduğu varsayılmıştır.

İlk olarak Hojati ve diğ. (2005) tarafından önerilen Hojati-Bector-Smimous (HBS2) yöntemi, çok değişkenli modelleme yapılabilecek şekilde genelleştirilmiş ve veriye uygulanmıştır. Elde edilen sonuçların doğrusal regresyon ile karşılaştırılması amacı ile veriye doğrusal regresyon analizi uygulanmıştır. Benzer şekilde, HBS2 yöntemi ve diğer bulanık yöntemlerin karşılaştırılması amacı ile Türkşen ve Çelikyılmaz (2006) tarafından geliştirilen Bulanık Fonksiyonlar (BF) yöntemi uygulanmıştır. Ayrıca Özer (2009) tarafından geliştirilen Parametrik Olmayan İyileştirilmiş Bulanık Sınıflandırma Fonksiyonları (POİBSF) yöntemi, bulanık doğrusal regresyon analizine uyarlanmıştır. Geliştirilen bu yöntem Parametrik Olmayan İyileştirilmiş Bulanık Fonksiyonlar (POİBF) yöntemi olarak adlandırılmış ve veriye uygulanarak sonuçlar karşılaştırılmıştır.

Anahtar Kelimeler: *Bulanık Regresyon, HBS2, Bulanık Fonksiyonlar, Parametrik Olmayan İyileştirilmiş Bulanık Fonksiyonlar, Döküm Süreci*

ACKNOWLEDGMENTS

I would like to express profound gratitude to my supervisor, Dr. Özlem Türker Bayrak, and my co-supervisor Prof. Dr. Gülser Köksal for their invaluable support, encouragement, supervision and useful suggestions throughout this work. Their moral support and continuous guidance enabled me to complete my work successfully. I am also highly thankful to Assoc. Prof. Dr. İnci Batmaz for her valuable suggestions throughout this study.

I also wish to thank Gizem Özer and Elçin Kartal not only for their support but also for their friendship and for enjoyable time we spent during the study. I would like to present my thanks to all my friends and my colleagues for their encouragement and moral support they provided during my study.

Most special thanks go to my family Abdul Kadir Kılıç, Senem Çelik and Tuğba Kılıç, for their everlasting love, moral support and valuable guidance. I would like to thank them for always being there whenever I need. I would never succeed without their love and support.

I also would like to thank to examining committee members Dr. Ferda Can Çetinkaya and Dr. Haluk Aygüneş for their instructive criticism during the examination of the study.

TABLE OF CONTENTS

STATEMENT OF NON-PLAGIARISM iii

ABSTRACT iv

ÖZ vi

ACKNOWLEDGMENTS viii

TABLE OF CONTENTS ix

LIST OF TABLES xi

LIST OF FIGURES xii

1.INTRODUCTION 1

2.LITERATURE SURVEY AND BACKGROUND 4

2.1. HBS2 Model 13

2.2. Fuzzy Functions 15

2.3. Fuzzy C-Means (FCM) 18

2.4. Nonparametric Improved Fuzzy Classifier Function 25

2.5. Linear Regression 26

3.DESCRIPTION OF THE METAL CASTING DATA 29

4.GENERALIZATION OF THE HBS2 MODEL TO THE MULTI VARIABLE CASE 34

5.NON PARAMETRIC IMPROVED FUZZY FUNCTIONS 41

6.PERFORMANCE ANALYSIS 44

6.1. Performance Analysis For HBS2 Model With Two Variables 44

6.2. Performance Analysis For LR, HBS2, HBS2-V2, FF And NIFF 45

6.3. Model Interpretation 49

6.3.1. NIFF Model Interpretation 49

6.3.2. HBS2 Model Interpretation 50

6.4. Performance Analysis For Different H Values of HBS2 Model 51

6.5. Performance Analysis For Different Spread Of Independent Variables 52

7.CONCLUSIONS & SUGGESTIONS FOR FURTHER STUDIES 53

REFERENCES R1

APPENDIX A A1

APPENDIX B A5

LIST OF TABLES

Table 1 Illustration of the Model Indices for Two Independent Variable Case............... 35

Table 2 Illustration of Indices for Eight Independent Variables Case.............................. 38

Table 3 Performance Measures for Two Variables 45

Table 4 Performance Measures of LR, HBS2, HBS2-V2, FF and NIFF.............................. 46

Table 5 ANOVA Results... 47

Table 6 Multiple Comparison Results.. 48

Table 7 Performance Measures of HBS2-V2 for Different H Values 51

Table 8 Performance Measures of HBS2 for Different Spread Levels............................ 52

LIST OF FIGURES

Figure 1 A Triangular Membership Function 5

Figure 2 Membership Functions of Fuzzy Sets, "Young", "Middle Aged" and "Old" 6

Figure 3 A Fuzzy Linear Relationship 8

Figure 4 H-Certain Observed Interval 8

Figure 5 Illustration of Hi value 10

Figure 6 Illustration of the Tanaka Model 11

Figure 7 Illustration of the Fuzzy x Case 15

Figure 8 Cluster Validity Index Diagrams 24

Figure 9 A Cylinder Head 29

Figure 10 A Graphical Representation of HBS2 Model 37

Figure 11 Summary of Improvements Involving Fuzzy Functions 41

CHAPTER 1

INTRODUCTION

The nature of the uncertainty in a problem is very important that engineers should ponder prior to their selection of an appropriate method to express the uncertainty (Ross, 2004). Uncertain information can take on many different forms. There is uncertainty that arises because of complexity; for example, the complexity in the reliability network of a nuclear reactor. There is uncertainty that arises from ignorance, from various classes of randomness, from the inability to perform adequate measurements, from lack of knowledge, or from vagueness like fuzziness (Ross, 2004). The idea proposed by Zadeh (1965) suggests that setting membership is the key to decision. Belman and Zadeh (1970) note that there is a need for differentiation between randomness and fuzziness. They argue that the premise of "imprecision can be equated with randomness" is questionable, and that fuzziness is a major source of imprecision in many decision processes.

In a competitive market place, it is essential to make effective data analysis for manufacturing companies in order to put themselves one step further from the competitors. Making effective data analysis improves ability to make right decisions at critical decision points. Therefore, nowadays, selecting more accurate methods and techniques to use during data analysis becomes more essential for the manufacturing companies since these methods help them to make important decisions.

The aim of this study is to model process quality in order to reduce number of defects in a metal casting process. Since reducing number of defects directly decreases the related cost, determining the process variables which have the most important effects on defect and making decisions regarding the process variables become important.

In order to reach more accurate conclusions, before selecting the methods, the nature of the uncertainty is considered in this study. Since the uncertainty type of the metal casting process is not only randomness, but also imprecision and vagueness is included and the deviations between observed and estimated variables are supposed to be due to the indefiniteness of the system structure as described by Tanaka et al. (1982), fuzzy regression techniques are used for modeling in the scope of the study.

Fuzzy Linear Regression developed by Tanaka et al. (1982) aims to model the vague and imprecise phenomena using the fuzzy functions defined by Zadeh (1975). Fuzzy regression is a non-statistical method where the deviations between observed and estimated values are assumed to depend on the indefiniteness/vagueness of the parameters which govern the system structure, not on its measurement errors (Kim et al. 1996). It gives rise to a possibility distribution that accounts for the imprecise nature or vagueness of our understanding of a phenomenon. Classical statistical regression makes rigid assumptions about the statistical properties of the model; e.g. the normality of error terms. These assumptions are difficult to justify unless a sufficiently large data set is available. The violation of such basic assumptions could adversely affect the validity and performance of statistical regression. It has been stated that fuzzy regression may be more effective than statistical regression when the assumptions are either violated or cannot be properly employed, for example crisp and imprecise data is available (Gharpuray et al. 1986).

Unlike statistical methods, fuzzy methods may work with fuzzy data as well as crisp data. Fuzzy data, the members of which are the fuzzy numbers, can be thought of as interval numbers, values within which have varying degrees of memberships.

Because of the nature of the metal casting process variables which varied in an interval with a certain degrees of membership, both dependent and independent variables are considered as fuzzy numbers during the study.

In this study, a metal casting data set which is decided to be handled as fuzzy data set is used to model the process quality by HBS2 fuzzy linear regression method developed by Hojati et al. (2005). In order to apply HBS2 method to the data set, the method is generalized to multi variable case in scope of this study. Difficulties that are met during the application of the multi variable HBS2 method are discussed, and a method is proposed to overcome these difficulties. Performance of the proposed method and possible reasons behind its low and high performances are discussed. Possible use of these approaches in other cases is also discussed.

In order to make a comparison between HBS2 and statistical regression techniques the Multi variable Linear Regression approach is used. And also in order to make a comparison between HBS2 and other fuzzy methods, Fuzzy Functions (FF) method which is developed by Türkşen and Çelikyılmaz (2006) is used. Furthermore, Non-Parametric Improved Fuzzy Classification Functions presented by Özer (2009), is adapted to the regression case. The method named as Non-Parametric Improved Fuzzy Functions is applied to the same data set for comparison of all solutions.

This thesis is organized as seven chapters. In the second chapter, literature survey and background information about the methods used is described. HBS2 Method, Fuzzy Functions, Fuzzy C-Means Method and Nonparametric Improved Fuzzy Classifier Function methods are discussed. Explanations regarding metal casting data set are given in Chapter 3. Generalization of HBS2 model is given in Chapter 4. Non-Parametric Improved Fuzzy Functions approach is represented in Chapter 5. In the sixth chapter, the performances of the models are compared with respect to comparison criteria. Finally conclusions and future study suggestions can be found in Chapter 7. In the appendix the comparison criteria which are used for comparison of the model performances are described.

CHAPTER 2

LITERATURE SURVEY AND BACKGROUND

In regression analysis, dependent variable, $\boldsymbol{y}$, is a function of the independent variables and the degree of contribution of each variable to the output is represented by coefficients of these variables.

A crisp linear regression model is given in Eq.(2.1) :

$$\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{a}) = a_0 + a_1 x_1 + a_2 x_2 + \ \dots + a_k x_k \tag{2.1}$$

where

$\boldsymbol{y}$: dependent variable

$\boldsymbol{x}_i$: independent variables

$\boldsymbol{a}_i$: coefficients.

In conventional regression techniques, the difference between the observed values and the values estimated from the model is assumed to be due to the observational errors and the difference is considered as a random variable (Coppi, 2007). Upper and lower bounds for the estimated value are established and the probability that the estimated value will be within these two bounds represents the confidence of the estimate. In other words, conventional regression analysis is probabilistic.

But in fuzzy regression, the difference between the observed and estimated values is assumed to be due to the ambiguity inherently present in the system (Kim et al. 1996). The output for a specified input is assumed to be a range of possible values, i.e., the output can take any of these possible values. Therefore, fuzzy regression is possibilistic in nature. Moreover, fuzzy regression analyses use fuzzy functions to represent the coefficients as opposed to crisp coefficients used in conventional regression analysis.

Equation (2.2) shows a typical fuzzy linear regression model,

$$\widetilde{\boldsymbol{Y}} = f(\boldsymbol{x}, \widetilde{\boldsymbol{A}}) = \tilde{A}_0 + \tilde{A}_1 x_1 + \tilde{A}_2 x_2 + \ \dots + \tilde{A}_k x_k \qquad (2.2)$$

where

$\widetilde{\boldsymbol{A}}_{\boldsymbol{k}}$: k^{th} fuzzy coefficient

$\boldsymbol{x}$: independent variables.

Fuzzy regression estimates a range of possible values that are represented by a possibility distribution (a membership function). Membership functions are formed by assigning a specific membership value (degree of belonging) to each of the estimated value.

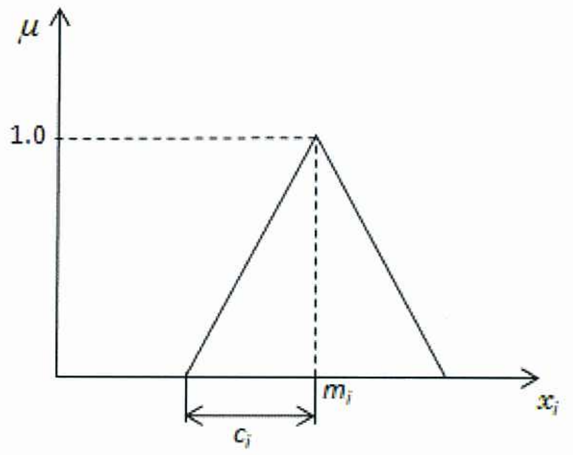

Figure 1 A Triangular Membership Function

Membership functions have been commonly formulated with straight lines. Among them, the simplest is the triangular or trapezoidal membership functions. Triangular membership functions, shown in Figure 1, allow for the solution to be found via a linear programming formulation whereas other membership functions for the coefficients require alternative approaches. In the figure m_i is the center and c_i is the spread value of the triangular fuzzy number.

The triangular membership function $\mu_{\tilde{A}}$ is expressed as;

$$\mu_{\widetilde{A_i}}(\boldsymbol{a}_i) = \begin{cases} 1 - \dfrac{|\boldsymbol{m}_i - \boldsymbol{c}_i|}{\boldsymbol{c}_i} & \boldsymbol{m}_i - \boldsymbol{c}_i \leq \boldsymbol{x}_i \leq \boldsymbol{m}_i + \boldsymbol{c}_i \\ 0 & Otherwise. \end{cases} \tag{2.3}$$

As an example, Figure 2 shows membership functions of fuzzy sets, "young", "middle aged" and "old".

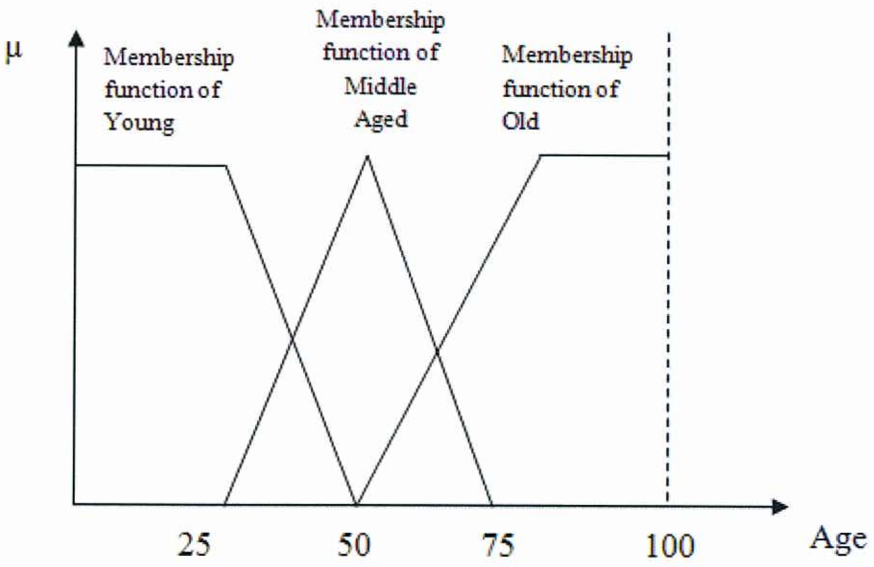

Figure 2 Membership Functions of Fuzzy Sets, "Young", "Middle Aged" and "Old"

(Source: Ross, 2004)

Horizontal axis of graph in Figure 2 represents "age" in years (the universal set X) and vertical axis represents the degree to which a person can be labeled as "young", "middle age" or "old". Hence, membership graph represents membership function of a fuzzy set, which represents "the group of people that can be considered young", "middle age" and "old". Further details can be found in Ross (2004).

The fuzzy function $\tilde{A}$ is a function of two parameters, m and c, known as the middle value and the spread, respectively. The spread denotes the fuzziness of the function. Therefore, the output is a revised version of Eq. (2.2)

$$\breve{\boldsymbol{Y}}_i = (\boldsymbol{m}_0, \boldsymbol{c}_0) + \sum_{j=1}^{k} (\boldsymbol{m}_{\mathrm{j}}, \boldsymbol{c}_{\mathrm{j}})\, \boldsymbol{x}_{ij} \quad i = 1,2, \dots, n \qquad (2.4)$$

where

n: number of observations

$\boldsymbol{x}_{ij}$: independent variables

$\boldsymbol{m}_{\mathrm{j}}$:midpoints of coefficients

$\boldsymbol{c}_{\mathrm{j}}$: spread of coefficients.

The output data, the input data, and the coefficients can be either fuzzy or crisp. In this thesis, because of the reasons explained in the Chapter 3, fuzzy input and fuzzy output dataset is used for modeling the system.

Similar to fuzzy coefficients, input variables' midpoints can be represented by x_{ij}, and spreads of the input variables can be represented by f_{ij}. For the fuzzy input, fuzzy output case, equation (2.4) can be re-written as follows:

$$\breve{\boldsymbol{Y}}_i = (\boldsymbol{m}_0, \boldsymbol{c}_0) + \sum_{j=1}^{k} (\boldsymbol{m}_{\mathrm{j}}, \boldsymbol{c}_{\mathrm{j}}) \left(x_{ij}, f_{ij}\right) \quad i = 1,2, \dots, n. \qquad (2.5)$$

where

n: number of observations

$\boldsymbol{x}_{ij}$:midpoints of independent variables

$\boldsymbol{f}_{ij}$: Spread of independent variables

$\boldsymbol{m}_{\mathrm{j}}$:midpoints of coefficients

c_j: spread of coefficients.

Fuzzy linear regression is proposed by Tanaka et al. (1982) to determine a fuzzy linear relationship as shown in Eq.(2.2). A simple fuzzy linear relationship (i.e. only one explanatory variable) can be represented by a band with a centre line as shown in Figure 3.

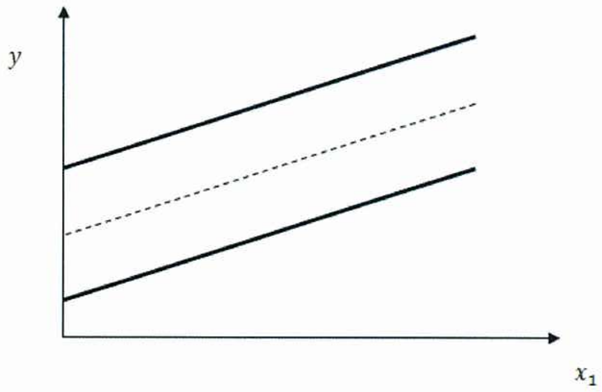

Figure 3 A Fuzzy Linear Relationship

Given a symmetric triangular fuzzy number for y_i , if we are only interested in that part of y_i which has a membership value of at least *H, 0 ≤ H ≤ 1,* we should use the interval $[\overline{\boldsymbol{y}}_i - (1-H)\boldsymbol{e}_i, \overline{\boldsymbol{y}}_i + (1-H)\boldsymbol{e}_i]$, where H represents the minimum degree of certainty acceptable, and we will refer to this interval as H-certain observed interval. This interval is the bold line segment in Figure 4.

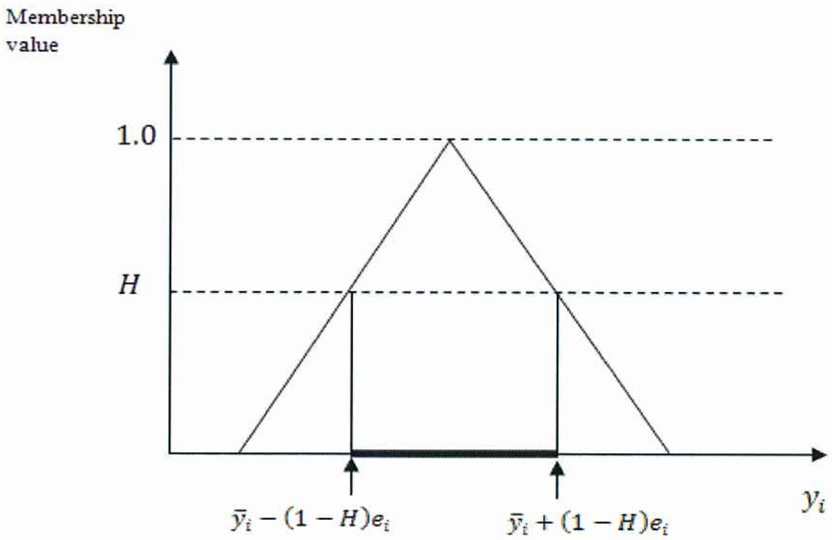

Figure 4 H-Certain Observed Interval

Similarly, the predicted interval corresponding to a specific set of x values having membership value of at least H is,

$$\begin{gathered}[\textstyle\sum_{j=0}^{k}(\boldsymbol{\alpha}_j - (1-H)\boldsymbol{c}_j)(\overline{\boldsymbol{x}}_{ij} - (1-H)\boldsymbol{f}_{ij}), \\ \textstyle\sum_{j=0}^{k}(\boldsymbol{\alpha}_j + (1-H)\boldsymbol{c}_j)(\overline{\boldsymbol{x}}_{ij} + (1-H)\boldsymbol{f}_{ij})]\end{gathered} \qquad (2.6)$$

where α_j's denotes the midpoints of the coefficients and $\bar{x}_{ij}$'s denotes the observations' midpoints. This interval is referred as H-certain predicted interval.

Fuzzy regression can be separated it into two cases;

Case 1: Independent variables are crisp, and the response variable is fuzzy.

Case 2: Independent variables are fuzzy and the response variable is also fuzzy.

Case 1 studies are initiated by the following model which is developed by Tanaka et al. (1982).

Tanaka Model:

$$Minimize \quad c_0 + c_1 + c_2 + \ldots + c_k \qquad (2.7)$$

$$Subject\ to \quad \sum_{j=0}^{k}(\alpha_j + (1-H)c_j)x_{ij} \geq \bar{y}_i + (1-H)e_i\ for\ i = 1,2,\ldots,n, \qquad (2.8)$$

$$\sum_{j=0}^{k}(\alpha_j - (1-H)c_j)x_{ij} \leq \bar{y}_i - (1-H)e_i\ for\ \ i = 1,2,\ldots,n, \qquad (2.9)$$

$$\alpha_j\ free \qquad c_j \geq 0 \qquad for \quad j = 0,\ldots,k \qquad (2.10)$$

where

$\boldsymbol{\alpha}_j$: midpoint of coefficient

$\boldsymbol{c}_j$: spread of coefficient

$\boldsymbol{x}_{ij}$: independent variable

$\overline{\boldsymbol{y}}_i$: midpoint of dependent variable

e_i : spread of dependent variable.

The model forces the H-certain predicted intervals (dotted vertical lines in Figure 6) to include H-certain observed interval (bold vertical lines). And the objective function of the model minimizes the total spreads of the coefficients. According to the Tanaka Model, membership value of an observed dependent variable and its estimated fuzzy dependent variable, H_i, must be at least H (Tanaka et al. 1982). H_i value for fuzzy observed dependent variables is illustrated in Figure 5 below.

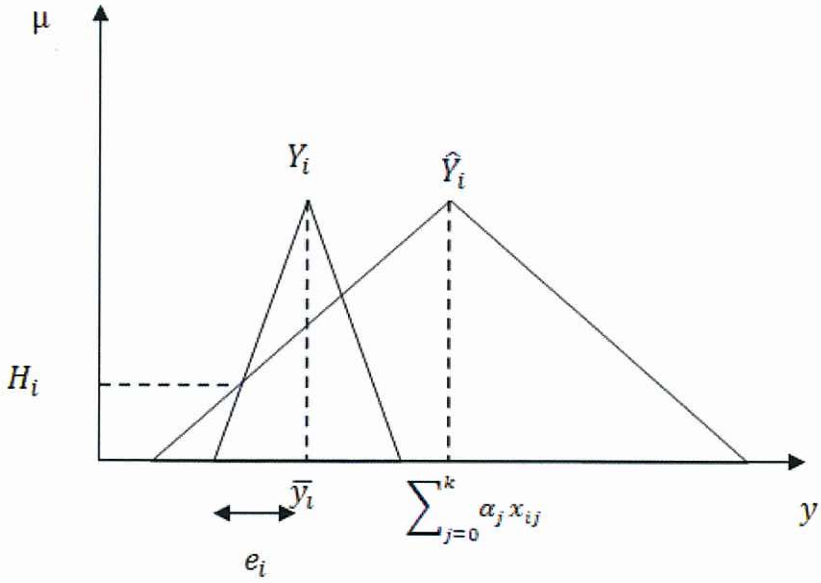

Figure 5 Illustration of H_i value
(Source: Özer, 2009)

As it can be seen, H_i is the maximum membership degree that predicted fuzzy interval contains the observed fuzzy interval. According to the model, midpoints of the predicted fuzzy regression coefficients are not affected by the H value, however, spread values of the predicted fuzzy regression coefficients increase with the increase in H value (Kim et al. 1996). Since H_i value increases when the midpoints of the predicted and the observed dependent variable get closer, H level can be seen as the level of credibility or level of confidence desired (Kim et al. 1996). Since it is determined by the user, proper selection of H level is important for the fuzzy regression model (Wang and Tsaur, 2000). It is suggested to determine H value according to the sufficiency of the data (Wang and Tsaur, 2000). If the data set collected is sufficiently large and reliable, then H level should be determined as 0 and it should be increased with the decreasing volume of the data set and the degree of reliability.

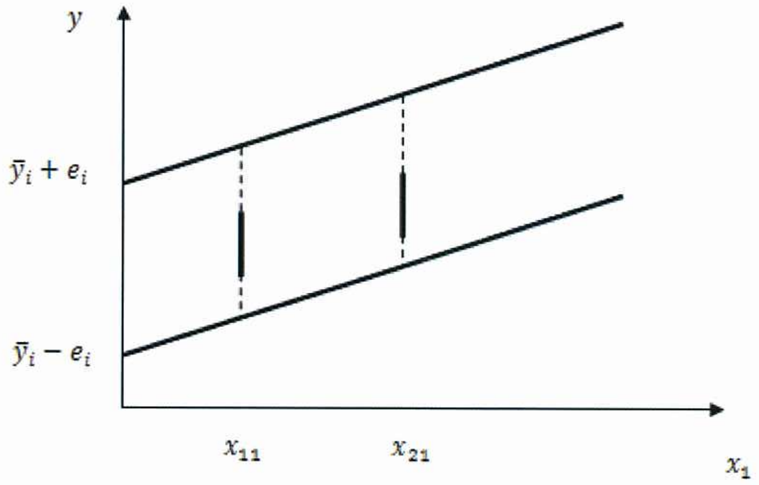

Figure 6 Illustration of the Tanaka Model

There have been a few criticisms of this approach. One shortcoming is that the solution is x_j-scale dependent and many c_j's turn out to be zero (Jozsef, 1992). To overcome this problem, instead of sum of half-width of regression coefficients in (2.7), sum of half-width of the predicted intervals can be used as the objective function (Tanaka and Ishibuchi, 1991). Thus the objective function becomes:

$$Minimize \qquad \sum_{i=1}^{n}\sum_{j=0}^{k} \mathrm{c_j}\, x_{ij}. \tag{2.11}$$

Some articles have proposed major changes to Tanaka et al. (1982) approach. Savic and Pedrycz (1991) suggest first to find the centers, α_j's, using ordinary least squares method and then to solve *Tanaka Model* with these α_j's.

Tanaka and Ishibuchi (1991) also suggest first to determine the centers α_j's using ordinary least squares method, but then to solve a quadratic programming version of the *Tanaka Model* with these α_j's. Celmins (1987) modified the least squares method to the case where both dependent and independent variables have triangular fuzzy number values in such a way that their joint membership function is a cone.

Another shortcoming of Tanaka et al. (1982) approach is that each H-certain predicted interval is required to contain the corresponding H-certain observed interval. This results in large coefficient half-widths, c_j , if any response 'value' has large half-width e_j or if there is an 'outlier' response. Tanaka et al.'s (1989) conjunctive model relaxes this requirement, only requiring that each H-certain predicted interval intersect the associated H-certain observed interval.

Regarding Case 2 approach, Sakawa and Yano (1992) classify the independent variables into three groups according to the expected range of values of fuzzy regression coefficients, A_j, and solve the problem based on these groups.

Peters (1994) developed a model in which predicted intervals are allowed to intersect observed intervals rather than including them in order to decrease the sensitivity of the outliers. Ozelkan and Duckstein (2000) proposed a similar formulation to that of Peter (1994), but have not required the prediction intervals to intersect the observed intervals.

Hojati et al. (2005) develop a method similar to Ozelkan and Duckstein (2000), but their method tries to obtain narrower intervals by minimizing the difference between the observed and the predicted dependent variable when the predicted dependent variable includes the observed dependent variable. Hojati et al. (2005) developed two models within this scope; fist they applied the model to crisp input and fuzzy output data set and called it as HBS1, then they expanded the model for the fuzzy input and fuzzy output data sets which is called as HBS2.

The performances of the Tanaka's, Peter's, Özelkan's and Hojati's first model are compared by the Hojati et al. (2005) and it is concluded that the Özelkan's model and Hojati's first model named HBS1 has better performance than the others. Furthermore regarding case 2 approach Hojati's second model named HBS2 and

Sakawa's model is compared and found that HBS2 has better performance than the Sakawa's model.

In this study, since input and output variables are both fuzzy in the metal casting data set, the method developed by Hojati et al. (2005) named as HBS2 is used to model the system.

2.1. HBS2 MODEL

In HBS2 Model, the fuzzy regression coefficients are chosen such that the total deviation of upper points of predicted and associated observed intervals and deviation of lower points of predicted and associated observed intervals are minimized at both lower points ("left") and upper points ("right") of each of the independent variable values (except x_0). For simplicity, the following LP is formulated for the case when there is only one independent variable (in addition to x_0):

$$Minimize \quad \sum_{i=1}^{n}[\, d_{ilU}^{+} + d_{ilU}^{-} + d_{ilL}^{+} + d_{ilL}^{-} + d_{irU}^{+} + d_{irU}^{-} + d_{irL}^{+} + d_{irL}^{-}\,] \tag{2.12}$$

$$Subject\ to: \quad \sum_{j=0}^{1}[(\alpha_j + (1-H)c_j)(\bar{x}_{ij} - (1-H)f_{ij})] + d_{ilU}^{+} - d_{ilU}^{-} = \bar{y}_i + (1-H)e_i \quad for\ i = 1,2,\dots,n. \tag{2.13}$$

$$\sum_{j=0}^{1}[(\alpha_j + (1-H)c_j)(\bar{x}_{ij} + (1-H)f_{ij})] + d_{irU}^{+} - d_{irU}^{-} = \bar{y}_i + (1-H)e_i \quad for\ i = 1,2,\dots,n. \tag{2.14}$$

$$\sum_{j=0}^{1}[(\alpha_j - (1-H)c_j)(\bar{x}_{ij} - (1-H)f_{ij})] + d_{ilL}^{+} - d_{ilL}^{-} = \bar{y}_i - (1-H)e_i \quad for\ i = 1,2,\dots,n. \tag{2.15}$$

$$\sum_{j=0}^{1}[(\alpha_j-(1-H)c_j)(\bar{x}_{ij}+(1-H)f_{ij})]+d_{irL}^{+}-d_{irL}^{-}$$

$$=\bar{y}_i-(1-H)e_i \quad for \ = 1,2,\dots,n. \tag{2.16}$$

$$d_{ilU}^{+}, d_{ilU}^{-}, d_{ilL}^{+}, d_{ilL}^{-}, d_{irU}^{+}, d_{irU}^{-}, d_{irL}^{+}, d_{irL}^{-} \geq 0$$
$$for \ i = 1,2,\dots,n. \tag{2.17}$$

$$\alpha_j \ free, \qquad c_j \geq 0, \qquad for \ \ j = 0,1 \tag{2.18}$$

where

d : distance between observations and estimations,

α_j : centers of the coefficients,

c_j : spreads of the coefficients,

$\bar{x}_{ij}$: centers of the independent variables,

f_{ij}: spreads of the independent variables,

$\bar{y}_i$: centers of the dependent variable,

e_i : spread of the dependent variable.

The indices "l" refers to the left (lower) point and "r" refers to the right (upper) point of the independent variable intervals, and "U" refers to the upper points and "L" refers to the lower points of the predicted interval.

The constraints of the model equalize the observed and estimated dependent variables. And sum of the distances between observed and estimated dependent variables are minimized in the objective function. Illustration of the "fuzzy x"case is given in Figure 7. The rectangle in the figure shows the observed area, and the parallelogram (in bold lines) shows the predicted area. Specifically, the x dimension of observed area for observation i ranges from $\bar{x}_{i1}-f_{i1}$ to $\bar{x}_{i1}+f_{i1}$, and the y dimension of observed area for observation i ranges from $\bar{y}_i - e_i$ to $\bar{y}_i + e_i$. That is, the observed area is a rectangle.

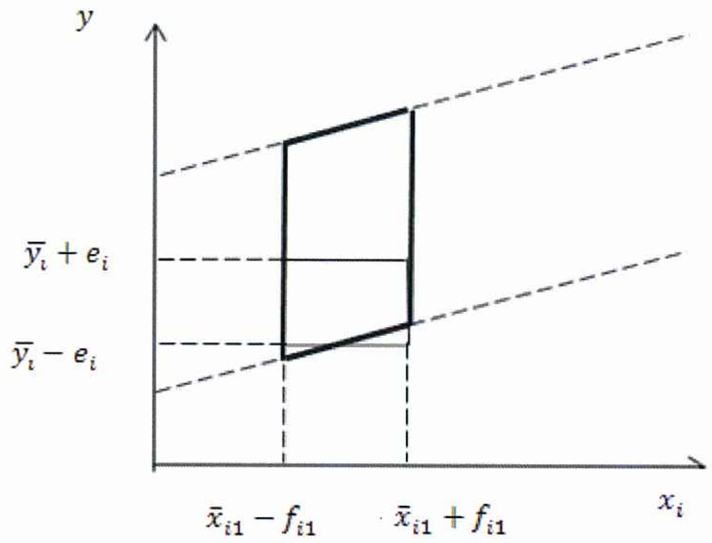

Figure 7 Illustration of the Fuzzy $\boldsymbol{x}$ Case

However, the predicted area is a parallelogram with the same x dimension as the observed area, and y dimension at $\bar{x}_{i1} - f_{i1}$ ranging from $(\alpha_0 - c_0) + (\alpha_1 - c_1)(\bar{x}_{i1} - f_{i1})$ to $(\alpha_0 + c_0) + (\alpha_1 + c_1)(\bar{x}_{i1} - f_{i1})$, called "left" and y dimension at $\bar{x}_{i1} + f_{i1}$ ranging from $(\alpha_0 - c_0) + (\alpha_1 - c_1)(\bar{x}_{i1} + f_{i1})$ to $(\alpha_0 + c_0) + (\alpha_1 + c_1)(\bar{x}_{i1} + f_{i1})$, called "right" .

HBS2 requires 2^{k+1} constraints for each observation i (k is the number of independent variables other than intercept). This is a shortcoming of HBS2 for high number of independent variable cases. In order to overcome this shortcoming, Hojati et al. (2005) suggests to eliminate some of the constraints resulting in unsatisfactory performance.

2.2. FUZZY FUNCTIONS

Fuzzy systems based on Fuzzy Rule Bases (FRB) are successfully used to support problem-solving and decision making. A classical way to represent the human knowledge is using "IF...THEN" fuzzy rules. The "IF" part represents the antecedents (input fuzzy set), and the "THEN" part represents the consequents (output fuzzy sets). In these systems, membership values of fuzzy sets represent degree of belongingness, degree of compatibility, weight of strength of an object.

The fuzzy function approach is initially introduced by Türkşen (2005). Recently the algorithm is extended and combined with other soft computing approaches for performance improvement, faster and more accurate parameter optimization and uncertainty reduction (Çelikyılmaz, 2005; Çelikyılmaz et al. 2007).

The standard fuzzy functions are multi variable crisp valued functions. Fuzzy Functions approaches have emerged from the idea of representing each unique rule of an FRB system in terms of the 'fuzzy functions'. One of their prominent feature is that the degree of belongingness of each sample vector in a fuzzy set has a direct effect on how the local fuzzy functions of the particular set is defined.

Fuzzy clustering methods are the core methods of the structure identification part of fuzzy functions systems. Fuzzy c-Means Clustering (FCM) (Bezdek, 1981) is generally used in fuzzy functions approaches to obtain membership values.

A common way to optimize the parameters of fuzzy systems is to separate the dataset into several randomly selected training and testing subsets. The training dataset is used to learn the system structure and optimize the parameters. Similarly, testing dataset is used to measure the modeling performance. This process is repeated several times with different random subsets to form training and testing datasets. Then the overall performance is calculated by the predetermined methods.

Multi-input, single output problems are the main interest of this thesis. Let $(\boldsymbol{x}^t, y^t) = \{(\boldsymbol{x}_1^t\ , y_1^t), (\boldsymbol{x}_2^t\ , y_2^t), \dots, (\boldsymbol{x}_n^t\ , y_n^t)\}$ represent an input-output dataset, where t denotes the training data set and every datum is composed of $(nv + 1)$ dimensions of input vectors, $\boldsymbol{x}_k^t = \left(x_{1,k}^t, \dots, x_{nv,k}^t\right), k = 1, \dots, n$, a total of n vectors, and an output, y_k^t is the same. *Z* is the $(n \times (nv + 1))$ input-output matrix, n is the total number of data vectors, $i = 1, \dots, c$ is the cluster identifiers where c

represents the total number of clusters identified, and m is the degree of fuzziness (i.e. overlapping degree) which are parameters of the FCM clustering method. Let $\mu_{ik} \in [0,1]$ represent membership value of the k^{th} datum in cluster i. Therefore the list of parameters of the training algorithm is

- Number of clusters of the system model, c, is a discrete value,
- Degree of fuzziness of the system model $m \in [1.1, \infty]$,
- Type of the system to be modeled, e.g. linear or non-linear.

Algorithm is as follows:

Step 1:

Choose FCM clustering parameters, $m \geq 1.1$ (degree of fuzziness) and $c>1$ (the number of clusters) and ε (a termination threshold)

Step 2:

Execute FCM using $\mathrm{Z}(\mathbf{x}^{\mathrm{t}}, \mathbf{y}^{\mathrm{t}})$ to find the cluster centers, and interactive (I/O) membership values for m and c by;

$$\mu_{ik}(x,y) = [d_{ik}(x,y) / \sum_{j=1}^{c} d_{jk}(x,y)]^{\frac{2}{1-m}} \quad , \qquad \forall 1 \leq i \leq c, 1 \leq k \leq n \tag{2.19}$$

where $d_{ik}(x,y) = \|(x_k, y_k) - \boldsymbol{v}_i(x,y)\|$.

Step 3:

Find membership values of the input space using

$$\mu_{ik}(x) = [\boldsymbol{d}_{ik}(x) / \Sigma_{j=1}^{c} \boldsymbol{d}_{jk}(x)]^{\frac{2}{1-m}} \quad , \qquad \forall 1 \leq i \leq c, 1 \leq k \leq n \tag{2.20}$$

where $d_{ik}(x) = \|\boldsymbol{x}_k^t - \boldsymbol{v}_i(x^t)\|$.

Step 4:

For each cluster i

4.1. Membership values of each input data sample, μ_{ik} and their transformations is augmented to the original input space to map the original input matrix.

4.2. Estimate the parameters.

Standard FCM is implemented to training input-output data to generate membership values $\mu(x^t, y^t)$ and cluster centers $\boldsymbol{v}_i(x^t, y^t), i = 1, \dots, c$. In step 3, membership values corresponding to input space, $\mu_i(x^t)$, and cluster centers of the given input space, denoted by $\boldsymbol{v}_i(x)$ are obtained.

2.3. *FUZZY C-MEANS (FCM)*

Fuzzy clustering algorithms can map a given dataset into overlapping clusters, while computing membership values that specify to what degree each object belongs to these captured clusters. The most commonly used type of fuzzy clustering algorithm, is FCM clustering algorithm (Bezdek, 1981). In FCM clustering algorithm, it is assumed that the number of clusters, c, is known or at least fixed; i.e., FCM algorithm partitions a given data set $\boldsymbol{X} = \{\boldsymbol{x}_{1,} \dots, \boldsymbol{x}_n\}$ into c clusters. Since the assumption of a known or a previously fixed number of clusters is not realistic for many data analysis problems, there are techniques such as cluster validity index (CVI) analysis to determine the number of clusters.

A fuzzy clustering algorithm partitions the given dataset $\boldsymbol{X}$ into c number of overlapping clusters, forming a fuzzy partition matrix U which is a matrix of degree of membership of every object x_k, *k=1,…,n* in every cluster $i, i = 1, \dots, c$:

$$\mathbf{X}=\begin{bmatrix} x_{1,1} & \cdots & x_{1,nv} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,nv} \end{bmatrix} \qquad \mathbf{U}=\begin{bmatrix} \mu_{1,1} & \cdots & \mu_{1,c} \\ \vdots & \ddots & \vdots \\ \mu_{n,1} & \cdots & \mu_{n,c} \end{bmatrix}.$$

In the fuzzy clustering algorithm, each cluster is represented by a vector called "cluster center", $\boldsymbol{V} = (v_{1,}v_{2,} \dots, v_n)$. Let each cluster represented by a cluster prototype, v_i. FCM clustering algorithm tries to minimize an objective function with two pieces of prior information: number of clusters and fuzziness constant as follows:

$$min\, J(\boldsymbol{X}; \boldsymbol{U}, \boldsymbol{V}) = \sum_{i=1}^{c} \sum_{k=1}^{n} (\mu_{ik})^m d^2\,(x_k, v_i) \tag{2.21}$$

where c is the number of clusters, n is the number of objects (observations), and $v_i's$ are the cluster center.

In equation (2.21), m ∈(1,∞) represents the "degree of fuzziness" of the fuzzy clustering algorithm and it determines degree of overlapping of the clusters. *m=1* means no overlapping which represents a crisp clustering structure. Here $d^2\,(x_k, v_i)$ is a measure of distance between *k*th object and *i*th cluster's center. Squared distances satisfy that the objective function is non-negative, *J>0*. The objective function will be 0 when all data objects are cluster centers. When data objects are far away from cluster centers, the objective function will get larger.

In order to avoid trivial solutions, two constraints are imposed, and the FCM algorithm can be displayed as follows:

$$min\, J(\boldsymbol{X}; \boldsymbol{U}, \boldsymbol{V}) = \sum_{i=1}^{c} \sum_{k=1}^{n} (\mu_{ik})^m d^2\,(x_k, v_i) \tag{2.22}$$

$$s.t. \quad 0 \leq \mu_{ik} \leq 1 \quad \forall\, i, k, \tag{2.23}$$

$$\sum_{i=1}^{c} \mu_{ik} = 1, \qquad \forall\, k > 0 \tag{2.24}$$

$$0< \textstyle\sum_{k=1}^{n} \mu_{ik} < n\,, \qquad \forall\, i > 0. \tag{2.25}$$

where

μ_{ik}: membership degrees

d : distance between the observation and the cluster center.

The constraint in (2.24) implies that each row of partition matrix adds up to 1. The constraint in (2.25) implies that the column total of membership values cannot exceed the number of data vectors, n, nor it can be zero. This means that there is at least one member assigned to each cluster. As the distance formula, Euclidean formula given below is used

$$d_2(a,b) = \left[\sum_{i=1}^{nv} (\boldsymbol{a}_i - \boldsymbol{b}_i)^2\right]^{1/2} . \tag{2.26}$$

FCM algorithm stops according to a termination criterion, e.g., either after certain number of iterations, or if magnitude of separation of two nearest clusters is less than a predetermined value (ε), etc.

Cluster centers and membership values are calculated according to the following formula:

$$\boldsymbol{\mu}_{\boldsymbol{ik}}{}^{(t)} = \sum_{j=1}^{c} \left[\frac{d(x_k, v_i{}^{(t-1)})}{d(x_k, v_{ij}{}^{(t-1)})}\right]^{\frac{2}{1-m}} , \tag{2.27}$$

$$\boldsymbol{v}_i{}^{(t)} = \frac{\left(\sum_{k=1}^{n} (\mu_{ik}{}^{t}) x_k\right)}{\left(\sum_{k=1}^{n} (\mu_{ik}{}^{t})\right)} , \forall i = 1, \dots, c. \tag{2.28}$$

General FCM algorithm is as follows: Given data vectors $\mathbf{X} = \{\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{n}\}$, number of clusters, c , degree of fuzziness, m , and termination constant, ε (maximum iteration number), initialize the partition matrix , **U** , randomly.

Step 1 : Find the initial cluster centers by (2.28) using membership values of the initial partition matrix as inputs.

Step 2 : Start iteration $\mathrm{t} = 1, \ldots, \text{max iteration number}$

Step 2.1: Calculate membership values of each input data object k, in cluster i, $\mu_{ik}^{(t)}$, using the membership function in (2.27), where x_k are input data objects as vectors and $v_i^{(t-1)}$ are cluster centers from $(t-1)^{th}$ iteration,

Step 2.2 : Calculate the cluster center of each cluster i at iteration t, $v_i^{(r)}$ using the cluster center function in (2.28),

Step 2.3 : Stop if termination condition is satisfied, e.g., $\left|v_i^{(t)} - v_i^{(t-1)}\right| \leq \varepsilon$. Otherwise go to step 1 .

Since parameter m represents the degree of overlap of clusters, as m gets larger, the results become fuzzier and overlapping becomes wider. As m gets smaller, fuzzy clustering results become closer to a crisp model. m=1 is the same as the crisp clustering where there is no overlapping between clusters, and all membership values are $\mu_{ik} \in \{0,1\}$.

Earlier research (Türkşen, 1999) indicates that, m=2 should be used in system modeling analysis. In a study done by Özkan and Türkşen (2004), maximum and minimum values of *m* are proven to be within [1.4,2.6] based on Taylor expansion of the membership function.

Cluster Validity Indexes

Cluster validity indexes are given in the Handbook of Fuzzy Clustering and Data Analysis Toolbox of MATLAB. Here they are only summarized. Different validity measures have been proposed in the literature. None of them is perfect by oneself, therefore, in this study several indexes which are described below are used:

a) **Partition Coefficient (PC):** Measures the amount of "overlapping" between clusters. It is defined by Bezdek (1981) as follows:

$$PC\ (c) = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^2 \qquad (2.29)$$

where μ_{ij} is the membership of data point j in cluster i.

b) **Classiffication Entropy (CE):** It measures the fuzziness of the cluster partition only, which is similar to the Partition Coefficient:

$$CE\ (c) = -\ \frac{1}{N}\textstyle\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}\log(\mu_{ij}). \qquad (2.30)$$

c) **Partition Index (SC):** Is the ratio of the sum of compactness and separation of the clusters.

$$SC\ (c) = \sum_{i=1}^{c}\frac{\sum_{j=1}^{N}(\mu_{ij})^m\left\|x_j - v_i\right\|^2}{N_i\sum_{k=1}^{c}\|v_k - v_i\|^2}. \qquad (2.31)$$

For the above mentioned formula v_i denotes the cluster centers and N_i is the number of clusters.

d) **Xie and Beni's Index (XB):** It aims to quantify the ratio of the total variation within clusters and the separation of clusters:

$$XB\ (c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^m\left\|x_j - v_i\right\|^2}{N_i min_{i,j}\ \|v_k - v_i\|^2}. \qquad (2.32)$$

e) **Dunn's Index (DI):** This index is originally proposed to be used at the identification of "compact and well separated clusters".

$$DI\ (c) = min_{i \in c}\ \{min_{i \in c, i \neq j}\{\frac{min_{x \in C_i, y \in C_j,} d(x,y)}{max_{k \in c}\{max_{x,y \in C} d(x,y)\}}\}\} \tag{2.33}$$

where $d(x, y)$ is the distance between x and y.

f) **Alternative Dunn Index (ADI):** The aim of modifying the original Dunn's index was that the calculation becomes more simple, when the dissimilarity function between two clusters is rated in value from beneath by the triangle-nonequality

$$d(x,y) \geq \left|d(y, v_j) - d(x, v_j)\right|$$

$$ADI\ (c) = min_{i \in c}\ \{min_{i \in c, i \neq j}\{\frac{min_{x_i \in C_i, x_j \in C_j,} \left|d(y, v_j) - d(x_i, v_j)\right|}{max_{k \in c}\{max_{x,y \in C} d(x,y)\}}\}\}\ . \tag{2.34}$$

As it is mentioned earlier, cluster validity indexes are used to identify the initial algorithm parameters. During parameter selection, diagrams are drawn using MATLAB 9.0 and breaking points of the diagrams are chosen as the parameters of the algorithms. In Table 8 example of the diagrams are given. Since all of them are hardly decreases or increases at *c*=3, number of cluster is chosen as 3 in this case. There is also other common breaking point however less number of partitioning is suggested in the Handbook of Fuzzy Clustering and Data Analysis Toolbox of MATLAB. The other parameters are chosen similarly.

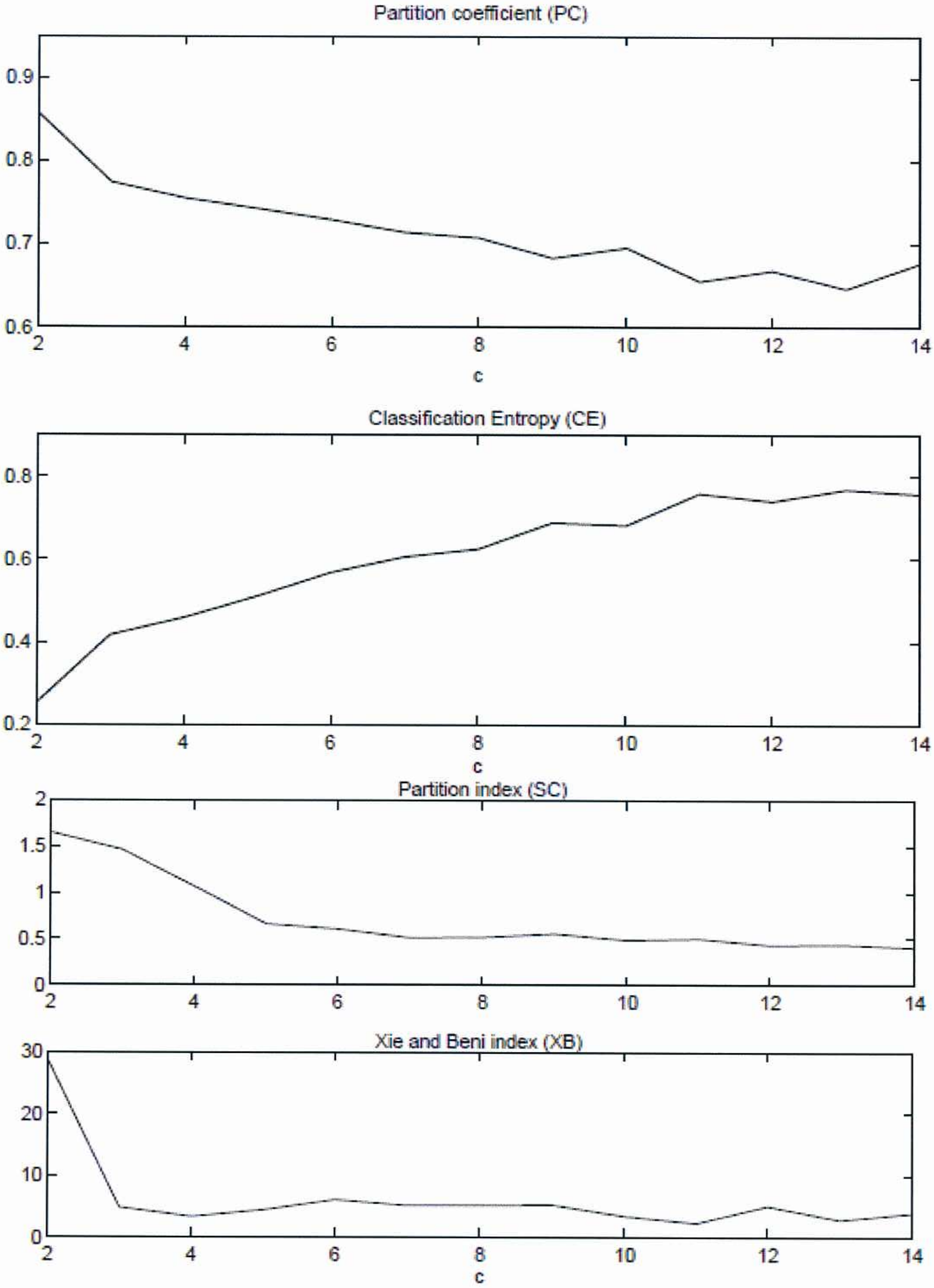

Figure 8 Cluster Validity Index Diagrams

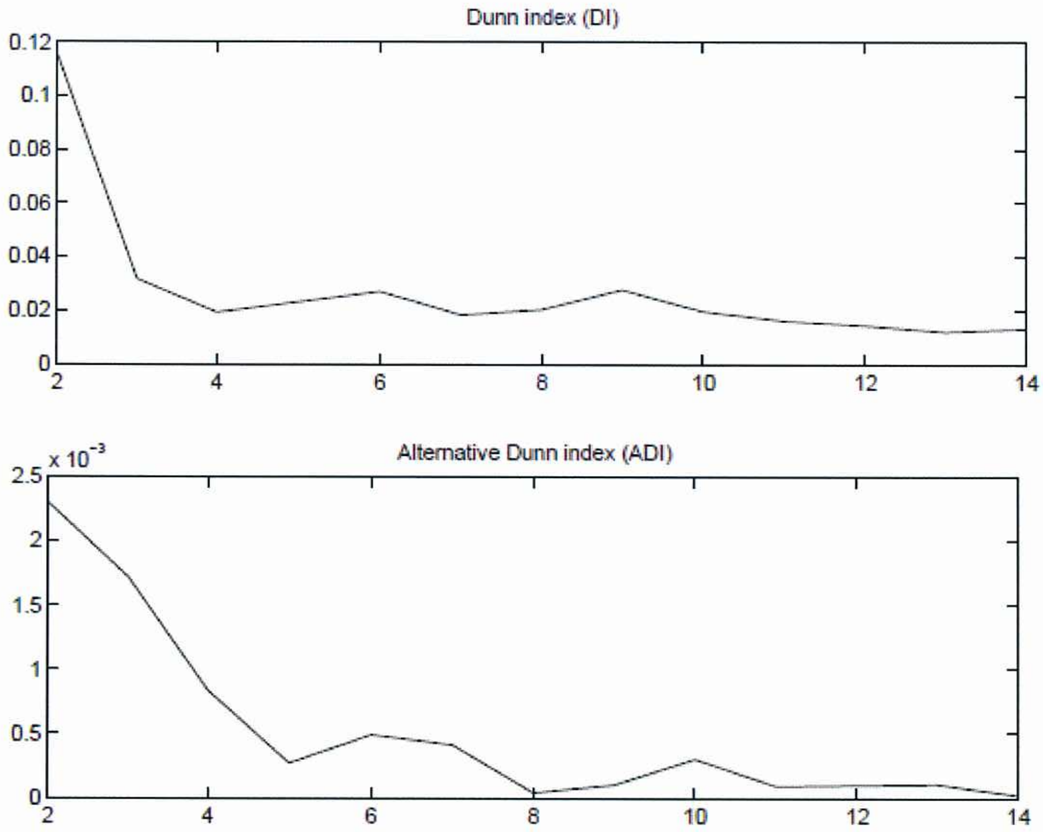

Figure 8 Cluster Validity Index Diagrams (cont.)

2.4. NONPARAMETRIC IMPROVED FUZZY CLASSIFIER FUNCTION

Nonparametric Improved Fuzzy Classifier Function developed by Özer (2009) is an improvement of the Improved Fuzzy Classifier Function (IFCF) approach. In Özer's (2009) study, performance of the NIFCF is compared with another fuzzy classification method, Fuzzy Classifier Function (FCF) and a statistical classification method, Logistic Regression (LGR). It is found that NIFCF method gives more satisfactory results compared with the other methods.

Improved Fuzzy Clustering (IFC) approach has been developed by Çelikyılmaz and Türkşen (2008) as an improvement of the FCM method. This proposed algorithm aims to transform membership values into powerful predictors to be used for approaches based on fuzzy functions. The prediction power of membership values are tried to be increased by using a function called interim fuzzy function, which is constructed to estimate output variable by using only membership values and their transformations (Çelikyılmaz and Türkşen, 2008). They compare the results of the FCM and IFC algorithms using an artificial data set. The comparison results indicate that membership values calculated by the IFC algorithm are better predictors of the output variable than the membership values calculated by the standard FCM algorithm.

FCF approach is the adaptation of FF approach to classification problems (Çelikyılmaz et al., 2007). This method is very similar to FF method, but a classification method is used for building a model for each cluster rather than a prediction method as in FF approach. IFCF method is the improved version of FCF method which uses the IFC algorithm rather than the FCM algorithm. Improved Fuzzy Functions (IFF) approach is developed by Çelikyılmaz and Türkşen (2008). This approach proposes to use IFC algorithm in the clustering phase of the FF approach.

In Özer's (2009) study, during the application of IFC algorithm using Logistic Regression as a classifier, a fitting problem has been met at the clustering phase of the algorithm. In order to overcome the fitting problem, a non-parametric method, MARS, in the clustering phase of the IFC method is proposed to use, which automates the model formation and selection of transformations of predictors as well as the selection of variables to find a best model fit. The clustering method, which proposes to use a nonparametric method, MARS, as a classifier, and the fuzzy classifier method, which proposes to use this method as a clustering algorithm are called Nonparametric Improved Fuzzy Clustering (NIFC) and Nonparametric Improved Fuzzy Classifier Function (NIFCF), respectively (Özer et al., 2009).

2.5. LINEAR REGRESSION

Regression analysis answers question about the dependence of a response variable on one or more predictors, including prediction of future values of a response, discovering which predictors are important, and estimating the impact of changing a predictor on the value of the response.

Linear regression was the first type of regression analysis to be studied rigorously and to be used extensively in practical applications. The reason is that models depending linearly on their unknown parameters are easier to fit than models related non-linearly to their parameters, and the statistical properties of the resulting estimators are easier to determine.

For estimating the unknown parameters in linear regression model, ordinary least squares (OLS) is used. This method minimizes the sum of squared distances between the observed responses in a set of data and the fitted responses from the regression model.

The goal of linear regression is to adjust the values of slope and intercept to find the line that best predicts Y from X. More precisely, the goal of regression is to minimize the sum of the squares of the vertical distances of the points from the line. The next values of the output variable can be estimated using linear regression by multiplying the observations with coefficients.

Assumptions of Linear Regression are:

- Homoscedasticity – the variance of the error terms is constant for each value of x. To check this, the residuals are plotted versus the X value(s). This plot has to be shapeless.
- Linearity – the relationship between each x and y is linear. To check this, again the residuals are plotted versus the X value(s). This plot has to be shapeless.
- Normally Distributed Error Terms – the error terms follow the normal distribution.
- Independence of Error Terms – successive residuals are not correlated.

For the variable selection stepwise regression is also used. Stepwise regression is a technique for choosing the variables to include in a multiple regression model.

Forward stepwise regression starts with no model terms. At each step it adds the most statistically significant term (the one with the highest F statistic or lowest p-value) until there are none left. Though there are two other stepwise approaches in the literature, in this study, forward stepwise regression is used.

CHAPTER 3

DESCRIPTION OF THE METAL CASTING DATA

A real life data set provided from a manufacturing company from the metal casting industry, which was studied by Bakır (2007) for the purpose of quality improvement is used in this study. The data for a particular product, the cylinder head, is studied, which is seen as an important part because of its effect on the performance of another part, the internal combustion engine. Figure 9 shows a typical cylinder head.

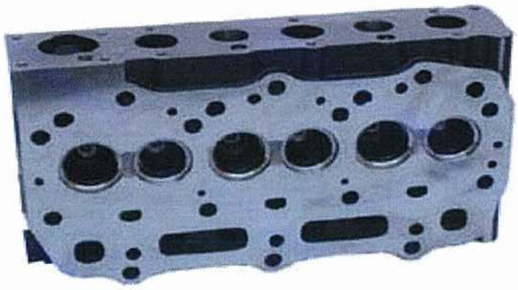

Figure 9 A Cylinder Head

(Source: Bakır, 2007)

Casting is the process of making product having complex shapes by pouring molten material into a mold and breaking out the solidified material from the mold.

The main reason of using this procedure to make products is the difficulty of other methods such as cutting from the metal. Other methods can also be not economical. Metal casting has three main subsequent processes: core making, molding, and melting.

The company performs batch production. A batch is a group of product produced in a certain day under the same process settings. Number of products produced in the batches is different. The company collects data to monitor batches and there is no way to know under which exact process values the individual items were produced. For that reason, process values of a batch represents all items belong to that batch. All of the process values are measured by sampling from products produced. Frequency of sampling varies among the variables because of the economical reasons or difficulties. Most of the measurements are taken during the production so that every batch has its separate value. However, some of them are taken once a week and considered as the values of batches performed during the following week. At the end of each batch production, all of the products are inspected for certain defect types and number of defective items is recorded according to the types of defect. Another problem arising here is the fact that if any defect type is observed on a product enough to reject it, no further analysis is performed to see the existence of other defect types. Only main cause to reject a product is recorded even more than one defect type are observed on the same product. Consequently, possible correlations between defect types are not provided by the data. Differences between frequencies of sampling results in lots of missing data and uncertain values of individual items forced us to aggregate data to the batch level. The initial data set is processed by Bakır (2007) and finally 36 process variables for 61 observations were gathered.

Because of the reasons mentioned above, uncertainty in the structure of the system is obvious. Most widely used conceptual basis for handling uncertainty is probability theory. On the other hand, when the systems in which imprecise data exist fuzziness is the source of uncertainty rather than randomness.

Because of the reasons mentioned above, the deviations between observations and estimations are supposed to be due to the indefiniteness of the system structure and fuzzy regression techniques is used for modeling in the scope of the study. In other words, in order to overcome all the indefiniteness, the fuzzy type of uncertainty is supposed to exist in the data, and fuzzy regression techniques are used for modeling.

Since the usage of high number of independent variables in fuzzy methods causes some shortcomings such as increasing spreads for estimated outputs (Nasrabadi and Nasrabadi 2004), collinearity between variables and increasing the calculation time (Wang and Tsaur, 2000), 36 process variables are decided to be reduced by a variable selection procedure. Using stepwise LR, 8 significant variables are chosen. During variable selection SPSS 9.0 is used. Variable selection results are given in appendix B.

Finally eight independent and one dependent variable is identified from the initial data set using variable selection. Since there is a confidentiality agreement with the metal casting manufacturing company, the names of the independent variables and the dependent variable cannot be given in the scope of the study. However, dependent variable is the defect rate determined by the manufacturing company. It is important to reduce the number of defects in order to improve the quality of the product and reduce the cost. And the independent variables are the variables which are already known as having important effects on the dependent variable such as oven temperature, humidity, gas permeability, viscosity, combustion loss, etc.

After variable selection procedure is carried out, metal casting data set is partitioned using a 3-fold and 3-replicate cross validation method in order to compare the performances of the methods. According to this approach, the data is randomly divided into three parts three different times (replicates). For each replicate, models are developed, each time using two different parts (folds) of the data, and the models are tested on the third part using several performance measures. Therefore, for the final representation, mean values of the performance parameter are calculated and located at the performance parameter tables.

Fuzzy methods may work with fuzzy data as well as crisp data. When the nature of the process variables is considered, it is realized that they can be thought of as interval numbers which have varying degrees of memberships. Although the names of the variables are confidential for the manufacturing company, oven temperature variable can be given as an example in order to explain the concept. Oven temperature changes in a certain interval. Even the set value remains the same, there are lots of factors affecting the current temperature of the oven. For this reason, there is another kind of indefiniteness which can be handled by possibility rather than probability. The temperature of the oven can be measured as 1420 ^{0} C, but it is known that the real temperature of the oven is between 1410 and 1430. Therefore, instead of identifying the temperature of the oven as crisp such as 1420, using fuzzy numbers which can be varied between 1410 and 1430 in a certain membership degree can be more appropriate. The fuzziness of the other variables can be explained similarly.

For that reason instead of using crisp values, fuzzy input and output observations are used. In order to provide fuzzy input and output observations, spread values of the variables are required. Since the data is prepared by taking averages over some number of daily observations, say n, one could think of using estimates of standard deviations, $\sigma_{\bar{x}}$, of these averages in the spread determination in the following manner:

$$\text{Spread} \cong k\sigma \cong k\sigma_{\bar{x}}, \sqrt{n}$$

Here k could be taken as 2 or 3. However, n is not the same for different days and variables. After consulting with the process engineers, it has been observed that their spread estimates are close to $\sigma_{\bar{x}}$ values. Hence, in the study, spreads are taken as the sample standard deviation of the observations. Yet, a sensitivity analysis is performed for some other values of them in Section 6.5.

The same data set is used by Bakır (2007) and Özer (2009). In Bakır's (2007) study, logistic regression and decision trees are used for modeling. At the end of the study, none of the final models were found to be significant. On the other hand, satisfactory results have been provided by the decision tree approach. Metal casting data set is also used by Özer (2009) in order to compare the results of the newly developed NIFCF model with FCF and Logistic Regression methods. At the end of the study, it is indicated that NIFCF has better performance than the Logistic Regression.

CHAPTER 4

GENERALIZATION OF THE HBS2 MODEL TO THE MULTI VARIABLE CASE

In this chapter, generalization of HBS2 model developed and expressed for one independent variable by Hojati et al. (2005) is presented. In order to use the method for multi variable metal casting data, a generalization is applied to the model. For simplicity, first, two independent variable generalization is done; then it is adapted to eight independent variables. If we write the model for two independent variables, the model is as follows:

$$\hat{Y}=A_0 + A_1 x_{i1} + A_2 x_{i2} \qquad i = 1, \dots, 61. \tag{4.1}$$

HBS2 requires 2^{k+1} constraints for each observation (where k is the number of independent variables other than intercept). Thus for this case, we have 2^{2+1} =8 constraints for each observation.

In the original model developed by Hojati et al. (2005), upper and lower point of the coefficient, and right and left point of the first independent variable are used for the purpose of determination of the observed fuzzy dependent variable. For current case, we have one new independent variable which is $\bar{x}_{i2}$ and we have to add two more points in order to determine new dimension added to the observation area to provide the minimum deviation.

Table 1 Illustration of the Model Indices for Two Independent Variable Case

Coefficients and Variables	Illustrations of the points	Abbreviations of the points
$(\alpha_0 + (1-H)c_0)$	Upper of α	$d_{i,k_{jh},U}$
$(\alpha_0 - (1-H)c_0)$	Lower of α	$d_{i,k_{jh},L}$
$(\bar{x}_{i1} + (1-H)f_{i1})$	Right of x_1	$d_{i,k_{1r}}$
$(\bar{x}_{i1} - (1-H)f_{i1})$	Left of x_1	$d_{i,k_{1l}}$
$(\bar{x}_{i2} + (1-H)f_{i2})$	Right of x_2	$d_{i,k_{2r}}$
$(\bar{x}_{i2} - (1-H)f_{i2})$	Left of x_2	$d_{i,k_{2l}}$

If we name the two new points as “outside” and “inside” and use the indices given in Table 1 the new model will be as follows;

HBS2 for two independent variables:

Minimize

$$\sum_{\substack{i=1 \\ \mathrm{m} \in \mathrm{M}}}^{61} (d^+_{im} + d^-_{im}), \tag{4.2}$$

$$M = \{(k_{1h}, k_{2h}, \dots, k_{8h}, v)\}, h \in \{r, l\}, v \in \{U, L\}$$

Subject to:

$$\sum_{j=0}^{2}[(\alpha_j + (1-H)c_j)(\bar{x}_{ij} - (1-H)f_{ij})] + d^+_{i,k_{1l},k_{2l},U} - d^-_{i,k_{1l},k_{2l},U} = \bar{y}_i + (1-H)e_i \tag{4.3}$$

$$for\ \ i = 1,2, \dots, 61$$

$$\sum_{j=0}^{1}[(\alpha_j + (1-H)c_j)(\bar{x}_{ij} - (1-H)f_{ij})] + (\alpha_2 + (1-H)c_2)(\bar{x}_{i2} + (1-H)f_{i2}) + \tag{4.4}$$

$$d^+_{i,k_{1l},k_{2r},U} - d^-_{i,k_{1l},k_{2r},U} = \bar{y}_i + (1-H)e_i \quad for\ \ i = 1,2, \dots, 61$$

$$\sum_{j=0}^{2}[(\alpha_j + (1-H)c_j)(\bar{x}_{ij} + (1-H)f_{ij})] + d^+_{i,k_{1r},k_{2r},U} - d^-_{i,k_{1r},k_{2r},U} = \bar{y}_i + (1-H)e_i \tag{4.5}$$

$$for\ \ i = 1,2, \dots, 61$$

$$\sum_{j=0}^{1}[(\alpha_j + (1-H)c_j)(\bar{x}_{ij} + (1-H)f_{ij})] + (\alpha_2 + (1-H)c_2)(\bar{x}_{i2} - (1-H)f_{i2}) + \\ d^{+}_{i,k_{1r},k_{2l},U} - d^{-}_{i,k_{1r},k_{2l},U} = \bar{y}_i + (1-H)e_i \qquad for \;\; i = 1,2,\dots,61 \tag{4.6}$$

$$\sum_{j=0}^{2}[(\alpha_j - (1-H)c_j)(\bar{x}_{ij} - (1-H)f_{ij})] + d^{+}_{i,k_{1l},k_{2l},L} - d^{-}_{i,k_{1l},k_{2l},L} = \bar{y}_i - (1-H)e_i \\ for \;\; i = 1,2,\dots,61 \tag{4.7}$$

$$\sum_{j=0}^{1}[(\alpha_j - (1-H)c_j)(\bar{x}_{ij} - (1-H)f_{ij})] + (\alpha_2 - (1-H)c_2)(\bar{x}_{i2} + (1-H)f_{i2}) + \\ d^{+}_{i,k_{1l},k_{2r},L} - d^{-}_{i,k_{1l},k_{2r},L} = \bar{y}_i - (1-H)e_i \qquad for \;\; i = 1,2,\dots,61 \tag{4.8}$$

$$\sum_{j=0}^{2}[(\alpha_j - (1-H)c_j)(\bar{x}_{ij} + (1-H)f_{ij})] + d^{+}_{i,k_{1r},k_{2r},L} - d^{-}_{i,k_{1r},k_{2r},L} \\ = \bar{y}_i - (1-H)e_i \\ for \;\; i = 1,2,\dots,61 \tag{4.9}$$

$$\sum_{j=0}^{1}[(\alpha_j - (1-H)c_j)(\bar{x}_{ij} + (1-H)f_{ij})] + (\alpha_2 - (1-H)c_2)(\bar{x}_{i2} - (1-H)f_{i2}) + \\ d^{+}_{i,k_{1r},k_{2l},L} - d^{-}_{i,k_{1r},k_{2l},L} = \bar{y}_i - (1-H)e_i \qquad for \;\; i = 1,2,\dots,61 \tag{4.10}$$

$$All \;\; d^{+} and \;\; d^{-} \geq 0 \qquad i = 1,2,\dots,61 \\ \alpha_j \;\; free \qquad c_j \geq 0 \quad for\, j = 0,1,2 \tag{4.11}$$

where

d : distance value between observations and estimations,

$\boldsymbol{\alpha_j}$: centers of the coefficients,

$\boldsymbol{c_j}$: spreads of coefficients,

$\boldsymbol{\bar{x}_{ij}}$: centers of the independent variables,

$\boldsymbol{f_{ij}}$: spreads of the independent variables,

$\boldsymbol{\bar{y}_\iota}$: centers of the dependent variable,

$\boldsymbol{e_i}$: spread of the dependent variable,

r : right,

l : left,

U : upper,

L : lower,

k_j: variable j.

While the original model developed by Hojati et al. (2005) indicates rectangular observation area, this model indicates a prisma since there is one more dimension added to the model, which is represented by indices "inside" and "outside". The graphical representation of the model can be seen in Figure 10.

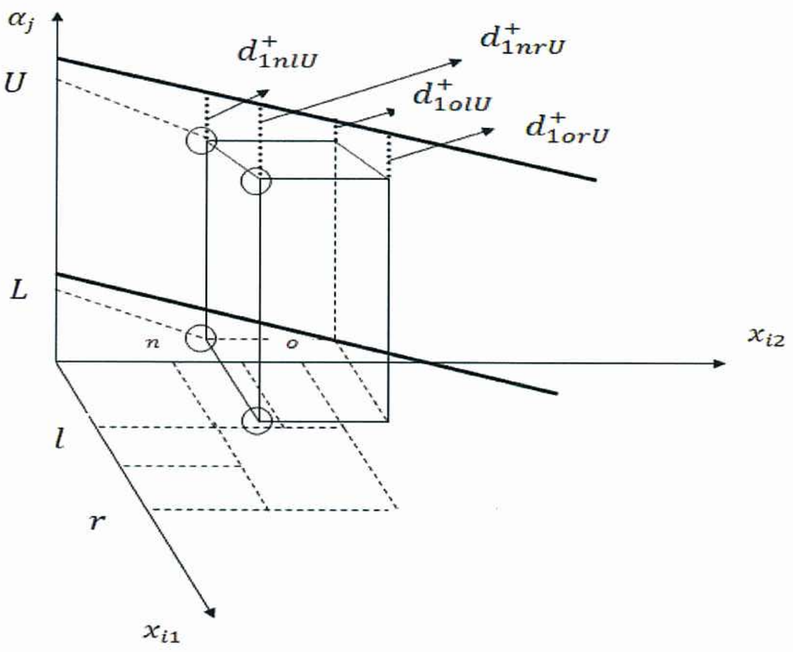

Figure 10 A Graphical Representation of HBS2 Model

In order to complete the main modeling problem, the HBS2 model is generalized to 8 independent variables and 61 observations that requires 2^{8+1} =512 constraints for each observation which is too much to solve and beside a main shortcoming of HBS2 model. For each independent variable, one new dimension is added to the model. The new dimensions are named as in Table 2.

Though it is not feasible to write 512 constraints manually, it is possible to write them accordingly to the index set *M* as defined in the model.

Table 2 Illustration of Indices for Eight Independent Variables Case

Coefficients and Variables	Illustrations of the points	Abbreviations of the points
$(\alpha_0 + (1-H)c_0)$	Upper of α	$d_{i,k_{jh},U}$
$(\alpha_0 - (1-H)c_0)$	Lower of α	$d_{i,k_{jh},L}$
$(\bar{x}_{i1} + (1-H)f_{i1})$	Right of x_1	$d_{i,k_{1r}}$
$(\bar{x}_{i1} - (1-H)f_{i1})$	Left of x_1	$d_{i,k_{1l}}$
$(\bar{x}_{i2} + (1-H)f_{i2})$	Right of x_2	$d_{i,k_{2r}}$
$(\bar{x}_{i2} - (1-H)f_{i2})$	Left of x_2	$d_{i,k_{2l}}$
$(\bar{x}_{i3} + (1-H)f_{i3})$	Right of x_3	$d_{i,k_{3r}}$
$(\bar{x}_{i3} - (1-H)f_{i3})$	Left of x_3	$d_{i,k_{3l}}$
$(\bar{x}_{i4} + (1-H)f_{i4})$	Right of x_4	$d_{i,k_{4r}}$
$(\bar{x}_{i4} - (1-H)f_{i4})$	Left of x_4	$d_{i,k_{4l}}$
$(\bar{x}_{i5} + (1-H)f_{i5})$	Right of x_5	$d_{i,k_{5r}}$
$(\bar{x}_{i5} - (1-H)f_{i5})$	Left of x_5	$d_{i,k_{5l}}$
$(\bar{x}_{i6} + (1-H)f_{i6})$	Right of x_6	$d_{i,k_{6r}}$
$(\bar{x}_{i6} - (1-H)f_{i6})$	Left of x_6	$d_{i,k_{6l}}$
$(\bar{x}_{i7} + (1-H)f_{i7})$	Right of x_7	$d_{i,k_{7r}}$
$(\bar{x}_{i7} - (1-H)f_{i7})$	Left of x_7	$d_{i,k_{7l}}$
$(\bar{x}_{i8} + (1-H)f_{i8})$	Right of x_8	$d_{i,k_{8r}}$
$(\bar{x}_{i8} - (1-H)f_{i8})$	Left of x_8	$d_{i,k_{8l}}$

In order to increase the performance of the reduced model some of the constraints selected randomly are added and the new model is written as follows:

HBS2-V2 (for eight independent variable)

Minimize

$$\sum_{\substack{i=1 \\ \mathrm{m} \in \mathrm{M}}}^{61} (d_{im}^{+} + d_{im}^{-}), \tag{4.12}$$

$$M = \{(k_{1h}, k_{2h}, \dots, k_{8h}, v)\}, h \in \{r, l\}, v \in \{U, L\}$$

Subject to:

$$\begin{gathered}\sum_{j=0}^{8}\left(\alpha_j+(1-H)c_j\right)\left(\bar{x}_{ij}-(1-H)f_{ij}\right)+\\ d^{+}_{i,k_{1l},k_{2l},k_{3l},k_{4l},k_{5l},k_{6l},k_{7l},k_{8l},U}-d^{-}_{i,k_{1l},k_{2l},k_{3l},k_{4l},k_{5l},k_{6l},k_{7l},k_{8l},U}=\bar{y}_i+(1-H)e_i\\ for\quad i=1,2,\dots,61\end{gathered}\tag{4.13}$$

$$\begin{gathered}\sum_{j=0}^{7}\left(\alpha_j+(1-H)c_j\right)\left(\bar{x}_{ij}-(1-H)f_{ij}\right)+(\alpha_8+(1-H)c_8)(\bar{x}_{i8}+(1-H)f_{i8})+\\ d^{+}_{i,k_{1l},k_{2l},k_{3l},k_{4l},k_{5l},k_{6l},k_{7l},k_{8r},U}-d^{-}_{i,k_{1l},k_{2l},k_{3l},k_{4l},k_{5l},k_{6l},k_{7l},k_{8r},U}=\bar{y}_i+(1-H)e_i\\ for\quad i=1,2,\dots,61\end{gathered}\tag{4.14}$$

$$\begin{gathered}\sum_{j=0}^{8}\left(\alpha_j+(1-H)c_j\right)\left(\bar{x}_{ij}+(1-H)f_{ij}\right)+\\ d^{+}_{i,k_{1r},k_{2r},k_{3r},k_{4r},k_{5r},k_{6r},k_{7r},k_{8r},U}-d^{-}_{i,k_{1r},k_{2r},k_{3r},k_{4r},k_{5r},k_{6r},k_{7r},k_{8r},U}=\bar{y}_i+(1-H)e_i\\ for\quad i=1,2,\dots,61\end{gathered}\tag{4.15}$$

$$\begin{gathered}\sum_{j=0}^{7}\left(\alpha_j+(1-H)c_j\right)\left(\bar{x}_{ij}+(1-H)f_{ij}\right)+(\alpha_8+(1-H)c_8)(\bar{x}_{i8}-(1-H)f_{i8})+\\ d^{+}_{i,k_{1r},k_{2r},k_{3r},k_{4r},k_{5r},k_{6r},k_{7r},k_{8l},U}-d^{-}_{i,k_{1r},k_{2r},k_{3r},k_{4r},k_{5r},k_{6r},k_{7r},k_{8l},U}\\ =\bar{y}_i+(1-H)e_i\\ for\quad i=1,2,\dots,61\end{gathered}\tag{4.16}$$

$$\begin{gathered}\sum_{j=0}^{8}\left(\alpha_j-(1-H)c_j\right)\left(\bar{x}_{ij}-(1-H)f_{ij}\right)+\\ d^{+}_{i,k_{1l},k_{2l},k_{3l},k_{4l},k_{5l},k_{6l},k_{7l},k_{8l},L}-d^{-}_{i,k_{1l},k_{2l},k_{3l},k_{4l},k_{5l},k_{6l},k_{7l},k_{8l},L}=\bar{y}_i-(1-H)e_i\\ for\quad i=1,2,\dots,61\end{gathered}\tag{4.17}$$

$$\begin{gathered}\sum_{j=0}^{7}\left(\alpha_j-(1-H)c_j\right)\left(\bar{x}_{ij}-(1-H)f_{ij}\right)+(\alpha_8-(1-H)c_8)(\bar{x}_{i8}+(1-H)f_{i8})+\\ d^{+}_{i,k_{1l},k_{2l},k_{3l},k_{4l},k_{5l},k_{6l},k_{7l},k_{8r},L}-d^{-}_{i,k_{1l},k_{2l},k_{3l},k_{4l},k_{5l},k_{6l},k_{7l},k_{8r},L}=\bar{y}_i-(1-H)e_i\\ for\quad i=1,2,\dots,61\end{gathered}\tag{4.18}$$

$$\begin{gathered}\sum_{j=0}^{8}\left(\alpha_j-(1-H)c_j\right)\left(\bar{x}_{ij}+(1-H)f_{ij}\right)+\\ d^{+}_{i,k_{1r},k_{2r},k_{3r},k_{4r},k_{5r},k_{6r},k_{7r},k_{8r},L}-d^{-}_{i,k_{1r},k_{2r},k_{3r},k_{4r},k_{5r},k_{6r},k_{7r},k_{8r},L}=\bar{y}_i-(1-H)e_i\\ for\quad i=1,2,\dots,61\end{gathered}\tag{4.19}$$

$$\sum_{j=0}^{7}\left(\alpha_j-(1-H)c_j\right)\left(\bar{x}_{ij}+(1-H)f_{ij}\right)+(\alpha_8-(1-H)c_8)(\bar{x}_{i8}-(1-H)f_{i8})+\tag{4.20}$$

$$d^{+}_{i,k_{1r},k_{2r},k_{3r},k_{4r},k_{5r},k_{6r},k_{7r},k_{8l},L} - d^{-}_{i,k_{1r},k_{2r},k_{3r},k_{4r},k_{5r},k_{6r},k_{7r},k_{8l},L}$$
$$= \bar{y}_i - (1 - H)e_i$$
$$for \quad i = 1,2, \dots ,61$$
$$All \;\; d_i{}^{+} and \;\; d_i{}^{-} \geq 0 \quad for \quad i = 1,2, \dots ,61$$
$$\alpha_j \;\; free, \qquad c_j \geq 0. \quad for \quad j = 1,2, \dots ,8 \qquad (4.21)$$

where

d : distance values between observations and estimations,

$\boldsymbol{\alpha_j}$: centers of the coefficients,

$\boldsymbol{c_j}$: spreads of coefficients,

$\boldsymbol{\bar{x}_{ij}}$: centers of the independent variables,

$\boldsymbol{f_{ij}}$: spreads of the independent variables,

$\boldsymbol{\bar{y}_i}$: center of the dependent variable,

$\boldsymbol{e_i}$: spread of the dependent variable,

r : right,

l : left,

U : upper,

L : lower,

$\boldsymbol{k_j}$: variable j.

CHAPTER 5

NON PARAMETRIC IMPROVED FUZZY FUNCTIONS

Nonparametric Improved Fuzzy Function (NIFF) method proposed and used in this chapter is inspired by the method developed by Özer (2009) named as Nonparametric Improved Fuzzy Classifier Function (NIFCF).

NIFF method is very similar to NIFCF method. In NIFF method, Linear Regression method is used as prediction method for building model rather than a classification method as in NIFCF method. The improvement phases of the NIFCF and NIFF is demonstrated in Figure 11.

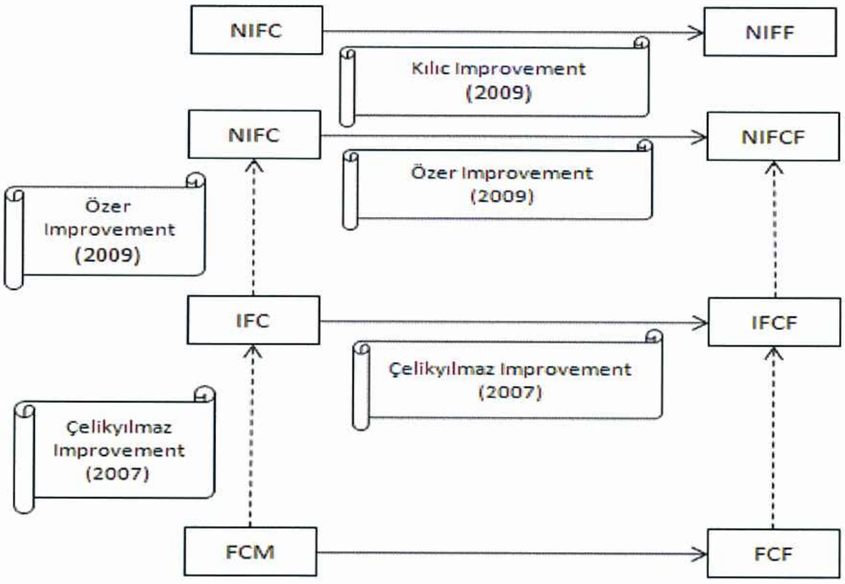

Figure 11 Summary of Improvements Involving Fuzzy Functions

NIFCF algorithm is given in the Özer's (2009) study. The steps of the algorithm taken from the Özer's (2009) study are shown as follows in order to explain the differences between NIFCF and NIFF steps:

Steps of training algorithm for NIFCF:

1. Set initial parameter, α, which is the level used for eliminating the points farther away from the cluster centers.
2. Calculate cluster centers for input-output variables, $\boldsymbol{v}(\boldsymbol{XY})_i$ and interim fuzzy functions for each cluster using NIFC algorithm.

   $$\boldsymbol{v}(\boldsymbol{XY})_i = \left\{v(x_1)_i, \dots, v(x_p)_i, v(y)_i\right\}$$

   where

   $v(x_j)_i$:cluster center of the j^{th} independent variable for the i^{th} cluster,

   $v(y)_i$:cluster center of the dependent variable for the i^{th} cluster,
3. For each cluster $i = 1, \dots, c$

   3.1. For each observation number $k = 1, \dots, n$

   Using cluster centers for input space, $\boldsymbol{v}(\boldsymbol{X})_i = \left\{v(x_1)_i, \dots, v(x_p)_i\right\}$

   3.1.1. Calculate membership values for input space, u_{ik}.

   $$u_{ik} = \left(\sum_{j=1}^{c} \left[\frac{\|X_k - \boldsymbol{v}(\boldsymbol{X})_i\|^2 + SE_{ik}}{\|X_k - \boldsymbol{v}(\boldsymbol{X})_j\|^2 + SE_{jk}} \right]^{\frac{1}{m-1}} \right)^{-1}$$

   where

   SE_{ik}: squared error term between the actual output and predicted output value of the k^{th} observation using interim fuzzy function calculated for the i^{th} cluster at step 2.

   3.1.2. Calculate alpha-cut membership values, μ_{ik}:

   $$\mu_{ik} = \{u_{ik} \geq \alpha\}$$

   3.1.3. Calculate normalized membership values,

   $$\gamma_{ik} = \frac{\mu_{ik}}{\sum_{j=1}^{c} \mu_{jk}}$$

3.2. Determine the new augmented input matrix for each cluster *i*, $\mathbf{\Phi_i}$, using observations selected according to α-cut level. $\mathbf{\Phi_i}$ matrix is composed of input variable matrix, $\mathbf{X_i^\alpha}$, vector of normalized membership values for the cluster *i*, γ_i, and the matrix composed of their selected transformations, $\gamma_i{}'$, such as $\gamma_i{}^2, \gamma_i{}^3, \gamma_i{}^m$, exp($\gamma_i$), log((**1**-$\gamma_i$)/ γ_i).

$\mathbf{\Phi_i}(\mathbf{X}, \gamma_i) = [\mathbf{X_i^\alpha} \quad \gamma_i \quad \gamma_i']$

where

$\mathbf{X_i^\alpha} = \{x_k \in \mathbf{X} \,|u_{ik}(x_k) \geq \alpha, k = 1, \dots, n\}$

3.3. Using Logistic Regression, calculate a fuzzy classifier function using new augmented matrix $\mathbf{\Phi_i}(\mathbf{X}, \gamma_i)$.

At the last step of the algorithm which is 3.3, in NIFCF method, logistic regression is used for the construction of the model. On the other hand, in NIFF method, linear regression is used rather than logistic regression for the construction of the model. This is the main and only difference between NIFF and NIFCF.

The motivation to propose the NIFF method is the same as NIFCF; i.e., in order to be able to partition data into clusters, a model should be fitted at each iteration. If a model cannot be formed at any iteration of the loop, the algorithm is terminated and the data cannot be clustered. Since it may not always be possible to fit a model (like as metal casting data set), in order to overcome all of these fitting problems, it is proposed to use a non-parametric method, MARS, in the clustering phase of the IFC method.

CHAPTER 6

PERFORMANCE ANALYSIS

In this chapter, the performance analysis of the HBS2, HBS2-V2, FF and NIFF methods are realized in terms of different variable case solutions which are two and eight variables. In order to compare the models' performances comparison criteria which are explained in Appendix A in detail are used.

6.1. PERFORMANCE ANALYSIS FOR HBS2 MODEL WITH TWO VARIABLES

The solutions obtained from HBS2 model which are gathered from complete constraints set and two variables are in Table 3. In order to compare the results, linear regression and Fuzzy Functions also applied to the metal casting data set. As it is mentioned in Chapter 2, in the first step of the FF algorithm, *m* and *c* values are determined by the user according to the values of the cluster validity indexes. In this study, all of the models are coded in MATLAB 9.0 using functions from its optimization toolbox.Also for the determination of the *m* and *c* values according to the breaking points of the cluster validity index lines, MATLAB 9.0 is used and the selection methodology is explained in detail in Chapter 2. All of the calculations are made for the H=0.5 value. The solutions obtained are given in Table 3. As it is seen from the Table 3, results obtained from HBS2, LR and FF are considerably close to each other. Nonetheless, it can be said that performance parameters of FF is better than the other models for some of the measures.

Table 3 Performance Measures for Two Variables

	LR	HBS2	FF	NIFF
Mean Absolute Error	0.029	0.025	0.022	0.020
Mean Square Error	0.003	0.003	0.003	0.003
Root Mean Square Error	0.050	0.050	0.050	0.050
r	0.541	0.521	0.560	0.560
R^2	0.297	0.279	0.314	0.314
ADJ R^2	0.246	0.226	0.263	0.263
PWI1	0.841	0.922	0.999	0.999
PWI2	0.967	0.955	0.999	0.999

On the other hand, since these models include only two of the eight variables found significant for this case, it can be seen that satisfactory results have not been provided in terms of R^2 and Adj R^2 for each model.

In the performance analysis LR is used to compare performance of the HBS2 model with conventional statistical regression techniques and FF is used to compare with fuzzy regression techniques. And according to the results, it can be seen that for the metal casting data set, HBS2 performs considerably close solutions to the conventional statistical regression. On the other hand performance of the FF method is superior.

This analysis is just made for seeing the performance of the HBS2 model for full constraints performance; therefore, further analysis is not realized since the aim of the study is related with whole casting data set.

6.2. PERFORMANCE ANALYSIS FOR LR, HBS2, HBS2-V2, FF AND NIFF

Since obtaining solutions for HBS2 model with high number of constraints like casting data is hard, Hojati et al. (2005) discussed elimination of some of the constraints and they eliminated left upper and right lower constraints for all independent variables. Similar to that discussion, metal casting data set is modeled using the incomplete constraints set as proposed by original study and the results obtained from average of nine replications are demonstrated at the HBS2 column of Table 4.

Table 4 Performance Measures of LR, HBS2, HBS2-V2, FF and NIFF

	LR	HBS2	HBS2-V2	FF	NIFF
Mean Absolute Error	0.022	0.049	0.036	0.024	0.024
Mean Square Error	0.001	0.005	0.004	0.002	0.002
Root Mean Square Error	0.030	0.065	0.057	0.043	0.041
r	0.800	0.558	0.607	0.650	0.659
R^2	0.645	0.323	0.385	0.429	0.448
ADJ R^2	0.514	0.074	0.122	0.263	0.274
PWI1	0.938	0.877	0.895	0.909	0.909
PWI2	0.989	0.938	0.939	0.971	0.964

As it is expected, the performance of the HBS2 model is considerably inadequate compared with LR and FF. In order to improve the performance of the HBS2 model for this case, it is expanded by the addition of new constraints as given in (5.12)-(5.21) and the performance obtained from the new model is demonstrated at the HBS2-V2 column of Table 4. Additional constraints which indicate the higher and lower points of the $^{-}$are chosen randomly. It can be seen from the table that the performance of the model is improved in terms of performance criteria. The other methods solutions obtained from eight independent variables are demonstrated in related columns.

These methods are statistically compared by using One-way ANOVA for each measure separately in order to see whether there is a statistically significant difference between them according to the performance measures mentioned above. One-way ANOVA test is performed using SPSS 9.0. Although, Repeated ANOVA is more appropriate to use for this hypothesis testing, the results of One-way ANOVA shown in **Table 5** indicate strong significance of the differences among the methods. Hence, it is chosen to continue with multiple comparison tests using

Tukey's approach. Although the independence assumption of One-way ANOVA is not satisfied, the constant variance and normality assumptions are generally satisfied for this data.

Table 5 ANOVA Results

ANOVA						
		Sum of Squares	**df**	**Mean Square**	**F**	**Sig.**
MAE	Methods	4.49E-03	4	1.12E-03	7.093	0
	Error	6.33E-03	40	1.58E-04		
	Total	1.08E-02	44			
MSE	Methods	8.55E-05	4	2.14E-05	6.401	0
	Error	1.34E-04	40	3.34E-06		
	Total	2.19E-04	44			
RMSE	Methods	6.93E-03	4	1.73E-03	6.722	0
	Error	1.03E-02	40	2.58E-04		
	Total	1.72E-02	44			
R	Methods	0.295	4	7.38E-02	6.183	0.001
	Error	0.477	40	1.19E-02		
	Total	0.773	44			
$\mathbf{R}^2$	Methods	0.528	4	0.132	6.738	0
	Error	0.783	40	1.96E-02		
	Total	1.311	44			
ADJ $\mathbf{R}^2$	Methods	1.063	4	0.266	9.904	0
	Error	1.073	40	2.68E-02		
	Total	2.136	44			
PWI1	Methods	1.84E-02	4	4.59E-03	2.972	0.031
	Error	6.17E-02	40	1.54E-03		
	Total	8.01E-02	44			
PWI2	Methods	1.72E-02	4	4.30E-03	5.566	0.001
	Error	3.09E-02	40	7.73E-04		
	Total	4.81E-02	44			

The results of Tukey's multiple comparison tests with family error rate 0.05 are shown in Table 6.

Table 6 Multiple Comparison Results

		HBS2	HBS2-V2	FF	NIFF
MAE	LR	0,001 (<)	0,177	0,997	0,997
	HBS2		0,206	0,002 (>)	0,002 (>)
	ImpHBS2			0,327	0,327
	FF				0,999
MSE	LR	0,001 (<)	0,033 (<)	0,794	0,876
	HBS2		0,629	0,016 (>)	0,01 (>)
	ImpHBS2			0,331	0,249
	FF				0,999
RMSE	LR	0 (<)	0,009 (<)	0,508	0,611
	HBS2		0,801	0,033 (>)	0,021 (>)
	ImpHBS2			0,321	0,244
	FF				0,999
R	LR	0 (<)	0,005 (<)	0,044 (<)	0,068 (<)
	HBS2		0,049>	0,391	0,296
	ImpHBS2			0,919	0,847
	FF				0,999
R^2	LR	0 (<)	0,003 (<)	0,018 (<)	0,036 (<)
	HBS2		0,049 (>)	0,505	0,343
	ImpHBS2			0,963	0,877
	FF				0,999
ADJ R^2	LR	0 (<)	0 (<)	0,018 (<)	0,027 (<)
	HBS2		0,171	0,123	0,091
	ImpHBS2			0,371	0,297
	FF				0,999
PWI1	LR	0,016 (<)	0,155	0,526	0,526
	HBS2		0,86	0,413	0,413
	ImpHBS2			0,938	0,938
	FF				0,999
PWI2	LR	0,003 (<)	0,003 (<)	0,634	0,313
	HBS2		1	0,116	0,317
	ImpHBS2			0,118	0,322
	FF				0,982

Table **6** shows that the LR model performance is better than the HBS2 model performance with respect to all performance measures at the significance level of α=0.05. Since the constraints of the HBS2 model are not complete, poor model performance is not unexpected here. In order to overcome the poor performance problem, proposed HBS2-V2 model has better performance than HBS2 with respect to r and R^2. According to these results, it can be said that adding constraints increases the performance of the HBS2 model. Since to write all the constraints are not always practical, the most suitable constraints can be added and the model performance can be increased.

When compared with other fuzzy models, Table **6** shows that FF and NIFF methods have significantly better performances than HBS2 according to MAE, MSE and RMSE for the metal casting data set. However, there is not any significant difference detected between fuzzy classification methods, FF and NIFF.

6.3. MODEL INTERPRETATION

In the previous section, different fuzzy regression models are constructed and performances of the models are compared. In order to make necessary interpretations regarding the relationships between dependent and independent variables, NIFF and HBS2 models formed are examined in this section.

6.3.1. NIFF Model Interpretation

NIFF model is as follows;

$$\begin{aligned}\widetilde{Y} = 0.000344 - 0.011595\, x_1 - 0.977558 x_2 + 0.111752\, x_3 \\ - 0.000554 x_4 - 0.185520 x_5 + 0.107427 x_6 \\ - 0.011623 x_7 - 0.000125 x_8 + 5.297141 x_m \\ - 18.808484 x_{t1} - 0.536847 x_{t2}\end{aligned} \qquad (4.22)$$

As it can be seen from the model, coefficient value of x_0 is 0.000344 , which means that when all the independent variables are zero, estimation of defect rate is close to zero percent. Besides x_2 has the maximum effect on the dependent variable compared with the other independent variables. If x_2 changes one unit and all the other variables remains the same, the defect rate decreases by 0.977558.

In the model, there are three more variables which are x_m, x_{t1} and x_{t2} indicating the normalized membership value and transformations of the independent variables, respectively. As it can be seen x_{t1} has the most negative effect on the dependent variable. Also x_m has a considerably higher positive effect on the dependent variable compared with the other independent variables.

When the model is used for the prediction purpose, after observation values are gathered, values of x_m, x_{t1} and x_{t2} variables must be calculated for the new observed data set. The calculation methods of these variables have already been mentioned in the algorithm in Chapter 5. The main and only difference to recalculate the variables is to use observed data set for estimation.

6.3.2. HBS2 Model Interpretation

HBS2 model is as follows;

$$\widetilde{Y} = (-0.034\,,0.000052) + (0.00010987\,,0.00004\,)\,x_1 + (-0.007, 0\,)x_2 + (-0.072, 0\,)x_3 + (0.028, 0.0007\,)x_4 + (-0{,}000507, 0.0003\,)x_5 + (-0.047, 0\,)x_6 + (0.114, 0.00102\,)x_7 + (-0.015, 0.000014\,)x_8$$

In the model the midpoints and the spread values of the coefficients of each independent variable are shown respectively. As it can be seen from the model, midpoint of the coefficient value of x_0 is -0.034 , which means that when all the independent variables are zero, estimation of defect rate will be that value. The spread value of x_0 is 0.000052 meaning that the value of the x_0 can change between -0.03405 and -0.033948.

It can be seen from the model that x_7 has the maximum effect on the fuzzy dependent variable compared with the other independent variables. If x_7 changes one unit and all the other variables remains the same, the defect rate increases between 0.11298 and 0.11502.

6.4. PERFORMANCE ANALYSIS FOR DIFFERENT H VALUES OF HBS2 MODEL

In Chapter 2, it is mentioned that, H value is determined by the user and proper selection of H value is important for the fuzzy regression model. In order to see the effects of H value selection, HBS2-V2 model is applied for different H values which are $H = 0.5$ and $H = 1$ and the performance of these models are compared in terms of the performance measures.

As it is seen from Table 7, HBS2-V2 model performs better when $H = 0.5$ according to MAE, r ,R^2 ,ADJR^2 , PWI1 and PWI2 and all the performance measure values are equal for $H = 0.5$ and $H = 0$. Although $H = 0$ produces same results with $H = 0.5$, since data set is not reliable we suggest $H = 0.5$ for this case. Besides when the spread values of the dependent variable is considered, it can be said that increasing H value increases the spread of the dependent variable.

Table 7 Performance Measures of HBS2-V2 for Different H Values

	HBS2-V2 **H=0**	**HBS2-V2** **H=0.5**	**HBS2-V2** **H=1**
Mean Absolute Error	0.036	0.036	0,038
Mean Square Error	0.004	0.004	0,003
Root Mean Square Error	0.057	0.057	0,054
r	0.607	0.607	0,576
R^2	0.385	0.385	0,349
ADJ R^2	0.122	0.122	0,109
PWI1	0.895	0.895	0,906
PWI2	0.939	0.939	0,950

6.5. PERFORMANCE ANALYSIS FOR DIFFERENT SPREAD OF INDEPENDENT VARIABLES

In Chapter 3, it is mentioned that in order to provide fuzzy observations, the crisp observations are identified as midpoints of the fuzzy observations and one standard deviation from the center is identified as the spread of the fuzzy observations. It is also mentioned that in order to validate the calculated intervals, confirmations are taken from the experts of the manufacturing processes. The performance of the models can be affected by this assumption. In order to analyze how model performance is affected by the spread level, spread of the independent variables is increased and this new data set is modeled.

Performance results are as shown in Table 8. It can be seen from the table that for the metal casting data set, model performance is getting worse by the increase of spread values. This result also supports the appropriateness of the fuzzification method used in the study.

Table 8 Performance Measures of HBS2 for Different Spread Levels

	HBS2 ($_{\bar{x}}$)	**HBS2 ($6_{\bar{x}}$)**
Mean Absolute Error	0.025	0.038
Mean Square Error	0.003	0.003
Root Mean Square Error	0.050	0.058
r	0.521	0.044
R^2	0.279	0.20
ADJ R^2	0.226	0.145
PWI1	0.922	0.876
PWI2	0.955	0.945

CHAPTER 7

CONCLUSIONS & SUGGESTIONS FOR FURTHER STUDIES

This study has two purposes. One of them is to model a metal casting process quality in order to reduce the number of defects by identifying the process variables which have the most important effects on a certain defect type. The second purpose is to test and develop certain fuzzy regression approaches for the case data set.

In order to do that, first the nature of the uncertainty in the problem is considered prior to selection of an appropriate method. It is decided to use Fuzzy regression methods for the modeling of the casting data. Since both dependent and independent variables of the casting data are considered fuzzy, HBS2 Fuzzy regression method is selected. In order to compare the performance of HBS2 method with a conventional statistical method, LR is used and to make comparison with fuzzy regression methods, FF is used.

Initially, the HBS2 method is applied with complete constraint set for two independent variables. It is seen that the performance results gathered from LR, HBS2 and FF are considerably close to each other but HBS2 and FF have better performance than LR. Moreover, FF is superior for some of the performance measures. Furthermore, HBS2 method is generalized to multi variable modeling case. It is concluded that the ease of use of the HBS2 model is closely related with the number of constraints included.

As it is expected for the metal casting data set, performance of the model is found to be inadequate. In order to overcome the performance problem, more constraints are added and the new version of the model is named as HBS2-V2. It is concluded that for some performance measures, HBS2-V2 has statistically better performance than the HBS2 model for the metal casting data. In other words, adding constraints increases the performance of the HBS2 model. It is obvious that the highest model performance will be obtained when all constraints are included. Since writing all the constraints is not efficient, the model performance can be increased by adding more constraints. Since the selection of the constraints becomes more important, as a further study, design of experiments methods can be used for the identification of the optimum constraint set. So maximum performance can be achieved with minimum number of constraints.

Furthermore, HBS2 method is also compared with fuzzy regression methods. As fuzzy regression method FF is selected. And it is concluded that for some performance measures FF method has statistically better performance than the HBS2 model for the metal casting data.

Finally, another fuzzy regression method which is triggered by the Özer's (2009) study named NIFF is proposed. This method is constructed by using Linear Regression method as prediction method for building model rather than a classification method as in NIFCF method. This method is also applied to the casting data and although the results gathered from performance measures indicates that the results are better than the FF, there is not any statistically significant difference detected between them.

As a result, for the metal casting data, it can be said that in case there is less number of variables and poor model performance which increases the fuzzy type of indefiniteness of the system, FF and NIFF can be used for modeling in order to increase the overall performance of the model. The disadvantage of the NIFCF method is identified as slower running of the program (Approximately 30 minutes for each FCM parameter identification). But by increasing the code efficiency, this disadvantage can be eliminated.

Since all of this analysis is realized in a single data set, it is not possible to conclude final decisions about the overall performances of the models. In other words, models may give different results on different data sets. For that reason as further study, it is useful to make similar analysis with different data sets.

REFERENCES

1. **Bakır, B.** (2007). “Defect Cause Modeling with Decision Tree and Regression Analysis: A Case Study in Casting Industry”. *M.S Thesis, Middle East Technical University, Ankara, Turkey.*

2. **Belman, R.E., Zadeh, L.A.,** (1970). “Decision Making In a Fuzzy Environment”. *Management Science 17, B141-B164.*

3. **Bezdek, J. C.** (1981). “Pattern Recognition with Fuzzy Objective Function Algorithms”. *New York: Plenum Press.*

4. **Celmins, A.** (1987). “Least Squares Model Fitting To Fuzzy Vector Data”. *Fuzzy Sets and Systems 22 245–269.*

5. **Coppi, R.** (2007). “Management of Uncertainty in Statistical Reasoning: The Case of Regression Analysis”. International Journal of Approximate Reasoning. Volume 47, 284-305

6. **Çelikyilmaz, A.** (2005). “Fuzzy Functions with Support Vector Machines”. *M.Sc Thesis, Information Science, Industrial Engineering Department, University Of Toronto.*

7. **Çelikyılmaz, A., Türkşen, I.B.** (2008). “Enhanced Fuzzy System Models With Improved Fuzzy Clustering Algorithm”. *IEEE Transactions On Fuzzy Systems, 16(3), 779-794*

8. **Çelikyılmaz, A., Türkşen, I.B., Aktaş, R., Doğanay, M.M. And Ceylan, N.B.** (2007). “A New Classifier Design with Fuzzy Functions”. *Rough Sets, Fuzzy Sets, Data Mining And Granular Computing, 4482, 136-143.*

9. **Gharpuray, M., Tanaka, H., Fan, L., and Lai, F.** (1986). "Fuzzy Linear Regression Analysis of Cellulose Hydrolysis". *Chemical Engineering Communications 41, 299-314*

10. **Hojati, M., Bector, C.R., Smimou, K.** (2005). "A Simple Method for Computation of Fuzzy Linear Regression". *European Journal of Operational Research, Computing, Artificial Intelligence and Computer Technology, Vol.166, Pp. 172-184.*

11. **Jozsef, S.** (1992). "On The Effect of Linear Data Transformations In Possibilistic Fuzzy Linear Regression". *Fuzzy Sets and Systems 45 185–188.*

12. **Kim, K. J., Moskowitz, H., Koksalan M.** (1996). "Fuzzy Versus Statistical Linear Regression" *European Journal of Operational Research 92417–434.*

13. **Nasrabadi, M.M., Nasrabadi, E.** (2004). "A Mathematical-Programming Approach to Fuzzy Linear Regression Analysis". *Applied Mathematics and Computation, 155, 873-881.*

14. **Ozelkan, E.C., Duckstein L.** (2000). "Multi-Objective Fuzzy Regression: A General Framework". *Computers And Operations Research 27 635–652.*

15. **Özer, G.** (2009). "Fuzzy Classification Models Based on Tanaka's Fuzzy Linear Regression Approach and Nonparametric Improved Fuzzy Classifier Functions". *M.S Thesis, Middle East Technical University, Ankara, Turkey.*

16. **Özer, G., Kılıç, T., Kartal, E., Batmaz, İ., Bayrak, Ö.T., Köksal, G. And Türkşen, İ.B.** (2009). "A Nonparametric Improved Fuzzy Classifier Function Approach For Classification Based On Customer Satisfaction Survey Data". *Book Of Abstracts Of The 23rd European Conference On Operational Research. Bonn, Germany.*

17. **Özkan, I., Türkşen, I.B.** (2004). "Entropy Assessment of Type-2 Fuzziness", IEEE International Conference on Fuzzy Systems 2, 1111–1115.

18. **Peters, G.** (1994). "Fuzzy Linear Regression with Fuzzy Intervals". *Fuzzy Sets and Systems 63, 45–55.*

19. **Ross, T.J.** (2004). "Fuzzy Logic with Engineering Applications". *John Willey & Sons, Inc. New York, NY.*

20. **Savic, D.A., Pedrycz, W.** (1991). "Evaluation of Fuzzy Linear Regression Models". *Fuzzy Sets and Systems 39 51–63.*

21. **Sawaka, M., Yano, H.** (1992). "Multiobjective Fuzzy Linear Regression Analysis for Fuzzy Input–Output Data". *Fuzzy Sets and Systems 47 173–181.*

22. **Tanaka, H., Uejima, S., Asai, K.** (1982). "Linear Regression Analysis with Fuzzy Model". *IEEE Transactions on Systems, Man, and Cybernetics 12 903–907.*

23. **Tanaka, H., Ishibuchi, H.** (1991). "Identification of Possibilistic Linear Systems by Quadratic Membership Functions of Fuzzy Parameters". *Fuzzy Sets and Systems 41 145–160.*

24. **Tanaka, H., Hayashi, I., Watada, J.** (1989). "Possibilistic Linear Regression Analysis for Fuzzy Data". *European Journal of Operational Research 40 389–396.*

25. **Türkşen, I.B.** (1999). "Type I And Type II Fuzzy System Modeling". *Fuzzy Sets and Systems, Vol. 106, Pp. 11-34.*

26. **Türkşen, I.B.** (2005). "Fuzzy Functions with LSE". *International Journal Of Fuzzy Systems, 6(2), 157-121*

27. **Türkşen, I.B., Çelikyılmaz, A.** (2006). "Comparison Of Fuzzy Functions With Fuzzy Rule Base Approaches". *International Journal of Fuzzy Systems, 8(3), 137-149.*

28. **Wang, H.F., Tsaur, R.C.** (2000). "Insight of a Fuzzy Regression Model". *Fuzzy Sets and Systems, 112(3), 355-369.*

29. **Zadeh, L.A.** (1965). "Fuzzy Sets". *Information and Control 338-353.*

30. **Zadeh, L.A.** (1975). "The Concept Of A Linguistic Variable And Its Application To Approximate Reasoning". *Information And Science 8,199-249*

APPENDIX A

COMPARISON CRITERIA

In order to compare the model performances, comparison criteria discussed below are used.

a) Mean Absolute Error

The mean absolute error is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The mean absolute error (MAE) is given by

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \qquad (6.1)$$

$$e_i = Y_i - \hat{Y}_i. \qquad (6.2)$$

As the name suggests, the mean absolute error is an average of the absolute errors.

b) Mean Square Error

The mean square error (MSE) of an estimator is one of many ways to quantify the amount by which an estimator differs from the true value of the quantity being estimated. MSE measures the average of the square of the "error" The error is the amount by which the estimator differs from the quantity to be estimated:

The difference occurs because of randomness or because the estimator doesn't account for information that could produce a more accurate estimate. The mean square error is calculated as

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e_i{}^2. \tag{6.3}$$

c) Root Mean Square Error

The root mean square error (RMSE) is a frequently-used measure of the differences between values predicted by a model or an estimator and the values actually observed from the thing being modeled or estimated. RMSE is a good measure of accuracy. These individual differences are also called residuals and the RMSE serves to aggregate them into a single measure of predictive power. It's calculated as:

$$RMSE = \sqrt{MSE} \tag{6.4}$$

d) Correlation Coefficient

Correlation (often measured as a correlation coefficient) indicates the strength and direction of a linear relationship between two random variables A and B. In general statistical usage, correlation refers to the departure of two random variables from independence. And the general formulation of correlation coefficient is

$$r = \frac{\sum_m \sum_n (\ A_{mn} - \bar{A}\ )(\ B_{mn} - \bar{B}\ )}{\sqrt{\sum_m \sum_n (\ A_{mn} - \bar{A}\ )^2 \sum_m \sum_n (\ B_{mn} - \bar{B}\ )^2}} \tag{6.5}$$

where A and B are the variables, and $\bar{A}$ and $\bar{B}$ are the corresponding mean values. The closer the coefficient is to either −1 or 1, the stronger the correlation between the variables.

e) Coefficient of Determination (R^2)

The coefficient of determination, R^2, is a measure of the global fit of the model. Specifically, R^2 is an element of [0, 1] and represents the proportion of variability in Y_i that may be attributed to some linear combination of the regressors (explanatory variables). More simply, R^2 is often interpreted as the proportion of response variation "explained" by the regressors in the model. Thus, R^2 = 1 indicates that the fitted model explains all variability in Y, while R^2= 0 indicates no 'linear' relationship between the response variable and regressors. An interior value such as R^2 = 0.7 may be interpreted as follows: "Approximately 70% of the variation in the response variable can be explained by the explanatory variables. The remaining 30% can be explained by unknown." It's calculated as

$$R^2 = 1 - \frac{SSE}{SST} \tag{6.6}$$

$$\text{where } SSE = \textstyle\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i), \tag{6.7}$$

$$SSR = \textstyle\sum_{i=1}^{n} \hat{e}_i^2\,, \tag{6.8}$$

$$SST = SSE + SSR. \tag{6.9}$$

f) Adjusted R^2

When a new variable is added to the model, R^2 increases. To overcome this feature, R^2 is adjusted for the number of explanatory terms in the model. Unlike R^2, the adjusted R^2 increases only if the new term improves the model more than would be expected by chance. The adjusted R^2 can be negative, and will always be less than or equal to R^2. The adjusted R^2 is calculated as

$$Adjusted\ R^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{6.10}$$

where p is the total number of regressors in the linear model (but not counting the constant term), and n is the sample size.

g) PWI 1

It is the ratio between the numbers of error values which are greater than two standard deviations to the total observation number. When performance of the model is improved, lower PWI 1 values are obtained.

h) PWI 2

It's similar to PWI1. It's the ratio between the number of error values which is greater than 3 standard deviations to the total observation number. Its' interpretation is exactly the same as PWI1.

APPENDIX B

VARIABLE SELECTION

Variables Entered/Removed(b)

Model	Variables Entered	Variables Removed	Method
1	X8, X3, X6, X2, X1, X5, X7, X4(a)		Enter

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,854(a)	0,73	0,632	3,78E-02

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8,51E-02	8	1,06E-02	7,439	,000(a)
	Residual	3,15E-02	22	1,43E-03		
	Total	0,117	30			

a
Predictors: (Constant), X8, X3, X6, X2, X1, X5, X7, X4

b
Dependent Variable: Y

Coefficients(a)					
		Unstandardized Coefficients		Standardized Coefficients	t
Model		B	Std. Error	Beta	
	(Constant)	8,594	14,114		2,735
	X1	2,95E-04	0,001	1,076	0,572
	X2	-9,85E-03	0,006	-2,801	-1,723
	X3	-15,184	5,488	-28,4	-2,767
	X4	1,528	0,536	28,345	2,854
	X5	-1,35E-04	0,001	-0,425	-0,191
	X6	-0,296	0,221	-0,158	-1,337
	X7	8,33E-02	0,092	7,448	0,908
	X8	-1,12E-02	0,015	-4,913	-0,749
	Dependent Variable: Y				