DAILY NATURAL GAS CONSUMPTION PREDICTION BY MARS AND CMARS MODELS FOR RESIDENTIAL USERS IN ANKARA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF APPLIED MATHEMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

YAVUZ YILMAZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
SCIENTIFIC COMPUTING

AUGUST 2015

Approval of the thesis:

DAILY NATURAL GAS CONSUMPTION PREDICTION BY MARS AND CMARS MODELS FOR RESIDENTIAL USERS IN ANKARA

submitted by **YAVUZ YILMAZ** in partial fulfillment of the requirements for the degree of **Master of Science in Department of Scientific Computing, Middle East Technical University** by,

Prof. Dr. Bülent Karasözen
Director, Graduate School of **Applied Mathematics** _____

Assoc. Prof. Dr. Ömür Uğur
Head of Department, **Scientific Computing** _____

Prof. Dr. Gerhard-Wilhelm Weber
Supervisor, **Scientific Computing, METU** _____

Examining Committee Members:

Prof. Dr. Gerhard Wilhelm Weber
Scientific Computing, METU _____

Assoc. Prof. Dr. Sevtap Kestel
Actuarial Sciences, METU _____

Assist. Prof. Dr. Fikriye Nuray Yılmaz
Applied Mathematics, Gazi University _____

Date: _____

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: YAVUZ YILMAZ

Signature :

ABSTRACT

DAILY NATURAL GAS CONSUMPTION PREDICTION BY MARS AND CMARS MODELS FOR RESIDENTIAL USERS IN ANKARA

Yılmaz, Yavuz

M.S., Department of Scientific Computing

Supervisor : Prof. Dr. Gerhard-Wilhelm Weber

August 2015, 73 pages

Energy efficient construction and operation of modern energy transmission and distribution systems is one of the major challenging problems in engineering area. Prior to every kind of natural gas related study, regardless of their financial or engineering features, demand forecast figures should be seen. Decision making of natural gas investment planning and operation in a city, region or country are highly important engineering problems that have very important economic effects. Determination of the total gas supply import expenditures, the tariffs, additional costs for the extra investments in order to provide safe and continuous gas supply to additional consumers are some of the other confronted problems. Additionally, predicting residential purpose users gas consumption is indispensable for efficient system operation and required for planning decisions at natural gas Local Distribution Companies (LDCs) and Transmission System Operator companies (TSOs). Residential users are major consumers that usually demand significant amount of total gas supplied in distribution systems especially in winter season. Due to the fact that all residential users should be satisfied and

the distribution systems have limited capacity for the gas supply, proper planning and forecasting in high seasons and whole year have become critical and essential.

This study is conducted for the responsibility area of Başkentgaz which is the local gas company of Ankara. As of gas year of 2014, Başkentgaz owns approximately 90% of overall maximum permissible residential consumption capacity of Ankara with its districts residential user gas distribution network. Within the scope of this work, MARS (Multivariate Adaptive Regression Splines) and CMARS (Conic Multivariate Adaptive Regression Splines) predictive models for one-day ahead natural gas consumption of residential users are formed. The models not only compare both methods, but they also analyze the effect of actual daily minimum and maximum temperatures versus the Heating Degree Day (HDD) equivalent of their average. Using the obtained one-day ahead models with daily data on 2009-2012, the daily consumption of each day in 2013 has been predicted and the results have been compared with the actual data obtained from Başkentgaz. The outcomes of the study present MARS and CMARS methods to the natural gas industry as two new competitive approaches. The thesis is ended with a conclusion and an outlook to the future studies.

Keywords : Natural Gas Consumption Forecast, Multivariate Adaptive Regression Splines, Conic Multivariate Adaptive Regression Splines, Conic Quadratic Programming, Energy.

ÖZ

ANKARA' DAKI EVSEL KULLANICILARIN MARS VE CMARS MODELLERİ İLE GÜNLÜK DOĞAL GAZ TÜKETİM TAHMİNİ

Yılmaz, Yavuz

Yüksek Lisans, Bilimsel Hesaplama Bölümü

Tez Yöneticisi : Prof. Dr. Gerhard-Wilhelm Weber

Nisan 2015, 73 sayfa

Modern enerji iletim ve dağıtım sistemlerinin enerji verimli bir şekilde kurulması ve işletilmesi, mühendislik alanındaki başlıca zorlu problemlerdendir. Finansal ya da mühendislik özellikleri olup olmadığına bakılmaksızın, doğal gaz ile ilgili çalışmalardan önce, gelecekteki tüketim rakamlarının görülmesi gereklidir. Toplam gaz ithalat harcamalarının, tarifelerin ve yeni eklenen kullanıcılara güvenli bir şekilde ve sürekli olarak gaz arzının sağlanabilmesi amacıyla gerekli ek yatırımların maliyetlerinin belirlenmesi, karşılaşılan diğer zorlu sorunlardır. Ayrıca, doğal gaz dağıtım ve iletim şirketleri tarafından, evsel tip kullanıcıların doğalgaz talep tahminlerinin yapılması, verimli bir sistem işletmeciliği için kaçınılmaz ve planlama kararlarının alınabilmesi için gereklidir. Evsel kullanıcılar, özellikle kış mevsiminde hayli yüksek miktarda gaz talebi olan ve dağıtım şebekelrinden beslenen en önemli müşterilerdir. Tüm evsel kullanıcıların gaz ihtiyaçlarının karşılanması gerekliliği ve dağıtım şebekelerine tedarik edilen gaz miktarının belirli sınırlı kapasitede olması nedeniyle, yüksek tüketimin olduğu dönemler ve tüm yıl için uygun planlamanın ve tahminlerin yapılması kritik ve gereklidir.

Bu çalışma Ankara ili yerel doğal gaz dağıtım şirketi olan Başkentgaz' ın sorumluluk alanı içersindeki bölgede yapılmıştır. 2014 yılı sonu itibarıyla, Başkentgaz Ankara ili ve ilçeleri' ndeki tüm evsel kullanıcılara gaz sağlayan dağıtım şebekesinin %90' ına sahiptir. Bu çalışma kapsamında, MARS (Çok Değişkenli Uyarlanabilir Regresyon Eğrileri) ve CMARS (Konik Çok Değişkenli Uyarlanabilir Regresyon Eğrileri) tahmin modelleri, evsel kullanıcıların günlük doğal gaz tüketimlerinin bir gün öncesinden tahmini için elde edilmiştir. Oluşturulan modeller yalnızca bu iki metodu karşılaştırmakla kalmayıp, günlük en düşük ve en yüksek sıcaklıkların gerçek değerleri veya günlük ortalama sıcaklığın ısıtma gün sayısı (HDD) değerinin alınması durumlarının etkilerinin bir mukayesesini sunmaktadır. 2009-2012 yıllarına ait günlük verilerle oluşturulan modellerden yararlanılarak, 2013 yılı için doğal gaz tahminleri oluşturulmuş ve sonuçlar Başkentgaz tarafından ölçülmüş gerçek verilerle karşılaştırılmıştır. Elde edilen sonuçlar, MARS ve CMARS metotlarını, doğal gaz endüstrisi için iki yeni ve diğer modellerle yarışabilir yaklaşımlar olarak tanıtmaktadır. Çalışmanın bitiminde, bir sonuç bölümü ve ileride yapılacak çalışmalara bir bakış sunan gelecekteki çalışmalar bölümü sunulmaktadır.

Anahtar Kelimeler: Doğal Gaz Tüketim Tahmini, Çok Değişkenli Uyarlanabilir Regresyon Eğrileri, Konik Çok Değişkenli Uyarlanabilir Regresyon Eğrileri, Konik Karesel Programlama, Enerji.

To My Family

ACKNOWLEDGMENTS

I would like to voice my tremendous thankfulness to my thesis supervisor Prof. Dr. Gerhard-Wilhelm Weber for his endless guidance and valuable advices during this study. It is my honor to be a student of such an outstanding and productive academician and great person.

I also would like to thank to Dr. Ayşe Özmen for her share of experiences and expertise and also many thanks for her patience and objective visions in all steps of the study.

I am also grateful to my friend Caner Fuad Yazıcı for his friendship. Also, his perspectives to natural gas markets helped me to evaluate the gas market together with both finance and engineering concerns.

I am grateful to Başkentgaz A.Ş. for sharing the consumption and meteorological data for academical purposes.

I would like to conduct my thanks to Assoc. Prof. Dr. Sevtap Kestel for her share of knowledge in natural gas and energy, and thanks to her encouragement of making additional studies.

I also would like to express my gratitudes to Institute of Applied Mathematics for their understanding, help and supply of effective study programs.

I thank a lot to my dear brother Oğuz Yılmaz. He is always my best friend, advisor and objective critic.

I would like to thank to Dr. Rainer Kurz, from Solar Turbines, for his support and patience on additional studies we have cooperated during this study.

I would like to thank to Salford-Systems and Asrad Mühendislik which are the manufacturer of the original MARS software SPM and its partner in Turkey, respectively, for the supply of SPM-M-64 for academic practice purposes.

I dedicate this study to the memories of my cousins Ayten Yılmaz and Emine Yılmaz.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

σ^2	Variance (Error Variance)
ε	Stochastic Component (noise)
B	Set of BFs
E	Conditional Expectation
r	Correlation Coefficient
R^2	Coefficient of Determination
R_{adj^2}	Adjusted Multiple Coefficient of Determination
M_{max}	Maximum Number of BF
$RMSE$	Root Mean Squared Error
AAE	Average Absoulte Error
ACCR	Average Correct Classification Rate
ACER	Acency for the Cooperation of Energy Regulator
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
bcm	billion cubic meter
BF	Basis Function
CART	Classification and Regression Trees
CMARS	Conic Multivariate Adaptive Regression Spline
cmy	cubic meters per year
CQP	Conic Quadratic Programming
DM	Data Mining
DSO	Distribution System Operator
EC	European Commission
FERC	US Federal Energy Regulation Commission
GA	Genetic Algorithm
GAM	Generalized Additive Model
GCV	Generalized Cross Validation
GPLM	Generalized Partial Linear Method
HDD	Heating Degree Day
IPM	Interior Point Method

LDC	Local Distribution Company
LOF	Lack-of-Fit
LSE	Least Squares Estimation
LTF	Linear Transfer Function
MARS	Multivariate Adaptive Regression Spline
MSE	Mean Squared Error
NG	Natural Gas
NLME	Nonlinear Mixed Effect
PRSS	Penalized Residual Sum of Square
RCMARS	Robust Conic Multivariate Adaptive Regression Spline
RM/A	Pressure Reduction and Measurement Station
RSS	Residual Sum of Square
SS-ANOVA	Smoothing Spline Analysis of Variance
TANAP	Trans Anatolian Pipeline
TAP	Trans Adriatic Pipeline
TR	Tikhonov Regularization
TSO	Transmission System Operator

CHAPTER 1

INTRODUCTION

The Distribution System Operators (DSOs), which are also called as Local Distribution Companies (LDCs), are the service providers of residential natural gas (NG) consumers at present transportation structure [80, 81]. NG is primarily used for the space heating and cooking of food by residential users, therefore, they are naturally non-interruptible. The certainty at consumption predictions for them is quite important concerning their continuous gas supply and many researches have been conducted upon this topic with various techniques.

1.1 Research Motivations

One motivation for forecast studies of DSOs is the legal obligation of prediction set by regulatory authorities. In EU and Turkish markets, information flow to predetermined market authorities is obligatory by the regulations and Network Codes of the countries. Forecast studies also have an influence as the demand of DSOs is supplied via spot markets, or the majority of the demand is by international supply contracts, whether pipeline or Liquefied Natural Gas (LNG) gas. The majority of the contracts needs long-term agreements and has the liability to Take-or-Pay terms for purchasing parties.

Price determination is another driving consideration to produce forecasts. Competition in the retail market forces for more competitive pricing. Due to the price formation structure of market, operation cost minimization while providing necessary consumption is a market requirement.

Consequently, using Data Mining (DM) tools, producing prediction techniques that provides quite accurate prediction values, is extremely important to analysts. Various mathematical, statistical and econometric models are applied for natural gas demand forecasting. Some models can be named as, Hubbert Curve, Statistical, Artificial Neural Networks, Grey Prediction, Conditional Demand Analysis, Econometric, Mathematical, Expert System, Stochastic Gompertz Innovation Diffusion, Dynamic System and Simulated Annealing Models [66].

DM, frequently called as *information disclosure in the database*, is extensively used for decision support, marketing strategy evaluation, financial forecasting, process con-

trol, classification, prediction, clustering, summarization, sequential analysis and some other fields. It is a methodology to determine embedded patterns and relationships in dataset. Visualization of data, machine learning, artificial neural networks, regression trees, genetic algorithms and nonlinear methods are within the scope of DM.

1.2 Contributions of the Study

In this study, as an innovative contribution to NG demand forecasting studies, well known Multivariate Adaptive Regression Splines (MARS) [20] and yet rather new Conic Multivariate Adaptive Regression Splines (CMARS) [74] algorithms are applied in prediction of daily gas consumption of residential users in Ankara City, where gas is distributed by the Başkentgaz DSO company.

MARS, presented by Friedman [20], is a powerful regression model to predict the generic functions of two or more dimensional instances using the predictors confronted. Since the selection of basis functions (BFs) is problem-specific, MARS is named as flexible (adaptive) method. It is an contemporary and responsive tool that automates by establishing definite forecast models valid for continuous or binary target variables. It surpasses at detecting optimum transformations of variables and probable interaction in regression solution and effortlessly administers the complicated data architecture that generally conceals in multi-variable data. Therefore, MARS, adequately uncovers extensive data models and exchanges that are not so easy for other traditional methods to explain.

A quite successful data match using nonlinear BFs is attained. Being an unsupervised learning, no initial condition on the basic functional link between the target and input variables is needed with the nonparametric regression approach. In addition, MARS also searches for possible interactions between independent variables, confirming any possible interaction obtained since the model can help to improve the approximation of the data. As a modern statistical learning methodology, MARS, being a nonparametric regression tool, is very important in classification and regression. Its capability to determine the contribution of BFs with additive and interaction properties of its predictors make MARS a preferred predictive tool especially for high-dimensional problems. In order to estimate the model function, MARS uses two step-wise algorithms, a *forward* and a *backward* one. In the first step, the model is generated by adding BFs up to a maximum level of complexity is reached. In the backward step, the basis functions having least contribution to the overall result are removed from the model.

In CMARS part of the study, the aforementioned backward stage is not used. However, a Penalized Residual Sum of Squares (PRSS) is engaged for MARS backward stage in order to formalize the problem as a Tikhonov Regularization (TR) type [3, 49, 88] and the TR problem is worked out by Conic Quadratic Programming (CQP). The CMARS model is acquired as the regression complexity and classification tool MARS when MARS is penalized via a carefully prepared TR. The boundaries of final TR problem is determined by multi-objective optimization approach and the two-objective optimization problem is solved using Conic Quadratic Programming (CQP). CMARS has

a model-based formulation and handles continuous, convex optimization by introducing Interior Point Methods [74] and their Matlab® add-on codes, e.g., MOSEK [43], into the problem.

Another inventive approach of this study is at the selection of the ambient temperature inputs to the model. In literature, the gas consumption models are realized in two different forms of the daily temperature values. In the first group of studies, the daily minimum and maximum temperatures of the objected day are used directly in the model formation phase [44], whereas in the second group of models, instead of taking the direct temperature values into account, the Heating Degree Day (HDD) value of the average of the minimum and maximum of the projected day is used [10]. HDD is a measure of the heat quantity required for a specific day. It represents the demand for energy to warm a house and the HDD is derived using the average ambient temperature. Both MARS and CMARS prediction tools are applied at two similar datasets differing only in temperature inputs being daily minimum and maximum temperatures or HDD values of the daily average temperature.

The study is made with a real dataset which is divided into training and testing data groups. The training set includes daily input variables of period 2009-2012. The test set has inputs for 2013. Predictor variables given to the model including daily meteorological data, previous-day daily gas consumption data, the number of residential users and other supplementary inputs. As aforementioned, the direct temperature values and HDD value of average daily temperature are applied in order to obtain two different datasets for both training and test sets. MARS and CMARS algorithms are used to obtain models to forecast daily consumption of Ankara City. Each sovereign model has the capability to provide the daily consumption of Ankara, for all four seasons without creating separate models for winter and summer terms. At the end, the models built by MARS and CMARS have been compared. Therefore, another contribution of the study is, with the models, not only the daily and annually demand, but also the monthly demand throughout 2013 are predicted.

1.3 Outline

This thesis study is structured as follows:

In Chapter 2, information about the NG markets and some previous basic analytic forecast models are given, respectively. Better understanding those techniques may also keep us to better assess the more rigorous mathematical methods MARS and CMARS which will be presented than.

Chapter 3 provides the theory and formulations of the more model-based models and techniques MARS and CMARS in detail, respectively.

Chapter 4 explains the real-world problem solved in our study. The obtained models of MARS and CMARS are also given in this section.

A conclusion and remarks for future studies are stated in Chapter 5.

CHAPTER 2

LITERATURE SURVEY AND FRAMEWORK

2.1 Natural Gas Markets

Considering the 2013 records [7], USA is the largest gas consumer. Following the USA, EU countries are the major gas users. Turkey is also one of the largest consumers of Europe as provided in Appendix A.

When the market gas delivery paths are considered, despite the varaitions in countries, a general structure can be outlined in Figure 2.1. In that path, LDCs are the only service providers of residential gas users for present transportation structure and may be represented as in Figure 2.1. Electricity is produced in various ways like coal-fired, hydroelectric, nuclear, etc., other than natural gas power plants. Additionally, some electrical power plants and some types of consumers other than residential users may also use alternative energy sources, such as fuel oil, temporarily in case of any emergency. It has, therefore a different characteristics than the residential users.

At that point, a market analysis of the USA, EU and Turkey will provide background information to the study.

2.1.1 US and EU Gas Markets

Being the largest gas user, in the USA, there are approximately 210 natural gas transmission pipeline networks and more than 506000 km of pipeline as it is illustrated in Appendix B. In addition, there are about 1.9 million km of distribution pipeline owned by 1200 DSOs. Additionally, a natural gas pipeline network is a highly integrated transmission and distribution grid and the gas offtake is provided from 49 entry points from which 33 active entry points of pipeline imports/exports [30] and 8 active entry points of LNG imports/exports as provided in Appendix C. The gas is delivered from approximately 11000 delivery points [71].

When the market properties are considered, US is characterized by high domestic production, a very complicated pipeline infrastructure and thousands of market players at every stage of the supply chain. Both conventional and non-conventional supply sources such as shale gas exist in the country.

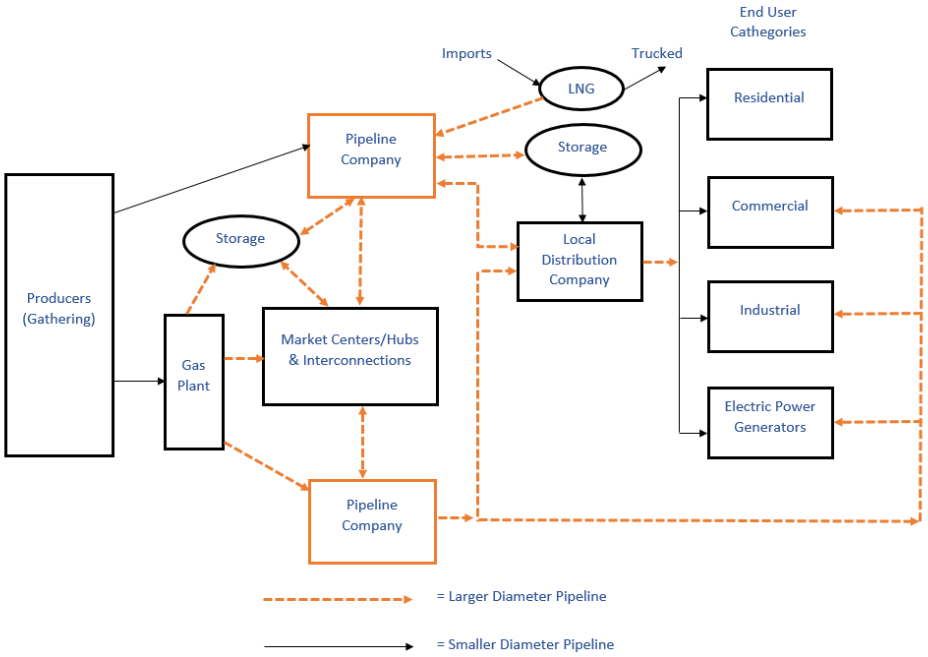


Figure 2.1: General NG Transportation Structure.

Within the NG value chain of the local distribution companies (LDCs), electrical and industrial consumers depend on trader companies and independent marketing companies, which trade pipeline gas capacity and negotiate directly with the producers. The prices for wholesale market, are not regulated and prices are formed by market competition. LDCs sale gas rates are generally regulated, but that allows them to transfer the prices to the end users. The sale prices are regulated so that LDCs allow large margin on gas purchases and sales. This action is compulsory to control the power of LDC due to their natural monopoly behavior against end users. Regulation aims the distribution sales price, but not the commodity price. In many cases, there exists single LDC in every local region, and the LDC typically possesses the distribution network. End-customers cannot shift one supplier to another LDC [71].

In the US market, regulation of markets always followed by developments of gas market. Once the gas is brought into states, regulation followed. The problems faced in long term contracting and common reorganization for distributing risks and organize the markets and some other regulatory and supply issues resulted in elimination of longterm contracts. In general, the state owned companies or entities and market institutions have impact on each other, they are not as effective as they are in the EU, where the public entities are trying to form markets with regulation and other administrative measures [71]. Again, historically, there are two important measures adopted by US Federal Energy Regulation Commission (FERC). The measure Order No. 436, dated 1985, allowed interstate pipelines transportation to *open access* and it limited the use of long-term contracts. Furthermore, with that measure, the LDCs and end users with high gas consumption were permitted to provide NG directly from the producer companies, bypassing the transmission system operators [33]. TSOs that accept to allow access to their transmission pipelines had right to put an *open access tariff* and that

is determined by FERC as the transportation services cost as shown in Figure 2.3. To promote the competition in the bulk supply, FERC allowed gas market players to make sales and buys of NG on behalf of other industry participants. A market operation scheme, representing before and after FERC Order No. 436, dated 1985, are given in Figures 2.2 and 2.4.

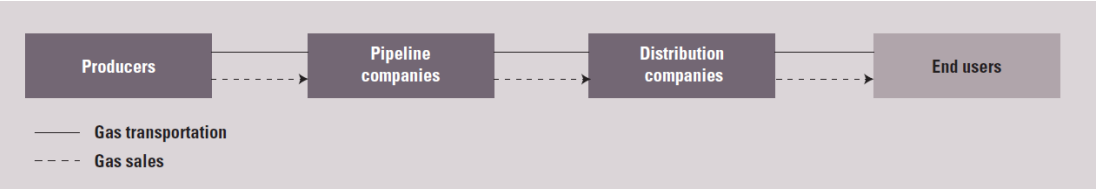


Figure 2.2: Traditional Structure of US Gas Market, before 1985 [33].

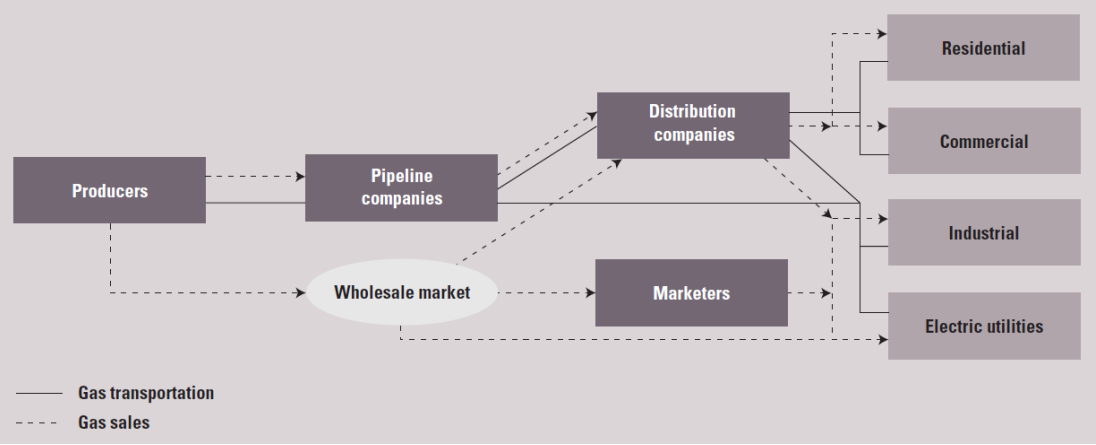


Figure 2.3: Start of Access to Pipeline Transport, 1985-92 [33].

The second important measure of FERC is with Order No. 636 in 1992 which brought the deregulation to the market and it entailed the unbundling of sales and transportation activities in transmission pipeline companies. This separation is formed by the formation of separate companies to manage these activities [33]. The Order No. 636 also improved the method for determining transportation tariffs and brought a capacity-release-program to resale of firm transportation contracts. This program, allows shipper companies to purchase pipeline capacity from other similar companies that have temporary or permanent excess reserved capacity. Such gas trade allows allocation of transportation contracts between shippers and enables gas market players to pair the transportation contracts to the gas contracts [33].

Accordingly, had big impact on U.S. gas industry. Until 1985, the production, pipeline transportation, and distribution were separated. However the market was highly regulated and long-term contracts were dominating. The open access brought competition into the wholesale gas market, and gas sale became a separate business in the market. Finally, in 1992 [33], the unbundling of transportation companies created a competitive gas market and a generic diagram is provided in Figure 2.4 for that new market structure.

In the deregulated US market, at present, there are three main pipelines on the trans-

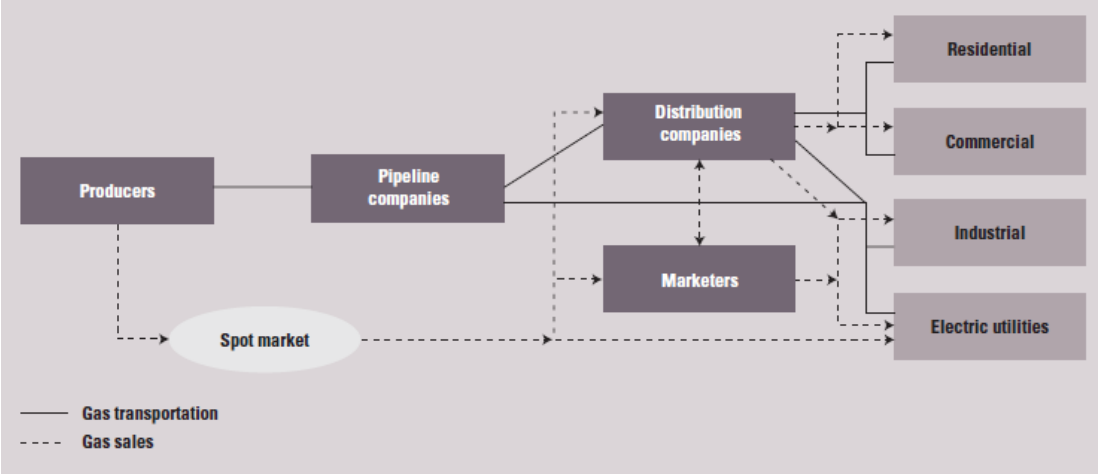


Figure 2.4: Unbundling of Gas Sales From Pipeline Transportation, after 1992 [33].

portation track, namely, Gathering Systems, Transmission Systems, and Distribution Systems [81]. Distribution Pipeline Systems carry the gas to residential gas users, some smaller scale industrial, various commercial and other types of users as specified in Appendix D.

One should also consider the soaring of shale gas production in the country, in order to fully understand and evaluate not only the US market conditions but also the European and Turkish markets. From 2011 to 2014, a 44% increase has been obtained. The increase is mostly from the shale gas production capacity which is predicted to reach at 4.53 trillion cubic metres per year (cmy) by 2040 [71]. In addition, the predictions envisage the demand of 2,3 bcmd compared to production of 2,7 bcms by 2030, which means much extra gas to export. Also, the difference in pricing existing between the US, Europe and Asian markets, the trade opportunities arises to drive investments with a concentration on expanding trade between the three markets, especially, from the US to Europe and to Asia. Those opportunities, for sure, make the US a significant exporter of liquefied natural gas (LNG), which is exceptional since just a decade ago the US was a substantial importer of LNG [41].

The gas markets in EU, however, are influenced by mainly imported gas. The most important of import gas supplies of EU are Norway, Russia and Algeria [71]. EU NG production is decreasing. Gas output in the European Union has been falling since the mid-1990s due to depleting resources. The production reached its peak in 1997 and has been declining since, with limited volatility. Natural gas production amounted to 173.7 bcm during 2012 [31], and the annual production quantities are detailed in Figure 2.5.

Another important figure in EU gas market is the shale gas. The role of shale gas is not only important for the US, but it may also have an impact on the EU market. Shale gas may tolerate some of the decrease at the already existing domestic producing regions as in the North Sea. Consequently, EU gas supply is quite dependent on imports from non-EU regions. Imports from non-EU countries are mostly made through high capacity NG pipelines that connects the EU to the producing and consuming regions [71].

At present, raise in the LNG imports protects security of supply however EU market is still dependent to its major supply sources in the upstream market. Majority of the producers of the upstream partners are generally state owned companies and may have effects of political decision of their governments.

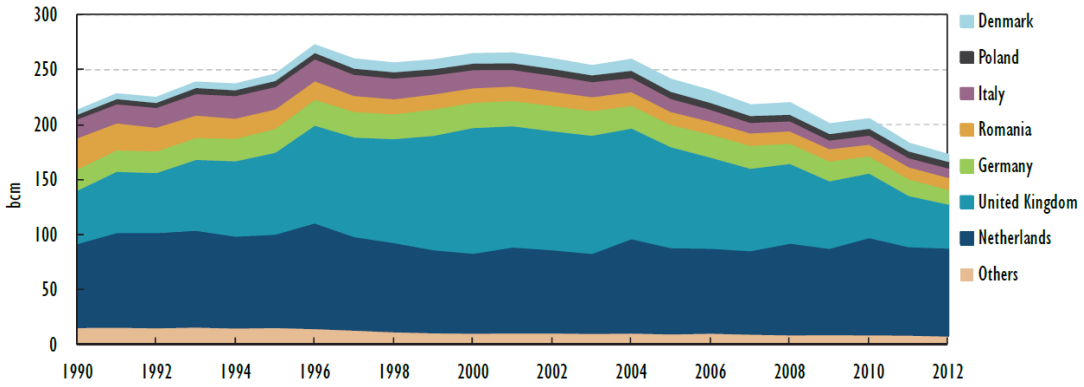


Figure 2.5: The Natural Gas Production at EU Countries, 1990-2012 [47, 48].

On the demand side, between 2008 and 2012, the gas demand has a trend to decline from its peak in 2010 446,9 millions tonnes of oil equivalent (mtoe), or 543 bcm, by about 10% to a level of 392,5 mtoe, or 477 bcm [31]. The fall of demand is mostly due to the global economic crisis therefore a demand decrease in industrial usage is main reason. Additionally, changing fuel use in the power generation sector [31] is another reason in the cutback of the gas utilization quantity. Demand in power generation usage of NG decreased 19.9% from 2007 to 2012, since a quite high quantity of electricity is produced from renewable energies to provide the EU electricity demand at present. Main reason for it is natural gas fueled power plants are less competitive against coal-fired plants despite their less CO_2 price. The EU consumption profile according to different sectors are supplied in Figure 2.6 for subsequent years.

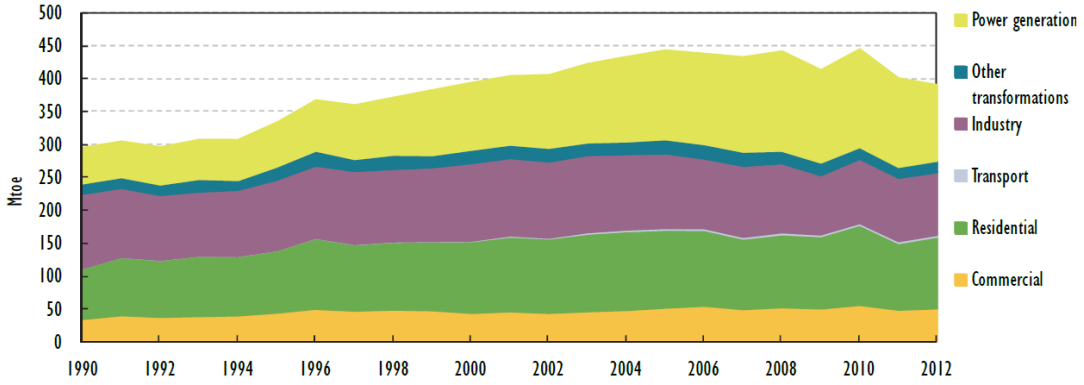


Figure 2.6: NG consumption, in mtoe, for different sectors in the European Union, 1990-2012 [47, 48].

In a historical perspective, energy in general was regulated in EU region. After 1980s, energy was the concern of two European treaties, which are the abolished European Steel and the Coal Community Treaty and the European Atomic Energy Treaty [71].

Within the third treaty, The European Community Treaty, today named as the *EU Treaty* or *Treaty on the Functioning of the European Union*, energy sector had not mentioned specifically, but it was a beginning.

Formerly, the European Union accepted the regulatory framework across its member states in order to shape the domestic energy market up to the end of 2014. As of 2008, the European gas market has a progress in market design and regulations with new-EU rules about *transparency*, *unbundling* and *entry-exit regimes*. The market liberalisation has added also the *market effectiveness*, *liquidity* and *cross-border trade*. With acceptance of the so-called *Third Internal Market Package*, EU countries are presently implementing a comprehensive framework that offers principles of complete and effective market opening, competition and liberalization [31]. The Third Package brought regulation of TSOs on cross-border issues into an EU framework that forces the independence and co-operation of the TSOs and the national regulatory authorities. Some important EU directives and regulations can be listed up as [31]:

- Directive 2009/73/EC of the European Parliament and of the Council of 13 July 2009 concerning common rules for the internal market in natural gas and repealing Directive 2003/55/EC ("Gas Directive").

- Regulation (EC) No 715/2009 of the European Parliament and of the Council of 13 July 2009 on conditions for access to the natural gas transmission networks and repealing Regulation (EC) No 1775/2005 ("Gas Regulation").

- Regulation (EC) No 713/2009 of the European Parliament and of the Council of 13 July 2009 establishing an Agency for the Cooperation of Energy Regulators ("ACER Regulation").

- Moreover, in 2011, a legislation governing the transparency and oversight of the wholesale energy market was adopted in the form of Regulation (European Union) No 1227/2011 of the European Parliament and of the Council on wholesale energy market integrity and transparency ("REMIT Regulation").

When the pipeline infrastructure is considered, the delivery points for the Western gas markets are placed at the border with Waidhaus and Mallnow at Germany and Baumgarten at Austria. On the other side, the delivery points for the Eastern European markets are located at the borders of Ukraine with Hungary, Poland, Romania and the Slovak Republic [71]. Since 2009, the only commissioned new pipeline is Nord Stream. In the Southern Gas Corridor to reach the production in Azerbaijan and bring it to the EU, the Trans Anatolian Pipeline Project (TANAP) is started to be constructed and it will pass from Turkey. The TANAP pipeline will be connected to the Trans-Adriatic Pipeline (TAP), which will cross Greece, Albania and Italy. Many other import projects are still in the planning stages. However they have experienced large delays (Galsi, White Stream). The overall NG infrastructure of EU countries are provided in Appendix E, and some network lengths of the EU countries are given in Table 2.1.

Table 2.1: Length of Member States gas grids by pipeline diameter [17].

	< 10" (km)	10"–24" (km)	> 24" (km)	Total (km)
AUSTRIA	4243	1398	1522	7163
BELGIUM	1912	479	1227	3618
BULGARIA	431	415	1758	2603
CROATIA	0	695	70	765
CZECH REPUBLIC	35	569	2753	3357
DENMARK	1078	324	1440	2841
ESTONIA	326	436	0	761
FINLAND	606	0	257	863
FRANCE	26799	476	6313	33588
GERMANY	34603	18187	14337	67127
GREECE	207	82	741	1029
HUNGARY	1021	2253	1925	5199
IRELAND	526	524	1057	2106
ITALY	10529	9039	9055	28623
LATVIA	403	184	520	1108
LITHUANIA	998	148	660	1806
LUXEMBOURG	41	239	0	280
NETHERLANDS	4063	1208	3144	8415
POLAND	5801	8668	1149	15618
PORTUGAL	168	225	738	1130
ROMANIA	1154	2405	1570	5129
SLOVAKIA	762	2888	1970	5621
SLOVENIA	752	6	0	758
SPAIN	908	4573	6627	12108
SWEDEN	965	0	20	985
UNITED KINGDOM	1637	3421	12771	17828

2.1.2 Turkish Gas Market

Turkey (including the state-owned energy company BOTAS¸ and other importer companies) is an entirely importer country, and it has an annual 52,2 Billion Cubic Meters (BCM) NG purchase contract. Approximately 57,5% of the contracts is supplied by Gazprom gas company. 14 BCM of the contracts with Gazprom is conveyed via a transit pipeline, which passes through Ukraine and some other countries, whereas, a 16 BCM contracted quantity is picked up from Blue Stream Pipeline. The remaining contracts are made as LNG contracts (4,4 BCM with Algeria, 1,2 BCM with Nigeria) and pipeline contracts (10 BCM with Iran and 6,6 BCM with Azerbaijan) [6]. Turkey is also one of the important customers of Gazprom in Europe, as given in Figure 2.7.

Moreover, Turkey is the fourth largest NG consumers of Europe considering the consumption in 2013, followed by Ukraine, as provided in Appendix A. The increasing demand of Turkey, can be explained by the increasing number of consumers especially by the new usage of NG in the cities previously not covered. Currently, NG is used in 74 of 81 cities as in Figure 2.8. Therefore, the only gas suppliers of residential users, LDCs, should forecast such demand increase in order to provide the continuous gas supply. As of 2014, there are approximately 65 LDCs in Turkey and they belong to

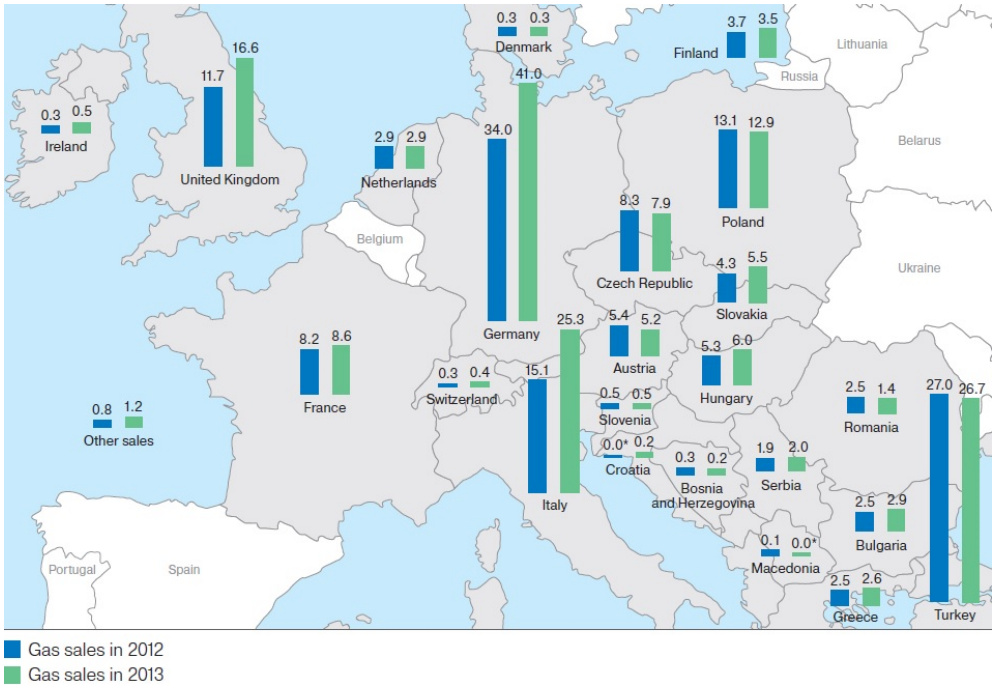


Figure 2.7: Gazprom Group Sales to non – Former Soviet Union (FSU) countries [46].

totally 23 different groups of companies [15].

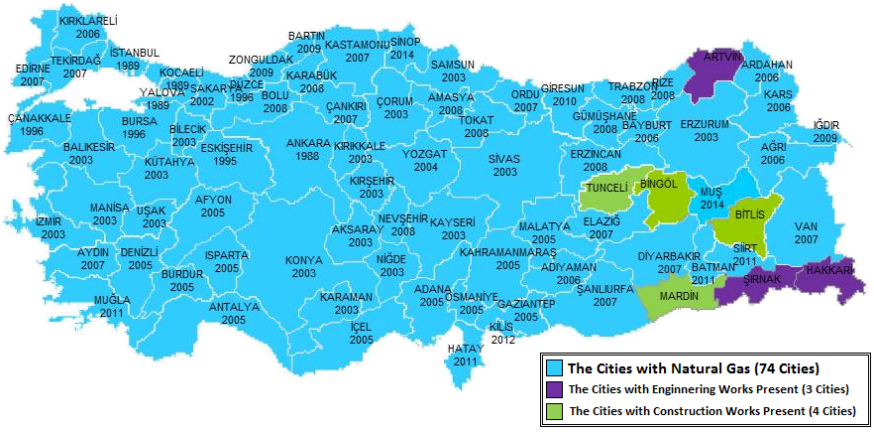


Figure 2.8: The Cities of Turkey with NG use infrastructure [6].

In such a large NG market, it is also important to evaluate the legislation that drives the trade. In 2001, the Natural Gas Market Law No. 4646, envisaged liberalization in the Turkish market and it enabled the private shippers to enter to the NG Market. The law also aimed unbundling of BOTAŞ and incorporation of an autonomous Transmission System Operator (TSO), increase in import competition, competition between wholesale suppliers, cost-based pricing for BOTAŞ, eligible consumer limit is set to zero, establishment of an NG spot market and Third-Party Access to gas market.

In Turkey, the LDCs like Başkentgaz should form their short-, mid- and long-term

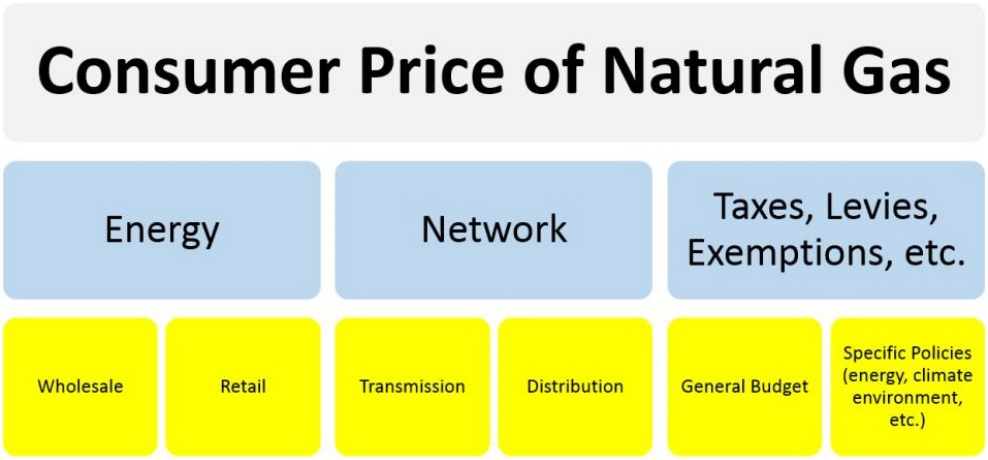


Figure 2.9: Price Formation Structure in EU [17].

planning as accurate as possible in order to provide security of supply for their end users. LDCs are also responsible for predicting the consumption quantities according to the Network Codes in EU and Turkish markets. They have to provide necessary information to regulatory entities routinely.

The information includes also short-term consumption forecasts for their consumers. In addition, in case that LDCs cause an unbalanced gas intake from TSO pipelines, the regulations and related laws prompt a quite high payment of penalties to them. That case brings LDCs under a pressure for forecast, especially, in winter periods, when the residential end-user consumption may reach to its highest and peak levels due to extreme cold conditions. Consumption predictions also determine the national necessary gas purchase quantities. Those gas supply quantities are provided by long-term contracts with Take-or-Pay terms or short-term spot gas purchases.

As in the EU market, in order LDCs to offer the minimum possible price to their consumers, the proper distribution system operating cost minimization is required according to price formation structure in retail level in Turkey as given in Figure 2.9.

For that reason, system operation margins have to be aligned with consumer's demand and the deviations over time. Operations must be quite flexible to compensate the demand fluctuations. Adjustment of the energy demand fluctuations wit the system operation constraints is achieved by prediction models formed for specific types of consumers like Residential End Users. From the LDC point of view, an accurate forecasting will enable to decrease the operation costs and eliminate the penalties that may arise due to unbalanced supply-demand quantities.

2.2 Forecasting Models Used for Consumption Prediction

Although the introduction of statistical and DM methods goes back to the end of 19th century as provided in Appendix G, consumption forecast studies started in the mid-

dle of the 20th century, and an enormous growth in researches is observed in the past decade. Previous academic studies, with various forecasting tools and techniques, investigated forecasting natural gas consumption demand [66]. In the literature, there are forecast models with different algorithms for various users in different geographical locations.

Distinct end users were studied throughout different geographical zones, on the country and city level. Spain's national industrial gas consumption for medium term horizon of 1-3 years [65] is one example of country level models. Consumption of Belgium [70], Poland [64], Taiwan [39], Canada [4] residential users consumption, and China's national NG supply and demand projections up to 2020 [40], and up to 2030 [38], are some of other countries' level studies. Using econometric models, Kuwait country's NG demand has been searched [16].

At the city level, many studies with various models exist in literature: in Beijing, Guangdong and Shanghai regional consumption projections of China using economic optimization model [32], and in Argentina, the greater Buenos Aires regional gas consumption for 1-5 days and long term of 1 to 5 years [21], Model determination of Slovenia's general city consumption [75], and Seoul of South Korea [94] were also studied.

Prediction algorithms also differ. The first tools that was established for modeling NG consumption is the Hubbert Curve model [29]. In subsequent studies, statistical models have been progressed and applied after 1960s. Time-series data in dynamic model for residential and commercial markets have been formed and the estimation has been made by ordinary least-squares model [63]. Then, the practical method of Linear Transfer Function (LTF) method has been introduced into Box-Jenkins ARIMA [75]. Another model, namely, Nonlinear Mixed Effects model (NLME) [8] has also been applied for forecast. This is a parametric model and built by several structural parts which provide an interpretation of the real cases. ANN has been introduced in 1988 with generalized back-propagation [10]. In the following studies, ANN has been studied extensively within many natural gas consumption forecast searches [4, 37, 66, 77]. Genetic Algorithms [58] and Nonlinear Regression Model [82] have been applied in modeling, too.

Turkey has also been studied in several previous works. Country's annual demand has been modeled with Simulated Annealing method [76]. At the city level, the major cities of Turkey have been investigated in bountiful investigations. Daily consumption data with Neural Network and Multivariate Time Series methods have been used for Istanbul city [14]. Artificial Neural Networks (ANN) method has been applied to other studies for Istanbul [22, 37]. The city of Ankara has been modeled using Artificial Neural Network (ANN) Method [22] and by Statistical Methods [23]. Multivariable Linear Regression Analysis is performed for the city of Ankara using a degree-day concept [24] and Autoregressive Integrated Moving Average (ARIMA) model has been used to predict the daily consumption of Sakarya City [1].

When the energy sector is considered, MARS and its variations have been employed in some areas such as electricity market modeling [92], energy price prediction [95] and gas turbines predictive maintenance scheduling [93].

Before going into the details of MARS and CMARS models and their regression predecessors, some other widely used prediction models of *Genetic Algorithms* and *Artificial Neural Networks* will briefly be introduced in the following sections.

2.2.1 Genetic Algorithms

The Genetic Algorithms (GA) method, developed by John Holland et al. [28] group at 1970s, offers rules for the solution of prediction problems. In the problems, the *genetic materials* are transmitted to *child* rules [78]. The genetic materials are the variables or categories in this method. Genetic algorithms apply the mechanisms of natural selection. The natural selection is applied by choosing the best adapted rules to prediction and by crossing and mutating them until a sufficiently predictive model is obtained [12, 56].

GA consists of operators and the three main operators are *selection* (choice) operator, *crossover* (overlapping) operator and *mutation* (variation) operator [59]. Although GA is used to improve othe predictive methods such as ANN, it has some disadvantages. They are mostly quite slow since the complexity raises dramatically due to the fact that when a function including some number of rules is used, then every rule after each generation must be considered. The resultant calculation may need to be done for thousands of rules.Therefore, GA is suitable for small size data. Additionally, this method is quite hard to implement since there are highly limited number of software packages to use [78].

2.2.2 Artificial Neural Network Method

After a neuron is defined by McCulloch and Pitts in 1943 [78], first neural network was built in 1958 by Rosenblatt. At present, *Artificial Neural Networks* (ANN) method or commonly referred as *neural networks* and its variations [66] became the most widely studied gas consumption forecast tool. There are 3 different classes are networks namely, *Single-Layer Feedforward Networks*, *Multilayer Feedforward Networks* and *Recurrent Networks* [27].

In ANN models, the *supervised learning* networks means the predictive models and *unsupervised learning* networks [78] are used for descriptive models. The minimum mean square error (MSE) [14] of the obtained supervised learning neural network assures the resultant input output relationship.

The attraction of ANN comes from its flexibility in formulating various functional inputs and output relationships. Complicated enough neural networks may formulate the arbitrary functions well [14].

ANN models have also various disadvantages [78]. The best global solution is not guaranteed to be converged. Secondly, there is a high risk of over-fitting especially when the number of data sample is very little compared with the number of input variables by even modeling minor fluctuations in the data. When there are lots of

input variables, ANN models may fail. The results are non-explicit therefore they maybe unacceptable for various applications, such that medical and automatic flight systems where the score formula is needed to be seen. To control of ANN model can be tedious. Finally, ANN methods are only applicable to continuous variables but not categoric variables and all variables should be transformed to the values in [0,1].

2.2.3 Linear Regression Models

In statistics and data mining, regression with its variations, has been extensively used in many applications at economics, science and engineering as predictive methods [49, 90]. It is a very effective and easy to implement tool to relate the dependent variables to a resultant model. In many real-world applications, the same independent variables (or regressors) do not yield to same results. Therefore, regression analysis arises as a strong tool to find the best correlation of the input variables and the target variable [90].

Before further investigation of MARS and CMARS, it will be beneficial to review some regression models used for various areas of applications.

In some engineering or scientific cases, an event or dependent variable can be explained by linear relationship of that event with some other independent variables or regressors which leads to a deterministic linear model.

Linear regression is used to pertain a continuous target variable *Y* to again a continuous n-dimensional input variable *X*. *Y* is assumed to be linearly dependent on *X* and independent, and that a knowledge of *X* enables us to improve our knowledge of *Y*.

The main hypothesis for linear regression is that the *conditional expectation* $E(Y|X = x)$ is linear function of x and it can be given as [78]

$$E(Y_i) = \alpha_0 + \alpha x_i, \quad \text{where} \quad \forall i = 1, 2, \ldots, N, \tag{2.1}$$

or, in the way of measurements or observations under noise, Equation (2.1) can be formulated as

$$y_i = \alpha_0 + \alpha x_i + \varepsilon_i, \quad \text{where} \quad E(\varepsilon_i) = 0 \quad \text{for} \quad i = 1, 2, \ldots, N. \tag{2.2}$$

In Equation (2.2), $\alpha_0 + \alpha x_i$ is the *deterministic component*, ε_i stands for the *stochastic component*. ε_i are also named as *errors*.

When the general case $n \in \mathbb{N}$ is considered, the existence of any linear relation of $E(Y_i) = a + \boldsymbol{b}^T \mathbf{x}_i$ is searched of that reason the estimators of *a* and $\boldsymbol{b}$ are defined as α and $\boldsymbol{\beta}$, respectively. The estimators are determined by *Ordinary Least-Squares Estimation* method that obtains a solution for the coefficients of α and $\boldsymbol{\beta}$ which minimizes the differences in the form of [78]

$$\vartheta = \sum_{i=1}^{N} (Y_i - \alpha - \boldsymbol{\beta}^T \mathbf{x}_i)^2. \tag{2.3}$$

Finally, the predictor equation is formulated as

$$Y = \alpha + \boldsymbol{\beta}^T \mathbf{X} + \varepsilon. \tag{2.4}$$

Since the parameters α and $\boldsymbol{\beta}$ coefficients minimize the difference given in Equation (2.3), both are obtained by equating the two partial derivatives of ϑ with respect to α and $\boldsymbol{\beta}$ [49] to 0. Therefore, the *unknown variance* or *error variance*, σ^2 is provided such that [85],

$$Var(\varepsilon) = \sigma^2, \tag{2.5}$$

which is the indicator how far the response data points have been distributed, enters the variance formula of the estimator as a factor, informing about the deviation from the mean value or regression line. Minimum variance means the data values are most concentrated around regression, whereas large variance means the data points are spread [42].

Linear regression is the fundamental of all other linear models and it is applicable almost all areas. Its *Simple Linear* and *Multiple Linear* types will be further investigated

The basic assumption in *Simple Linear Regression*, is that the continuous independent variable $\mathbf{X}$ and continuous Y are not independent. It is assumed that the true correlation between Y and $\mathbf{X}$ is a hyperplane in general. In linear regression, the mean value $E(Y|\mathbf{X} = \mathbf{x})$, the conditional expectation of Y given that $\mathbf{X} = \mathbf{x}$ for some $\mathbf{x} \in \mathbb{R}^n$, is investigated. The basic postulate of linear regression is that $E(Y|\mathbf{X} = \mathbf{x})$ is a linear function of $\mathbf{x}$, and for the *i*th observation, it can be given as [78]

$$E(Y_i) = a + \mathbf{b}^T \mathbf{X}_i \quad (i = 1, 2, \ldots, N) \tag{2.6}$$

or, likewise,

$$E(Y_i) = \alpha + \boldsymbol{\beta}^{\boldsymbol{T}} \mathbf{X}_i + \varepsilon_i, \text{ with } E(\varepsilon_i) = 0, i = 1, 2, \ldots, N. \tag{2.7}$$

Here, $\alpha + \boldsymbol{\beta}^T \mathbf{X}_i$ is the deterministic component, and ε_i is the stochastic component of the model; the values ε_i represent *noise*.

A simple regression model, such that $Y = \alpha + \beta X + \varepsilon$ fits, is obtained by using observed data points similar to Figure 2.10.

In fact, for $N = 1$, the estimators of b is the slope of the regression line, and the estimation a of the constant are [78]

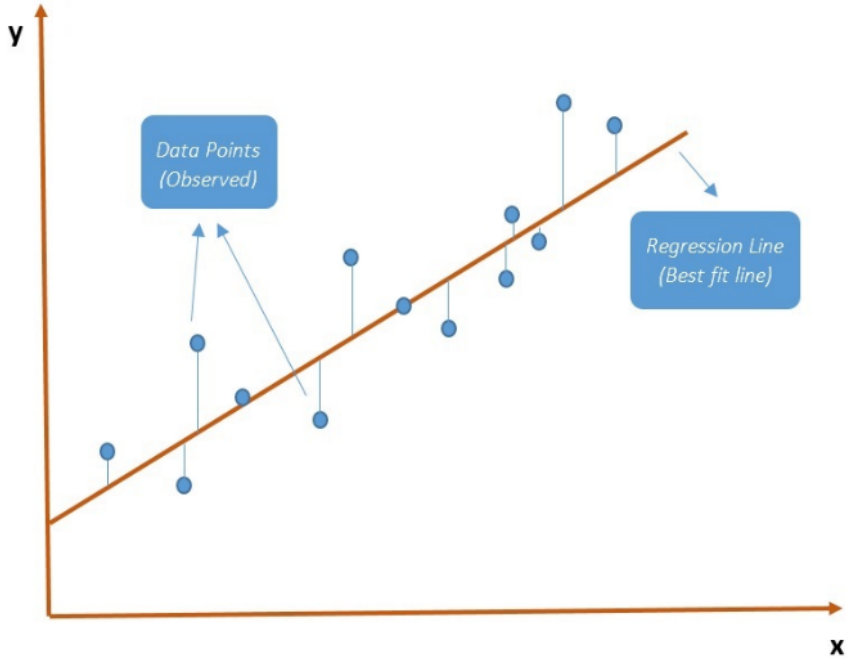


Figure 2.10: Regression line for Simple Linear Regression [49].

$$b = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{cov(X,Y)}{{\sigma_x}^2}, \tag{2.8}$$

$$a = \bar{y} - b\bar{x}, \tag{2.9}$$

where $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$ and $\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$.

Generally, in engineering, economic or scientific application problems, more than one predictor variable exists. Multiple linear regression models are the generalized version of simple linear regression models with several independent variables X_i [78]:

$$Y = a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n + \varepsilon. \tag{2.10}$$

Accepting the linear independence on X_i the model of Equation (2.10) turns into a *system* in matrix form for N many observations:

$$\mathbf{Y} = \mathbf{Xa} + \boldsymbol{\varepsilon}, \tag{2.11}$$

where the $\mathbf{Y}$, $\mathbf{X}$, and $\boldsymbol{\varepsilon}$ can be specified in general form as [83]:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1n} \\ 1 & x_{21} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nn} \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix}, \tag{2.12}$$

and $\boldsymbol{\varepsilon}$ is the residual vector. Therefore, Equation (2.10) becomes estimated by

$$\mathbf{y} = \boldsymbol{\alpha}^T \mathbf{x} + \boldsymbol{\varepsilon}, \tag{2.13}$$

where $\mathbf{x} \in \mathbb{R}^n$. As in the simple linear regression case, $\mathbf{a} = (a_1, a_2, \ldots, a_k)^T$ is the least-squares estimator of the vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)^T$, minimizing the sum

$$\sum_{i=1}^{N} (y_i - a_0 - \sum_{p=1}^{n} x_{ip} a_p)^2. \tag{2.14}$$

Using the assumptions of $N \geq n + 1$ and that $\mathbf{X}$ has full rank, we get [78]

$$\mathbf{a} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}. \tag{2.15}$$

The coefficient of determination is used to check the validity of a multiple linear regression model with the correlation of the variances of the fitted values and observed values of the dependent variable. For y_i representing the *observed values* of the dependent variable, $\bar{y}$ for its mean, and $\hat{y}_i$ for the fitted value, then the coefficient of determination is formulated as:

$$R^2 = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}. \tag{2.16}$$

2.2.4 Nonlinear Regression Models

In previous section, the multivariable linear regression models are defined in Equation (2.10) [61], which can also be specified as:

$$Y = g(x_1, x_2, \ldots, x_n; \mathbf{a}) + \varepsilon, \tag{2.17}$$

where $\mathbf{a} = (a_1, a_2, \ldots, a_n)^T$ is an unknown parameter vector. A model is called *nonlinear* if it includes at least one nonlinear parameter regardless of nonlinearity of the predictor variable [61]. Therefore, the following examples are also classified as linear despite the nonlinearity of the predictor variables:

$$Y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 {x_1}^2 + \gamma_4 {x_2}^2 + \gamma_5 x_1 x_2 + \varepsilon \tag{2.18}$$

and

$$Y = \gamma_0 + \gamma_1 \text{sin} x_1 + \gamma_2 \text{sin} x_2 + \varepsilon; \tag{2.19}$$

both are linear regression models. However, the following model is nonlinear, since it includes a nonlinear parameter term:

$$Y = \gamma_0 + \gamma_1 e^{\gamma_2 x} + \varepsilon. \tag{2.20}$$

A nonlinear model is represented as

$$Y_i = f(X_i, \boldsymbol{\gamma}) + \boldsymbol{\varepsilon}. \tag{2.21}$$

where f is the model function and X_i is the vector of related independent variables concerning the ith case. The model is similar to linear regression model, however, the outputs of the model are nonlinear functions. The difference from the linear case is one or more derivatives of the output or target function with respect to the variables are dependent on one or more other input variable [61].

In order to deal with specific dataset, we can form the N- dimensional vector, $\mathbf{a}(\boldsymbol{\gamma}) = (f(\mathbf{X}_i, \boldsymbol{\gamma}) : i = 1, 2, \ldots, N)^T$, and the nonlinear regression system is [61]

$$\mathbf{Y} = \mathbf{a}(\boldsymbol{\gamma}) + \mathbf{Z}. \tag{2.22}$$

Here, $\mathbf{Z}$ has a spherical normal distribution such that $E[\mathbf{Z}] = \mathbf{0}$, and

$$\text{Var}(\mathbf{Z}) = E[\mathbf{Z}\mathbf{Z}^T] = \sigma^2 \mathbf{I}. \tag{2.23}$$

Some nonlinear regression models are found in literature and used in many areas like population biology, growth occurring in organisms, plants, animals, etc. [49]. Some examples for nonlinear regression model are Malthus Model, Monomolecular Model, Logistic Model, Gompertz Model and Michaelis-Menten Model [49, 61].

Newton's Method, Gauss-Newton and Levenberg-Marquardt Methods are some of the nonlinear regression fit methods used in literature [3].

2.2.5 Generalized Additive Models

Ihe generalized additive model is founded by Hastie and Tibshirani approximately in 1990s and it is one of the latest DM tools in the literature as provided in Appendix G, Generalized Additive Models (GAMs) are classified as modern techniques from statistical learning, and they are applied in many prediction problems like financial mathematics, computational biology, medicine, chemistry and environmental protection. They are implemented by a local scoring algorithm with a scatter-plot smoother as building blocks proposed by [72].

In general, the GAMs have the formulation of $G(\mu(\mathbf{X})) = \Omega(\mathbf{X}) = \theta_0 + \sum_{j=1}^{n} f_j(X_j)$, where Ω is the function of predictors. The functions f_j are nonparametric and $\boldsymbol{\mu} = (\theta_0, f_1, f_2, \ldots, f_n)^T$ is the vector of unknowns to be estimated. The including of θ_0 as an average output enables the assumption of $E(f_j(X_j)) = 0$ for $j = 1, 2, \ldots, n$ [50, 72].

For identifying and characterizing nonlinear regression effects, they provide flexible solution techniques. The probability distribution target variable should be specified and therefore GAMs are parametric [49]; consequently, they can be classified as semi-parametric models.

It is highly important in GAMs to choose the suitable level of the *smoother* for a predictor. It can be obtained by determining the smoothing level with the introduction of adequate degrees of freedom. The proper balance ought to be established between the total observations and the total degrees of freedom used for model fit [50].

2.2.6 Nonparametric Regression Models

Although linear correlation between the response variable and the covariates is a strong assumption that is not necessarily valid for each model in practice. If we still fit real-life problems by linear method, the prediction can be misleading and wrong model formation occurs. Then, a nonparametric model would be more appropriate. A general nonparametric model takes the following form [86]:

$$Y_i = g(x_{i1}, \ldots, x_{ik}) + \varepsilon_i \quad \text{for} \quad i = 1, 2, \ldots, N, \tag{2.24}$$

where Y_i is the ith observation, N is the number of observations, and g is a nonspecified function. In nonparametric regression, the function of g is to be determined, whereas in parametric regression, the model parameters are searched. In order to find the functions g, there are mainly three widely used methods, which are: *Kernel Estimations*, *Regression Splines* and *Smoothing Splines* [90].

Kernel estimation methods, use various linear predictors to estimate the value at a specific point x. One of the famous linear estimators is *Nadaraya-Watson Estimator* and can be expressed as [90]:

$$\hat{g}_h(x) = \frac{\sum_{i=1}^{N} K_h(x_i - x) y_i}{\sum_{i=1}^{N} K_h(x_i - x)}. \tag{2.25}$$

Here, K is a kernel function and h is an interval length, which is a smoothing parameter that controls the size of the local neighborhood for N many data.

Second nonparametric regression model is the regression splines model with a nonparametric function including a set of basis functions. The model can be characterized for $N = 1$ as:

$$\hat{g}_h(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_r x^r + \sum_{i=1}^{M} A_i [x - k_i]_+^r, \tag{2.26}$$

where r is the order of the regression spline, k_i is the ith knot, $(a_0, \ldots, a_r, A_0, \ldots, A_M)^T$ is the coefficient vector and $[x - k_i]_+$ is a piecewise linear extension such that

$$[x - k_i]_+ = \begin{cases} x - k_i, & \text{if } x > k_i, \\ 0, & \text{otherwise.} \end{cases} \tag{2.27}$$

B-Spline Basis, *Natural Splines* and *Radial Basis* are other examples of basis functions.

The third well-known nonparametric regression model is called as smoothing spline and offered by Wahba [84]; and it is given as smoothing spline analysis of variance (SS-ANOVA) [86]. The smoothing regression function is written as

$$g(\mathbf{x}) = a + \sum_{i=1}^{n} g_i(x_i) + \sum_{1 \le i < k \le r} g_{ik}(x_i, x_k) + \ldots + g_{1\ldots r}(x_1, \ldots, x_r), \tag{2.28}$$

where a is a constant, the function g_i gives main effects, and g_{ik} are two-way interactions, and so on. In fact, SS ANOVA is a generalization of additive models since it enables the interaction of terms and the regression function g is decomposed into

several orthogonal functional components. When the interactions in Equation (2.25) are eliminated, then the model is reduced to an additive model.

There are also other nonparametric models such as projection-pursuit regression and Classification and Regression Trees (CART). MARS and CMARS functions are nonparametric, too, being nonsmooth and of a multiplicative nature [49].

CHAPTER 3

MARS AND CMARS MODELS

3.1 MARS Model

3.1.1 Procedure

MARS® is the short form of Multivariate Adaptive Regression Splines and it is a multivariate nonparametric regression procedure introduced by physicist and statistician, Jerome Friedman in 1991. Salford Systems Software, SPM (Salford Prediction Modeler), uses the original MARS algorithm of Friedman [20].

MARS is a data-driven model, and unlike other extensively used model-driven or supervised learning methods and algorithms, it is basically a regression model. Basis Functions (BFs), also called *splines*, are included as predictors based on the original data which enables quite flexible regression models. In MARS, all possible knot positions and all predictors are tracked and found together with every possible interaction in the model. The determination of the interactions is performed by the use of combinations of BFs. After MARS having determined the optimum quantities of basis functions and knot locations, the least-squares estimator method is applied in order to form the final model that gives the best approximation of the dataset with the remaining basis functions. Consequently, the final MARS additive model is determined with a two-phase process including **Forward** and **Backward Stages** [20, 62].

By the initial *forward stage*, often an over-fitting model is raised by including a rather large set of BFs. The model is reached by using a fast searching algorithm and the progress continues until the model constructs the user-defined maximum number $M_{\max}$ of basis functions. The model obtained at the initial stage has all of the possible BFs regardless of whether their assist to the overall performance is much or least. Therefore, the model in the forward stage needs to be cleared from the needless BFs and that obviously requires another stage, which is the backward stage of MARS.

In the *backward stage*, on the other hand, the over-fit model is trimmed to diminish the complexity of the model. Nevertheless, the model still administers the overall performance with the fit to the data. At the backward stage, the BFs that provides the smallest increase in the Residual Sum of Square (RSS) are selected from the model at every stage and as a result, an optimal model is obtained [36, 49, 60]. BFs are taken off

in order to get the optimum necessary quantities considering their minimum donation to the model. The stopping criterion for the backward stage aims to achieve in optimal balance between bias and variance.

MARS adopts piecewise linear expansions of BFs constituted by dataset. The form of the BFs is [26]:

$$[x-t]_+ = \begin{cases} x-t, & \text{if } x > t, \\ 0, & \text{otherwise,} \end{cases} \qquad [t-x]_+ = \begin{cases} t-x, & \text{if } x < t, \\ 0, & \text{otherwise,} \end{cases} \tag{3.1}$$

where t is a univariate knot determined using the dataset. The two *mirrored* functions are known as *truncated functions*. Figure 3.1 shows basis function pairs for $t = 0.5$ as an example.

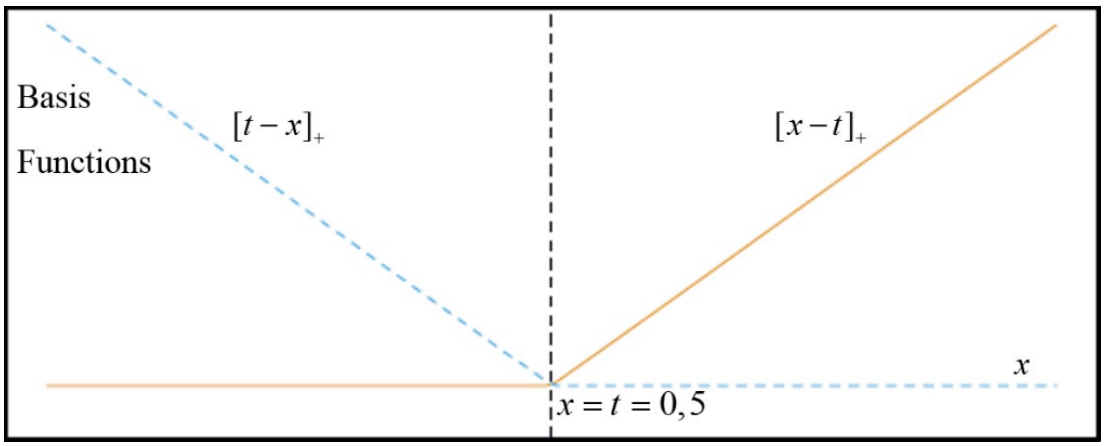


Figure 3.1: Sample Truncated Function.

In MARS algorithm, each function is written as piecewise linear; having a knot at the value t, and that is called a reflected pair. The objective is to model reflected pairs for each input dimension x_j $(j = 1, 2, \ldots, n)$ with n dimensional knots $\mathbf{t}_i = (t_{i1}, t_{i2}, \ldots, t_{in})^T$ at input data vectors $\mathbf{t}_i = (x_{i1}, x_{i2}, \ldots, x_{in})^T$, where $i = 1, 2, \ldots, N$. Accordingly, the set B of BFs is written by

$$B = \{[x_j - t]_+, [t - x_j]_+ \quad | \quad t \in \{x_{1j}, x_{2j}, \ldots, x_{Nj}\}, j = 1, 2, \ldots, n\}, \tag{3.2}$$

where N is the number of observations, p stands for the dimension of the input space so that, $2Nn$ BFs appears in case that all of the input values are distinct.

At forward stage of MARS, the model that adapts the data is constructed with BFs of the set B and their product. Consequently, the model here is

$$Y = E(\alpha_0 + \sum_{m=1}^{M} \alpha_m Q_m(\mathbf{X}) | \mathbf{X} = \mathbf{x}) + \varepsilon, \tag{3.3}$$

with vector of random variables and a vector. Here, ε is additive stochastic "noise" component that is supposed to have zero mean and constant variance. Furthermore, M is given as the number of BFs in the present model, $Q_m(\mathbf{x})$ are BFs included in the set B or products of two or more such functions, and α_m represents the unknown coefficients for the constant $1(m = 0)$ or for the mth BF. The form of the mth BF is specified as

$$Q_m(\mathbf{x}) = \prod_{k=1}^{K_m} [s_{km}.(x_{v(k,m)} - t_{km})]_+, \tag{3.4}$$

where K_m is the number of truncated linear functions multiplied in the mth basis function, $x_{v(k,m)}$ is the input variable for the kth truncated linear function in the mth basis function, t_{km} is the knot value conforming to the variable $x_{v(k,m)}$ and $s_{km} = \pm 1$. The lack-of-fit criterion is implemented to analyze the likely BFs.

In order to obtain the model, MARS forward stage is initiated with the $constant function$ $Q_0(\mathbf{x}) = 1$ to estimate, and all other functions in the set B are candidate functions. The attainable forms of the BFs $Q_m(\mathbf{x})$ are given as:

- 1,
- x_k,
- $[x_k - t_i]_+$,
- $x_k x_l$,
- $[x_k - t_i]_+ x_l$,
- $[x_k - t_i]_+ [x_l - t_j]_+$.

In the MARS algorithm, at every BF, predictor (input) variables may not be the equal. Q, the aforementioned BFs utilize various predictor variables, x_k and x_l with their knots being t_i and t_j. For each step of the forward stage, with a selection of the reflected pair in the BFs set of B, all multiples of the function $Q_m(\mathbf{x})$ in the model set are deemed. They are set as new function pair and included to the model set. The term which builds the maximum drop in the training error is in the form of

$$\alpha_{M+1} Q_k(\mathbf{x}) \cdot [x_j - t]_+ + \alpha_{M+2} Q_k(\mathbf{x}) \cdot [t - x_j]_+. \tag{3.5}$$

Due to the forward stage run, a large, over-fitting model is accessed very often. Then, the backward stage starts. Here, the terms that donate the smallest increase in the residual squared error are abolished at every step, and this repetitive procedure is followed until an optimum number of effective terms is ready at the end. Accordingly, a best model $\hat{f}_\mu$ for which each number of terms μ is formed with this process. In the MARS model, to find the optimal number of terms μ, the *generalized cross-validation* (GCV)

is used. Additionally, it represents the *lack of fit* in MARS model. The GCV formula introduced by Friedman [20] is:

$$LOF(\hat{f}_\mu) = GCV(\mu) = \frac{\sum_{i=1}^{N}(y_i - \hat{f}_\mu(\mathbf{x}_i))^2}{(1 - \frac{M(\mu)}{N})^2}, \tag{3.6}$$

where $M(\mu)$ stands for the effective number of parameters in the model, and N is the number of observations.

3.1.2 MARS vs. other Methods

MARS allows the user to determine possible nonlinearities in the link between target and predictor variables [13, 49]. Regression and other well-known statistical prediction methods require trying multiple combinations of the predictors in the dataset. On the other hand, MARS automatically searches the interactions between independent variables, and that makes it highly suitable, especially for large datasets. MARS identifies interactions with graphs which enables analysts to understand interactions [60].

Furthermore, solution times are short for medium- to high-dimensional problems, and it has the best prediction success when it is compared to linear models, principal component regression or CART, based on efficiency, accuracy, and implementation. It gives successful continuous results in predictive modeling [11].

The algorithm is formed on a methodology of modified recursive partitioning [20, 49]. It is an extension of Classification and Regression Tree (CART) [9]. However, MARS is not classified as decision tree technique, since CART uses indicator functions causing discontinuity that influences the model accuracy, whereas MARS uses piecewise continuous linear functions that produces a more effective way to model nonlinearities [79]. Additionally, it is compared to various parametric and nonparametric approximations routine in terms of its accuracy, efficiency, robustness, model transparency, and simplicity due to its inception [11].

Most recent gas consumption forecast algorithms are based on Neural Network models [66]. In some previous works, MARS has been referred to as a competitor to neural networks and it does not suffer from the limitations of neural networks [19]. Similarly, as in Neural Networks, MARS is highly effective when analyzing complex structures in the data. However, unlike neural networks, MARS is not a “black box” method so that it produces explainable results.

3.1.3 MARS Software

Salford Predictive Modeler (SPM) version SPM-64bit 7.0 version is used in the study. SPM software suit uses the original code generated by Freidman [20]. It provides a

very accurate and quick analytics and DM platform for various applications. SPM provides highly quick results. The suite version of SPM, which is also used in the study, includes Salford Systems' other products of CART, TreeNet, and Random Forests in addition to MARS. In suite version, the software also includes the *automation* feature that quickens the course of model formation by implementing substantial portions of the model exploration and refinement step for the user. Together with the automation option, SPM allows analyst to complete the work in one day that normally would be completed in a week even more with other softwares.

SPM software makes quite easy to load the data into the software. The data input formats. *csv*, *.xlsx* and some other types are allowed to import with a user friendly interface.

Following the data import, control of the operations are made from another user friendly main control screen as seen in Figure 3.2. There are some important control tabs user should be familier with.

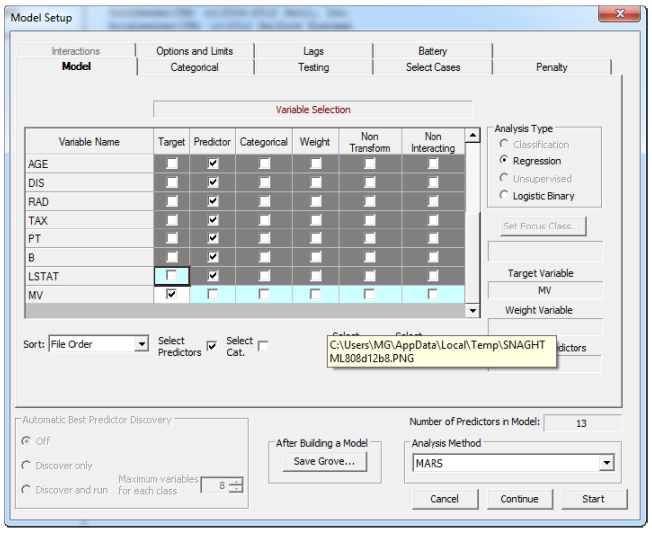


Figure 3.2: SPM Operations Main Screen [68].

Using the main screen, from the *Model* settings, the type of analysis, which is MARS for our study, selection of target variable and the dependent variables can be determined. One of the important features of MARS is its allowance to the processing of categoric data.

MARS, conducts categorical variables by building dummy variables for every possible groupings of levels and thoroughly looking for the best possible dummy variable to be placed into the model. The MARS type of categorical data handling s superior to any other traditional methods that construct dummy variables only for distinct categorical stages. The target variable can also be selected as categorical value by determining the *Analysis Type* as *Logistic Binary*. In the logistic binary mode, MARS treats the target variable coding as 0 or 1 although it is a regression procedure and handles the target variables as continuous.

Another important settings tab is *Options and Limits*. From here, maximum number of BFs, $M_{\max}$ is selected for the forward stage of MARS. The default value for $M_{\max}$ is 15, however the general procedure to set the maximum number of BFs is at least two to four times of the *truth*. Here, the truth is determined according to the previous experience of the analyst [68]. From the stability point of view, the maximum number of BFs is obtained at around $M_{\max} = 250$, while it is possible for even higher numbers with powerful workstations.

Using the *Options and Limits* tab, the maximum number of interaction is also controlled. Although the default setting is 1 which means no interaction, a setting of 2 will allow a 2-way interaction and so on. The restriction for maximum number of BFs and the largest degree of interaction should be evaluated together by the user in order to include the proper main effects.

Following the model run, one of the most important feature is obtained when the *Variable Importance* tab is checked for which a sample screen is given in Figure 3.3. The tab provides variables according to their importance scores. The variables are calculated on the 100% scale such that the most important variable always gets the full rate of 100%.

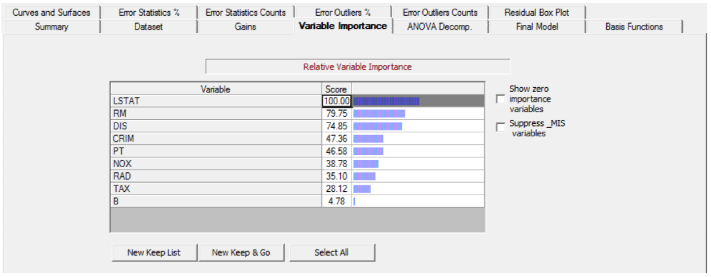


Figure 3.3: SPM Variable Importance Tab [68].

As a result of the backward stage, the resultant number of BFs allowing for minimum GCV is searched, as provided in Figure 3.4.

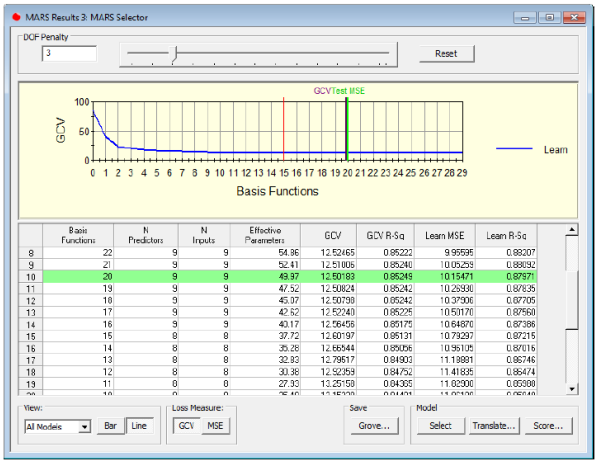


Figure 3.4: Minimum GCV track screen [68].

The results of the model can be visualized in 2D plots as in Figure 3.5 and in 3D plots as in Figure 3.6 if the number of interactions are 2-way or more. The visualizing of the variables contributions is the one of the most important features of MARS that eases modeling for the analyst.

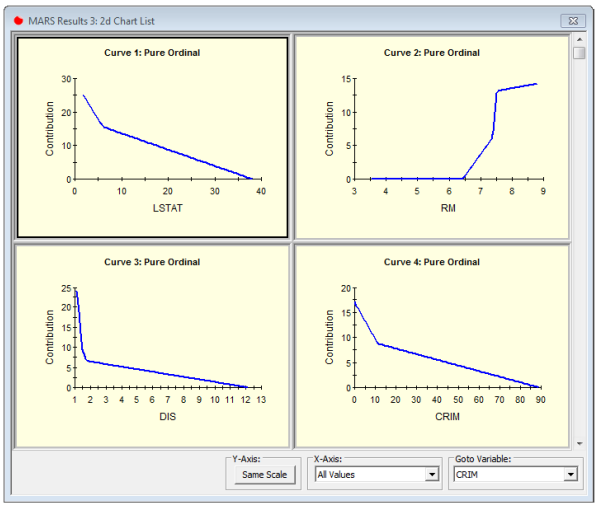


Figure 3.5: 2D Plots of Variable Contributions [68].

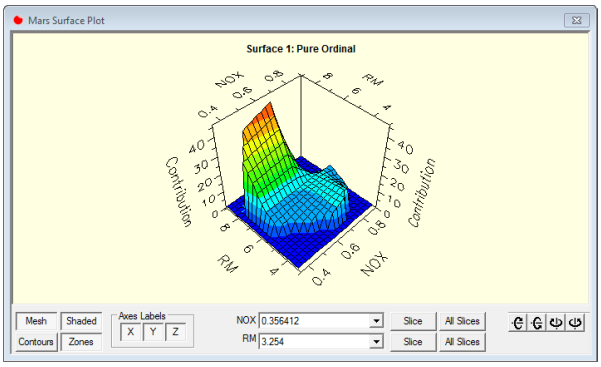


Figure 3.6: 3D Plots of Variable Contributions [68].

Finally, the resultant equation with optimum number of BFs and possible interactions is formulated in the *Basis Functions* tab. Examples of outputs for no-interaction case and 2-interaction case are provided as in Figures 3.7 and 3.8, respectively. The results of the model such that, R^2, *Mean Square Error (MSE)*, *ANOVA*, *f-value*, *t-value*, *p-value*, *gain and lift charts*, etc., can be viewed from other result tabssuch as *Summary*, *Gains*, *ANOVA Decomp* [49].

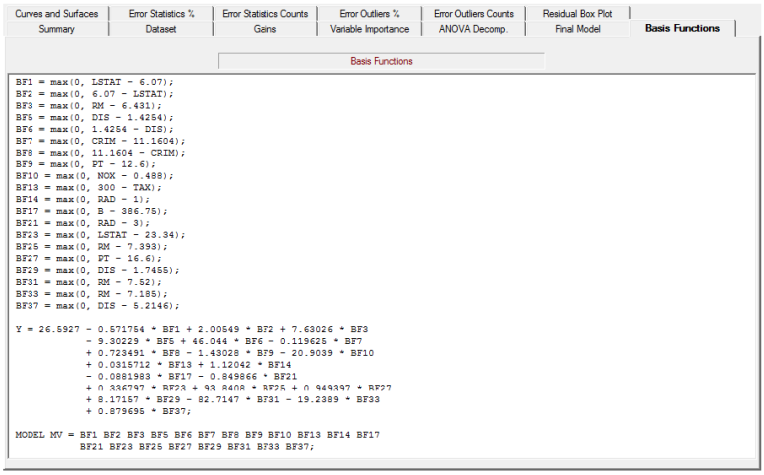


Figure 3.7: SPM output BFs, with no-interaction [68].

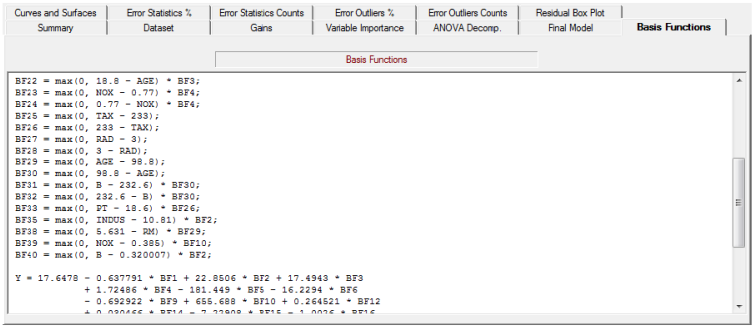


Figure 3.8: SPM output BFs, with 2 interactions [68].

3.2 CMARS Model

3.2.1 Introduction

Due to its flexible operation process, MARS algorithm provides highly successful implementations in many application fields. Following the proved success of MARS, several academic researches on MARS algorithm have been carried on and some of the studies are provided to alternatively improve its capability. CMARS is developed as a model based substitute to the backward stage of the MARS algorithm. Here, the letter “C” symbolize the terms of Conic, Convex, or Continuous. In CMARS, the adverse tradeoff between accuracy and stability is actually focused. Stability can also be named as “less complexity”. The objective to limit the complexity “under control” is attained with two different ways in [53]:

(i) preserving the first- and second-order derivatives for discretized integral of the BFs below an upper bound certain tolerance;

(ii) employing the state-of-the-art optimization theory. That allows a more unified processing of the forward stage and the backward stages of MARS.

There are characterizing milestones of CMARS and they must be clearly stated prior to applications with it.

3.2.2 Tikhonov Regularization

CMARS aims to minimize the *Penalized Residual Sum of Squares* (PRSS) by using all the BFs obtained in the forward stage. The Backward Stage of CMARS algorithm proposes Tikhonov regularization (or *ridge regression*) (TR) for PRSS and it is followed by conic quadratic programming. The resulting optimization problem is solved by interior point methods [74]. A problem is named as ill-posed when it has no unique or stable solution with some perturbations on data [49]. TR is the most popular method to make these problems regular and stable. When considering the general PRSS equation [73],

$$PRSS = \sum_{i=1}^{N}(y_i - f(\mathbf{x}_i))^2 + \sum_{m=1}^{M_{\max}} \phi_m \sum_{\substack{|\boldsymbol{\theta}|=1 \\ \boldsymbol{\theta}=(\theta_1,\theta_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V(m)}} \int_{W^m} \alpha_m^2 [D_{r,s}^{\boldsymbol{\theta}} \Psi_m(\mathbf{t}^m)]^2 d\mathbf{t}^m. \tag{3.7}$$

where $\boldsymbol{\Psi}_m(\mathbf{x}^m)$ is the mth BF and it is defined as [88];

$$\boldsymbol{\Psi}_m(\mathbf{t}^m) = \prod_{j=1}^{K_m} [s_{K_j^m} \cdot (t_{K_j^m} - \tau_{K_j^m})]_+. \tag{3.8}$$

Here, K_m stands for the total number of truncated linear functions multiplied in the mth BF, $t_{K_j^m}$ is the jth truncated linear function's input variable for the mth BF, $\tau_{K_j^m}$ is the knot value of the variable $t_{K_j^m}$ and $s_{K_j^m}$ provides the $\pm$ signs. Additionally, in general $PRSS$ Equation (3.7), $D_{r,s}^{\boldsymbol{\theta}} \Psi_m(\mathbf{t}^m) = \frac{\partial \Psi}{\partial^{\theta_1} t_r^m \partial^{\theta_2} t_s^m}(\mathbf{t}^m)$ for $|\boldsymbol{\theta}| = \theta_1 + \theta_2$ and $\theta_1, \theta_2 \in \{0, 1\}$. The $PRSS$ equation will also be detailed in Section 3.2.4.

This problem maybe ill-posed, therefore, TR follows the $PRSS$ equation [3, 25]. In order to make $PRSS$ simpler, for every derivative term the uniform penalization parameter ϕ_m is introduced. Then the PRSS problem becomes TR problem. Solution with TR can be stated easily with Singular Value Decomposition (SVD) of the coefficient matrix), of a related linear system of equations. Equation (3.7) is first formulated as follows:

$$
\begin{aligned}
PRSS \approx & \| \mathbf{y} - \boldsymbol{\Psi}(\mathbf{b})\boldsymbol{\alpha} \|_2^2 + \sum_{m=1}^{M_{max}} \phi_m \sum_{i=1}^{(N+1)^{K_m}} L_{im}\alpha_m^2 \\
& = \sum_{m=1}^{M_{max}} \phi_m [(L_{1m}\alpha_m)^2 + (L_{2m}\alpha_m)^2 + \ldots + (L_{(N+1)^{K_m}}\alpha_m)^2] \\
& = \| \mathbf{y} - \boldsymbol{\Psi}(\mathbf{b})\boldsymbol{\alpha} \|_2^2 + \sum_{m=1}^{M_{max}} \phi_m \| \mathbf{L}_m\alpha_m \|_2^2,
\end{aligned}
\tag{3.9}
$$

where $\mathbf{L}_m = (L_{1m}, L_{2m}, \ldots, L_{N+1^{K_m}m})^T \quad (m = 1, 2, \ldots, M_{\max})$. However, there is a finite number of tradeoff or penalty parameters $\phi_1, \phi_2, \ldots, \phi_{M_{\max}}$. But the problem requires a uniform penalization in order to become a TR problem. Then, by taking a single for all derivative terms, the $PRSS$ problem can be rearranged such that Equation (3.10) is obtained as a classical Tikhonov Regularization problem [3, 73]. The PRSS problem can be stated as in Equation (3.10):

$$
PRSS \approx \| \mathbf{y} - \boldsymbol{\Psi}(\mathbf{b})\boldsymbol{\alpha} \|_2^2 + \phi \| \mathbf{L}\boldsymbol{\alpha} \|_2^2 . \tag{3.10}
$$

In Equation (3.10), there are two objective functions to be optimized via linear combination. The solution consists of both objectives one by one in terms of a trade-off solution and that process makes the problem a multi-objective optimization problem. In the solution, TR combines both objective functions into a single functional form with a penalty term However, following the TR application, some complicated combinations of weighted linear sums of objectives can be obtained. In that point at the following stage, the problem is represented by conic quadratic programming and its solution completes the CMARS model [34, 74].

The penalty parameters provided in Equation (3.9) in the PRSS are not easily providing a parametric upper bound in a constraint of the conic quadratic problem, in general. Therefore, minimizing $PRSS$ is given by Tikhonov regularization, as the easiest way where there is only one penalty parameter The parameters and *K*, being the upper bound, are found in some reliance and equivalence [3]. Using Tikhonov regularization, logarithmic scales are used such that a "kink" point exists on the efficiency boundary. Such logarithmic scales are generally called to provide an *L-curve*. That particular point is the closest to the origin and it is selected with its penalty parameter [3].

The L-curve is an *efficiency curve* (also called *efficiency frontier*) and used to keep track of optimal solutions for a larger finite points in the two axes coordinate scheme. The complexity is at one axis and the length of the residual vector is placed at the other axis.

3.2.3 Conic Quadratic Programming Problem and Its Solution

CMARS employs a $PRSS$ for the MARS backward stage as a TR problem and the resultant two-objective optimization problem is solved using the continuous optimization method called Conic Quadratic Programming (CQP) [49]. Well structured CQPs are, herewith, resembling linear programs, hence, their solution is obtained permitting the use of interior point methods [57]. A Conic Quadratic Programming (CQP) is the problem of minimizing a linear objective function subject to the intersection of an affine set and the Cartesian product of quadratic (or second-order or Lorentz or ice-cream) cones such as [49]

$$\{\mathbf{x} \in \mathbb{R}^{m+1} \mid x_{m+1}^2 \geq \sum_{j=1}^{m} x_j^2,\ x_{m+1} \geq 0\}. \tag{3.11}$$

Therefore, a general formulation for conic optimization problem can be given as [5, 49]:

$$\begin{array}{ll} \text{minimize} & c^T\mathbf{x} \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \in K, \end{array} \tag{3.12}$$

where the cone K is a Cartesian product of several ice-cream cones [73],

$$K = L^{m_1} \times L^{m_1} \times \ldots \times L^{m_k}, \tag{3.13}$$

so that problem given in Equation (3.12) becomes [49]:

$$\begin{array}{ll} \text{minimize} & c^T\mathbf{x} \\ \text{subject to} & \mathbf{A}_i\mathbf{x}_i - \mathbf{b}_i \geq_{L^{m_i}} \mathbf{0} \quad (i = 1, 2, \ldots, r). \end{array} \tag{3.14}$$

Here, $\mathbf{A}_i\mathbf{x}_i - \mathbf{b}_i \geq_{L^{m_i}} \mathbf{0}$ $(i = 1, 2, \ldots, r)$ represents finitely many *ice-cream cone constraints* of the CQP, $[\mathbf{A}; \mathbf{b}]$ is defined as the data matrix, and its partition is shown as [5]

$$[\mathbf{A};\mathbf{b}] = \begin{bmatrix} [A_1;b_1] \\ [A_2;b_2] \\ \cdots \\ [A_r;b_r] \end{bmatrix}, \tag{3.15}$$

where the whole data matrix is subdivided by sub-matrices as follows:

$$[\mathbf{A}_i;\mathbf{b}_i] = \begin{bmatrix} \mathbf{D}_i & \mathbf{d}_i \\ \mathbf{p}_i^T & q_i \end{bmatrix}. \tag{3.16}$$

Finally, the most explicit form of CQP problem is formulated as follows [49, 74]:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \mathbf{c}^T\mathbf{x} \\ \text{subject to} \quad & \parallel \mathbf{D}_i\mathbf{x} - \mathbf{d}_i \parallel_2 \leq \mathbf{p}_i^T\mathbf{x} - q_i \ (i = 1, 2, \ldots, r). \end{aligned} \tag{3.17}$$

Equations (3.16), (3.17) and, $\mathbf{D}_i$ in Equation (3.17) are of the format $(m_i - 1) \times dim\mathbf{x}$, having the same row dimensions. The vector $\mathbf{d}_i$ has same dimensions with the columns of the matrices $\mathbf{D}_i$, the $\mathbf{p}_i$ vector has the same dimension with $\mathbf{x}$ and q_i are real numbers.

After the conic quadratic programming problem is formed, for its solution, an Interior Point Method (IPM) is applied. IPM was introduced by the seminal work of [35] and it was first offered for the solution of linear programming [5]. A general approach for classical interior point scheme can be started with the general problem definition of

$$\text{minimize} \quad \mathbf{c}^T\mathbf{x} \quad \text{such that} \quad \mathbf{x} \in S \subset \mathbb{R}^n, \tag{3.18}$$

where S is a closed convex set for which the assumption of having nonempty interior holds. It is a general idea that, continuous unconstrained minimization problems are easy, therefore the widely accepted aim is to reduce Equation (3.18) to a series of smooth unconstrained optimization problems. Then, the *barrier (or interior penalty)* function of $F(\mathbf{x})$ for the feasible set of S is selected. The barrier function $F(\mathbf{x})$ is

smooth and convex on the interior. It "explodes" from interior int$S$ and approaches a boundary point of S, such that [5]:

$$(x_j) \in (\text{int}S)^{\mathbb{N}},\ x = \lim_{j\to\infty} \mathbf{x}_j \in \partial S \ \Rightarrow F(\mathbf{x}_j) \to \infty\ (j \to \infty). \tag{3.19}$$

The one-parametric family of barrier functions generated by the objective is obtained as

$$F_t(\cdot) = t\mathbf{c}^T \cdot + F(\cdot) \mid \text{int}S \to \mathbb{R}. \tag{3.20}$$

In Equation (3.20), t stands for the penalty parameter and assumed to be nonnegative. The results obtained from mild regularity assumptions (e.g., S is bounded) are:

(i) Every function $F_t(\cdot)$ achieves its minimum for the interior of S, the minimizer $\mathbf{x}_*(t)$ being unique.

(ii) The central path $\mathbf{x}_*(t)$ is a continuous curve and its all limiting points, where $t \to \infty$ belongs to the set of optimized solutions of the problem given in Equation (3.18).

As a result of the interior point method, the initial situation is restored in a very close point to the initial point; however, the latest point has been carried along the central path towards the optimum set of given by Equation (3.18) [5].

3.2.4 Procedure

In CMARS, the penalized residual sum of squares (PRSS) for MARS with a TR problem approach in the backward stage is set up. The CMARS algorithm, penalizes a model not only with respect to "steepness" but, especially, its "curvature" ("energy" or "complexity"), and the procedure is called *regularization* [52, 72, 88]. Indeed, that is a first-order and, especially, second-order regularization, in terms of integrals and, then, following the discretizing the integrals, with sums of squares.

In CMARS algorithm, the backward stage of MARS is not exploited. The PRSS method is utilized with a number of $M_{\max}$ BFs in the forward stepwise algorithm of MARS. Penalty terms and Least-Squares Estimation (LSE) are applied as well in order to control the lack of fit from the tradeoff perspective between complexity (accuracy) and stability. As a result for MARS applications, PRSS gets the form [51, 49, 90].

$$PRSS = \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i))^2 + \sum_{m=1}^{M_{\max}} \phi_m \sum_{\substack{|\boldsymbol{\theta}|=1 \\ \boldsymbol{\theta}=(\theta_1,\theta_2)^T}}^{2} \sum_{\substack{r<s \\ r,s\in V(m)}} \int_{W^m} \alpha_m^2 [D_{r,s}^{\boldsymbol{\theta}} \Psi_m(\mathbf{t}^m)]^2 d\mathbf{t}^m. \tag{3.21}$$

In this representation, where the elements of the set $V(m) = \{(K_j^m)|j = 1, 2, \ldots, K_m\}$ enumerate the variables with respect to the m th BF of Ψ_m, and $\mathbf{t}^m = (t_{m_1}, \ldots, t_{m_{K_m}})^T$ represents the vector of variables that donates to BF Ψ_m . The terms $\phi_m \geq 0$ are the *penalty parameters*, where $m = 1, 2, \ldots, M_{\max}$. Besides, W^m is a sufficiently large K_m -dimensional parallel pipe, containing the data which occur in the regarded subspaces. Following a careful discretization and approximating of the the multivariate integral $\int_{W^m} \alpha_m^2 [D_{r,s}^{\boldsymbol{\theta}} \Psi_m(\mathbf{t}^m)]^2 d\mathbf{t}^m$, the approximate relation in Equation (3.7) can be rearranged [72] so that

$$PRSS \approx \| \mathbf{y} - \boldsymbol{\Psi}(\mathbf{b})\boldsymbol{\alpha} \|_2^2 + \sum_{m=1}^{M_{max}} \phi_m \sum_{i=1}^{(N+1)^{K_m}} L_{im}^2 \alpha_m^2, \tag{3.22}$$

where

$$L_{im} = \sum_{\substack{|\boldsymbol{\theta}|=1 \\ \boldsymbol{\theta}=(\theta_1,\theta_2)^T}}^{2} \sum_{\substack{r<s \\ r,s \in V(m)}} [D_{r,s}^{\boldsymbol{\theta}} \Psi_m(\hat{\mathbf{x}_i^m})]^2 \Delta \hat{\mathbf{x}}_i^m]^{\frac{1}{2}}. \tag{3.23}$$

Here, $\boldsymbol{\Psi}(\mathbf{b}) = (\boldsymbol{\Psi}(\mathbf{b}_1), \boldsymbol{\Psi}(\mathbf{b}_2), \ldots, \boldsymbol{\Psi}(\mathbf{b}_N))^T$ is an $(N \times (M_{max} + 1))$-matrix, and $\sum_{i=1}^{N} (y_i - \boldsymbol{\alpha}^T \boldsymbol{\Psi}(\mathbf{b}_i))^2 = (\mathbf{y} - \boldsymbol{\Psi}(\mathbf{b}_i)\boldsymbol{\alpha})^T (\mathbf{y} - \boldsymbol{\Psi}(\mathbf{b}_i)\boldsymbol{\alpha}) = \| \mathbf{y} - \boldsymbol{\Psi}(\mathbf{b})\boldsymbol{\alpha} \|_2^2)$. Finally, the $PRSS$ approximation can be specified as

$$PRSS \approx \| \mathbf{y} - \boldsymbol{\Psi}(\mathbf{b})\boldsymbol{\alpha} \|_2^2 + \phi \| \mathbf{L}\boldsymbol{\alpha} \|_2^2 . \tag{3.24}$$

In Equation (3.24), $\mathbf{L}$ is a diagonal $(M_{\max} + 1) \times (M_{\max} + 1)$-matrix and $\boldsymbol{\alpha}$ is an unknown vector which is resolved with data. Then the $PRSS$ problem is transformed into a classical TR problem with $\phi > 0$, $\phi = \lambda^2$ for some $\lambda \in \mathbb{R}$. TR problem as given in Equation (3.24) is solved through a continuous optimization technique, CQP [49, 50, 54, 72] Therefore, $PRSS$ can be reorganized as a problem of CQP. When focusing on an efficient choice of a bound $K > 0$, the problem is rearranged as

$$\begin{aligned} \underset{t,\boldsymbol{\alpha}}{\text{minimize}} \quad & t, \\ \text{subject to} \quad & \|\boldsymbol{\Psi}(\boldsymbol{b})\boldsymbol{\alpha} - \mathbf{y}\|_2 \leq t, \\ & \|\mathbf{L}\boldsymbol{\alpha}\|_2 \leq \sqrt{K}. \end{aligned} \tag{3.25}$$

It should be underlined that, the choice of K according to statistical performance or comparison criteria has to be the conclusion of a deep and broad learning process. After that point, the problem formulation is similar to the form provided in Equation (3.17), therefore problem solution is obtained by solution of CQPs.

In CMARS applications, the backward stage is solved with the MOSEK add-on to the MATLAB software [43]. MOSEK is a widely used for its excellent interior-point optimizer properties and it can be run with different platforms like Python, Matlab or R [43].

Following the detailed explanations of MARS and CMARS methods, it is important to provide their general modeling attributes.

3.2.5 CMARS vs. MARS

MARS is very sensitive to sampling size and design of the experiment, which has led, mainly, development of robust CMARS. In addition, MARS is generally lack of providing good results when compared with other approximation techniques in small datasets, although it obtains top results with medium or large ones with controlled experimentation [11].

Since CMARS is developed as an alternative to the backward part of the MARS algorithm, the performance of CMARS is mostly compared to that of MARS. Some of the studies are made for classification problems, and the others are used for predictions [90].

As of a classification problem solution technologies, MARS and CMARS have been applied for developing classification models to determine if a person has diabetes or not [74, 90]. The hold-out validation technique (75% of observations used to train data) is used, and *Average Correct Classification Rate* (ACCR) measures are calculated. Both MARS and CMARS method provide similar results considering the ACCR measures. On the other hand, in terms of the true diagnose of the disease, CMARS is better than MARS. For both training and test sets, the MARS and CMARS scoring models are close to each other; therefore, for the mentioned studies, CMARS gives a superior estimation compared to MARS [51, 90, 89].

CMARS and MARS both perform good mostly on large training/test samples, and as the training/test sample size decreases from medium to small, MARS and CMARS performance decrease. For large training and test samples, CMARS is better than MARS considering MSE. For small datasets, CMARS is more stable than MARS. However, on medium to large datasets, MARS is more stable than CMARS [90]. If the performance with respect to scale is considered, almost for all performance measures, MARS and CMARS perform the same as the scale changes from small to medium and medium to large.

When the computational run times (in seconds) for both MARS and CMARS are considered, they seem to be related to the sample size not the problem scale. As a result, the least amount of run times for both methods is obtained for small size samples regardless of the scale. MARS gives solutions very fast compared to CMARS, since its applications are based on the professional software SPM-M-64 by Salford-Systems. The elapsed time of CMARS method almost increases up to three to five times compared to that of MARS, as the sample size increases [2, 55, 90, 91].

As a new application area of both models, MARS and CMARS will be applied in NG consumption prediction for a real-time problem in the next chapter, and a comparison will be provided.

CHAPTER 4

REAL-WORLD APPLICATION WITH MARS AND CMARS

4.1 Introduction

This study has been made for the responsibility area of Başkentgaz. Ankara, with over 5 millions of population is the second most crowded city of Turkey after Istanbul with a population of over 14 millions as of 2014, Ankara is the second largest city with its population over 5 millions, according to Turkish Statistics Institute records [69]. Ankara is also one of the most crowded cities of Europe considering 2014 records as symbolized in Table 4.1.

In this study, the daily consumption of Ankara City residential users of Başkentgaz is modeled using MARS and CMARS algorithms. Meteorological and consumption datasets together with supplementary inputs like the unit cost of gas for the residential users, the exchange rate of USD and Turkish Liras have also been used as inputs.

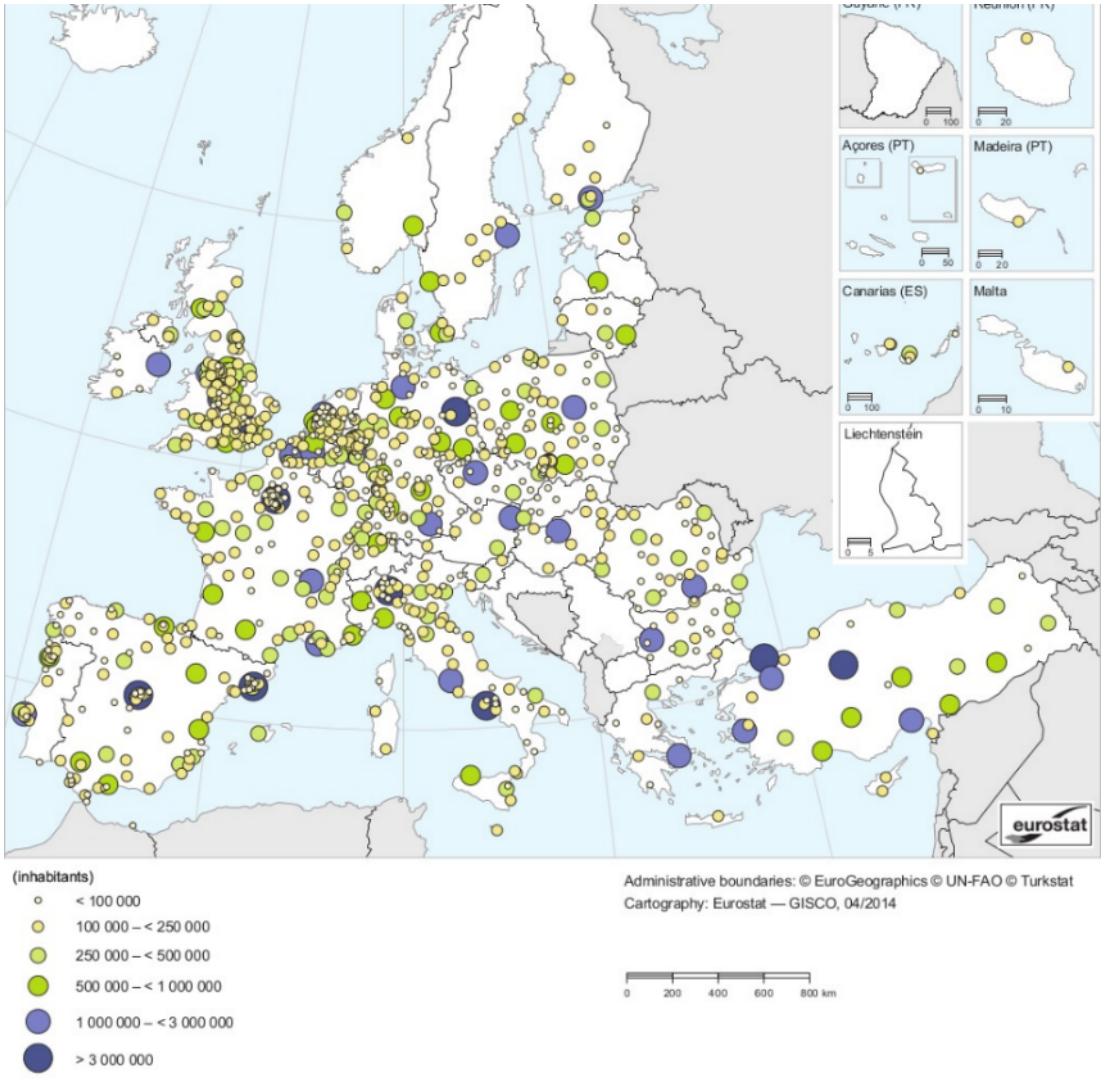
4.2 Description of Datasets

In the study, the period of daily training and test sets has been selected for the interval of 2009-2013. The training set has been formed for the period of 2009-2012, whereas the test set belongs to 2013. Meteorological, NG consumption and supplementary inputs are used in this application. Here, meteorological input data have been gathered from Turkish State Meteorological Service. Training and test sets are formed with the dataset given in Table 4.1.

The calculation method of Heating Degree Day (HDD) varies for different countries. However, Turkish State Meteorological Service accepts the HDD definition administered by Eurostat. Consumption of energy, specifically NG, depends strongly on weather conditions. When the temperature falls down below some value of heating threshold, consumers use more energy due to an increased requirement for space heating. Therefore, consumption data with corrected temperatures help to interpret energy consumption trends.

HDD is formulated as HDD $= 22 - T$, where 22 specifies the temperature at which the user heats the house. Eurostat, however, normally sets this temperature as 18° C . The

Table 4.1: Residential population in major cities of European countries [18].



experiences of analysts of Başkentgaz showed that 22° C reflects more accurately the user NG consumption behavior in Ankara City. The HDD formula is applied when the temperature is below the heating threshold being 15° C as given by Eurostat. For the cases, where the temperature is over 15° C, the HDD is taken as zero, showing that no heating is required. Eurostat defines the average temperature T as $T = (T_{\text{min}} + T_{\text{max}})/2$, where T_{min} is the Daily Minimum Temperature in $^\circ$ C and T_{max} is the Daily Maximum Temperature in $^\circ$ C.

Another subgroup of input variables is consumption. NG is taken from BOTAŞ, which is the transmission system operator of Turkey, to the LDC of Başkentgaz A.Ş. via 5 delivery points which are called Pressure Reduction and Measurement Stations (RM/A). Those stations are Yapracık RM/A and Başkentgaz Gölbaşı RM/A, and the residential users in Ankara City receives the consumption mainly from those delivery points. Therefore, the total daily NG consumptions from aforementioned stations provided as

input to the models for the period of 2009-2013. In MARS and CMARS models, when the day-ahead natural gas consumption is taken as target variable, the previous day consumption is received as a predictor variable introducing the time-series approach into the model, too.

In addition to meteorological and NG consumption inputs, some supplementary inputs are granted to the models. One such input is the currency exchange rate being the Turkish Liras / USD. One of the important inputs is the number of residential users for each year's January. The number of users has a certain effect on the NG consumption. Therefore, instead of supplying the number of residential users as a separate predictor variable. The natural gas consumptions have been implemented into the algorithms as ratio of Daily NG consumption to the number of residential users for that year, instead of keeping the consumption as it is. The final input to the model is the retail cost of the NG for the residential users in the form of cost in Turkish Liras/ Sm^3.

4.3 Applications of MARS and CMARS Methods

First, using the training dataset explained above, several MARS models were developed using Salford System's MARS software [67]. After selecting the best two among them, the CMARS models were constructed as defined in Subsection 3.2. In the CMARS algorithm, the MARS models were obtained by using Salford System's SPM software MARS modeling tool. Then, the maximum number of BFs $M_{\max}$ and largest degree of interaction were defined. For the first model, $M_{\max}$ is 20, and the highest degree of interaction is 2. For the second model, $M_{\max}$ is 25, and the highest degree of interaction is 2.

MARS and CMARS methods are applied to two different cases. The first model obtained in both MARS and CMARS includes the daily average temperature converted to HDD. However, the second model includes the minimum and maximum daily temperature without being converted to HDD.

For each model, $M_{\max}$ number of BFs and the optimally estimated models with the reduced number of BFs are constructed after the forward and the backward step of MARS by its software. At the end, the final models used for MARS algorithm and the largest models used for CMARS algorithm are found and represented in Subsections 4.3.1 and 4.3.2, respectively.

4.3.1 MARS Models

Following the backward stage of MARS, for two alternative models, the numbers of BFs are shortened to 12 and 19, respectively. Therefore, the final MARS models are obtained in the following form:

$$
\begin{aligned}
\hat{Y_1} &= \alpha_0 + \sum_{m=1}^{M} \alpha_m \Psi_m(\mathbf{x}^m) \\
&= \alpha_0 + \alpha_1 \text{max}\{0, x_1 - 1.3717\} + \alpha_2 \text{max}\{0, 1.3717 - x_1\} \\
&+ \alpha_3 \text{max}\{0, x_5 - 2.1282\} \cdot \text{max}\{0, x_1 - 1.3717\} \\
&+ \alpha_4 \text{max}\{0, 1.1138 - x_1\} \cdot \text{max}0, x_2 + 0.1815 \\
&+ \alpha_5 \text{max}\{0, x_1 + 0.6248\} \\
&+ \alpha_6 \text{max}\{0, x_5 - 1.0580\} \cdot \text{max}0, x_2 + 0.1815 \\
&+ \alpha_7 \text{max}\{0, 1.0580 - x_5\} \cdot \text{max}\{0, x_2 + 0.1815\} \\
&+ \alpha_8 \text{max}\{0, x_5 - 0.2085\} \cdot \text{max}\{0, x_2 + 0.1815\} \\
&+ \alpha_9 \text{max}\{0, x_5 - 2.1282\} \cdot \text{max}\{0, x_2 + 0.1815\} \\
&+ \alpha_{10} \text{max}\{0, x_2 - 1.0682\} \cdot \text{max}\{0, 1.3717 - x_1\} \\
&+ \alpha_{11} \text{max}\{0, x_2 - 0.1424\},
\end{aligned}
\tag{4.1}
$$

$$
\begin{aligned}
\hat{Y_2} &= \alpha_0 + \alpha_1 \text{max}\{0, x_1 - 1.3717\} + \alpha_2 \text{max}\{0, 1.3717 - x_1\} \\
&+ \alpha_3 \text{max}\{0, x_3 - 0.3140\} + \alpha_4 \text{max}\{0, 0.3140 - x_3\} \\
&+ \alpha_5 \text{max}\{0, x_5 - 0.1739\} \cdot \text{max}\{0, x_1 - 1.3717\} \\
&+ \alpha_6 \text{max}\{0, -0.5836 - x_1\} \cdot \text{max}\{0, 0.3140 - x_3\} \\
&+ \alpha_7 \text{max}\{0, -0.5836 - x_1\} \cdot \text{max}\{0, 0.3140 - x_3\} \\
&+ \alpha_8 \text{max}\{0, x_1 - 1.5815\} \\
&+ \alpha_9 \text{max}\{0, x_5 + 0.7140\} \cdot \text{max}\{0, 1.5815 - x_1\} \\
&+ \alpha_{10} \text{max}\{0, -0.7140 - x_5\} \cdot \text{max}\{0, 1.5815 - x_1\} \\
&+ \alpha_{11} \text{max}0, x_3 + 1,6124 \\
&+ \alpha_{12} \text{max}\{0, x_3 + 0.1260\} + \alpha_{13} \text{max}\{0, 0.1634 - x_5\} \\
&+ \alpha_{14} \text{max}\{0, x_5 - 0.9595\} \\
&+ \alpha_{15} \text{max}\{0, x_5 - 2.1282\} \cdot \text{max}\{0, 1.3717 - x_1\} \\
&+ \alpha_{16} \text{max}\{0, x_5 - 2.1312\} \cdot \text{max}\{0, 1.3717 - x_1\} \\
&+ \alpha_{17} \text{max}0, x_1 - 1.0946 \cdot \text{max}\{0, x_5 - 0.1634\}.
\end{aligned}
\tag{4.2}
$$

Here, x_1, x_2, x_3, x_5 are the standardized Previous Day NG Consumption, Daily Average Temperature in HDD, Daily Maximum Temperature and NG Residential User Unit Price, respectively. For our alternative models, unknown parameters are found and presented in Tables 4.2 and 4.3.

4.3.2 CMARS Models

For CMARS algorithm, to avoid non-differentiability for the optimization problem of Equation (3.5), the knot values selected are different from but very close to the

Table 4.2: Parameter values of MARS algorithm for the first model.

α_0	α_1	α_2	α_3	α_4	α_5
1.209	0.608	-0.93	-1.461	0.2151	-0.119
α_6	α_7	α_8	α_9	α_{10}	α_{11}
0.385	-0.047	-0.24	-0.442	0.7384	0.477

Table 4.3: Parameter values of MARS algorithm for the second model.

α_0	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
0.267	1.016	-0.817	-0.710	0.987	-0.178	-0.151	-0.819	-0.259
α_9	α_{10}	α_{11}	α_{12}	α_{13}	α_{14}	α_{15}	α_{16}	α_{17}
-0.044	0.642	0.336	0.362	-0.148	0.379	-26.824	25.828	-0.109

related input data. For the forward stage, before starting to the optimization problem in Equation (3.22), with the $M_{\max}$ number of BFs represented in Subsection 4.3, the greatest models become

$$
\begin{aligned}
\hat{Y_1} &= \alpha_0 + \sum_{m=1}^{M} \alpha_m \Psi_m(\mathbf{x}^m) \\
&= \alpha_0 + \alpha_1 \text{max}\{0, x_1 - 1.3718\} + \alpha_2 \text{max}\{0, 1.3718 - x_1\} + \alpha_3 \text{max}\{0, x_2 + 0,1816\} \\
&+ \alpha_4 \text{max}\{0, -0,1816 - x_2\} + \alpha_5 \text{max}\{0, x_5 - 2,1283\} \cdot \text{max}\{0, x_1 - 1,3718\} \\
&+ \alpha_6 \text{max}\{0, 2,1283 - x_5\} \cdot \text{max}\{0, x_1 - 1.3718\} \\
&+ \alpha_7 \text{max}\{0, x_1 - 1.1139\} \cdot \text{max}\{0, x_2 + 0.1816\} \\
&+ \alpha_8 \text{max}\{0, 1,1139 - x_1\} \cdot \text{max}\{0, x_2 + 0.1816\} \\
&+ \alpha_9 \text{max}\{0, x_1 + 0,6249\} + \alpha_{10} \text{max}\{0, -0,6249 - x_1\} \\
&+ \alpha_{11} \text{max}\{0, x_5 - 1,0581\} \cdot \text{max}\{0, x_2 + 0.1816\} \\
&+ \alpha_{12} \text{max}\{0, 1,0581 - x_5\} \cdot \text{max}\{0, x_2 + 0.1816\} \\
&+ \alpha_{13} \text{max}\{0, x_5 - 0,2086\} \cdot \text{max}\{0, x_2 + 0.1816\} \\
&+ \alpha_{14} \text{max}\{0, 0,2086 - x_5\} \cdot \text{max}\{0, x_2 + 0.1816\} \\
&+ \alpha_{15} \text{max}\{0, x_5 - 2,1283\} \cdot \text{max}\{0, x_2 + 0.1816\} \\
&+ \alpha_{16} \text{max}\{0, 2,1283 - x_5\} \cdot \text{max}\{0, x_2 + 0.1816\} \\
&+ \alpha_{17} \text{max}\{0, x_2 - 1,0683\} \cdot \text{max}\{0, 1,3718 - x_1\} \\
&+ \alpha_{18} \text{max}\{0, 1,0683 - x_2\} \cdot \text{max}\{0, 1,3718 - x_1\} \\
&+ \alpha_{19} \text{max}\{0, x_2 - 0,1425\} + \alpha_{20} \text{max}\{0, 0,1425 - x_2\},
\end{aligned}
$$

$$
\begin{aligned}
\hat{Y_2} = {} & \alpha_0 + \alpha_1 \text{max}\{0, x_1 - 1.3718\} + \alpha_2 \text{max}\{0, 1.3718 - x_1\} \\
& + \alpha_3 \text{max}\{0, x_3 - 0,3141\} + \alpha_4 \text{max}\{0, 0,3141 - x_3\} \\
& + \alpha_5 \text{max}\{0, x_5 - 0,1740\} \cdot \text{max}\{0, x_1 - 1,3718\} \\
& + \alpha_6 \text{max}\{0, 0,1740 - x_3\} \cdot \text{max}\{0, x_1 - 1,3718\} \\
& + \alpha_7 \text{max}\{0, x_1 + 0,5837\} \cdot \text{max}\{0, 0,3141 - x_3\} \\
& + \alpha_8 \text{max}\{0, -0,5837 - x_1\} \cdot \text{max}\{0, 0,3141 - x_3\} \\
& + \alpha_9 \text{max}\{0, x_1 - 1,5816\} + \alpha_{10} \text{max}\{0, 1,5816 - x_1\} \\
& + \alpha_{11} \text{max}\{0, x_5 + 0,7141\} \cdot \text{max}\{0, 1,5816 - x_1\} \\
& + \alpha_{12} \text{max}\{0, -0,7141 - x_5\} \cdot \text{max}\{0, 1,5816 - x_1\} + \alpha_{13} \text{max}\{0, x_3 + 1,6125\} \\
& + \alpha_{14} \text{max}\{0, -1,6125 - x_3\} + \alpha_{15} \text{max}\{0, x_3 + 0,1261\} + \alpha_{16} \text{max}\{0, -0,1261 - x_3\} \\
& + \alpha_{17} \text{max}\{0, x_5 - 0,1635\} + \alpha_{18} \text{max}\{0, 0,1635 - x_5\} + \alpha_{19} \text{max}\{0, x_5 - 0,9596\} \\
& + \alpha_{20} \text{max}\{0, 0,9596 - x_5\} + \alpha_{21} \text{max}\{0, x_5 - 2,1283\} \cdot \text{max}\{0, 1,3718 - x_1\} \\
& + \alpha_{22} \text{max}\{0, 2,1283 - x_5\} \cdot \text{max}\{0, 1,3718 - x_1\} \\
& + \alpha_{23} \text{max}\{0, x_5 - 2,1313\} \cdot \text{max}\{0, 1,3718 - x_1\} \\
& + \alpha_{24} \text{max}\{0, 2,1313 - x_5\} \cdot \text{max}\{0, 1,3718 - x_1\} \\
& + \alpha_{25} \text{max}\{0, x_1 - 1,0947\} \cdot \text{max}\{0, x_5 - 0,1635\}.
\end{aligned}
$$

Following the discretized form of multi-dimensional integrals in Equation (3.7) as finally expressed by $\mathbf{L}$, for the second part of the optimization solution in Equation (3.24), the $\mathbf{L}$ matrices become diagonal (21×22)- and (26×26)- matrices and the elements of the first column of $\mathbf{L}$ are all zero. From Equation (3.9), for our alternative models, $\|\mathbf{L}_1\boldsymbol{\alpha}\|_2^2$ and $\|\mathbf{L}_2\boldsymbol{\alpha}\|_2^2$ are defined as

$$
\begin{aligned}
\|\mathbf{L}_1\boldsymbol{\alpha}\|_2^2 = {} & (1.142\alpha_1)^2 + (1.574\alpha_2)^2 + (1.697\alpha_3)^2 \\
& + (1.912\alpha_4)^2 + (0.028\alpha_5)^2 + (0.04\alpha_6)^2 \\
& + (0.562\alpha_7)^2 + (0.109\alpha_8)^2 + (1.818\alpha_9)^2 \\
& + (0.693\alpha_{10})^2 + (0.263\alpha_{11})^2 + (0.022\alpha_{12})^2 \\
& + (0.308\alpha_{13})^2 + (1.018\alpha_{14})^2 + (0.165\alpha_{15})^2 \\
& + (0.205\alpha_{11})^2 + (0.018\alpha_{12})^2 + (0.267\alpha_{13})^2 \\
& + (1.595\alpha_{14})^2 + (1.080\alpha_{15})^2.
\end{aligned}
$$

$$
\begin{aligned}
\|\mathbf{L}_2\boldsymbol{\alpha}\|_2^2 = &(1.142\alpha_1)^2 + (1.574\alpha_2)^2 + (1.349\alpha_3)^2 \\
&+ (1.655\alpha_4)^2 + (0.072\alpha_5)^2 \\
&+ (0.040\alpha_7)^2 + (0.135\alpha_8)^2 + (1.047\alpha_9)^2 \\
&+ (1.639\alpha_{10})^2 + (0.178\alpha_{11})^2 + (0.032\alpha_{12})^2 \\
&+ (1.938\alpha_{13})^2 + (0.896\alpha_{14})^2 + (1.506\alpha_{15})^2 \\
&+ (1.513\alpha_{16})^2 + (1.483\alpha_{17})^2 + (0.979\alpha_{18})^2 \\
&+ (1.144\alpha_{19})^2 + (1.359\alpha_{20})^2 (0.265\alpha_{22})^2 \\
&+ (0.265\alpha_{24})^2 + (0.186\alpha_{25})^2,
\end{aligned}
$$

When the greatest models are obtained and the $\mathbf{L}$ matrices in Equation (3.24) are evaluated, the $PRSS$ is reformulated as a problem of CQP for each model. Here, the values $\sqrt{K}$ in Equation (3.24) are determined by train and error method. CMARS provides several solutions and each of them is based on 20 and 25 BFs, when MOSEK [43] is applied for the CMARS code. To treat the CQP problems, MOSEK employs an interior point optimizer. After obtaining MOSEK formats, we solve our models by using MOSEK software [43]. Consequently, for our two models, the unknown parameters are defined and represented as given in Tables 4.4 and 4.5.

Table 4.4: Parameters of CMARS algorithm for the first model.

α_0	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}
0.203	0.090	-0.490	0.005	-0.188	-1.363	0.007	0.018	0.212	0.331	-0.497
α_{11}	α_{12}	α_{13}	α_{14}	α_{15}	α_{16}	α_{17}	α_{18}	α_{19}	α_{20}	
0.022	0.316	0.134	-0.376	-0.462	0.012	0.779	0.023	0.143	0.165	

Table 4.5: Parameters of CMARS algorithm for the second model.

α_0	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
1.202	0.441	0.039	0.066	0.244	-0.280	0.127	-0.095	-0.895
α_9	α_{10}	α_{11}	α_{12}	α_{13}	α_{14}	α_{15}	α_{16}	α_{17}
0.232	-0.084	-0.294	0.944	-0.079	0.314	-0.002	0.237	0.052
α_{18}	α_{19}	α_{20}	α_{21}	α_{22}	α_{23}	α_{24}	α_{25}	
-0.232	0.307	0.035	-21.340	-0.133	20.775	-0.129	-0.089	

4.4 Performance Comparison Methods

In order to evaluate and compare the results, some prediction performance criteria and related measures have been determined numerically. Mostly used accuracy measures for MARS and CMARS are tabulated in Table 4.6.

Table 4.6: Accuracy measures [49].

MEASURE	**FORMULA**
Adjusted Multiple Coefficient of Determination (R^2_{adj})	$R^2_{adj} = 1 - \frac{(N-1)}{(N-p-1)}(1 - \frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y}_i)^2})$
Average Absolute Error (AAE)	$AAE = \frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}_i\rvert$
Root Mean Squared Error ($RMSE$)	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$
Correlation Coefficient (r)	$r = \frac{\frac{\sum_{i=1}^{N}[(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})]}{(N-1)}}{\sqrt{sd(y)^2 sd(\hat{y})^2}}$

Here, N is the number of sample observations and p is the number of predictor variables, In aforementioned equations, y_i being the observed value (in the study corresponds to standardized daily consumption), with $\bar{y}$ denoting the mean of the set of the values y_i and $sd(y)$ stands for the standard deviation of the set for y_i. Additionally, $\hat{y}_i$ is the predicted value for the ith observed data, and $\bar{\hat{y}}$ is the mean value of the set of predicted values. The accuracy measures of R^2_{adj} (Adjusted Multiple Coefficient of Determination), AAE (Average Absolute Error), $RMSE$ (Root Mean Squared Error) and r (Correlation Coefficient) are used in the present study in order to compare methods of MARS and CMARS. It should be reminded that, in any model, for R^2_{adj} and r the values closer to 1 are better, whereas for AAE and $RMSE$, smaller values provide better results.

4.5 Comparison and Results

In order to compare the accuracy of MARS and CMARS models, the regression coefficients and estimation errors are determined based on the Adjusted Multiple Coefficients of Determination (R^2_{adj}), Average Absolute Error (AAE), Root Mean Squared

Error ($RMSE$) and Correlation Coefficient (r) , and the results are shown in Table 4.7.

Table 4.7: Comparison of the models.

	First Model				Second Model			
	training		test		training		test	
	MARS	**CMARS**	**MARS**	**CMARS**	**MARS**	**CMARS**	**MARS**	**CMARS**
R^2_{adj}	0.994	0.994	0.988	0.988	0.994	0.994	0.974	0.972
AAE	0.053	0.053	0.071	0.071	0.054	0.054	0.103	0.103
RMSE	0.080	0.080	0.107	0.108	0.080	0.079	0.156	0.160
r	0.997	0.997	0.995	0.995	0.997	0.997	0.988	0.987

The training data have been gathered for the daily consumption values of 2009-2012, and the test data have been provided for daily consumption values of 2013. For both MARS and CMARS models, the predicted and observed values are provided in standardized form and given at Figures 4.1-4.4 on the same graphs.

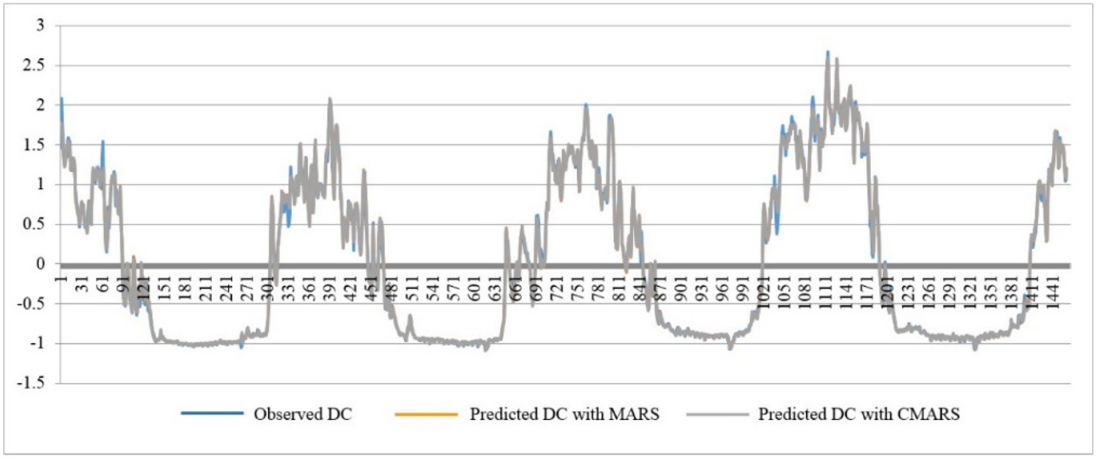


Figure 4.1: Real and Forecast Values of the First Model for Training Data.

As we may deduce from Figures 4.1-4.4, when the real and forecast values of the target variables for MARS and CMARS algorithms are taken into account, these can provide adequate results. Considering the modeling phases, MARS may obtain less costly solutions in NG consumption forecast, since it utilizes a smaller number of BFs. When the two models are compared, for the test data, the first model performed better than the second model when all measures are considered.

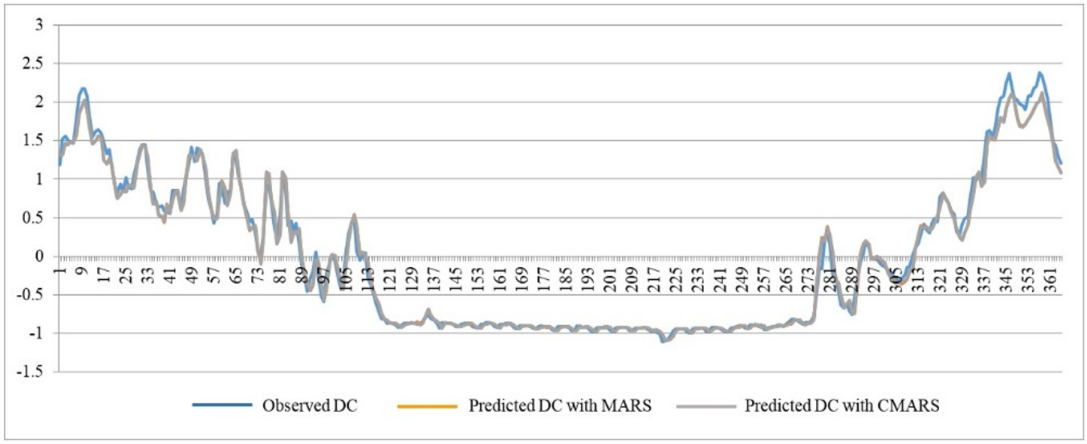


Figure 4.2: Real and Forecast Values of the First Model for Test Data.

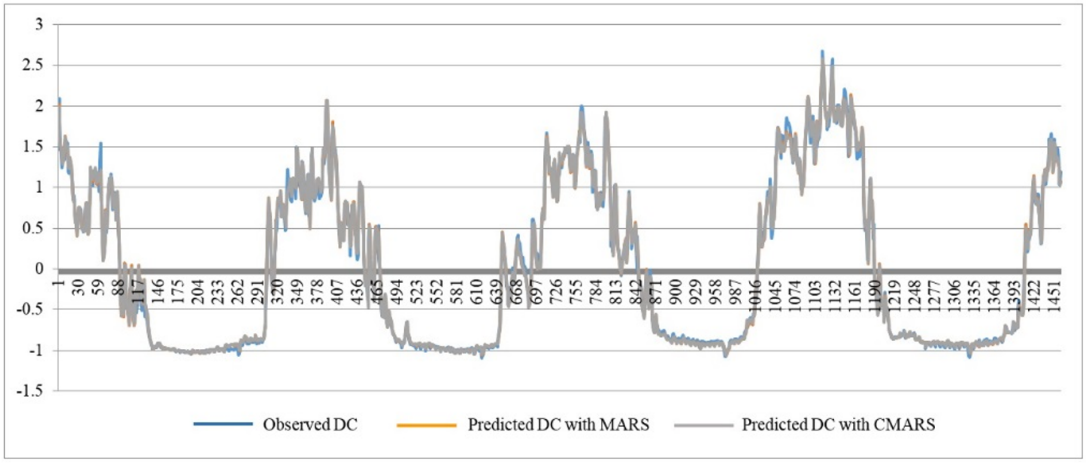


Figure 4.3: Real and Forecast Values of the Second Model for Training Data.

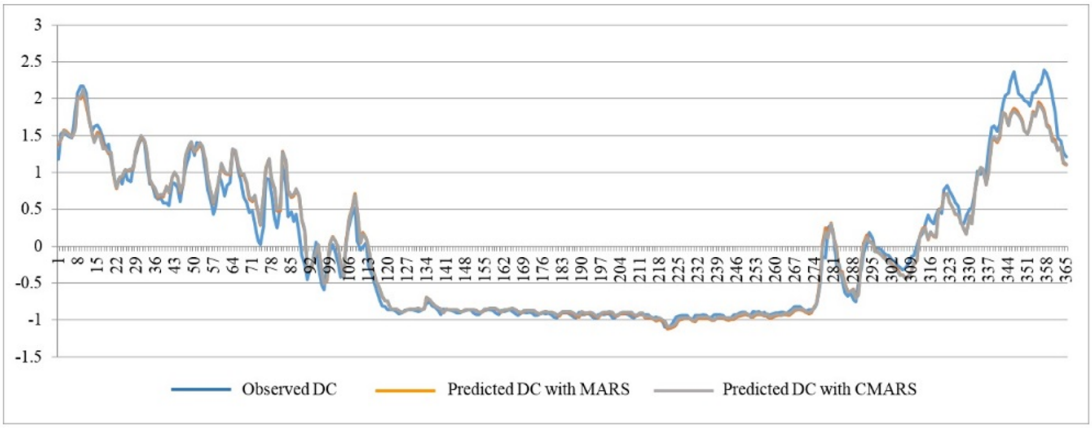


Figure 4.4: Real and Forecast Values of the Second Model for Test Data.

CHAPTER 5

THE OUTCOMES OF THE STUDY AND RECOMMENDATIONS FOR PROSPECTIVE WORKS

5.1 Conclusion

In this study, MARS and CMARS predictive algorithms are studied on the NG consumption forecasting of residential end users of Başkentgaz LDC. Consumption models have been produced with the training dataset, including the years of 2009-2012. The test dataset includes daily data for 2013. CMARS does not promise, generally, but often yields a superior "stabilization" over MARS. It provides better results than MARS in terms of accuracy. CMARS is even prevailing over MARS, in some cases, by properties of CMARS such as the dampening and smoothing effect on probable outliers exploited by the model-based behavior, and the existing regularization in the CMARS approach [53].

Both MARS and CMARS provide important results on modeling the residential consumption daily forecasts in a one-year period. They can enable us to obtain valuable information on monthly total NG consumption having monthly average consumption per day and peak demand consumption per day for each month. This information is quite important for LDCs and for TSOs analytic studies for short- and long-term planning. Since the forecasting studies are very closely related with meteorological input and predictions, the results are highly dependent on the accuracy of the meteorological data forecasts. Therefore, any analyst using prediction model results on consumption forecast should update the consumption results with amended meteorological inputs. In addition, the forecast models should be revised using the extended dataset by an additional number of data gathered.

Following the modeling practices of residential end-user consumptions with MARS and CMARS, it is observed that, one has to solve an additional problem of knot selection while using SPM v64 software to solve MARS algorithm. The number of variables used in CMARS is a display of some still existing complexity of CMARS. As the direct measurements and HDD values of temperatures compared, the models obtained with HDD values provide better results for both MARS and CMARS methods, which is also obtained in for Ankara [24]. In demand forecast studies of LDCs, both methods may be utilized to produce annual consumption figures.

5.2 Recommendation for Future Studies

The study introduces two innovative approaches to NG consumption forecast problems in energy industry. In the future studies, the study can be extended with applications of various other traditional and recent methods. ANN [45] may be applied to the Ankara data with Conic Generalized Partial Linear Model (CGPLM), its robustified variation (RCGPLM) and robust version of CMARS (RCMARS) [87]. The CGPLM and RCGPLM methods may provide some decrease of some complexity and by using robustification in optimization, the variance can be decreased [49, 50, 53, 54, 87]. As a result of that extensive study, broad comparison of the conventional and state-of-the-art methods will be obtained.

In addition to the consumption prediction study, MARS and CMARS methods are considered to have applications at other areas of engineering. As an example, concurrent with the present study, MARS application in gas turbines predictive maintenance has been implemented, successfully [93]. The study, may be extended to include CMARS, RCMARS, RCGPLM and ANN methods in order to provide detailed comparative research.

The comparative studies of MARS and CMARS methods with traditional ANN methods may also take part in energy industry with various datasets. Renewable energy supply and demand forecasts, the energy machineries' (i.e. pumps, turbines, wind mills, compressors etc.) performance and degradation prediction studies are evaluated to have significant contributions to academical researches by introducing new effective and fast predictive DM tools. The methods are estimated to have improvement in energy efficiency by operation costs minimization.

REFERENCES

[1] M. Akpinar and N. Yumusak, Forecasting household natural gas consumption with arima model: A case study of removing cycle, in *Application of Information and Communication Technologies (AICT), 2013 7th International Conference on Application of Information and Communication Technologies*, pp. 1–6, Oct 2013.

[2] O. S. Alp, E. Buyukbebeci, A. Iscanoglu Çekiç, F. Y. Ozkurt, P. Taylan, and G.-W. Weber, CMARS and GAM: CQP—Modern optimization methods applied to international credit default prediction, Journal of Computational and Applied Mathematics, 235(16), pp. 4639–4651, 2011, ISSN 0377-0427, congressional Contributions to Computational and Applied Mathematics: ICCAM2009.

[3] R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter Estimation and Inverse Problems*, Academic Press, Boston, 2013, ISBN 978-0-12-385048-5.

[4] M. Aydinalp-Koksal and V. I. Ugursal, Comparison of neural network, conditional demand analysis, and engineering approaches for modeling end-use energy consumption in the residential sector, Applied Energy, 85(4), pp. 271–296, 2008, ISSN 0306-2619.

[5] A. BenTal, Lectures on Modern Convex Optimization, 2013, http://www2.isye.gatech.edu/nemirovs/LectModConvOpt.pdf.

[6] BOTAŞ Petroleum Pipeline Corporation, Natural Gas Purchase Agreements, 2008, August 2014, http://www.botas.gov.tr/index.asp.

[7] BP, Statistical Review of World Energy, June, August 28, 2014, http://www.bp.com/en/global/corporate/press/speeches/bp-statistical-review-of-world-energy-2014.html.

[8] M. Brabec, O. Konár, E. Pelikán, and M. Maly, A nonlinear mixed effects model for the prediction of natural gas consumption by individual customers, International Journal of Forecasting, 24(4), pp. 659–678, 2008, ISSN 0169-2070, Energy Forecasting.

[9] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Chapman & Hall, 1984.

[10] R. Brown, P. Kharouf, X. Feng, L. Piessens, and D. Nestor, Development of feed-forward network models to predict gas consumption, in *Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on*, volume 2, pp. 802–805 vol.2, June 1994.

[11] S. Crino and D. Brown, Global optimization with multivariate adaptive regression splines, Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 37(2), pp. 333–340, April 2007.

[12] H. Dawid, *Adaptive Learning by Genetic Algorithms Analytical Results and Applications to Economic Models*, Springer, Berlin Heidelberg, 1999, ISBN 978-3-642-62106-2.

[13] J. Deichmann, A. Eshghi, D. Haughton, S. Sayek, and N. Teebagy, Application of multiple adaptive regression splines MARS in direct response modeling, Journal of Interactive Marketing, 16(4), pp. 15–27, 2002, ISSN 1520-6653.

[14] O. F. Demirel, S. Zaim, A. Çalışkan, and P. Ozuyar, Forecasting natural gas consumption in Istanbul using neural networks and multivariate time series methods, Turkish Journal of Electrical Engineering & Computer Science, 20(5), 2012.

[15] Distribition Companies of Turkey, Turkish Union of Natural Gas Distribution Companies (GAZBİR), 2015, http://www.gazbir.org.tr/Default.aspx?lang=1.

[16] M. N. Eltony, Demand for natural gas in Kuwait: An empirical analysis using two econometric models, International Journal of Energy Research, 20(11), pp. 957–963, 1996, ISSN 1099-114X.

[17] European Commission, Energy Prices and Costs Report, Brussels, August, 2014, http://ec.europa.eu/energy/sites/ener/files...pdf.

[18] Eurostat, Regional Statistics Illustrated, European Commission, August, 2014, Brussels, http://ec.europa.eu/eurostat/cache/RSI/#?vis=city.statistics.

[19] L. Francis, Is MARS better than neural networks?, Martian Chronicles, 2011.

[20] J. H. Friedman, Multivariate adaptive regression splines, Annals of Statistics, 19(1), pp. 1–67, 03 1991.

[21] S. Gil and J. Deferrari, Generalized model of prediction of natural gas consumption, Journal of Energy Resources Technology, 126(2), pp. 90–98, 2004, ISSN 0195-0738.

[22] F. B. Görucu, Artificial neural network modeling for forecasting gas consumption, Energy Sources, 26(3), pp. 299–307, 2004.

[23] F. B. Görucu, Evaluation and forecasting of gas consumption by statistical analysis, Energy Sources, 26(3), pp. 267–276, 2004.

[24] F. Gümrah, D. Katircioglu, Y. Aykan, S. Okumus, and N. Kilinçer, Modeling of gas demand using degree-day concept: Case study for Ankara, Energy Sources, 23(2), 2001.

[25] P. C. Hansen, *Rank-deficient and discrite ill-posed problems: numerical aspects of linear inversion*, SIAM, Philadelphia, 1998.

[26] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York, 2009, ISBN 978-0-387-84857-0.

[27] S. Haykin, *Neural Networks and Learning Machines*, Pearson Education, Inc., Upper Saddle River, New Jersey 07458, 2008, ISBN 978-0-13-147139-9.

[28] J. H. Holland, *Adaptations in Natural Artificial Systems*, University of Michigan Press, Michigan, 1975.

[29] M. K. Hubbert, Energy from fossil fuels, Science, 109(2823), pp. 103–109, 21949.

[30] International Energy Agency, Energy Policies of IEA Countries, The United States Review, May 10, 2015, http://www.iea.org/bookshop/489-EnergyPoliciesofIEACountries-TheUnitedStates.

[31] International Energy Agency, Energy Policies of IEA Countries, The European Union Review 2014, May 10, 2015, http://www.iea.org/bookshop/486-EnergyPoliciesofIEACountries-TheEuropeanUnion.

[32] B. Jiang, C. Wenying, Y. Yuefeng, Z. Lemin, and D. Victor, The future of natural gas consumption in Beijing, Guangdong and Shanghai: An assessment utilizing MARKAL, Energy Policy, 36(9), pp. 3286–3299, 2008, ISSN 0301-4215.

[33] A. Juris, Public policy for the private sector, development of competitive natural gas markets in the United States, March 1998, 17728, The World Bank Group, Finance, Private Sector and Infrastructure Network.

[34] R. Karbauskaitė, G. Dzemyda, and V. Marcinkevičius, Dependence of Locally Linear Embedding on the Regularization Parameter, TOP: An official journal of the Spanish Society of Statistics and Operations Research, 18, 2(12), 354–376, Springer, ISSN:1134-5764, 2010, http://www.econbiz.de/Record/dependence-of-locally-linear-embedding-on-the-regularization-parameter-karbauskait%C4%97-rasa/10009885676.

[35] N. Karmarkar, A new polynomial-time algorithm for linear programming, Combinatorica, 4(4), pp. 373–395, 1984, ISSN 0209-9683.

[36] M. Kriner, *Survival Analysis with Multivariate Adaptive Regression Splines*, Phd thesis, Fakultat fur Mathematik, Informatik und Statistik, der Ludwig-Maximilians-Universitat Munchen, Fakultat fur Mathematik, Informatik und Statistik, der Ludwig-Maximilians-Universitat Munchen, 2007.

[37] R. Kızılaslan and B. Karlık, Comparison neural networks models for short term forecasting of natural gas consumption in istanbul, in *Applications of Digital Information and Web Technologies, 2008. ICADIWT 2008. First International Conference on the*, pp. 448–453, Aug 2008.

[38] G. J. Li Junchen, Dong Xiucheng, Dynamical modeling of natural gas consumption in China, Natural Gas Industry, 30(4), 127, 2010.

[39] L.-M. Liu and M.-W. Lin, Forecasting residential consumption of natural gas using monthly and quarterly time series, International Journal of Forecasting, 7(1), pp. 3–16, 1991, ISSN 0169-2070.

[40] Y. Ma and Y. Li, Analysis of the supply-demand status of china's natural gas to 2020, Petroleum Science, 7(1), pp. 132–135, 2010, ISSN 1672-5107.

[41] K. B. Medlock, A. M. Jaffe, and M. O'Sullivan, The global gas market, LNG exports and the shifting us geopolitical presence, Energy Strategy Reviews, 5(0), pp. 14–25, 2014, ISSN 2211-467X, US energy independence: Present and emerging issues.

[42] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, John Wiley and Sons., New York, 2013, ISBN 978-1118539712.

[43] MOSEK Software, Copenhagen: MOSEK ApS, 1997, August 2014, www.mosek.com.

[44] P. Musilek, E. Pelikán, T. Brabec, and M. Simunek, Recurrent neural network based gating for natural gas load prediction system, in *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, pp. 3736–3741, 2006.

[45] R. Neuneier and H. Zimmermann, How to train neural networks, in G. Montavon, G. Orr, and K.-R. Müller, editors, *Neural Networks: Tricks of the Trade*, volume 7700 of *Lecture Notes in Computer Science*, pp. 369–418, Springer Berlin Heidelberg, 2012, ISBN 978-3-642-35288-1.

[46] OAO Gazprom, Gazprom Corporate Website, August 26, 2014, http://www.gazprom.com/f/posts/60/660385/gazprom-annual-report-2013-en.pdf.

[47] OECD/IEA, Energy Balances of OECD Countries, Report, 2014, http://wds.iea.org/wds/pdf/OECDBALDocumentation.pdf .

[48] OECD/IEA, Energy statistics of Non-OECD Countries, Report, 2014, http://www.oecd-ilibrary.org/energy-statistics-of-non-oecd-countries-2014.

[49] A. Özmen, *Robust Conic Quadratic Programming Applied to Quality Improvement – A Robustification of CMARS*, MSc Thesis, Scientific Computing, Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey, September 2010, Supervisor: Gerhard-Wilhelm Weber.

[50] A. Özmen, *Advences in Robust Identification of Spline Models and Networks by Robust Conic Optimization, with Applications to Different Sectors*, Phd thesis, Scientific Computing, Institute of Applied Mathematics, Middle East Technical University, Scientific Computing, Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey, 2015.

[51] A. Özmen, İ. Batmaz and G.-W. Weber, Precipitation modeling by polyhedral RCMARS and comparison with MARS and CMARS, Environ Model Assess, 19(2), p. 425–435, 2014.

[52] A. Özmen and G.-W. Weber, A new robust optimization tool applied on financial data, Pacific Journal of Optimization, 9(3), pp. 535–552, 2013.

[53] A. Özmen and G.-W. Weber, RMARS: Robustification of multivariate adaptive regression spline under polyhedral uncertainty, Journal of Computational and Applied Mathematics, 259, Part B(0), pp. 914–924, 2014, ISSN 0377-0427, Recent Advances in Applied and Computational Mathematics: ICACM-IAM-METU On the occasion of 10th anniversary of the foundation of Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey.

[54] A. Özmen, G.-W. Weber, Z. Çavuşoğlu, and Ö. Defterli, The new Robust Conic GPLM method with an application to finance, Prediction of credit default, Journal of Global Optimization, 56(2), pp. 233–249, June 2013.

[55] A. Özmen, G. W. Weber, İ. Batmaz, and E. Kropat, RCMARS: Robustification of CMARS with different scenarios under polyhedral uncertainty set, Communications in Nonlinear Science and Numerical Simulation, 16(12), pp. 4780–4787, 2011, ISSN 1007-5704, sI:Complex Systems and Chaos with Fractionality, Discontinuity, and Nonlinearity.

[56] C. R. Reeves and J. E. Rowe, *Genetic Algorithms - Principles and Perspectives, A Guide to GA Theory*, Kluwer Academic Publishers, New York, Boston, Dordrecht, London, Moscow, 2003, ISBN 1-4020-7240-6.

[57] C. Roos, T. Terlaky, and J. P. Vial, *Interior Point Methods for Linear Optimization*, Springer US, 2005.

[58] C. Rui, W. Jian, W. Li, Y. Ningjie, and Z. Pengyan, The forecasting of china natural gas consumption based on genetic algorithm, in *INC, IMS and IDC, 2009. NCM '09. Fifth International Joint Conference on*, pp. 1436–1439, Aug 2009.

[59] C. Rui, W. Jian, W. Li, Y. Ningjie, and Z. Pengyan, The Forecasting of China Natural Gas Consumption Based on Genetic Algorithm, Networked Computing and Advanced Information Management, International Conference on, 0, pp. 1436–1439, 2009.

[60] P. Sasieni, Generalized additive models, Statistics in Medicine, 11(7), pp. 981–982, 1992, ISSN 1097-0258.

[61] G. Seber and C. Wild, *Nonlinear Regression*, John Wiley & Sons, Inc., New Jersey, 2005.

[62] P. Sephton, Forecasting recessions: Can we do better on MARS?, Review, Federal Reserve Bank of St. Louis, pp. 39-49, 2001.

[63] J. K. Shim, Pooling cross section and time series data in the estimation of regional demand and supply functions, Journal of Urban Economics, 11(2), pp. 229–241, 1982, ISSN 0094-1190.

[64] J. Siemek, S. Nagy, and S. Rychlicki, Estimation of natural-gas consumption in poland based on the logistic-curve interpretation, Applied Energy, 75(1–2), pp. 1–7, 2003, ISSN 0306-2619, energex 2002 - Oil and Gas - Topic III and Nuclear Energy - Topic IV.

[65] E. F. Sánchez-Úbeda and A. Berzosa, Modeling and forecasting industrial end-use natural gas consumption, Energy Economics, 29(4), pp. 710–742, 2007, ISSN 0140-9883, Modeling of Industrial Energy Consumption.

[66] B. Soldo, Forecasting natural gas consumption, Applied Energy, 92(0), pp. 26–37, 2012, ISSN 0306-2619.

[67] SPM Software for MARS by Salford-Systems, August 24, 2014, http://www.salfordsystems.com.

[68] SPM Users Guide, Introducing MARS, August 24, 2014, http://media.salford-systems.com/pdf/spm7/IntroMARS.pdf.

[69] T. Statistical Institute, Adrese Dayalı Nüfus Kayıt Sistemi (ADNKS) Veri Tabanı, August 28, 2014, Ankara, http://tuikapp.tuik.gov.tr/adnksdagitapp/adnks.zul.

[70] J. Suykens, P. Lemmerling, W. Favoreel, B. de Moor, M. Crepel, and P. Briol, Modelling the Belgian gas consumption using neural networks, Neural Processing Letters, 4(3), pp. 157–166, 1996, ISSN 1370-4621.

[71] K. Talus, United States natural gas markets, contracts and risks: What lessons for the European Union and Asia-Pacific natural gas markets?, Energy Policy, 74(0), pp. 28–34, 2014, ISSN 0301-4215.

[72] P. Taylan, G.-W. Weber, and A. Beck, New approaches to regression by generalized additive models and continuous optimization for modern applications in finance, science and techology, Optimization, 56(5-6), 2007, ISSN 02331934.

[73] P. Taylan, G.-W. Weber, and F. Yerlikaya Özkurt, Continuous optimization applied in MARS for modern applications in finance, science and technology, Preprint 2007-19, Institute of Applied Mathematics, Middle East Technical University, 06800 Ankara, Turkey, October 2007.

[74] P. Taylan, G.-W. Weber, and F. Yerlikaya Ozkurt, A new approach to Multivariate Adaptive Regression Splines by using Tikhonov Regularization and Continuous Optimization, TOP, 18(2), pp. 377–395, 2010, ISSN 1134-5764.

[75] M. Thaler, I. Grabec, and A. Poredoš, Prediction of energy consumption and risk of excess demand in a distribution system, Physica A: Statistical Mechanics and its Applications, 355(1), pp. 46–53, 2005, ISSN 0378-4371.

[76] M. Toksari, Predicting the natural gas demand based on economic indicators: Case of turkey, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 32(6), pp. 559–566, 2010.

[77] Z. Tonkovic, M. Zekic-Sušac, and M. Somolanji, Predicting natural gas consumption by neural networks, Tehnicki Vjesnik, 16(3), pp. 51–61, 2009.

[78] S. Tufféry, *Data Mining and Statistics for Decision Making*, John Wiley & Sons, Ltd, 2011, ISBN 9780470979174.

[79] R. D. Veaux, D. Psichogios, and L. Ungar, A comparison of two nonparametric estimation schemes: MARS and Neural Networks, Computers & Chemical Engineering, 17(8), pp. 819–837, 1993, ISSN 0098-1354, an International Journal of Computer Applications in Chemical Engineering.

[80] US Energy Information Administration, Natural Gas, August 28, 2014, https://www.eia.gov/pub/oilgas/naturalgas/../ngpipeline/process.html.

[81] US Department of Transportation, Natural Gas Pipeline Systems, Pipeline Safety Stakeholder Communications, 2014.

[82] J. Vondrácek, E. Pelikán, O. Konár, J. Cermáková, K. Eben, M. Maly, and M. Brabec, A statistical model for the estimation of natural gas consumption, Applied Energy, 85(5), pp. 362–370, 2008, ISSN 0306-2619.

[83] H. M. Wadsworth, *Handbook of Statistical Methods for Engineers and Scientists*, R. R. Donnelley & Sons Company, Chicago, 1998.

[84] G. Wahba, *Spline models for observational data*, volume 59, SIAM, 1990.

[85] R. Walpole, R. Myers, S. Myers, and K. Ye, *Probability and Statistics for Engineers and Scientists*, Pearson Education Inc., Boston, 2013, ISBN 9781292023922.

[86] D. Wang, *Model Selection and Estimation in Generalized Additive Models and Generalized Additive Mixed Models*, Phd thesis, Statistics, Graduate Faculty, North Carolina State University, Statistics, Graduate Faculty, North Carolina State University, Raleigh, North Carolina, the USA, 2013.

[87] G.-W. Weber, Z. Çavuşoğlu, and A. Özmen, Predicting default probabilities in emerging markets by new conic generalized partial linear models and their optimization, Optimization, 61(4), pp. 443–457, 2012.

[88] G.-W. Weber, İ. Batmaz, G. Köksal, P. Taylan, and F. Yerlikaya Özkurt, CMARS: A new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimisation, Preprint 2009-16, Institute of Applied Mathematics, Middle East Technical University, 06800 Ankara, Turkey, September 2009.

[89] G.-W. Weber, A. Özmen and Y. Yilmaz, A Natural Gas Consumption Forecasting Model for Residential User, conveyed at 9th International Summer School, AACIMP-2014, National University of Technology of the Ukraine, Kyiv, Ukraine, August 1-15, 2014.

[90] F. Yerlikaya Ozkurt, İ. Batmaz, and G.-W. Weber, A review of conic multivariate adaptive regression splines CMARS: A powerful tool for predictive data mining, Preprint 2012-15, Institute of Applied Mathematics, Middle East Technical University, 06800 Ankara, Turkey, 2012.

[91] F. Yerlikaya-Ozkurt, C. Vardar-Acar, Y. Yolcu-Okur, and G.-W. Weber, Estimation of the hurst parameter for fractional brownian motion using the CMARS method, Journal of Computational and Applied Mathematics, 259, Part B, pp. 843–850, 2014, ISSN 0377-0427, Recent Advances in Applied and Computational Mathematics: ICACM-IAM-METU: On the occasion of 10th anniversary of the foundation of Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey.

[92] M. H. Yıldırım, *Electricity Market Modeling Using Stochastic and Robust Optimization*, Phd thesis, Scientific Computing, Institute of Applied Mathematics, Middle East Technical University, Scientific Computing, Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey, September 2014.

[93] Y. Yılmaz, R. Kurz, A. Özmen, and G.-W. Weber, A new algorithm for scheduling condition-based maintenance of gas turbines, GT2015, to appear in proceedings of the ASME Turbo Expo 2015: Turbine Technical Conference and Exposition, 2015.

[94] S.-H. Yoo, H.-J. Lim, and S.-J. Kwak, Estimating the residential demand function for natural gas in Seoul with correction for sample selection bias, Applied Energy, 86(4), pp. 460–465, 2009, ISSN 0306-2619.

[95] H. Zareipour, K. Bhattacharya, and C. Canizares, Forecasting the hourly Ontario energy price by Multivariate Adaptive Regression Splines, in *Power Engineering Society General Meeting, 2006. IEEE*, 2006.

APPENDIX A

Annual NG Consumption of Some Countries

COUNTRIES	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	Profile from 2003 to 2013	Change from 2012 to 2013
United States	540.8	526.4	511.1	524	545.6	570.8	584	603.6	648.5	681.2	687.6		2.40%
Russian Federation	379.5	389.3	394.1	415	422	416	389.7	414.2	424.6	416.3	413.5		-0.40%
Iran	85	98.7	102.8	112	125.5	134.8	143.2	152.9	162.4	161.5	162.2		0.70%
China	33.9	39.7	46.8	56.1	70.5	81.3	89.5	106.9	130.5	146.3	161.6		10.80%
Japan	79.8	77	78.6	83.7	90.2	93.7	87.4	94.5	105.5	116.9	116.9		0.20%
Germany	85.5	85.9	86.2	87.2	82.9	81.2	78	83.3	74.5	78.4	83.6		7.00%
United Kingdom	95.3	97.4	94.9	90	91	93.4	87	94.2	78.1	73.7	73.1		-0.60%
Italy	71.2	73.9	79.1	77.4	77.8	77.8	71.5	76.2	71.4	68.7	64.2		-6.20%
Turkey	20.9	22.1	26.9	30.5	36.1	37.5	35.7	39	44.7	45.3	45.6		1.10%
Ukraine	69	68.5	69	67	63.2	60	46.8	52.2	53.7	49.5	45		-8.90%
France	43.2	45.1	44.8	43.7	42.4	43.8	41.8	46.9	40.5	42.2	42.8		1.70%
Netherlands	40	40.9	39.3	38.1	37	38.6	38.9	43.6	38.1	36.4	37.1		2.00%
Spain	23.6	27.4	32.4	33.7	35.1	38.6	34.6	34.6	32.2	31.3	29		-7.20%
Belgium	16	16.2	16.4	16.7	16.6	16.5	16.8	18.8	16.6	16.9	16.8		-0.40%
Poland	12.5	13.2	13.6	13.7	13.8	14.9	14.4	15.5	15.7	16.6	16.7		1.10%
Romania	18.3	17.5	17.6	18.1	16.1	15.9	13.3	13.6	13.9	13.5	12.5		-7.50%
Hungary	13.2	13.1	12.2	10.8	13.1	14	12.7	12.6	10.3	10.2	8.6		-16.20%
Austria	9.4	9.5	10	9.4	8.9	9.5	9.3	10.1	9.5	9	8.5		-5.90%
Czech Republic	8.7	9.1	9.5	9.3	8.7	8.7	8.2	9.3	8.4	8.2	8.4		3.20%
Slovakia	6.3	6.1	6.6	6	5.7	5.7	4.9	5.6	5.2	4.9	5.4		11.50%
Republic of Ireland	4.1	4.1	3.9	4.4	4.8	5	4.7	5.2	4.6	4.5	4.5		-0.10%
Norway	4.3	4.6	4.5	4.4	4.3	4.3	4.1	4.1	4.3	4.4	4.4		1.40%
Portugal	3	3.7	4.2	4.1	4.3	4.7	4.7	5.1	5.2	4.5	4.1		-9.60%
Denmark	5.2	5.2	5	5.1	4.6	4.6	4.4	5	4.2	3.9	3.7		-4.10%
Greece	2.4	2.7	2.7	3.1	3.7	3.9	3.3	3.6	4.4	4.1	3.6		-11.50%
Switzerland	2.9	3	3.1	3	2.9	3.1	3	3.3	3	3.3	3.6		12.40%
Finland	4.5	4.3	4	4.2	3.9	4	3.6	3.9	3.5	3.1	2.8		-6.70%
Lithuania	3.1	3.1	3.3	3.2	3.6	3.2	2.7	3.1	3.4	3.3	2.7		-18.30%
Bulgaria	2.8	2.8	3.1	3.2	3.2	3.2	2.3	2.6	2.9	2.7	2.6		-3.10%
Sweden	0.8	0.8	0.8	0.9	1	0.9	1.1	1.6	1.3	1.1	1.1		-1.80%

Figure A.1: Annual NG Demand Figures in Billion Cubic Meters (BCM) per year [7].

APPENDIX B

Map of US Natural Gas Transmssion and Distribution Pipelines

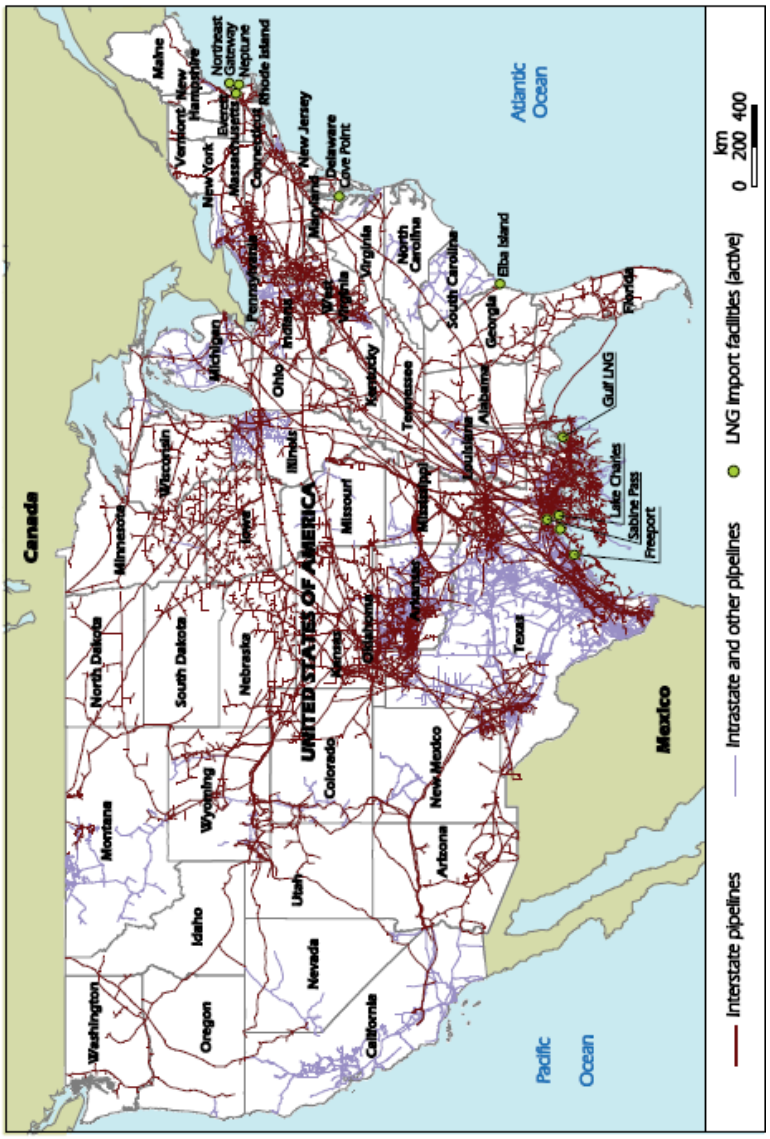
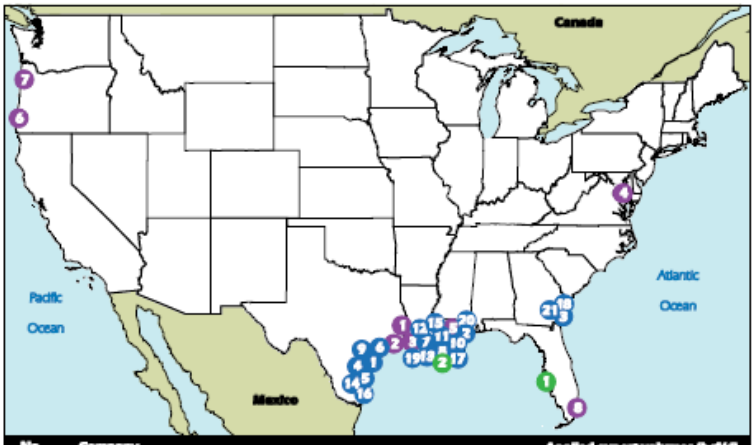


Figure B.1: US Natural Gas Pipelines and LNG Terminals [30].

APPENDIX C

Map of US LNG Terminals



No.	Company	Applied export volumes (bcf/d)
	Applications to export LNG to both FTA and non-FTA countries approved by DOE	
1	Sabine Pass	2.2
2	Freeport LNG	1.8
3	Lake Charles	2
4	Cove Point LNG	0.77
5	Cameron LNG	1.7
6	Jordan Cove Energy	0.8
7	Oregon LNG	1.25
8	Carib Energy	0.04
	Applications to export LNG to FTA countries approved by DOE but under DOE review for non-FTA countries	
1	Gulf Coast LNG	2.8
2	Gulf LNG	1.5
3	Southern LNG	0.5
4	Lavaca Bay LNG	1.38
5	Golden Pass	2,0
6	Corpus Christi LNG	2.1
7	CE FLNG	1.07
8	Waller LNG Services	0.19
9	Pangea LNG	1.09
10	Magnolia LNG	1.08
11	Gasfin Development	0,2
12	Main Pass - Freeport-McMoRan	3.22
13	Venture Global	0.67
14	Eos LNG & Barca LNG	3.2
15	Delfin LNG	1.8
16	Texas LNG	0.27
17	Louisiana LNG Energy	0.28
18	Trunkline LNG	2.0
19	Freeport McMoRan	3.22
20	Sabine Pass	1.38
21	Venture Global	0.67
	Applications pending approval to export LNG to FTA countries and under review for export to non-FTA countries	
1	Strom	0.02
2	SCT&E LNG	1.6

Figure C.1: US LNG Terminals [30].

APPENDIX D

Present US Market Structure

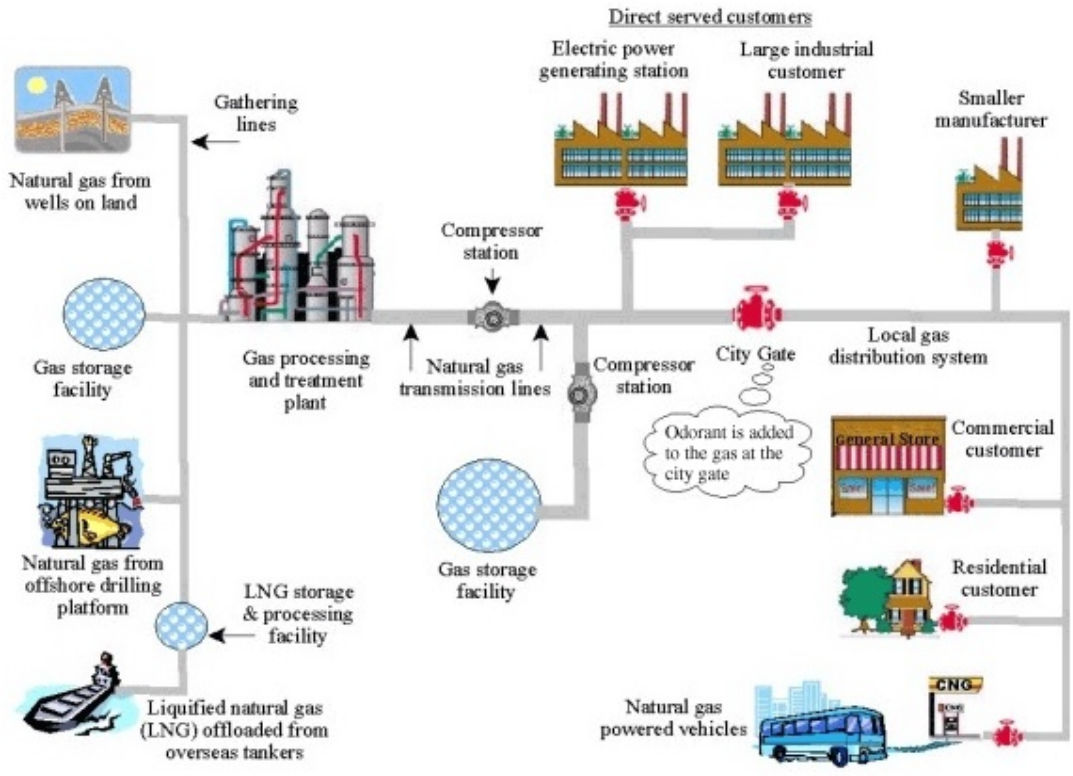


Figure D.1: NG Market Structure Present in the US [81].

APPENDIX E

Map of EU Natural Gas Transmssion and Distribution Pipelines

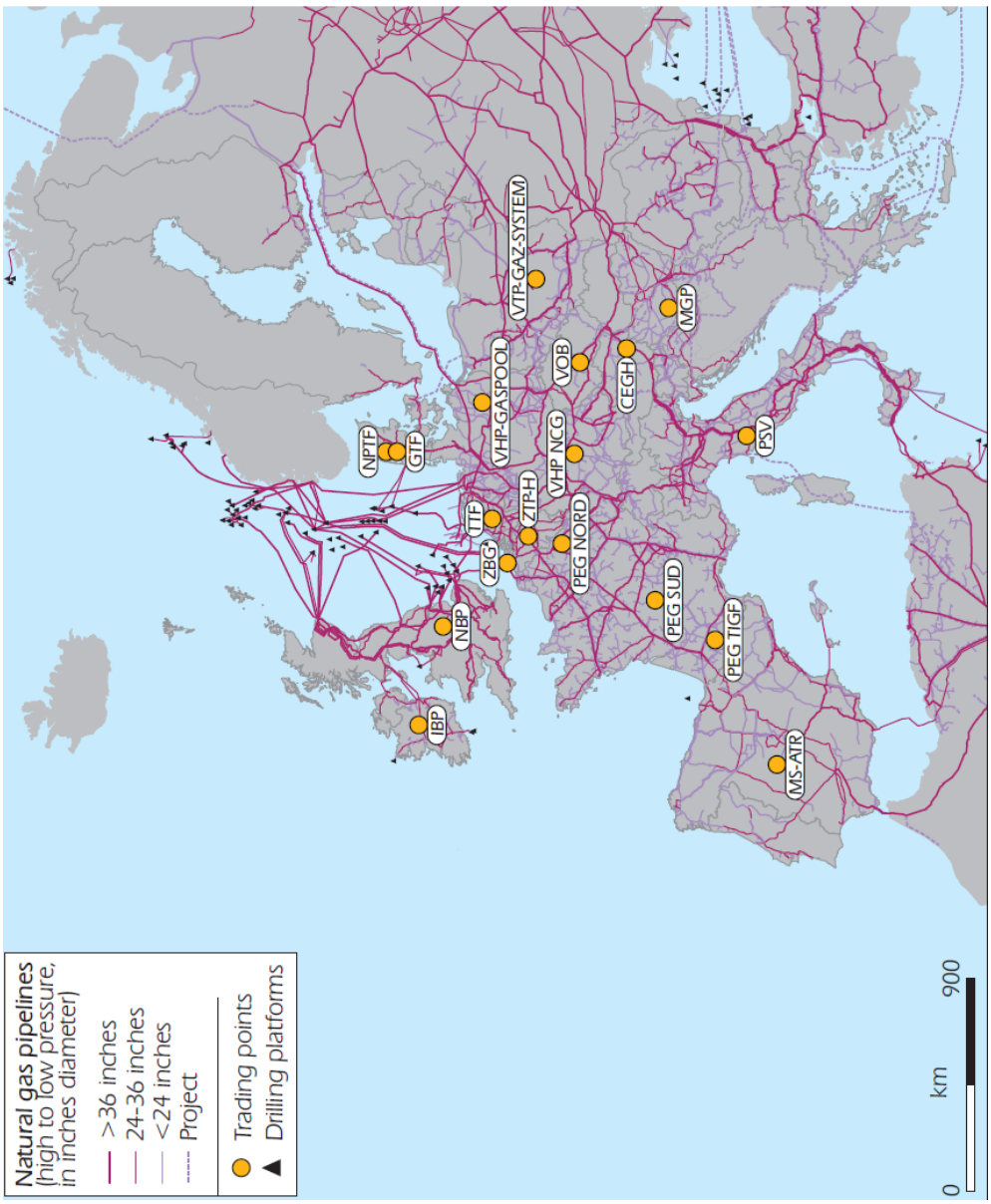


Figure E.1: EU Natural Gas Pipeline Network [31].

APPENDIX F

BOTAŞ Transmission System

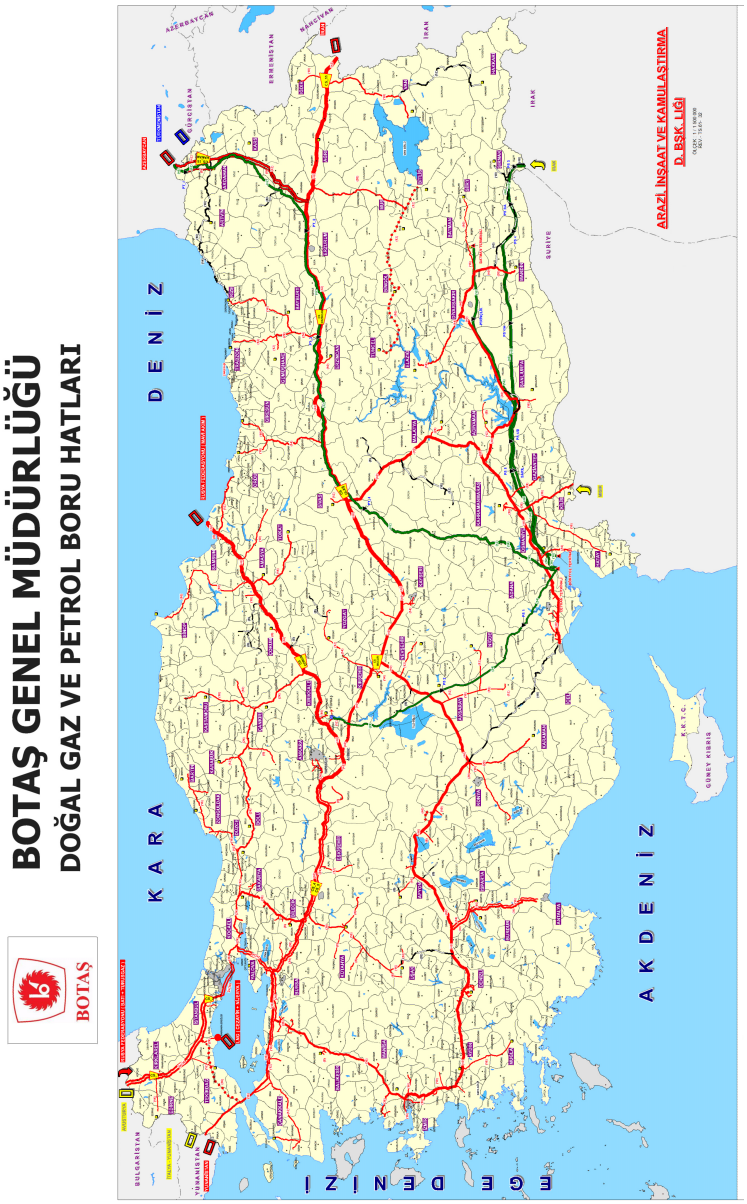


Figure F.1: BOTAŞ Transmission System Infrastructure Map [6].

APPENDIX G

Data Mining Methods Historical Progress

1875	Francis Galton's linear regression
1888	Francis Galton's correlation
1896	Karl Pearson's formula for the correlation coefficient
1900	Karl Pearson's w2
1933	Harold Hotelling's factor analysis
1934	Chester Bliss's probit model
1936	Discriminant analysis, developed by Ronald A. Fisher and Prasanta Chandra Mahalanobis
1936	Harold Hotelling's canonical correlation analysis
1941	Guttman's correspondence analysis
1943	Formal neuron invented by the neurophysiologist Warren McCulloch and the logician Walter Pitts
1944	Joseph Berkson's logistic regression
1958	Frank Rosenblatt's perceptron
1960	Appearance of the concept of exploratory data analysis in France (Jean-Paul Benz_ecri) and the USA (John Wilder Tukey)
1962	Jean-Paul Benz_ecri's correspondence analysis
1964	AIDdecision tree (precursor ofCHAID) invented by J.P. Sonquist and J.-A. Morgan
1965	E. W. Forgy's moving centres method
1967	J. MacQueen's k-means method
1970	Ridge regression proposed by Arthur E. Hoerl and Robert W. Kennard
1971	Edwin Diday's dynamic cloud method
1972	Generalized linearmodel formulated by JohnA.Nelder and RobertW.Wedderburn
1972	David Cox's proportional hazards regression model
1975	John Holland's genetic algorithms
1975	Gilbert Saporta's DISQUAL classification method
1979	Bootstrap method proposed by Bradley Efron
1980	CHAID decision tree developed by Gordon V. Kass
1982	Teuvo Kohonen's self-organizing maps (Kohonen networks)
1983	Herman and Svante Wold's PLS regression
1984	CART tree proposed by Leo Breiman, Jerome H. Friedman, R.A. Olshen and Charles J. Stone
1986	Multilayer perceptron invented byDavid E.Rumelhart and James L.McClelland
1990	Generalized additive model proposed by Trevor Hastie and Robert Tibshirani
1990	First appearance of the data mining concept
1991	Jerome H. Friedman's multivariate adaptive regression splines (MARS)
1993	J. Ross Quinlan's C4.5 tree
1993	Apriori algorithm proposed by R. Agrawal et al. for detecting association rules
1995	Vladimir Vapnik's learning theory and support vector machines
1995	Robert Tibshirani's lasso method of linear regression
1996	DBSCAN clustering algorithm proposed by M. Ester, H.-P. Kriegel, J. Sander and X. Xu
1996	Leo Breiman's bagging method
1996	Yoav Freund's and Robert E. Shapire's boosting method
1998	Leo Breiman's arcing method
2000	PLS logistic regression formulated by Michel Tenenhaus
2001	Leo Breiman's random forests
2005	Elastic net linear regression proposed by Zou and Hastie
2007	Grouped lasso method proposed by Yuan and Lin

i

Figure G.1: Statistical and Data Mining Methods Historical Timeline [78].