Data-based Dynamic Modeling for Refinery Optimization

Vahedi, Vahid; Jørgensen, Sten Bay

Publication date:
2001

Document Version
Publisher's PDF, also known as Version of record

Data-based Dynamic Modeling for Refinery Optimization

Ph.D. Thesis

Vahid Vahedi

Center for Computer Aided Process Engineering
CAPEC
Department of Chemical Engineering
Technical University of Denmark
2800 Lyngby, Denmark

ISBN 87-90142-78-0
Printed by Bookpartner, Nørhaven Digital, Copenhagen, Denmark

.... and if one does not continue correcting the book for the rest of one's life, it is because the same ironhard discipline, which is required to begin it, is also necessary to complete it.

Gabriel García Márques

Preface

This thesis is written as partial fulfillment of requirements for the Ph.D. degree. The project has been carried out as an industrial research project and in collaboration between Statoil Refinery in Denmark and Department of Chemical Engineering at Technical University of Denmark (DTU).
The supervisory group of the project are :

Professor Sten Bay Jørgensen,
Center for Computer Aided Process Engineering (CAPEC)
Department of Chemical Engineering, DTU,

Senior Engineer, M. Sc. Lars Erik Ebbesen,
Statoil Refinery, Denmark

Associated professor Carsten Aamand,
IVC-SEP Engineering Research Center
Department of Chemical Engineering, DTU,

This project has been carried out as industrial research project, and financed by the Danish Academy of Technical Science (ATV) and Statoil A/S Denmark.

April 2000

Vahid Vahedi

Acknowledgments

First of all, I would like to thank professor Sten Bay Jørgensen for his professional guidance, and supervision. I wish to thank the other members of the supervisory group, Senior Engineer Lars Erik Ebbesen, and associated professor Carsten Aamand for their guidance, and useful discussions.

A great part of the project was carried out at Statoil Refinery in Kalundborg. Many of the Statoil Refinery employees were involved by their discussions, and guidance. I would like to acknowledge their helpful support.

The last part of the project was carried out during my time at Danisco Sugar and Sweeteners Development Center (DSSD). I would like to thanks for their great support and encouragement.

Special thanks goes to the members of Center for Computer Aided Process Engineering (CAPEC). I would like to thank Lars Gregersen for his inspiring ideas and many helpful comments along the way, and John Bagtrup Jørgensen for his helpful effort in optimization software.

I wish to thank all my family and friends for their encouragement. Last but not least, I wish to thank my wife Minoo for her tremendous encouragement, and understanding.

Summary

This thesis deals with development of data-based dynamic models for refinery processes by using the methods in Process Chemometrics. The models are developed in order to predict the qualities of intermediate product streams in gasoline processing area. Multivariate predictive models are developed for prediction of Research Octane Number (RON), Reid Vapor Pressure (RVP), and concentration of aromatic compounds, e.g. benzene, in the product streams of catalytic reformer and isomerization units sent to blendstock tanks which are used for gasoline blending.

The chemometric models are applied in a multiperiod nonlinear optimization problem for the gasoline blending in order to provide prediction of previous, present and future values of the qualities in blendstocks tanks based one the variation of the upstream process. The optimization goal is to produce the required amount of high quality final gasoline products at the required time and to minimize the production and inventory costs. Solution of the optimization problem determines the optimum value of quality and amount of the blend components used in the gasoline blending in such a way that the needed quantities of the different final gasoline products can be produced on-time with the desired specifications with minimum operation and inventory cost.

In this work the available historical data is used to develop models based on information and knowledge obtained from the data. This is *data-based* modeling and the purpose is to predict quality variables which are expensive or difficult to measure as frequently as it is desired for control and optimization applications.

The general principle for data-based predictive modeling in this work is the methods in *Process Chemometrics*. These methods are divided in four general categories according to the linear, nonlinear, static, and dynamic characteristics of the system under study.
A brief review of the methods used in this work for development of data-based dynamic model is presented. This review include the essence of process chemometrics in order to be able to discuss the multivariate modeling techniques applied for development of process models in the subsequent chapters of this thesis.
In the class of static linear methods Principal Component Analysis (PCA), Principal Component Regression (PCR), and Partial Least Squares Regression (PLS) are discussed. PCA is used in data assessment, dimensional reduction through extracting the latent variables and applied mostly for process monitoring. PLS and PCR are used for developing input-output regression models. In the class of static nonlinear approaches Artificial Neural Networks (ANNs) exhibit a strong ability to nonlinear functional approximation. Nonlinear PLS regression in which nonlinear function is defined for the inner relationship of the PLS is another approach in this class of chemometrics methods. A Nonlinear Principal Component Analysis (NLPCA) model is developed based on Input Training Neural Networks (ITNN) which is used for data rectification. The method in the class of dynamic linear includes the methods in System Identification. System Identification deals with knowledge based predictive modeling using linear time series regression. The linear methods include ARX and ARMAX (Auto Regressive Moving Average with Exogenous input), which are linear models based on

parametric input output representations. A short description for the dynamic nonlinear methods is presented, in which the time-series type of model can be integrated in a nonlinear PLS model.

Different criteria in model validation is discussed in which two different reference models as average-model and zero-model are presented in order to assess the ability of prediction of the developed chemometric model. The concept of informative data set and persistence of excitation is presented, and the issue concerning the impact of closed-loop control on persistence of excitation of input is discussed

A description of the preliminary steps in the model development work in this thesis is presented. These preliminary steps concerns mainly with definition of the system limit, description of the output and selected input variables, assessment of data, data scaling and sampling, and description of data treatment. The outliers are found first by visualization of data in respective plots, and then a PCA model is performed in order to assess the representability of the data, discover any collinearity in the selected inputs, detection of distinct clusters of data due to different operation of the plant.

It has been observed that the quality variables are dependent on the previous value of themselves and the input variables. This means that a dynamic, time-series modeling approach is a suitable choice in this application. The method applied for model development in this work is ARX (Auto Regressive with Exogenous input) type of model in System Identification, in which Partial Least Squares Regression (PLS) method is used for its parameter estimation.
The advantage of developing a linear time-series model by a PLS regression is that the variation and structure of the output variable is directly used in PCA decomposition of the input variables. Applying PLS will use the strength of PCA in dimensional reduction of the data set and hence more effective modeling of the output.
Since the quality variables are either expensive or time consuming to measure, there are only a limited number of them available. A solution for the problem of missing output data is proposed by a suitable structure of ARX model.

An optimization model for gasoline processing area of the refinery has been developed. The model concerns prediction of the qualities of the products from reforming and isomerization processes and gasoline blending over multiple periods.
A decomposition of this model yield in a multi-period optimization model for gasoline blending unit. The objective is to minimize the cost of operation for gasoline production such that the quality and quantity demands are satisfied. The optimization model assumes that the qualities of final gasoline product is a linear function of the qualities of the blend component streams sent to the blending unit.
The objective function is a cost function which represent the cost of operation for production of gasoline products plus the inventory cost. This objective function is minimized subject to a set of constraints which represent the demands for quality and quantity of final gasoline products. The optimum solution yields in quality and quantity of the blend components needed to produced the desired products.
A case study is considered and the results are discussed. The results of testing the model during the case study indicate that the solution is a feasible, local optimum solution, and there is good agreement with the demands.

Resumé

Nærværende afhandling handler om udvikling af data-baserede dynamiske modeller for raffinaderi processer ved anvendelse af metoder i Proces Kemometri. Modellerne er udviklet til forudsigelse af kvalitetsvariable for mellem-produkt strømme i benzin-produktion sektionen i et olie raffinaderi. Multi-variable prædiktive modeller er udviklet til forudsigelse af oktan tal Research Octane Number (RON), damp-trykket Reid Vapor Pressure (RVP), og koncentrationen af aromatisk komponenter (benzen) i produkt-strømme fra en katalytisk reformer, og et isomerizerings anlæg. Disse strømme sendes til mellem-produkt tank og bruges som komponenter i benzin blandingen.

Kemometriske modeller bruges i et multi-periode ikke-lineært optimeringsproblem i benzin-blandingen til forudsigelse af foregående, nuværende, og fremtidige kvaliteter af indholdet af mellem-produkt tankerne baseret på proces variationen i reformer og isomerizering anlæg. Formålet med optimeringen er at producere de ønskede mængder af høj kvalitets benzinprodukter på en bestemt tid og samtidig minimere produktions og lagrings omkostninger.

Historiske data fra processen udnyttes til at udvikle modeller baseret på informationen gemt i data. Dette kaldes data-baseret modellering og formålet er at forudsige de variabler som er kostbare eller tidskrævende til at måle.

De generelle principper for data-baseret prædiktiv modellering der anvendes i dette projekt er metoder i Proces Kemometri. Metoderne i proces kemometri er opdelt i fire forskellige kategorier efter lineære og ikke-lineære såvel som statiske og dynamiske egenskaber af systemet.
En kortfattet gennemgang af metoder der anvendes til udvikling af modellerne i dette projekt er præsenteret. Denne gennemgang omfatter de væsentlige emner i proces kemometri og formålet er at kunne diskutere de anvendte fremgangsmåder i modeludviklingen i de efterfølgende kapitler i denne afhandling.
Blandt statiske lineære metoder er "Principal Component Analysis (PCA)", "Principal Component Regression (PCR), og "Partial Least Squares Regression (PLS)" diskuteret. PCA bruges til kvalitetsvurdering af data, og dimension reduktion af data og anvendes hovedsageligt til visualisering af processens opførsel. PLS og PCR anvendes til at udvikle input-output regressions modeller.
I den modelklasse der omfatter statiske og ikke-lineære metoder, anvendes "Artificial Neural Networks (ANNs)" der omfatter en god evne til approksimation af ikke-lineære funktioner. Ikke-lineær PLS regressions modeller hører også til denne klasse af kemometriske metoder, idet relationen mellem score matricer i input og output er defineret som en ikke-lineær funktion.
En ikke-lineær PCA, "Nonlinear Principal Component Analysis (NLPCA)" model er udviklet baseret på "Input Training Neural Networks (ITNN)", som kan anvendes til at rektificere data.
Metoder i "System Identification" anvendes til dynamiske lineære modeller i proces kemometri som er specielt egnet til at udvikle prædiktions modeller for dynamiske systemer. System Identifikation omhandler data-baseret prædiktiv modellering ved brug af lineær tids-serie

regression metoder, såsom ARX og ARMAX (Auto Regressive Moving Average with Exogenous input), som er lineær, tids-serie regressions modeller baseret på en parametrisk input-output repræsentation. Dynamiske ikke-lineære metoder er beskrevet kort. Denne type model opnås ved at integrere tids-serie modeller i en for eksempel ikke-lineære PLS regression modeller.

Forskellige kriterier til model-validering er diskuteret og to reference-modeller er defineret, en gennemsnit-model og en nul-model, for at vurdere prædiktionsevnen hos de udviklede kemometriske modeller. Koncepterne i "persistence of excitation" og "informative data" er præsenteret og virkning af lukket-sløjfe regulering på "persistence of excitation" er diskuteret.

De forskellige trin i udvikling af kemometriske modeller i dette projekt er beskrevet. Disse består hovedsagelig af system definition og begrænsning, beskrivelse af input og output data, data skalering, og data behandling. De data som falder langt udenfor normale data områder, de såkaldte outliers, er fundet ved hjælp af først en visualisering af data og dernæst i en PCA analyse som også bruges til at afsløre lineære sammenhænge mellem input variable, og grupperinger i data som kan være tegn på forskellige typer af proces operation.

Det er observeret at output variablene er relateret til de tidligere værdier af variablene selv og input variblene. Dette bevirker at en dynamisk tids-serie modellering kan anvendes. Kemometriske modeller er udviklet ved brug af ARX (Auto Regressive with Exogenous input) i system idenfikation, idet PLS modellen er anvendt til parameter estimationen.
Fordelen ved at bruge PLS regressionen i en tids-serie model er at informationen i output bruges direkte i PCA dekomponering af input variable, samt at PCA modellens evne til dimensionsreduktion bruges hvorved effektiv modellering kan opnås.
På grund af at måling af kvalitetsvariablene er både kostbar og tidskrævende er der sparsomme mængde af output data til rådighed. En måde at behandle problemet med manglende output data er at anvende en passende struktur af ARX modellen.

En optimerings model er udviklet til benzin produktions sektionen på raffinaderiet. Modellen dækker produktionen af både blandingskomponenter og de færdige benzinprodukter over multiple tids-perioder.
Dekomponering af denne optimerings model resulterer i en multi-periode optimerings model til benzin-blandingen. Formålet er at minimere operationsomkostningerne for benzin produktionen således at mængde og kvalitets kravene overholdes. I optimerings modellen er det antaget at selve blandingen er en lineær proces, og kvaliteten i det færdige benzin produkt er en lineær funktion af kvaliteten i blandings-komponent strømme som sendes til benzin blander enheden.
Mål-funktionen er en "Variable Cost" funktion som repræsenterer processens produktions og lagringsomkostninger. Denne mål-funktion minimeres. Begrænsningerne er specifikationen for produktet kvalitet og kravet for færdig benzin produktmængde, samt bånd på variablene. Den optimale løsning indeholder optimale værdier for komponenternes mængde og kvalitet som er nødvendigt for at producere de ønskede produkter. Disse optimale værdier sendes videre til avancerede proceskontrol for implementering.
Et eksempel på en produktion plan er betragtet og resultaterne er diskuteret. Resultaterne for test af modellen viser at løsningen er en realisabel, lokalt optimal løsning, og der er gode overensstemmelser med kravet. Modellen svaghed er at priserne for blandings komponenter er uafhængige af proces betingelserne.

Table of Contents

Chapter 1

Introduction

1.1 Background

The control of a typical refinery operation from management down to the smallest process unit can be hierarchically classified in the following four levels:

1 Planning & Scheduling
2 Optimization
3 Advanced Control
4 Regulation

The highest level is *Planning & Scheduling* which is responsible for short and long term planning and scheduling for manufacturing different products in order to fulfill the refinery's obligation and meet its engagements on time. The lowest level, *Regulation*, covers the conventional PID controllers used in different unit operations. These two levels, the highest and the lowest, have existed for many years and have continuously been under development. During the last decades *Advanced Control* has been developed intensively and has caused a remarkable progress in process control engineering.

The missing link between the planning & scheduling level and advanced control is *Optimization.* The optimization system receives goals and constraints from the higher level, which can be for instance the specifications for high quality gasoline. Also, it receives information and constraints from the lower level, for instance the quality of naphtha products from each production unit or capacity of inventory tanks. Based on these information, the optimization system computes an improved operating point, and the targets for reaching that point. The targets are sent to the advanced control system, that implements the targets by computing the appropriate set points for the controllers. Figure 1.1 shows the control hierarchy along with the different actions at each level. From a topological point of view, the operations in a refinery are normally divided into the following hierarchical structure, as it is also shown on the right hand side of figure 1.1:

1 Complete Refinery
2 Processing Area
3 Production Unit
4 Unit Operation

A Processing Area consists of one part of a refinery which has a close economical and functional coherence. An example for this is gasoline processing area which consists all those units and sections of the refinery which are directly involved in gasoline production, covering naphtha products from crude oil distillation column down to gasoline blending unit and final product tanks. A processing area include several production units

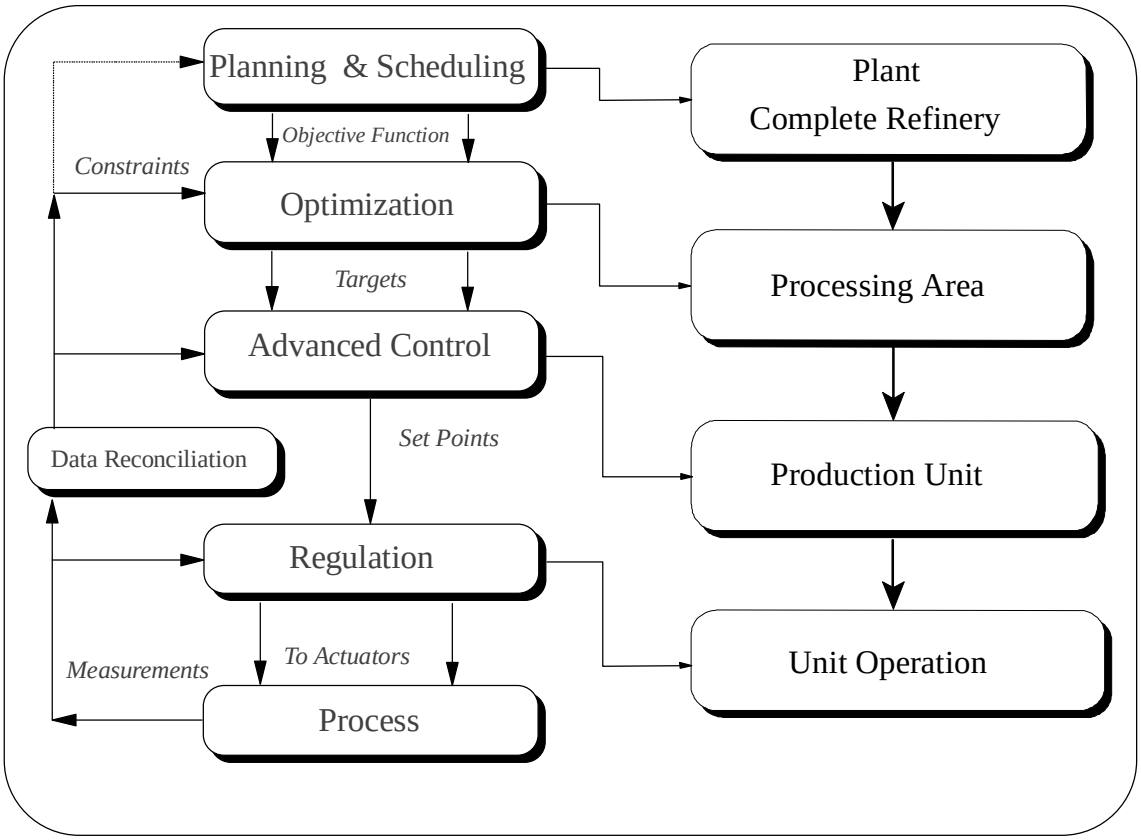

Figure 1.1: The Control Hierarchy, and Process Plant Hierarchy

A *Production Unit* is a part of processing area in the refinery which is responsible for a specific product improvement or production of a particular type of product in the processing area. Examples of production units are catalytic reformer, isomerization unit, and desulphurization plant. A production unit consists of a sequence of U*nit Operations*. A unit operation is the lowest level in the process plant hierarchy for instance a single distillation column.

As it is suggested in figure 1.1, there is direct connections between levels in the control and process plant hierarchies. Each of the four levels in the process plant hierarchy determines the functional domain and delimitation for the respective level in the control hierarchy.
Planning & scheduling at the top of the hierarchy is related to the whole refinery, a processing area, and even a production unit according to its time horizon. A long-term planning with a time horizon of one or several months is related to the whole refinery. For a specific processing area an intermediate-term planning for several weeks of operation is used to check the feasibility of the long-term planning. A short-term planning, usually a few days, concerns with a production unit (Singh et. al. 2000, Sullivan 1990, Agrawal 1995). Advanced control is related to a production unit in order to control some unit-operations which have close coherence. The time horizon for advanced control and conventional regulation is normally hours, minutes, or even seconds for PID controllers. Finally, at the end of the process plant hierarchy, a unit operation is subject to the conventional process control.

The long-term planning and scheduling is performed by using an off-line optimization and forecasts for crude oil prices, product demands, and process units performances. In intermediate-term planning, the information of quantities and qualities of refinery feedstocks, and intermediate products are used to revise and check the feasibility the long-term planning.
Optimization is the link needed to close the connection between the short-term planning and control on production unit level in order to produce the required amount of high quality products at the required time and to minimize the production cost.

Gasoline is one of the most important refinery products. The specifications for high quality gasoline products includes antiknock property, volatility, sulfur and aromatics contents. The antiknock property is expressed as octane number of the gasoline. The octane number of a fuel is defined as the percentage of iso-octane in a blend with n-heptane that exhibits the same resistance to knocking as the test fuel under standard condition in a standard engine. Isooctane and n-heptane are assigned to octane number of 100 and 0 respectively. (Palmer et. al. 1985). There are two standard test procedure in order to characterize the antiknock property. These are defined by American Society for Testing and Materials (ASTM). One definition is designated by ASTM D-908 and is called Research Octane Number (RON), and the other is Motor Octane Number (MON) under designation ASTM D-357 (Garry et al. 1994). RON represents antiknock property under the condition of low speed and frequent acceleration, normally during city driving, and MON represents the engine performance under heavy load and high speed condition, which is normally the condition of highway driving.
The vapor pressure of the gasoline is expressed in Reid Vapor Pressure (RVP) which is given by ASTM D-323. RVP together with gasoline boiling ranges represent the characteristics of gasoline like ease of starting, quick warm-up, tendency to vapor lock. High RVP improve engine economics and starting characteristics, and low RVP prevent vapor lock and reduce evaporation losses. The correct RVP is a compromise between high and low vapor pressure and depends very much to ambient temperature, climate, and season of the year and varies between 49 kPa in the summer and 93 kPa in the winter (Gary et. al., 1994).

The demands for environmental friendly gasoline products includes for instance the limit for aromatic compounds, lead, and sulfur contents in gasoline. A type of hydrocarbon compound containing at least one benzene ring is called aromatic compound.
In addition, regulations imposed by the governments in different countries place maximum restrictions on RVP to limit the emission of volatile organic compound in to the atmosphere.
Furthermore, the global efforts for reducing the consumption of fossil oil products, encouragement to use alternative form of energy, and improving the engine efficiency by car manufacturing company, have caused a downward tendency of gasoline consumption which consequently caused an over-capacity situation for the gasoline market.

This situation encourages the refiners to effectively reduce the amount of give-away for their products by employing optimization. The give-away situation can basically occur when several quality specifications have to be met at the same time and one or two of them become better than the desired. Re-blending in the gasoline blending system can also cause a significant reduction in refinery revenue by taking valuable tank space and blending time. Re-blending will become necessary if a blend does not meet the specification of the final gasoline product.

1.2 Motivation

The optimization level is an important step in the control hierarchy which is inevitable in fulfillment of the following essential requirements:

- Maintain product quality
- Meet the environmental demands
- Reducing the amount of giveaway
- Eliminate re-blending

Increasing the profitability and flexibility of the refinery operation is related to produce basic intermediate streams that can be blended to produce a variety of more specified final product. This concept is widely use in gasoline processing area. The gasoline blending challenge is to produce final gasoline product in such a way as to maximize profit while meeting all the specifications for the final gasoline products. Optimization of the gasoline blending process is thus an important issue considering that gasoline can yield 60-70% of total revenue of a typical refinery (Singh et al. 2000).

Gasoline blending can be considered as a batch process in which the quality and volume of the products are fixed by the refinery production schedule. If the blendstock tanks can be considered as so called standing tanks, in which there is no feed to the tank during the period of blending, then the measured or predicted qualities of the blend components is constant during the blending and a Linear Programming (LP) approach in optimization of blending process would be successful (Singh et al. 2000). The prediction of the qualities is performed by using multivariate regression methods and then applying a bias updating. The bias updating involves comparing measured blend component qualities with those predicted by the model and then the difference is added as a constant in order to appropriate the prediction model. The qualities are normally measured by daily laboratory analyses. This approach has been the existing practice in most refineries in which the quality variation of the blend component is assumed to be unchanged during the period of bias updating.

However, the current trend in the process of gasoline blending is based on a continuous feed to the component blend tanks, i.e. so called running tank. In this situation applying the bias-updated regression model would not be adequate since the qualities of the blend component will change due to the upstream process variation. The LP plus bias-updating formulation will not handle such time-varying feedstock qualities in order to find the optimal solution for the blending problem.
Thus, improved and advanced prediction model is needed for on-line prediction of the quality in the blendstock tank based on the variation of the upstream process.

The intermediate product used for gasoline blending are the products of different production units in the gasoline processing area. The products from catalytic reformer and isomerization units are the most important blendstocks.
The demand for high octane quality of gasoline has stimulated the use of catalytic reformer and isomerization unit. In the reformer process the hydrocarbon molecule structure is changed to form higher octane aromatics with a minor amount of cracking. In the isomerization process the isomers are formed from paraffins by catalytic reactions.
The qualities for some other blendstocks can be calculated or estimated more easily. For instance, oxygenate, butane and isopentane can be assumed to be pure components and thus their qualities can be reasonably estimated based on pure component property. The qualities of some blend components like Light Virgin Naphtha (LVN) can also be calculated or estimated since LVN contain light hydrocarbon components which can be identified by chromatographic analysis.
However, for the reformate and isomerate products from catalytic reformers and isomerization unit it is not possible to estimate or calculate the qualities easily. Besides, the variation of the RVP, and RON qualities, in the product streams of the catalytic reformers and isomerization units, i.e. reformate and isomerate, will particularly provide the possibility of producing different gasoline products with more definite octane number and vapor pressure specifications, and thus larger optimization potential. It is essentially important to have accurate information of the RON quality for reformate and isomerate products, since these are the only high octane number blendstocks applied for gasoline blending.
Furthermore, it is expensive to have on-line quality measurements for these intermediate products in order to have the same sampling frequency as the other process variables. The only existing measurement is laboratory analyses which are available only one per day, i.e. a sample rate of 24 hours, for each quality.

Hence, the above mentioned reasons form the foundation of the strong motivation for developing multivariate prediction models for quality variables of the blendstock.

1.3 Purpose

An important basis for optimization of the gasoline blending process is accurate predictive models for qualities of the blendstocks, especially reformate and isomerate products from catalytic reforming and isomerization processes.
The main purpose of this work is to develop data-based dynamic models in order to be able to predict the qualities of the blend components and supply the optimization system by the past, present and predicted future values of the qualities. The developed model are then used in a multiperiod nonlinear optimization problem for the gasoline blending.
The models are mainly for prediction of Research Octane Number (RON), Reid Vapor Pressure (RVP), concentration of aromatic compounds, e.g. benzene, in the blendstocks.

The optimization is concentrated around the gasoline blending unit of the refinery, and the objective is to determine the *targets* for the advanced control and conventional process control system by minimizing a cost function subject to a set of process and quality constraints in such a way that the needed quantities of the different final gasoline products can be produced on-time, with the desired specifications. The objective function represent the cost of operation for production of blending components plus the inventory cost, which is minimized subject to a set of constraints which represent the demands for quality and quantity of final gasoline products, provided the prediction of the qualities of the blend components.
The methods used in predictive quality modeling and optimization are discussed in the next section.

1.4 Method

The new developments in computer technology in general and specifically the developments in chemical engineering sciences made it possible for chemical engineers to handle the problems concerning process monitoring, evaluation, modeling, control and optimization more efficiently. Chemical engineers often need to extract the useful information from a large volume of data obtained from mostly poorly-known chemical processes. The obtained data from a chemical process is often noisy and faulty. Usually, in a control and optimization application, the data must be rectified before it is used in both calibration and validation of the process and prediction models.
Using first principal methods for prediction of quality variables for oil refinery processes are very difficult. For example in a catalytic reformer process dehydrogenization, cyclization and isomerization are the desired reactions in which the octane number will be improved (Garry et al. 1994). However hydrocracking and condensation reactions are not desired in this process in which the first one will produce light hydrocarbon and the second one will cause formation of coke. Controlling these reactions and estimating reaction kinetic parameters is a very challenging job, since the heavy naphtha feed is made up of a complex hydrocarbon mixture of C_7 to C_{10} .
The alternative to first principal models is *Data-based Based* modeling, in which available historical data is used in order to develop parametric models based on input-output data set. The methods in *Process Chemometrics* are applied for model development in this work.

Chemometric methods have their background in statistic analysis. Principal Component Analysis (PCA) is used in the data assessment, dimensional reduction, and extracting the latent variable. Partial Least Squares Regression (PLS), and Principal Component Regression (PCR) are commonly used for developing input-output regression models (Wise 1991, Esbensen et al. 1994, and Wise et al. 1996).
Since the qualities of the blendstocks depends on the past values of process variables, a *dynamic* modeling approach is used for model development. In this work most quality prediction models are developed mainly by ARX (Auto Regressive with Exogenous input) type of models, which are linear models based on time-series parametric input-output representations. This method has its background in *System Identification* theory (Ljung, 1987).
Artificial Neural Network (ANNs) are also used in developing predictive models for the case of static nonlinear models. ANNs show great ability in nonlinear functional approximation, because of their inherent nonlinarity. A neural network model, applying nonlinear sigmoid transfer function, can be trained to learn input-output data matching by recursive updating and training the internal model parameters, i.e. weights and biases (Haykin, 1994). A multilayer

feedforward network can approximate any continuous function with arbitrary accuracy (Hornik, et al., 1989, Cybenko, 1989). Choosing suitable inputs, which are derived from a basic chemical process knowledge is crucial for a successful ANN modeling. In this sense neural networks should not be considered as a black box, and effective implementation always requires a minimum degree of process knowledge to identify the relevant inputs (R. Braratti, et al., 1995).

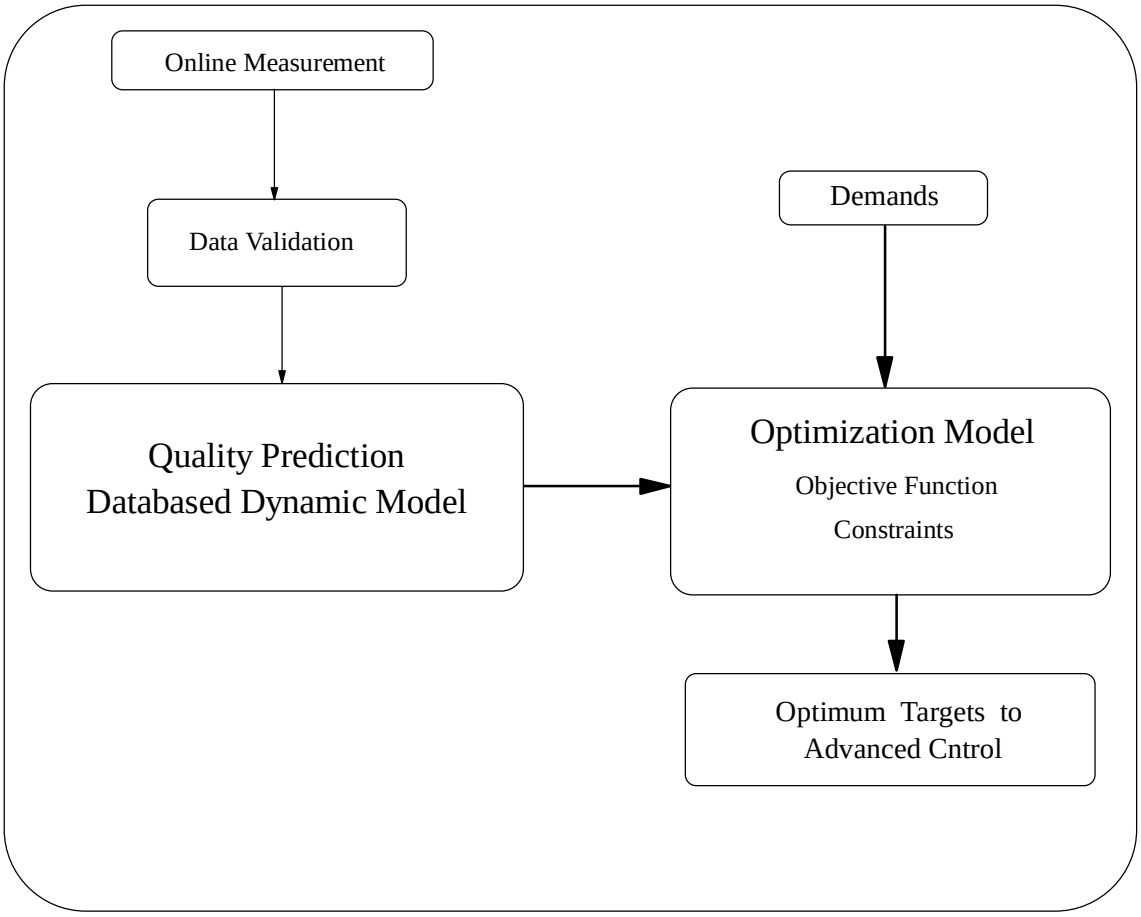

Figure 1.2: Optimization model, objective function, and constraints

Figure 1.2 shows a schematic diagram of data flow for optimization model; i.e. objective function and constraints. Data from on-line measurement of the variables are available to be applied in process and quality prediction models after appropriate pre-treatment by removing the *outliers* and performing *autoscaling*.

The developed dynamic models are used for prediction of the blend component qualities and used as constrained in optimization model. The blending problem is a multiperiod nonlinear optimization problem.

The demands and specification of final gasoline product are included and expressed as the constraints. The objective function is a cost function. The optimum values of variables are sent to the advanced control level to be implemented in the control of the gasoline unit.

1.5 Outline

In chapter 2 the process plant relevant for this work is described in order to provide a general knowledge about the mainstream flow and operation of different production units in the refinery. The feed streams to gasoline processing area are three naphtha streams. A set of stabilizer/splitter followed by catalytic reforming and isomerization processes are the main production units in this processing area that ends with a gasoline blending system and the final inventory product tanks.

Chapter 3 reviews briefly the methods used in multivariate predictive modeling . This review include the essence of process chemometrics in order to be able to discuss the multivariate modeling techniques applied for development of process models in this thesis.

In chapter 4 a description of the preliminary steps in the model development work in this thesis is presented. These preliminary steps concern mainly with definition of the system limit, description of the output and selected input variables, assessment of data, data scaling and sampling, and description of data treatment. Furthermore, it is attempted to present a scope for model development and to describe a procedure and different steps in the process chemometric approach of modeling. The described procedure in this chapter can be used as guidelines to model development

In chapter 5 the structure, calibration, validation, and performance of the chemometric models developed for prediction of RON, RVP, benzene contents of reformate and isomerate products are presented and discussed.

A multi-period optimization model for optimization of gasoline blending is presented chapter 6. The optimization model assumes that the prediction models for the streams sent to gasoline blending are available. A case study is considered as a scenario in production planning and scheduling and the optimum solution for this case is discussed.

The conclusions and suggestion for future work for process chemometric modeling and optimization are presented in chapter 7.

Chapter 2

Plant Description

2.1 Introduction

2.1.1 Purpose

The purpose of this chapter is to provide a general knowledge about the mainstream flow and operation of different production units in the refinery. This description focuses only on the main objective, and function of each production unit. The level of detail in this chapter is based on confidential consideration. Besides the aim is merely to provide the reader with a process knowledge enough to understand the optimization and quality prediction models discussed in the following chapters. Hence, the detail in control loop, flow diagram, and operation in some units are omitted.

2.1.2 Overview

The first major step in refining crude oil is a distillation process to separate the crude oil into 3-4 major products. This is a very important process and normally considered as the heart of a refinery. The crude oil distillation column products are, starting from the top of the column,

naphtha, kerosene, Light Gas Oil (LGO), Heavy Gas Oil (HGO), and finally the bottom product; fuel oil.

The gasoline processing area of the refinery receives three naphtha feed streams and produces the gasoline products into the final product tanks. The three naphtha feed streams are naphtha products of crude oil distillation column, condensate fractionator, and main fractionator in visbreaking/thermal cracking sections. The main production units in this area are three sets of naphtha stabilizer/splitter, two catalytic reformer, and one isomerization unit. Light and heavy naphtha, after splitters, are sent through desulfurization and hydro treating processes for removing sulfur and mercaptanes before sending to the isomerization and two catalytic reformer units. The desulfurization and hydro treating processes are beyond the scope of this work, and it is assumed the yield of production is close to 100% in these units.

Naphtha consists basically of hydrocarbon molecules from C_4 to C_{10} . Besides, depending on type of crude oil, a few percent of naphtha contents will be light gas, i.e. butane, propane, ethane, and methane, and also H_2S, and mercaptanes; i.e. RSH. (Gary et al, 1994).

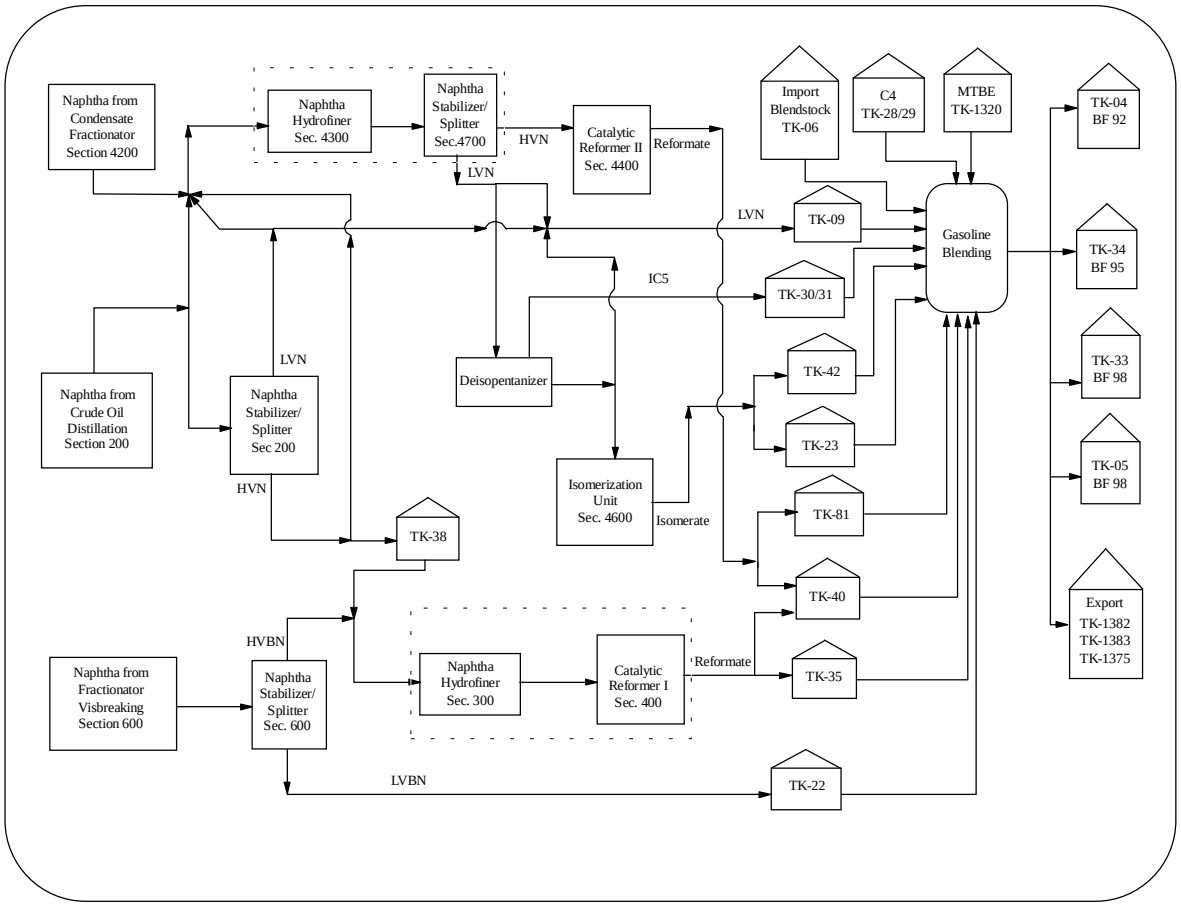

Figure 2.1 : A simplified flow diagram of gasoline processing area.

A set of stabilizer/splitter system split the naphtha into a Heavy Virgin Naphtha (HVN) and a Light Virgin Naphtha (LVN). The HVN streams are sent to two catalytic reformers after a desulfurization process, and then the reformate products are sent to tank. LVN is sent first to a deisopentanizer to separate isopentane (IC_5). The top product of the deisopentanizer is IC_5 which is sent to tank to be used as a blend component for gasoline blending. The bottom product of the deisopentanizer is partly sent to an inventory tank and used as a blend

component and the other part is sent to the isomerization unit, in which the isomerate product from this is also accumulated in tank and used as a blend component in gasoline blending.
As shown in figure 2.1, the products from isomerization unit; isomerate, catalytic reformer units; reformate, LVN, and IC_5 products, along with butane, purchased oxygenate, purchased blend stock, are sent to intermediate storage tanks which are later sent through the gasoline blending system, in which the final gasoline product is produced. The LVBN product shown in figure 2.1 is Light Virgin visBroken Naphtha, which is LVN from a visbreaker process.

2.2 Stabilizer/Splitter

As mentioned before, the three naphtha feed streams to the gasoline processing area are sent from crude oil distillation column section 200, condensate fractionator section 4200, and main fractionator in visbreaking/thermal cracking section 600. Condensate is a product from gas refinery and contain lighter hydrocarbons than crude oil. The visbroken naphtha contains normally more olefins than the other two.

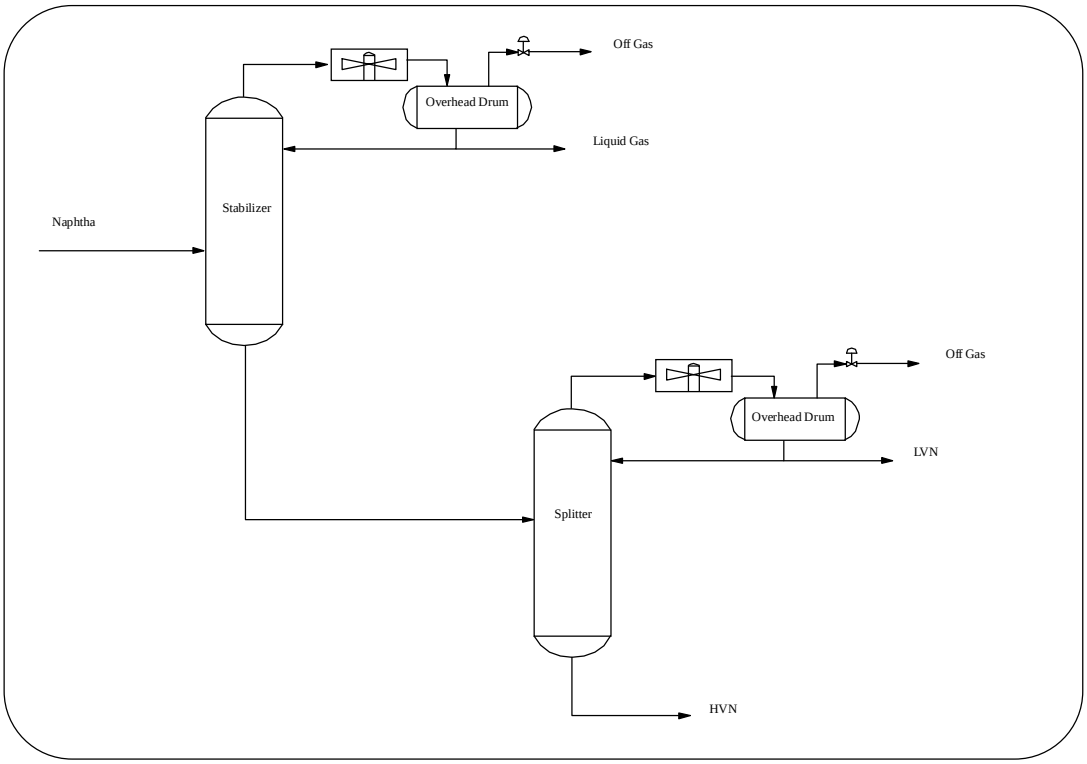

Figure 2.2: Stabilizer/Splitter; Principal sketch.

The general overview of operation in the three stabilizer/splitters are as follows. Figure 2.2 shows a principal sketch of this process. The first step is to separate the light gas from naphtha by distillation. The process is called stabilization. Exceeding concentration of light gas, i.e. methane, ethane, and propane in naphtha will cause formation of emulsion in naphtha and gasoline inventory tanks. The aim of stabilization is to remove the light gases and prevent formation of emulsion. The top product of the naphtha stabilizer consists mostly of butane and other hydrocarbons lighter than C_4. The bottom product of the stabilizer is stabilized naphtha, which consists of a small amount of C_4, C_5 and mostly higher hydrocarbons up to C_{10} . Stabilized naphtha is then sent to naphtha splitter distillation column. Hence, the second step is to split the stabilized naphtha into a Heavy Virgin Naphtha (HVN) and a Light Virgin Naphtha

(LVN). LVN is the top product of the splitter and contains mostly of hydrocarbon molecules between C_5 to C_7 . HVN is the bottom product and consists mostly of C_7 to C_{10}. The True Boiling Point (TBP) of LVN and HVN from a typical crude oil are in the range of 32-88 ^{0}C and 88-194 ^{0}C respectively (Gary and Glenn 1994).
The stabilizer can be one single column as shown in figure 2.2 or a series of distillation columns for removing ethane, propane, and butane which are normally called deethanizer, depropanizer, and debutanizer respectively.
The liquid top product of the splitter in crude oil distillation, i.e. LVN from section 200 in figure 2.1, is mixed with naphtha from condensate fractionator and sent first to a desulfurization process, i.e. hydro treating process, for removing H_2S and to convert the mercaptanes, RSH, to disulfides; RS-SR. Disulfides are insoluble in water and caustic solution. Then, LVN is sent further to the stabilizer/splitter system in section 4700 as shown in figure 2.1.
The bottom product of the naphtha splitter, HVN, in section 200 is sent partially to tank and later mixed with HVBN and sent to a naphtha hydro treating process. The sulfur and mercaptanes are removed and olefins are converted to paraffin by this hydro treating. The desulfurized naphtha is sent further to catalytic reformer in section 400.
The products from the stabilizer/splitter system in visbreaking/thermal cracking; section 600, are LVBN and HVBN, both referring to light and heavy visbroken naphtha respectively. LVBN is sent directly to tank and used as a blend component. HVBN is mixed with HVN from section 200, and sent through naphtha hydro treating process to the catalytic reformer in section 400 as mentioned above.
The HVN from splitter in section 4700 is sent directly to catalytic reformer in section 4400, respectively. The LVN is then sent to deisopentanizer in section 250 and further to isomerization unit.

2.2.1 Deisopentanizer

In this section isopentan, i.e. IC_5 , is separated from LVN. The feed to this section is LVN supplied by splitter in section 4700. The liquid top product is IC_5 and sent to tank as a blend stock for gasoline blending, which contains the maximum possible IC_5 and has an octane number of approximately 89.
A part of the bottom product of deisopentanizer is sent to LVN tank, and the other part is sent to isomerization unit.

2.3 Catalytic Reformers

The purpose of operation in catalytic reformer is to produce high octane number reformate from low octane desulfurized HVN and HVBN in order to provide the blend stock for gasoline blending. Hydrogen is produced in this process and later used in hydro treating processes and isomerization unit.
Dehydrogenization, cyclization and isomerization are the desired reactions in which the octane number will be improved and hydrogen will be produced. However hydrocracking and condensation reactions are not desired in this process in which the first one will produce light hydrocarbons and the second will cause coke formation. The catalytic reactions are mainly endothermic. More detail canbe found in Gary et al, 1994.
Generally, the process consists of a heater, a sequence of fixed bed reactors, a gas-liquid separator and finally a stabilizer distillation column.

There are two catalytic reformers in gasoline processing area. We just call them by I and II, or section 400 and section 4400 respectively. The strucure of the two reformers are principally the same, and minor details are out of scope of this work. The reformers are described in the following sections.

2.3.1 Catalytic Reformer I

Figure 2.3 shows a schematic diagram of this production unit. The desulfurized HVN feed stream is mixed with a recycle H_2 gas stream. Mixing H_2 with HVN is mainly for preventing undesired hydrocracking and condensation reactions. The combined gas and liquid stream is sent through a heat exchanger system before entering heater.

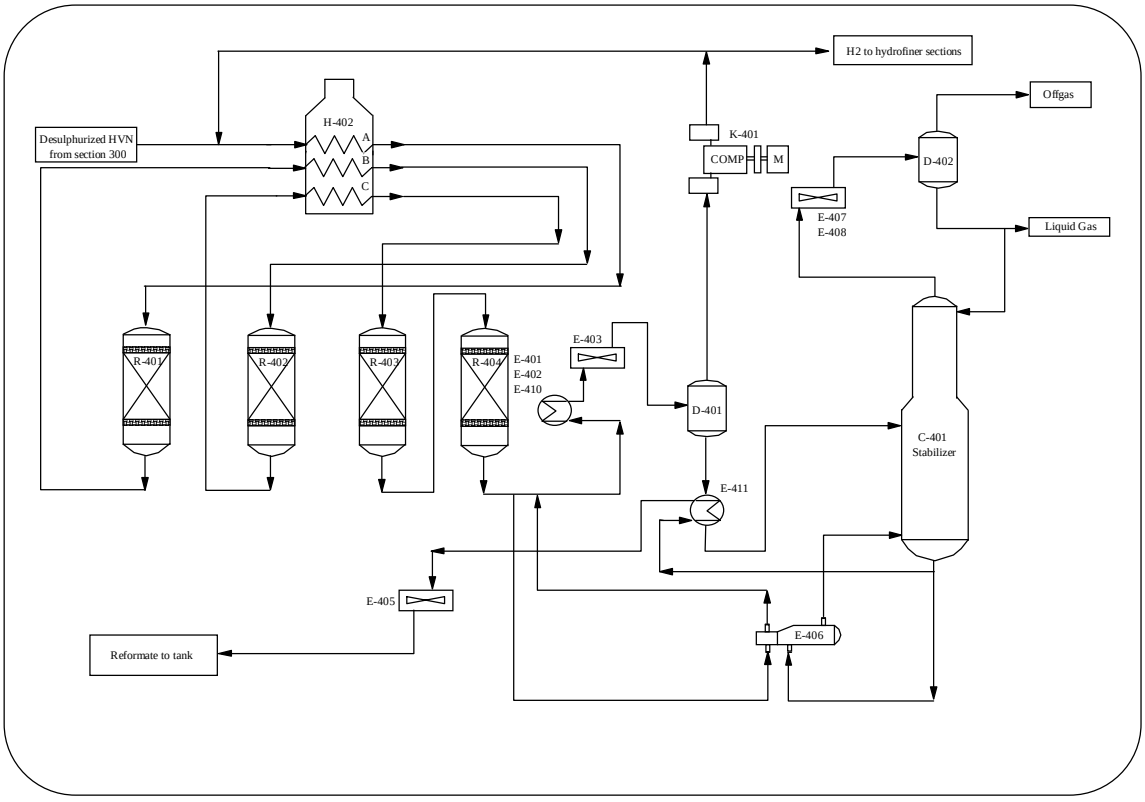

Figure 2.3: A simplified flow diagram of catalytic reformer I.

The feed stream enters in coil A of the heater and continue to the first reactor; R-401. Because of the endothermic reactions, the outlet stream of the R-401 is sent back to the heater through coil B and further to reactor number 2; R-402. Again the output of the R-402 is warmed up in the heater by coil C and continue to the third reactor R-403. The outlet stream of the R-403 is then sent to reactor R-404. The product stream of the R-404 is passed through a set of heat exchanger where the heat from the product stream is transferred to the feed stream to the coil A of the heater, and reboiler E-406 of reformate stabilizer distillation column.
The product stream is then cooled in cooler and then sent to product separator drum D-401 where the gas is separated from the liquid. The gas consists mostly of hydrogen. The liquid from the separator drum is sent to reformate stabilizer.
The stabilizer column C-401 produces a liquid bottom product, which is reformate, a gas top product and a liquid top product. Reformate product is sent to reformate tank.

2.3.2 Catalytic Reformer II

Desulfurized naphtha is sent to this section and mixed with recycle H_2 and then sent through a series of heat exchangers before entering the heater H-4401, as shown in figure 2.4.
In H-4401, HVN is warmed up first in convection zone and then in coil A from which it is sent to the first reactor R-4401. The outlet of the R-4401 is sent back to the heater through coil B and further to second reactor R-4402, and again product of this sent to heater and then to third reactor R-4403. This extensive heating is mainly due to the endothermic catalytic reactions and the necessity for heating the streams before entering each reactor.

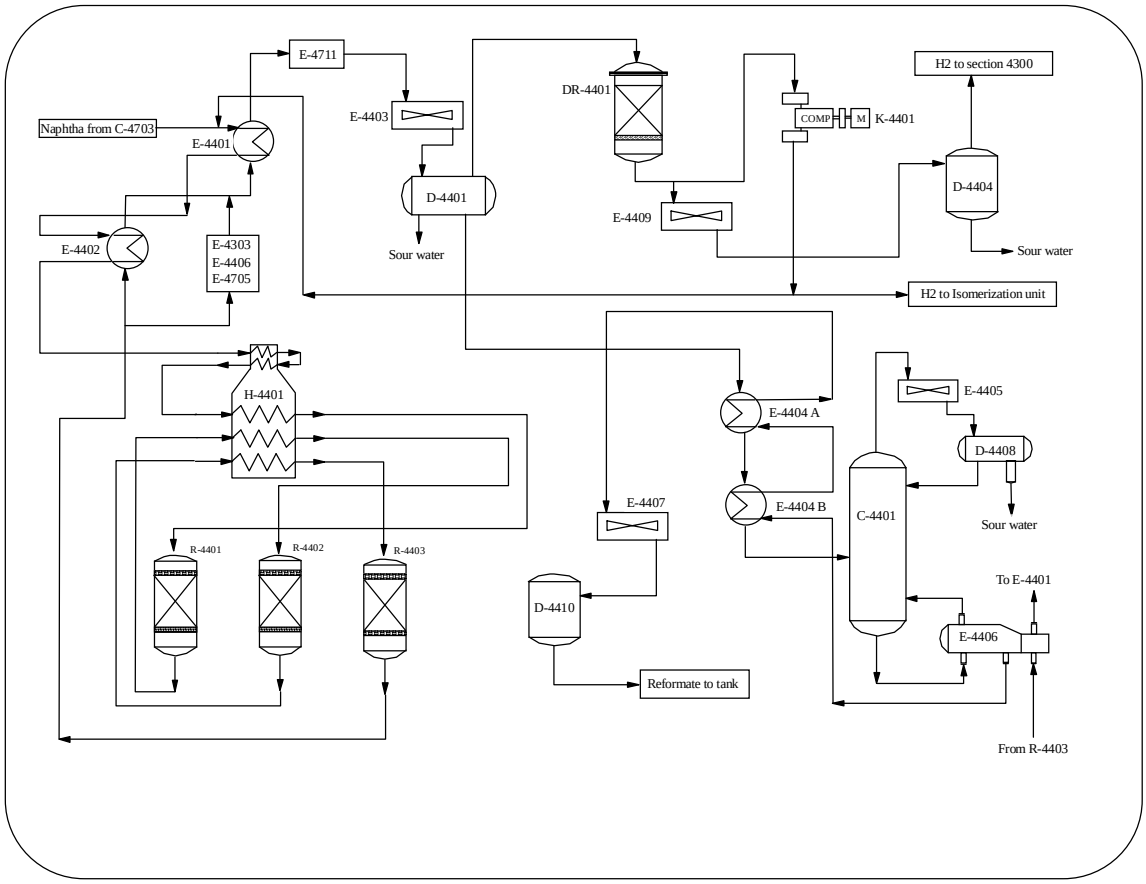

Figure 2.4: A simplified flow diagram of catalytic reformer II.

The product from the last reactor is sent through a series of heat exchangers for heat recovery and then to gas-liquid separator drum D-4401. The gas from separator drum is sent to a dryer where water and H_2S is removed from the gas. A part of gas from the dryer is recycled and mixed with the feed. The liquid from the separator D-4401 is sent to reformate stabilizer C-4401.
The gas top product of the stabilizer is sent to gas plant and the bottom product, which is reformate product, is sent to tank as a gasoline blend stock.

2.4 Isomerization Unit

The purpose of operation in this section is to convert low octane number LVN to high octane number by catalytic isomerization process. The reactions in this process are mainly exothermic.
The feed to the isomerization unit is LVN from deisopentanizer. The isomerization unit is made up of three parts, namely **Penex** unit where conversion of LVN takes place, **Molex** unit where separation of isomers takes place, and **Hot oil system** which is responsible for the necessary energy supply of the whole unit.

2.4.1 Penex Unit

Figure 2.5 shows a simplified schematic diagram of the Penex unit. LVN feed to the isomerization unit is mixed with extract from the Molex unit, described in section 2.4.2, and hydrogen from catalytic reformer II, section 4400. Then the feed is sent for preheating to E-4608 A/B, E-4609 and E-4610 where the feed is warmed up by the reactor product of R-4601 A, R-4601 B, and the hot oil system respectively.

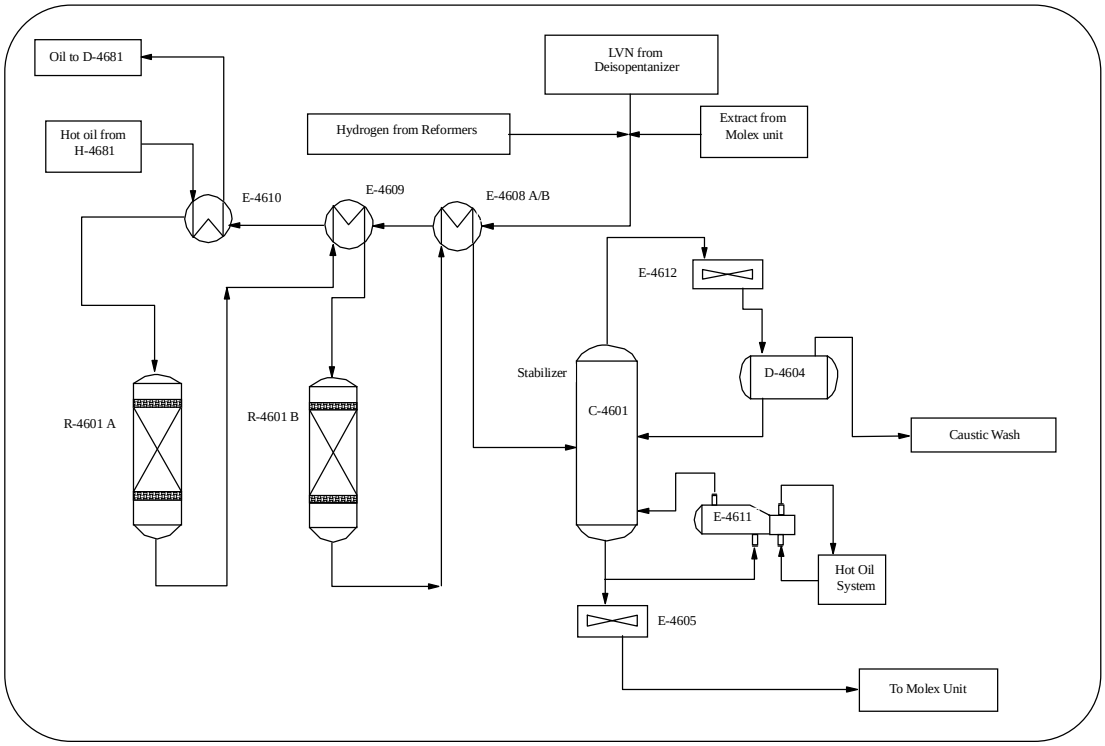

Figure 2.5 : A simplified flow diagram of Penex unit.

The feed is then sent to reactors R-4601 A/B. The chemical reactions of isomerization process take place in R-4601 A/B. The reactor product is sent for exchanging the heat with the feed in E-4609 and then back to R-4601 B. The product of the R-4601B is then sent for cooling in E-4608 A/B and further to stabilizer C-4601.
In stabilizer C-4601 the light hydrocarbons are removed from LVN as off gas in the top. The stabilizer overhead gas is cooled in E-4612 and accumulated in drum D-4604. The liquid from D-4604 is sent back to C-4601 as reflux. LVN is sent to Molex unit from the bottom of the stabilizer.

2.4.2 Molex Unit

In this section the branched; i.e. isomers, hydrocarbon molecules are separated from the other. Figure 2.6 shows Molex part of the isomerization unit.

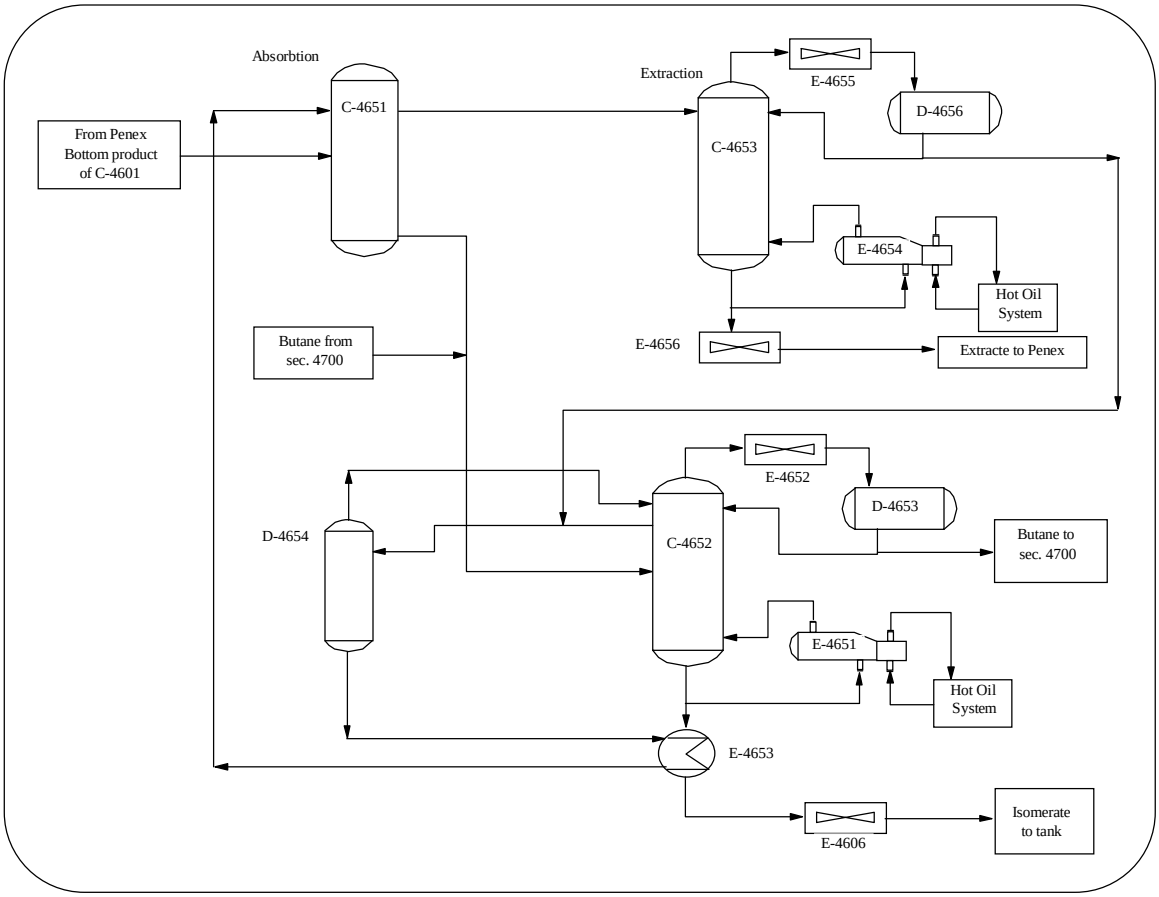

Figure 2.6 : A simplified flow diagram of Molex unit.

The bottom product of stabilizer C-4601 in the Penex unit is sent to absorbtion column C-4651. In this column the separation of isomeric compound takes place by an absorbent material, and butane as a desorbent liquid.

The non-chained hydrocarbon molecules is sent from C-4651 to extraction column C-4653 where the desorbent is separated. Desorbent is sent from top of the extraction column to desorbent drum D-4654. The bottom product of the extraction column C-4653 is cooled in E-4656 and sent back to Penex unit for isomerization.

The isomeric compounds is sent as the bottom product of C-4651 to isomerate column C-4652. Butane is mixed with the bottom product before entering to C-4652. The top product of C-4652 is sent to overhead drum D-4653 via cooler E-4652. A part of liquid from D-4653 is sent back to C-4652 as reflux and the rest is sent to C_3/C_4 splitter C-4705 in section 4700. A side stream of C-4652 is sent to desorbent drum D-4654. The desorbent from D-4654 is recirculated to the absorption column C-4651 via heat exchanger E-4653.

The bottom product of C-4652, which is isomerate, is cooled in E-4653 by desorbent from D-4654 and cooler E-4606. Isomerate is then sent to tank.

2.5 Gasoline Blending

The purpose of this operation is simply to produce final gasoline products by mixing the blend components. These blend components are mainly produced in the previous sections of refinery. Figure 2.1 includes also the gasoline blending unit.

Blend Component	Tank no.
Oxygenate	TK-20
Butane	TK-28+29
Import Naphtha	TK-06
LVN	TK-09
IC5	TK-30+31
Isomerate	TK-23 + 42
Reformate 4400	TK-81
Reformate 400	TK-35
LVBN	TK-22

Table 2.1 : Gasoline Blending Components

Tank no.	Final Gasoline Product
TK-04	Danish unleaded octane 92 (BF 92)
TK-34	Danish unleaded octane 95 (BF 95)
TK-05	Danish unleaded octane 98 (BF 98)
TK-82	Swedish unleaded octane 95 (SV 95)
TK-33	Swedish unleaded octane 98 (SV 98)
TK-83	German unleaded octane 91 (TYSK 91)
TK-75	German unleaded octane 95 (TYSK 95)

Table 2.2 : Final Gasoline Products

The blending components are listed in table 2.1. The oxygenate is an additive used for increasing the octane number, it can be MTBE, i.e. Methyl Tertiary Buthyl Ether, ETBE, i.e. Ethyl Tertiary Buthyl Ether, or ethanol. Import naphtha is also used if the produced blend components do not fulfill the desired specifications. The purpose is to minimize the consumption of oxygenate and import naphtha, since the price of these two components are high..

Table 2.2 shows the final gasoline products. The product qualities are well specified. Although there are several important properties of gasoline, the three significant qualities that have the greatest effects on engine performance are Reid Vapor Pressure (RVP), boiling range, and antiknock characteristic. Antiknock characteristic is measured and represented by octane number. There are two type of octane number; Research Octane Number (RON) and Motor Octane Number (MON) which are described in chapter 1.

2.6 Summary

In this chapter a general process knowledge about the mainstream flow and operation in gasoline processing area is presented. The level of detail in process description is chosen based on confidential agreement.

Among the production units in this area, we have focused on the following major units: three sets of naphtha stabilizer/spilltters, two catalytic reformers, one isomerization unit, and finally gasoline blending system. Three naphtha streams are sent to the gasoline processing area. These three naphtha feeds are sent from crude oil distillation, condensate fractionator, and main fractionator column in visbreaking/thermal cracking section. The naphtha feeds are first separated into a Heavy Virgin Naphtha (HVN) and a Light Virgin Naphtha (LVN). The HVN streams are sent to two catalytic reformers after a desulfurization process. The reformate products from catalytic reforming processes are sent to storage tank. LVN is sent first to deisopentanizer to separate isopentane (IC_5) and then to isomerization unit. The isomerate product is then sent to storage tank. The two reformate, isomerate, LVN, and IC_5 products mention above along with oxygenate, butane, import naphtha, and LVBN make totally nine blend components. The blend components are kept in intermediate storage tanks and used for producing final gasoline products in gasoline blending section. It is desired to minimize the consumption of oxygenate and import naphtha and reduce the give-away in final products, i.e. minimize the cost of operation, and meet the demand specifications for final gasoline products.

Chapter 3

Methods in Process Chemometrics

3.1 Introduction

3.1.1 Purpose

The purpose of this chapter is to present a review of methods used in process chemometrics. This review include the essence of process chemometrics in order to be able to discus the multivariate modeling techniques applied for development of process models in this thesis.

3.1.2 Background

The background, and basis principles in this research come from many areas. As far as the scope of this work is allowed, it is attempted to include the theory of multivariate modeling techniques so that the reader does not need to go to many other references in order to understand the development of the models in this work.

There are different approaches for development of models depending upon the purpose of the model. This purpose may be prediction of a process or quality variable, description of a phenomenon, or assessment of data obtained from chemical processes for process monitoring purposes. In this work we focus mostly on predictive modeling. This means that we take advantage of available historical data to develop models based on information and knowledge obtained from the data. This is *data-based* modeling and the purpose here is to predict process or quality variables which are expensive or difficult to measure as frequently as it is desired for control and optimization applications.
It is hardly possible to obtain a complete first principles model covering the reformation and isomerization reactions in the respective units, since there are numerous hydrocarbon components in the feed stream to these units and also because of high number of reactions occurring in the reactors. However, as a general discipline we consider the first principle model as a basis for selecting the relevant inputs for prediction models. It is crucial to select suitable input variables which contain major variables affecting the variation of the model output.
When suitable input variables are chosen, the next step is to estimate the model parameters, and in this sense estimation of model parameters can be defined as an approximation to input-output functional relationship, in which the best linear or nonlinear relation between input and output variables are found.
The choice of different approaches depends on the modeling objective and degree of non linearity. The complex refinery processes we are dealing with in this work are indeed highly nonlinear. However, it is possible to predict a single process or quality variable by making a linear approximation.

The general principle selected for knowledge based predictive modeling in this work is the methods in *Process Chemometrics*. A definition of Chemometrics is given by Wise et al. 1996 as follows: "Chemometrics is the science of relating measurements made on a chemical system to the state of system via application of mathematical or statistical methods".
Hence, the methods are based on data obtained from the system, and the purpose is to develop an empirical model for estimation of one or more properties of the system.
Process chemometrics includes both linear and nonlinear approaches. Moreover, it is important to consider the dynamic characteristics of the system. If the variables change with time, or their current value depends on the earlier values, then an appropriate dynamic model should bed used. Here, the time-series type of model is a good candidate.
Based on this consideration, the methods in process chemometrics are divided in four general categories according to the linear, nonlinear, static, and dynamic characteristics of the system under study.

3.1.3 Outline

In the class of static linear methods Principal Component Analysis (PCA), Principal Component Regression (PCR), and Partial Least Squares Regression (PLS) are discussed in this chapter. PCA is used in data assessment, dimensional reduction through extracting the latent variables and applied mostly for process monitoring.. PLS and PCR are used for developing input-output regression models. These are all presented and discussed in section 3.2.
In the class of static nonlinear approaches Artificial Neural Networks (ANNs) exhibit a strong ability to nonlinear functional approximation. Nonlinear PLS regression in which nonlinear

function is defined for the inner relationship of the PLS is another approach in this class of chemometrics methods. A discussion of ANN modeling and nonlinear PLS is presented in section 3.3. Furthermore, in section 3.3 a description of Nonlinear Principal Component Analysis (NLPCA) is presented. A NLPCA model is developed based on Input Training Neural Networks (ITNN) which is used for data rectification.

Section 3.4 deals with the method in the class of dynamic linear methods. In this section the methods used in System Identification are described System Identification deals with knowledge based predictive modeling using linear time series regression. The linear methods include ARX and ARMAX (Auto Regressive Moving Average with Exogenous input), which are linear models based on parametric input output representations

In section 3.5 a short description for the dynamic nonlinear class of methods is presented, in which the time-series type of model can be integrated in a nonlinear PLS model.

A discussion about different criteria in model validation is presented in section 3.6. Two different reference models as average-model and zero-model are presented in order to assess the predictability of the developed chemometric model.

In section 3.7 the concept of informative data set and persistence of excitation is presented.

A summary of the methods discussed in this chapter is given in section 3.8.

3.2 Static Linear Methods

3.2.1 Principal Component Analysis

Principal Component Analysis (PCA) is a method used for dimensionality reduction of data in which the data is decomposed to detect the underlying multivariate correlation structure which is also called hidden phenomena. PCA is a linear approach for decomposition of the original data into *structure* and *noise* parts, as it is expressed in equation (3.1).
The original data is usually made up of a set of several observations of variables. Each observation is called an *object* and consists of measurement of all variables at the same time. An object has the dimension of the number of the variables, which makes the *superficial dimensionality* of the data. Discovering the significant variation in the data is the first and important step in approaching an understanding of the process. The *intrinsic dimensionality* is the number of independent variables underlying the significant nonrandom variation in the data. These independent variables, also called *Principal Components* (PCs) or Latent Variables (LV), describe the properties of the original data by discovering the underlying correlation structure.
By using PCA an optimal transformation of the data from the original variable space to a principal component space, also called *factor* space, is made in which the essential information in the data is preserved. There will be a minimum sum of squares difference between the original data and the reconstructed data in PCA.
This method is basically a linear method for reduction of data dimensionality with minimum loss of information. Let X represent a (n *x* m) data matrix, in which n is the number of the observations and m is the number of variables. A PCA model is an approximation to the data matrix X, and can be described by the following model:

$$X = TP^T + E = Structure + Noise \qquad (3.1)$$

where T(n *x* f) and P(m *x* f) are *Score* and *Loading* matrices respectively, E(n *x* m) is residual or noise, and f is number of principal components . It is useful to formulate the PC model in equation (3.1) as an outer product of individual PC contributions:

$$X = t_1p_1^T + t_2p_2^T + ... + t_ip_i^T + ... + t_fp_f^T + E \qquad (3.2)$$

where t_i is the score vector for PC_i , p_i is the corresponding loading vector, and f is the number of PCs, which must be less than or equal to the smallest dimension of X, i.e. $f \leq \min\{n, m\}$.
The loading matrix is a transformation matrix between the original variable space and the PC space spanned by the principal components. The columns in P are called loading vectors and are orthonormal in which:

$$p_i^Tp_j = 0 \quad for\ i \neq j, \quad p_i^Tp_j = 1 \quad for\ i = j$$

Loading vectors give us information about the relationship between the original variables and the PCs. The columns in T are called the score vectors for each component and are orthogonal in which:

$$t_i^Tt_j = 0 \quad for\ i \neq j$$

The scores are the effect of observations on each PC.
The concept of principal components is related to eigenvectors of covariance or correlation matrix of X. The covariance matrix is defined by the following equation:

$$cov(X) = \frac{X^T X}{n-1} \tag{3.3}$$

The loading vectors are eigenvectors of cov(X), in which for each p_i

$$cov(X)\, p_i = \lambda_i\, p_i \tag{3.4}$$

where λ_i is the eigenvalue associated with the eigenvectors p_i . Thus, in PCA the eigenvector is called principal component, and the associated eigenvalue is a measure of the captured variance for each pair of score and loading vector.
In equation 3.3, it is assumed that the data X is adjusted to have a zero mean by subtracting off the original mean, and hence the data is mean centered. This type of data scaling is used in order to remove the effect of different dimensions in the data. If the mean centered data is additionally adjusted to unit variance by dividing each column in data matrix by its standard deviation, then the data is called *Autoscaled*. Applying autoscaled data in equation 3.3 will give the correlation matrix of the original data.

PCA model is based on projection of the original data matrix X on to a number of principal components along the direction of the *maximum variance* or *minimum squared projection distance*. That means that the first principal component (PC1) lies along the direction of the maximum variance, the second principal component (PC2) lies along the direction of the next maximum variance orthogonal to the first PC, and so on.

The maximum number of principal components can be either number of variables or number observations; i.e. number of objects, depending on which is the smallest, . The effective full dimension of the PC space is given by the rank of the X matrix. A full model is the case when number of PCs is the maximum, i.e. f = min(m,n). In this case the residual E is equal to zero and the decomposition of X only change the original coordinate system, i.e. the variable space, to the new coordinate system, i.e. the PC space, which is not optimal for separating process structure from noise since no separation between the structure and noise part of the data is accomplished. Thus, number of PC must be chosen for an optimum fit so that TP^T contains the relevant structure and then noise is collected in E. By this choice we obtain a principal component model as a transformation in which many original dimensions are transformed into another coordinate system with fewer dimensions. The transformation is achieved through projection or eigenvector decomposition.
PCA model involves only with one set of data. Methods relating two sets of data, input and output, i. e. X and Y, are generally called *multivariate calibration*, multivariate regression, or simply multivariate modeling.

3.2.2 Multivariate Modeling

Multivariate modeling is to establish or find a model for the connection between input and output; X and Y. The output (Y) matrix consists of *dependent* variables and the input (X) matrix contains the *independent* variables. The multivariate model is simply the regression relationship between the empirical input and output. Development of a model implies

establishment or in fact *estimating* the relationship between X and Y. This process is called *calibration*, training, or model parameter estimation, and the X-Y data used for this purpose is called calibration or training data set. Statistically, it means that we estimate the parameters in a regression model. The model is then used on a new set of X data in order for *prediction* of unknown Y.

3.2.3 Multi Linear Regression, MLR

Let start with a classical example; Multi Linear Regression MLR. The model is expressed mathematically in equation 3.5. This method combines a set of X or input variables in a linear combination that correlate closely to the corresponding output or Y values.

$$y = a_1 x_1 + a_2 x_2 + ... + a_n x_n + E \qquad (3.5)$$

where a_0 , a_1 , ..., a_n are constants, and called model parameters. Y and X are output and input variables, respectively, and E is the residual or error.
Equation 3.3 can be reformulated by defining the vectors Y and X representing the outputs and inputs, and vector B for the model parameters.

$$Y = X B + E \qquad (3.6)$$

It is now desired to determine the model parameters B so that error E is minimized. A common procedure is to use the least squares criteria for minimization of E^TE in order to find the optimum model parameters B. An estimate for B parameters can be found by the following equation (Esbensen et al. 1994):

$$\hat{B} = (X^T X)^{-1} X^T Y \qquad (3.7)$$

As it can be seen estimation of B involves a matrix inversion, $(X^TX)^{-1}$. If the X variables are inter correlated; i.e. approximately linearly dependent, matrix inversion in equation (3.7) becomes increasingly difficult and in worst case MLR will not work due to the linear dependency. To avoid this unfortunate numerical instability matrix X must have full rank, and this means some of the variables which correlate with each other must be omitted, which may result in loosing information.
Another problem that may cause failure of MLR method is error or existence of high level of noise in the X data. The MLR solution is represented by a least square plane optimally fitted to all data and implicitly assumed that the X variable are noisfree. It is assumed that only Y variables is affected by error and not the X variables.
To avoid these two problems Principal Component Regression (PCR) model is a good candidate in which bilinear projection methods are employed.

3.2.4 Principal Component Regression

A PCA model relies on the projection of the original data matrix X on to a number of principal components along the direction of the maximum variance in the X matrix. This concept is used in Principal Component Regression (PCR) in order to remove the effect of linear dependency and high level of noise in the X data.

PCR performs first a principal component decomposition exactly as PCA and then Y variable is regressed onto the decomposed X matrix. The score matrix T is used in PCR model instead of original X data, which is mathematically expressed as follows:

$$Y = T B + E \quad .(3.8)$$

in which number of columns in T is equal to the number of PCs retained by the PCA model.
By this choice we will obtain a model which is stable and robust against collinear X data, and if the data are defective or noisy. Furthermore, the concept of score and loading matrices can be used in order to interpret the result.
The resulting vector of regression coefficient B, which relate the scores in X to the output Y, is expressed as the following:

$$\hat{B} = (T^T T)^{-1} T^T Y \quad (3.9)$$

The regression vector can be obtained by multiplying the coefficient B by the loading matrix P.

$$r = P B \quad (3.10)$$

The estimate of the Y dependent variables can be obtained by multiplying the X matrix by the regression vector.

$$\hat{Y} = X r \quad (3.11)$$

In calculation of regression coefficient B the inverse of the scores covariance $(T^T T)^{-1}$ is used which is perfectly conditioned since the scores are orthogonal.
Despite the positive advantages mentioned above, PCR model is still not an optimal solution for multivariate calibration. The reason is that all the variation in X data will not necessarily create an optimal model to predict Y. In another word, there may easily be structured information in X that have nothing to do with Y. This problem can be avoided by applying Partial Least Squares (PLS) model in which the regression is performed in order to relate the variation of the independent variable directly to the variation of the dependent variables.

3.2.5 Partial Least Squares Regression

In PLS regression the variation and the data structure in the dependent variables Y is directly used in PCA decomposition of the independent variables X. We may think of PLS as a simultaneous decomposition of X and Y are performed using PCA.
By this an optimal regression is achieved with less principal components and more prediction ability, that also can handle noise, error, and collinearity in X data.
In order to explain how PLS works, it is easier to make a simplification and look at PLS as simply two simultaneous PCA analyses. T and P are score and loading matrices related to X and U and Q are score and loading matrices related to Y as it is shown in equation 3.12, and 3.13. Furthermore, one more loading matrix is calculated for X. This extra loading is called W loadings or PLS-weights.

$$X = T P^T + E \quad (3.12)$$

$$Y = U Q^T + F \quad (3.13)$$

PLS does not really perform two independently PCA analyses but in reality connect the scores in PCAX and PCAY models and by this let the structure in output data Y, directly affect the decomposition procedure in input X. Besides, principal components are not the same as in PCA, and they represent only the correlation between Y and X, and thus reduce the influence of large X variation which in fact does not correlate with Y. Therefore, they are called PLS components rather than PCA components.
The relationship between the scores U for to the dependent variables Y and the scores T for the independent variables X is expressed in equation 3.14, in which h denotes residuals.

$$U = f(T) + h \quad (3.14)$$

In linear PLS, it is assumed that the function f is defined by a simple linear equation as follows:

$$u = b\,t + h \quad (3.15)$$

The coefficient b is called the inner relationship, or internal regression coefficient.
The PLS algorithm can be sketched briefly in a simplified summary as follows. First a PCA analysis is performed on Y data, and the score for the first PLS-component U_1 is used as the starting value for T_1 in PCAX. So T_1 is replaced by U_1 in PLS algorithm, and decomposition of X data is then performed. By this the PCA model for X data is affected by the structure in Y data. After performing PCA on X data, the calculated loading matrix, P, is saved as W loading weights, and the score for the first PLS-component T_1 in X-space is immediately used as the starting value for the U_1 vector. By this we let the structure in X data also affect the PCA analysis on Y data. This procedure of calculation and substitution of U and T continues, also for other PLS-components, in an iterative manner until the convergence is reached. A set of T, W, U, Q matrices are calculated.
The PLS regression results in two loading matrices for X data. They are called loadings P and loading weights W or effective loading. The P loadings are the same as obtained in ordinary PCA and express the relationships between X data and the scores T. The W loadings express the relationship between X and Y data, and the columns in W matrix are in fact PLS-components. Both P and W matrices are important and may be used for interpretation of the PLS model or inspection of the model ability.
In practical application, it is preferable to apply a MLR type of model. In PLS regression, the matrices W, Q, and P are used for calculation of a set parameters which correspond to parameters B in equation 3.6. The estimation of the B parameters is performed by using the following equation.

$$\hat{B} = W (P^T W)^{-1} Q^T \quad (3.16)$$

PLS can also handle several covarying output variables. Its ability to extract the useful information from collinear, noisy, input data which is relevant for modeling the prediction of the output variables makes the PLS a powerful tool for linear regression modeling.

3.3 Static Nonlinear Methods

3.3.1 Nonlinear PLS Model

As it is described earlier, in linear PLS the relationship between the scores U for to the dependent variables Y and the scores T for the independent variables X is defined by a simple linear function. In many application of multivariate calibration the relationship between X and Y variables are indeed nonlinear. One method to capture the nonlinear correlation is to define a nonlinear function in equation 3.14 for the relationship between the scores U and T. This function can be defined as a polynomial of arbitrary order as it expressed in equation 3.17.

$$U = C_0 + C_1T + C_2T^2 + C_3T^3 + .. + h \qquad (3.17)$$

The other method is to describe this functionality by using Artificial Neural Networks (ANNs) in order to approximate the nonlinear relationship between the scores in X and Y.

3.3.2 Artificial Neural Networks

The excellent ability of Artificial Neural Networks (ANN) to consider nonlinearity in functional approximation problems makes it to a powerful tool for application in process industry. This ability of ANNs is due their inherent nonlinearity, as it will be described in the following. A multilayer feedforward neural network can approximate any continuos function with arbitrary accuracy (Hornik, et al., 1989, Cybenko, 1989).

3.3.2.1 Model Structure and Algorithm

The internal structure of a feed forward network consists of three major parts, each made up of layer(s) of neurons. Figure 3.1 shows a schematic diagram of internal structure of a typical neural network. The inputs to ANNs are provided in input layer; i.e. layer number one, which has the number of neurons equal to the number of input variables. The same is for the last layer; output layer, which also has the same number of neurons as the number of output variables. Between these two layers, there is one (or more) layer(s) called hidden layer(s), and contains the most important part of the model parameters which are developed during the calibration also called the training of the model. The question is now how many neurons should be chosen for the hidden layer in order to obtain a robust model with an acceptable model performance.
The ability of the neural networks to fit arbitrary nonlinear functions depends on the presence of a hidden layer with nonlinear nodes (Kramer, 1991). A suitable nonlinear function is the sigmoid, which is a continuous , smooth, and monotonically increasing function of the form:

$$\phi(x) = \frac{1}{1+e^{-x}} \qquad (3.18)$$

$$\begin{aligned} \phi(x) &\rightarrow 1 \quad for \quad x \rightarrow +\infty \\ \phi(x) &\rightarrow 0 \quad for \quad x \rightarrow -\infty \end{aligned} \qquad (3.19)$$

In the example suggested in figure 3.1, we have an input vector with p variables for p inputs, a hidden layer with s1 neurons, and an output layer consisting of q neurons for q output variables. Each input is weighted with an appropriate weight W. The elements in weight matrix W1 (s1 x p) are the corresponding weight for s1 neurons in the hidden layer and p input variables. Furthermore B1(s1) is a bias vector for the neurons in the hidden layer, which has s1 element for the neurons in the hidden layer. In the same manner, a weight matrix W2 of size (q x s1) and a bias vector B2(q) is defined for the output layer.
The internal activity level of a neuron is defined by an *Activity Function* as a dot product of the weight matrix and input to each layer. For instance the activity of each neuron in the hidden layer is defined as the following:

$$v_j = \sum_{i=1}^{m} W_{ji} x_i \qquad (3.20)$$

This activity function forms the input to the j'th *Sigmoid Transfer Function* defined in equation (3.18) to calculate the output of the neuron. The output matrix of hidden layer in this example can be expressed as follows:

$$y1 = \phi(W1 \bullet X + B1) \qquad (3.21)$$

In the same manner, the output of the last layer of the network is calculated. The network's output is often called the *predicted* value, which is compared with the *measured* value and an error is calculated. The network error is calculated as the difference between predicted and measured output.

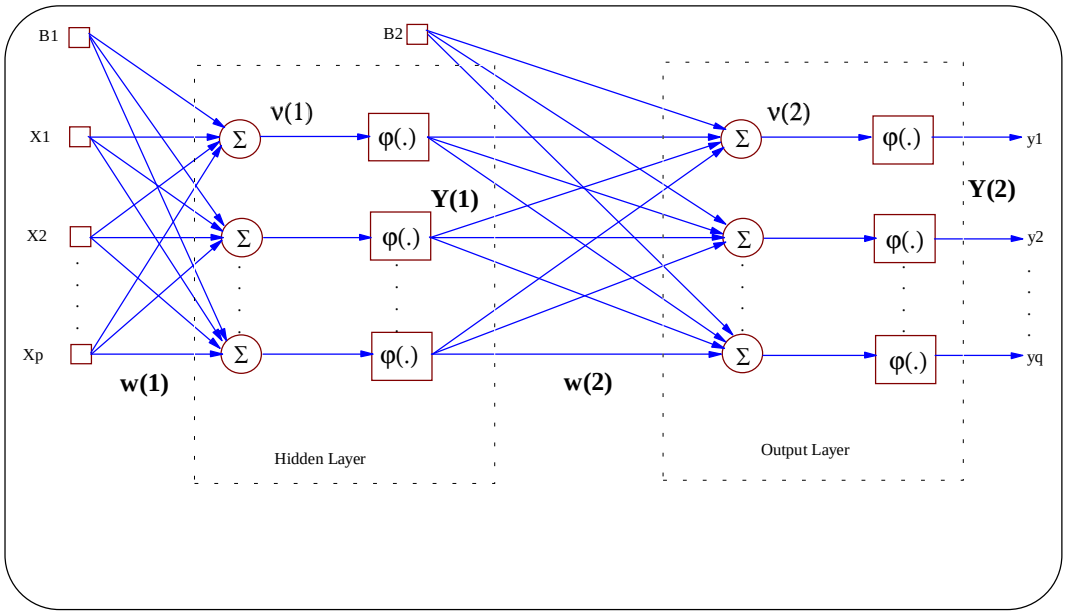

Figure 3.1 : A typical Structure of Neural Network.

In this example, for simplicity, we assumed that the input X in figure 3.1 is a vector of only one measurement for each input variable and also the output Y contains one corresponding output measurement, i.e. only one object. Normally, in a supervised learning using *Back Propagation* learning algorithm (Simon Haykin, 1994) , the network is trained on a batch of samples. Thus, the input, output and error E of the network are matrices of n number of samples.

The *Root Mean Sum Square of the Error* (RMSSE) is another representation for the error and defined as the following:

$$RMSSE = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{q}(y_{ij} - \bar{y}_{ij})^2}{nq}} \qquad (3.22)$$

where $\bar{y}$ is the average value for the output.
Back Propagation is a learning method in which the internal weights and biases in neural network is adjusted by minimizing the RMSSE. The error is propagated backward through the network to adjust the weights and biases in order to make the actual response of the network closer to *desired* or the *target* response.
In this work the Levenberg-Marquardt (LM) method is used for minimizing the RMSSE and updating the internal parameter of the network. This method is an approximation to Newton's method based on the following :

$$\Delta W = (J^T J + \alpha I)^{-1} J^T E \qquad (3.23)$$

where J is the Jacobean matrix of derivatives of each error with respect to each weight, E is the error matrix, and α is a scalar. For larger α equation (3.23) approximates a gradient descent approach and for smaller α it approaches the Gauss-Newton method.

3.3.2.2 Calibration and Validation

Training an ANN model is actually updating the internal weights and biases by presenting the training input-output data set (i.e. calibration set) to the network, and minimizing the error in the output. The training set consists of a number of input-output batches, which is introduced to the network repeatedly.
Generally, the total number of model parameters, i.e. number of weights and biases, should not exceed the number of input-output data batches. If the number of internal parameters exceed the number of batches, there will be a possibility of obtaining an over-fitted model in which the model will show poor prediction ability and the model performance will not be satisfactory.
The number of neurons in both input and output layer are fixed upon the number of input and output variables respectively. Hence, there is only number of neurons in the hidden layers which will eventually determine the total number of model parameters.
If there are few input-output data set; i.e. few batches, available, there will be a maximum limit for the number of neurons which can be chosen for the hidden layer. This will naturally make the upper limit for the number of neurons in the hidden layer. The lower limit is of course only one single neuron. The optimum number of neurons in the hidden layer is determined by using the prediction ability of neural model through a validation procedure in which the number of neurons is determined by minimum prediction error in the validation.
Cross validation is used for the test of model performance. A separate set of test (i.e. validation) data is chosen and introduced to the network. The model is simulated by freezing the last internal parameters and calculation of predicted output and also the prediction variance is performed.
A recursive method can be used in order to determine the number of hidden nodes (i.e. neurons). We start with only one neuron in the hidden layer, train the network by using the training data set until the calibration variance (i.e. RMSSEC) is minimum or as low as

possible. Then we simulate the ANN model and perform the validation and calculate the prediction variance (i.e. RMSSEP). This will be continued by choosing 2, 3, and more neurons and plot RMSSEP versus number of nodes. We expect that RMSSEP decrease as the number of the neurons increase until a certain optimum number is found. This is displayed schematically in figure 3.2.

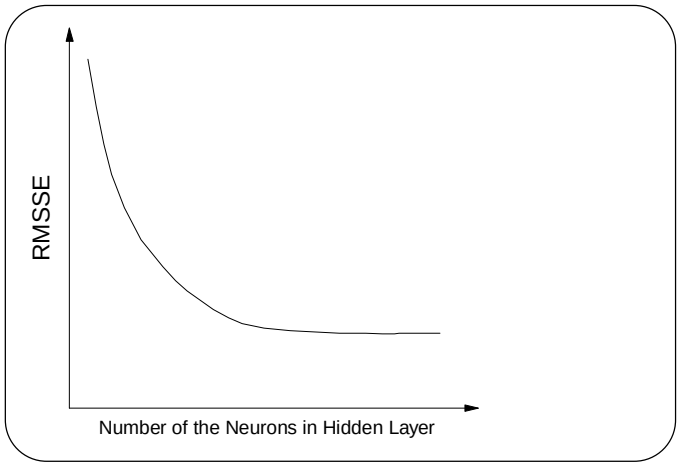

Figure 3.2 : RMSSEP versus number of neurons

An example of ANN modeling will be presented in the next subsection. This model is a quality prediction of final gasoline product after gasoline blending.

3.3.2.3 Example, Prediction of RON for Final Gasoline Product

A series of ANN models are developed for prediction of qualities of final gasoline products. These are prediction of RON, MON, RVP, benzene contents of the gasoline product, and prediction of D100 and D70 distillation points. D100 and D70 are percent gasoline evaporated at 100 and 70 degree Celsius respectively.
In this section we will present only one of them as an example which is prediction of RON for final gasoline product.
The process in gasoline blending unit is described in chapter two. The general principal in quality calculation of the gasoline product in this unit is a simple linear model based on the quality of the blend component. It is expressed mathematically as the following

$$Q = \sum_{i=1}^{n} v_i \, q_i \qquad (3.24)$$

where:

- Q is the quality of the product
- n is number of blend components
- v_i is the volume fraction of each blend component
- q_i is the corresponding quality of the blend component

With the exception of RON and MON all other qualities are directly calculated using equation (3.24). For calculation of RON and MON a nonlinear model is used since the octane quality

of the final product is a nonlinear function of qualities of blend components. Description of this nonlinear model is out of the scope of this thesis.
The objective of developing a model is to predict RON by applying ANNs techniques to cover the nonlinearity in the octane blending.

3.3.2.4 Model Structure and performance

The training data set contains data for 245 blends, which covers almost a year. The inputs to the model are 22 measurements of the flow rate and RON quality for 11 streams of blend components from the inventory tanks. The output is only one which the measured RON for the final gasoline product. Thus, the structure of ANN model is 1 neuron and 22 neurons in output and input layers respectively. The hidden layer consists of only two neurons.
It is interesting to compare the obtained ANN model with the calculated output using equation 3.24 which is a linear model.
Cross validation of the ANN model is performed. Figure 3.3 shows the result for comparison of measured RON, calculated by equation 3.24, and ANN predicted RON. Table 3.1 shows also the calculated average, and standard deviation for measured, calculated and, ANN predicted RON respectively. Furthermore, the prediction error, calculated as the difference between the measured and ANN predicted output, is shown in table 3.1.
This result shows that the ANN model is able to capture the nonlinear relation in RON prediction, since the variance of the prediction error from the developed ANN model is lower than the variance for both measurement and calculated linear model.

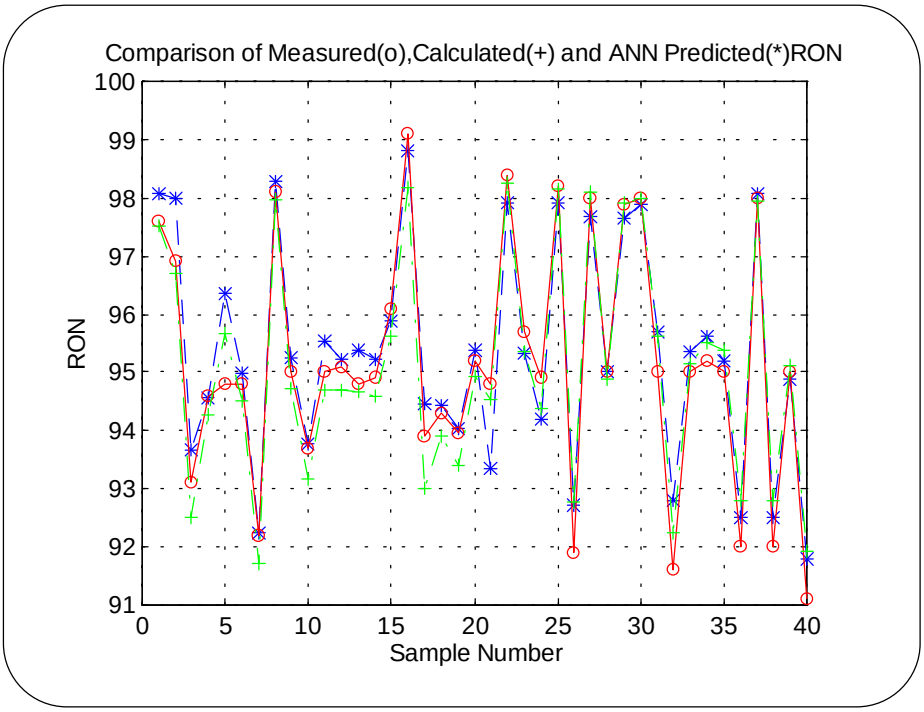

Figure 3.3 : Comparison of measured, calculated and ANN predicted RON

It is noteworthy to mention that the training data set cover production of all types of gasoline qualities from octane number 92 to 98 for the official Danish gasoline products. Hence, the standard deviation reported in table 3.1 is related to these three gasoline products. Standard

deviation for RON measurement at laboratory is 0.6, in which RON is measured by applying NIR techniques.

	Validation			
	Measured	Calculated	ANN Predicted	Prediction Error
Average	95.15	95.35	95.34	-0.0019
STD	2.06	2.17	1.90	0.0053
Max.	99.10	100.59	98.80	0.0144
Min.	91.10	90.43	91.80	-0.0155

Table 3.1 : Statistical data for measured, calculated and ANN predicted RON

3.3.2.5 Discussion

The described ANN model exhibit a good performance for prediction ability. In this case the system is static, in which there are direct, instantaneous, links between input output variables (Ljung, et al, 1994). The data used for the training of the models are not time series representation of the process. There is no dynamic behavior; i.e. change in state variables over the time, in the process . The time lag is a few seconds. These characteristics are important for a successful development of ANNs model as a static nonlinear model.

However, when the variables change with time, the system is dynamic and the described static ANN model will not work. The solution is to use a dynamic time-series model which is the subject of discussion in the following sections of this chapter.

3.3.3 Nonlinear Principal Component Analysis

3.3.3.1 Introduction

A *Nonlinear Principal Component Analysis (NLPCA)* model is proposed for reconciliation of data from a refinery naphtha splitter process. The NLPCA model is based on the well known method of *Principal Component Analysis (PCA)* used for dimensionality reduction in order to discover the significant variation in the data.
The proposed NLPCA model uses the inherent nonlinearity of Artificial Neural Networks (ANNs). The model is based on *Input Training Neural Network (ITNN),* in which the inputs is trained and adjusted along with the weight and biases of the network. Only one hidden layer is used in the internal structure of the neural network. When the ITNN model is properly trained, the trained input provides the nonlinear *factors*, which correspond to the principal components in the linear PCA model.
Input training network is based on *Autoassociative Network*, which consists of three hidden layers, i.e. a mapping layer, a bottleneck layer, and a demapping layer. Only the demapping part of the network is used in ITNN model
To achieve a better performance, the nonlinear PCA model starts from a linear PCA approach for initialization of the inputs to ITNN. The inputs, weights and biases of the network are then trained to reproduce the corresponding output pattern, which is the rectified data.

3.3.3.2 NLPCA

Nonlinear Principal Component Analysis (NLPCA) is used to uncover both linear and nonlinear significance variation in the data matrix, when nonlinear correlation exist among the variables. NLPCA has the same criterion of optimality as PCA, in which the sum of squared errors between the original variables and the NLPCA prediction is minimized.
The NLPCA method uses *Artificial Neural Networks* (ANNs). The nonlinear feature extraction can be performed by *Autoassociative* neural networks (Kramer, 1991). Autoassociative neural net is a feed forward network made up of three hidden layers; a *mapping layer*, a *bottleneck layer* and a *demapping layer* respectively. The dimensionality reduction is achieved in the hidden layer number two which has a small number of nodes. This method uses back propagation learning algorithm for training the network to perform identity mapping between the input and the output of the network.
Another method of NLPCA is an *Input Training Neural Network* (ITNN) proposed by Tan and Mavrovouniotis, 1995, in which only the demapping layer of Autoassociative neural network is used and the inputs are trained along with the network parameters. In a properly trained ITNN, the input layer provides the *nonlinear factors* or latent variables obtained from nonlinear dimensional reduction of the data.

3.3.3.3 Data Reconciliation

Data obtained from measurement of process variables are often noisy. In order to apply process measurements in modeling, control and optimization of the process, it is often necessary to rectify the data by performing a data reconciliation.
Traditional data reconciliation involved with minimization of the errors between the measured and the predicted variables from a rigorous mathematical model. This is in fact a nonlinear

optimization problem. Application of rigorous mathematical model is difficult for some chemical processes, especially for refinery processes in which the components and their compositions in the feed streams are unknown.
This is a strong motivation for using a statistical approach or neural network modeling for poorly unknown and highly nonlinear chemical processes.

3.3.3.4 Combining PCA and NLPCA

The purpose of this work is to use the concept of the NLPCA in order to perform data reconciliation of a refinery naphtha splitter process. A PCA model provides a first linear approach to determine the latent variables. The results from PCA is used as initialization for ITNN as a NLPCA to capture nonlinearity in the data pattern. The ITNN reproduce the inputs to the PCA in its output layer. Only one hidden layer is used in the ITNN model.
Using the information from PCA model the optimum number of the latent variables, i.e. number of inputs to ITNN, is determined.
A total number of fourteen variables are measured for the naphtha splitter process which implicitly represent the total mass and energy balance of the distillation column.
An ITNN is trained by back propagation using Levenberg-Marquardt learning method, which is an approximation to Newton's method.

3.3.3.5 Autoassociative Network

This method is used for identity mapping in which the network's inputs are produced at the output layer. The architecture of the neural network is made up of three hidden layers, as shown in figure 3.4. The first hidden layer is called *mapping layer*. The original data matrix is projected into the *feature* space, in which the output of the mapping layer represent the nonlinear principal components and therefore has f sigmoid nodes as the number of nonlinear PC's. These f nodes, containing sigmodal transfer functions, make the hidden layer number two which is called the *bottleneck* layer. Note that the number of nodes in the bottleneck is less than nodes in the mapping layer as a result of dimension reduction of the data. The third hidden layer is the *demapping* layer which represent the inverse mapping function and produce the reconstructed data in the output layer.

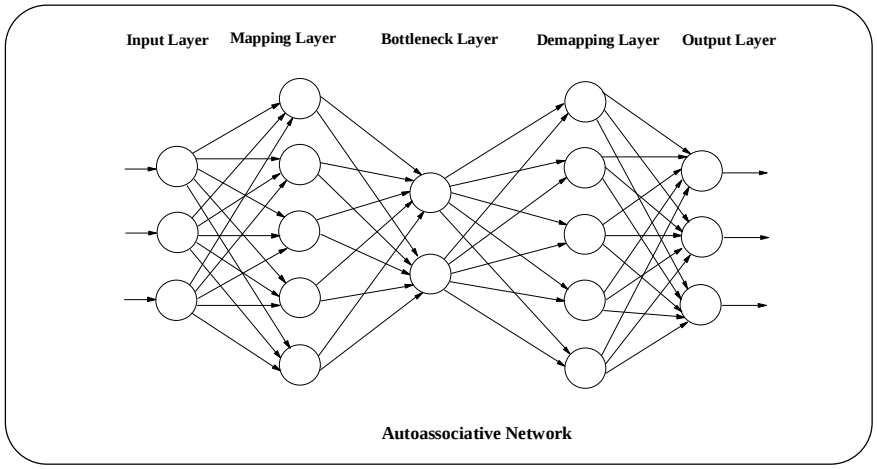

Figure 3.4 : An Autoassociative Neural Network.

The basic principal of the autoassociative neural network is analogous to the PCA. Based on equation (3.1) and (3.8) and using $P^TP = I$, the following equation can be written for the score matrix without loss of generality:

$$T = X P \tag{3.25}$$

In the nonlinear case, we are looking for score matrix T as a nonlinear function of the X as the following form:

$$T = G(X) \tag{3.26}$$

Cybenko (1989) has shown that a feed forward neural network with one hidden layer containing sigmodal transfer function can approximate any function with arbitrary accuracy.
Hence, the first layer in autoassociative neural network, i. e. the mapping layer, is used for approximate the G function in equation (3.26).

In analogy to the linear PCA, for the demapping of the data from the factor space, i. e. bottleneck layer, to the variable space the demapping layer of the network is used for approximation of the following H function:

$$\hat{X} = H(T) \tag{3.27}$$

n which the predicted X, i.e. the reconstructed data is produced in the output layer.

3.3.3.6 Input Training Neural Network

In input training neutral network only the demapping part of the autoassociative network is used. The input to the ITNN is trained by extending the back propagation algorithm to update the input as well as the network parameters, i. e. weights and biases. An example of ITNN architecture is shown in figure (3.5).
Tan and Mavrovouniotis (1995) have shown that training an ITNN with one input node and no hidden layer is equivalent to the linear PCA with one PC. Adding a hidden layer of sigmoid transfer function can basically capture both the linear and nonlinear variation in the data matrix and store the nonlinear PC's in the input layer.
Updating the input matrix is based on the extension of the back propagation learning algorithm. The steepest descent direction is derived as expressed in the equation (3.28) for minimizing the errors between the network's output and the desired output. Let use the same nomenclature as we used in section 3.3.2, and figure (3.1) for ANN modeling. Furthermore, let the desired output be data matrix Y of n samples and m variables, and the output of the network be Y2. The sum squared of errors is calculated by:

$$E = \sum_n \sum_m (Y2 - Y)^2 \tag{3.28}$$

The steepest direction for updating the new inputs X matrix is:

$$\Delta X = -\frac{\partial E}{\partial X} = -2 \sum_m (Y2 - Y) \frac{\partial Y2}{\partial X} \tag{3.29}$$

In this model, which we have linear nodes in the input and output layers and sigmodal nodes in the hidden layer, the output of the network Y2 is calculated as the following:

$$Y2 = \phi(W1 \cdot X + B1) \cdot W2 + B2 \tag{3.30}$$

where ϕ is sigmodal transfer function as defined in equation (3.18). The output from the hidden layer A1 is calculated as follows:

$$A1 = \phi(W1 \cdot X + B1) \tag{3.31}$$

The first derivative of a sigmoid function of the form ϕ [f(x)] can be calculated by the following equation :

$$\frac{\partial \phi[F(X)]}{\partial X} = \frac{\partial F(X)}{\partial X} \phi[F(X)] \{1 - \phi[F(X)]\} \tag{3.32}$$

Combining equations (3.29) through (3.32) yields:

$$\Delta X = -2 [W1^T \cdot A1 \cdot * (1 - A1) \cdot * (W2^T \cdot e)] \tag{3.33}$$

where the error e is equal to (Y2 - Y).
ITNN shows good ability of data rectification and converges much faster than autoassociative networks.

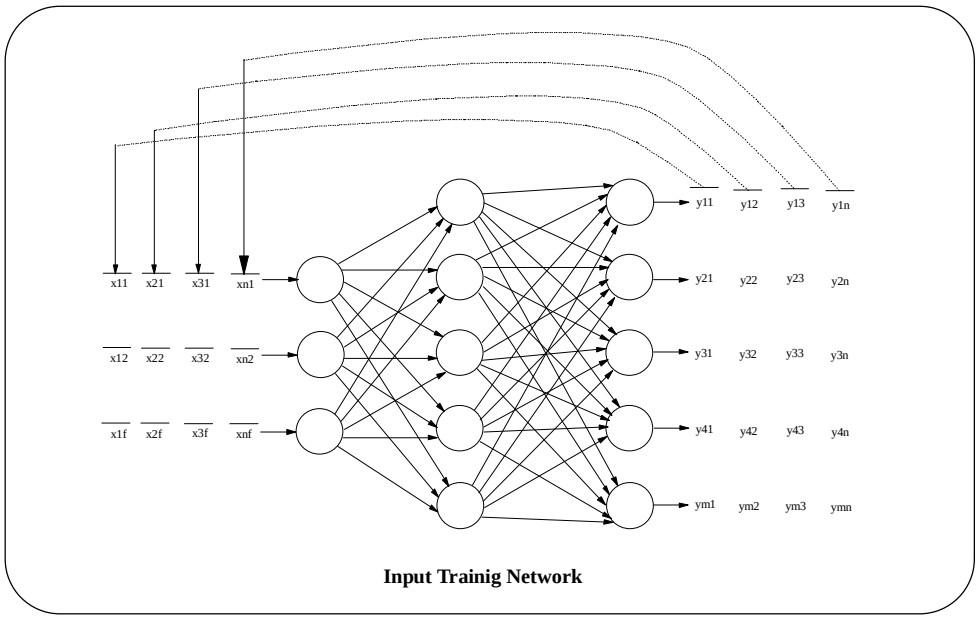

Figure 3.5: A typical Structure of Input Training Neural Network.

3.3.3.7 Combination of Linear PCA and ITNN

To obtain better and faster result, linear PCA and NLPCA is combined in one model.
The data matrix is first mean centered, i.e. the columns in the original data matrix are subtracted from their mean values, and then variance scaled, i.e. the columns are divided by their standard deviation. The data matrix which is mean centered and scaled to unit variance is also called *autoscaled* data.

The autoscaled data is then used for a PCA model. The number of PCs used in the model is based on the percentage captured variance by each PC.
The score matrix T from the PCA model is used for initialization of the input matrix X to the ITNN, as is shown in figure 3.6. Then the network is trained by adjusting the input, weights, and biases of the network.

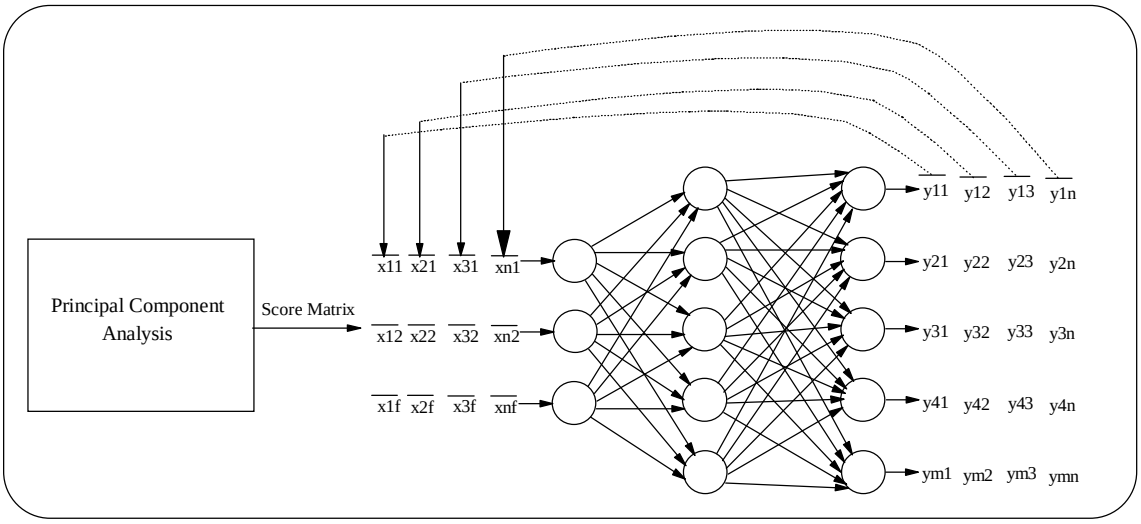

Figure 3.6: Combining PCA and ITNN.

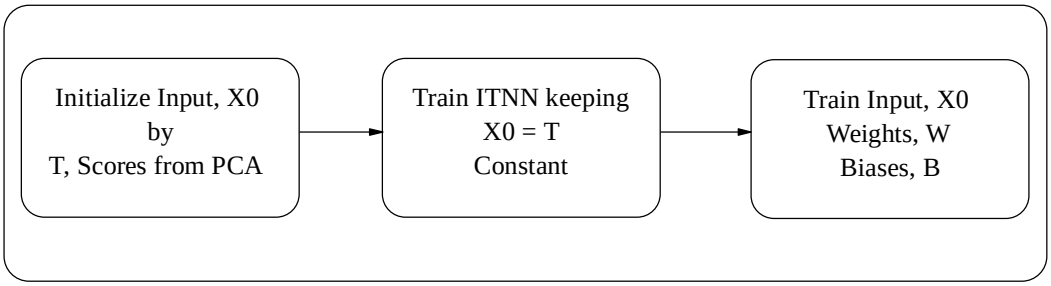

Figure 3.7: The Three steps in combining PCA and ITNN.

Training an ITNN by applying PCA initialization is carried out in three steps. First; the initial inputs is set equal to the score matrix from a PCA model. Second; the ITNN network is trained by freezing the inputs and updating weights and biases. Third; the network is trained by updating inputs, weights, and biases. This procedure is summarized in figure 3.7.

3.3.3.8 Example; Rectification of Splitter Data

ITNN model is used as a NLPCA method for rectifying data obtained from naphtha splitter distillation column. The process diagram is shown in figure 3.18. A total number of fourteen variables are measured around the column. These are listed in Table 3.2. The flow rate of the feed stream, distillate, and bottom product can be used for a total mass balance. A small amount of gas will be produced at the top of the column if the light gases are not completely removed by stabilizer distillation column before the splitter.
There is no measurement for the flow rate of the gas at the top. However, the total mass balance of the column can be approximately estimated by using feed, top product and bottom product flow rates.

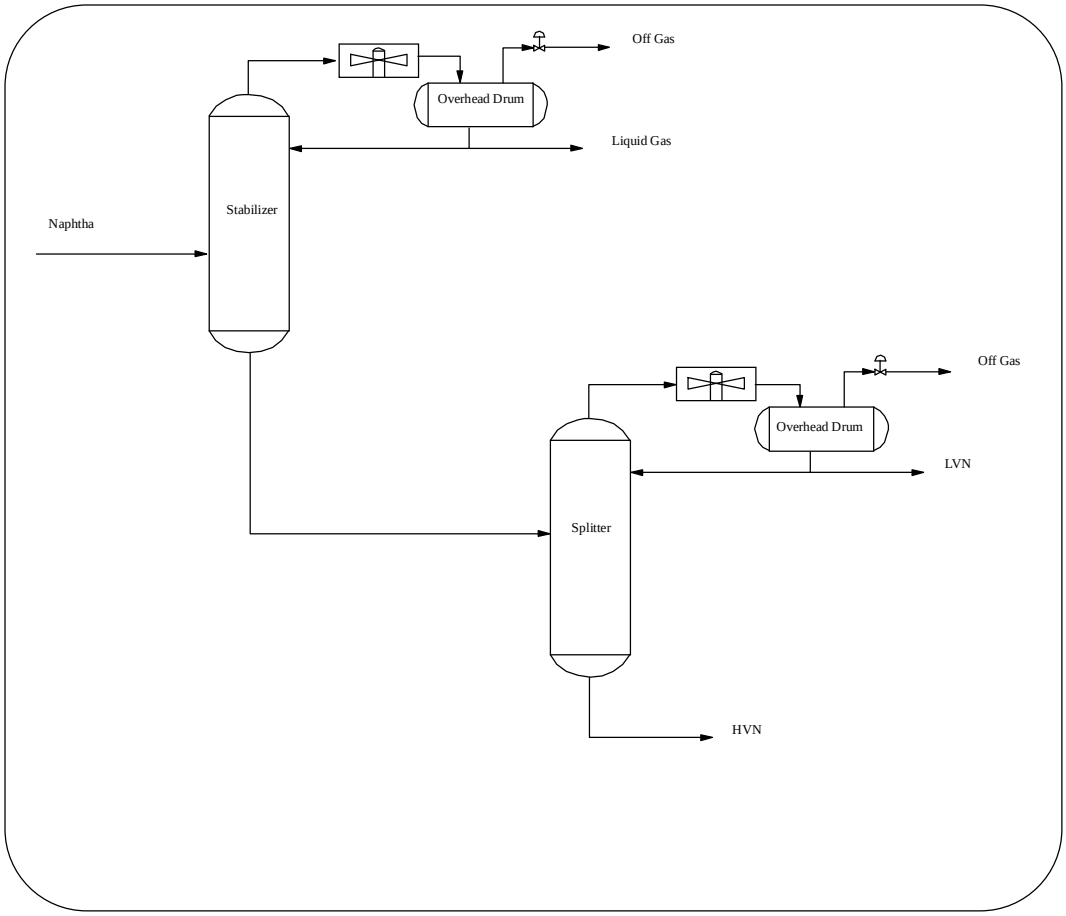

Figure 3.8: A schematic diagram of stabilizer/spiltter system.

No.	Description	Tag	Unit
1	Top Temperature	TT	^{0}C
2	Tray 23 Temperature	T23	^{0}C
3	Tray 18 Temperature	T18	^{0}C
4	Tray 9 Temperature	T9	^{0}C
5	Tray 3 Temperature	T3	^{0}C
6	Reflux Temperature	RT	^{0}C
7	Reflux Flow Rate	RF	m^3/hr
8	Feed Temperature	FT	^{0}C
9	Feed Flow Rate	FF	m^3/hr
10	Top Pressure	P	Bar
11	Top Product Flow Rate	LVN	m^3/hr
12	Bottom Product Flow Rate	HVN	m^3/hr
13	Reboiler Duty	QR	MW
14	Naphtha Cut Point	CP	^{0}C

Table 3.2: Description of the variables.

There are five measurements of temperature profile inside the column. Besides, temperature of the reflux and the feed streams are measured. These variables along with a calculated reboiler duty, i.e. QR, can represent the energy balance of the column.
Hydrocarbon components and the composition of the components in the feed stream are unknown. Generally, naphtha is a hydrocarbon mixture with a true boiling point range of between 30-40 ^{0}C to around 150-180 ^{0}C. A calculated naphtha cut point variable, which is a pressure corrected temperature of the naphtha product inside the atmospheric crude distillation column, can be used for a rough characterization of the naphtha stream produced in the column.
A set of data containing total number of 573 samples each with 14 measurements, with a sampling interval of one hour, is chosen for the NLPCA model of the naphtha splitter process. The data correspond to almost 24 days of operation. Figure 3.9 shows the variables value vs. sample number. It is obvious from the figure that the column was operating under different operation conditions during that period of sampling.

3.3.3.8.1 PCA model

Using the information from scores for PC1 and PC2, as shown in figure 3.10, we can detect three major clusters of data that represent three operation regions. These regions can be explained by two *pseudo steady states*, and one *transient state*, which is the transient from pseudo steady state region number one to the number two.
We define a *pseudo steady state* to be the state of the operation in which the changes in state variables are in relatively lower frequency. We can recognize two clearly pseudo steady state regions in the data matrix shown in figure 3.7.
We can roughly assume that the data from sample number 50 to 280 cover the pseudo steady state number one and the data from sample number 360 to 573 cover the pseudo steady state number two and the rest belong to the transient region. Hence, we split the data in three parts and focus on pseudo steady states.
The first step in NLPCA modeling is initialization of the inputs by a linear PCA, as it is shown in figure 3.7. Choosing the number of factors, or number of PCs is an important issue. Table 3.3 shows the percent variance captured by each PC.

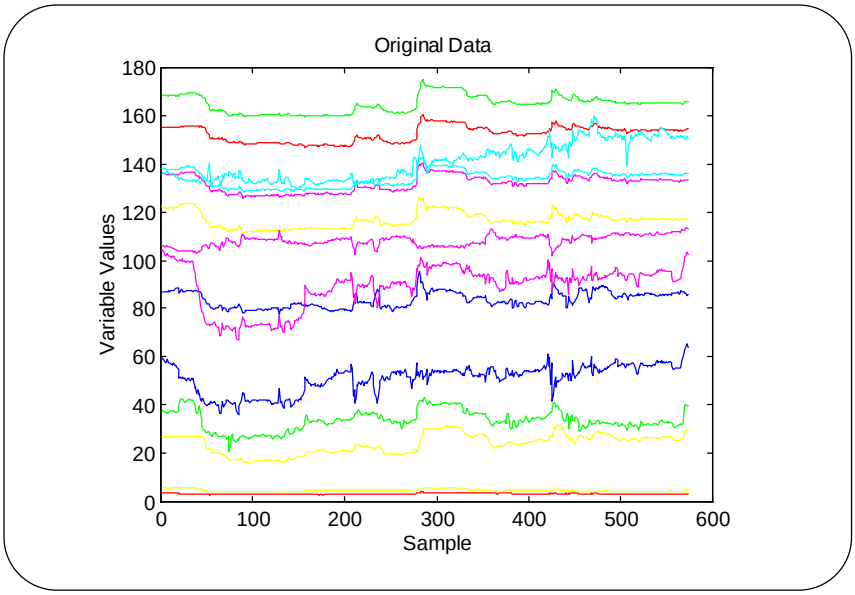

Figure 3.9: The Original Data

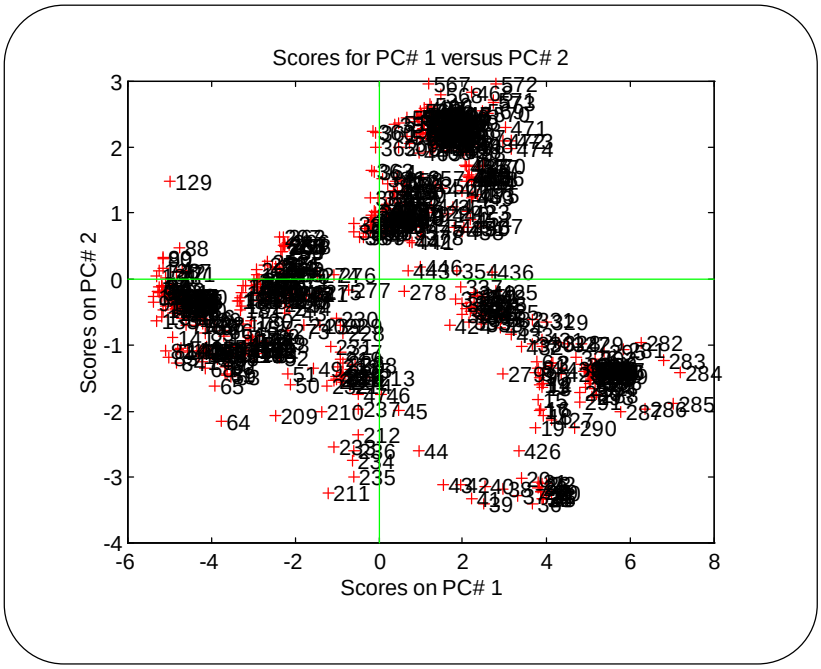

Figure 3.10: Scores for PC1 vs. PC2.

Percent Variance Captured by PCA Model

Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	7.36e+000	52.59	52.59
2	3.19e+000	22.76	75.35
3	1.72e+000	12.29	87.64
4	9.01e-001	6.44	94.07
5	5.99e-001	4.28	98.35
6	9.68e-002	0.69	99.04
7	5.26e-002	0.38	99.42
8	3.91e-002	0.28	99.70
9	1.44e-002	0.10	99.80
10	1.10e-002	0.08	99.88
11	8.46e-003	0.06	99.94
12	5.31e-003	0.04	99.98
13	1.95e-003	0.01	99.99
14	9.92e-004	0.01	100.00

Table 3.3: Percent Variance Captured by PCs.

If we choose too many factors we achieve a model close to full model in which the noise and the structure part are not separated and we have still a significant amount of noise in data. However, if we choose too few factors we lose a part of information, probably both the linear and nonlinear information, left in the noise part.

Since we are going to use the scores for initialization of the inputs to the NLPCA model, we choose five PCs in order to include nonlinear information.

3.3.3.8.2 ITNN Model

One of the important issue in ANN modeling is the internal architecture of the network, i. e. choosing the number of nodes in the hidden layer. The original data set is used to check the performance of the network for different inputs, i.e. number of factors, which is indeed number of the nodes in the hidden layer in ITNN. The experiences, so far, has shown that, for this application, choosing more nodes in the hidden layer will not improve the performance. However, increasing the number of factors can significantly reduce the network error.
It is important to remember that the number of the internal parameters of the network, i.e. weights and biases, should be less than number of the samples. Number of the internal parameters N_E is defined as follows:

$$N_E = f*S1 + S1*S2 + S1 + S2 = \text{Number of Internal Parameters}$$

where f, S1, S2 are number of nodes in the input, hidden, and output layers respectively.
When N_E is larger than the number of samples, ANN model may results in an "over-fitted", or "over-parametrized" network, which is poor in generalization characteristic. By using *Root Mean Sum Squared Error* (RMSSE) as defined in equation (3.22), we can compare the performance of the network for different factors (f) and number of nodes in the hidden layer (S1) as shown in table 3.4. As we expect, the RMSSE decreases, both in linear and nonlinear models, for increasing the number of nodes in input and hidden layers.

3.3.3.8.3 Results

As it is shown in table 3.4, a number of 5 principal component is the optimum choice in this example.
The transient region is omitted from the data, and we focus on the pseudo steady states. Just for the matter of curiosity we develop model for each pseudo steady states separately, and then combine these two in one model. Hence, two models are developed for the two pseudo steady states. These are called model no. 1 and model no. 2 respectively. Additionally a third model is developed by training an ITNN using a training data set which is a combination of the training data for model no. 1 and 2. This third model is just called model no. 3.

PCA		NLPCA		No. of PC f	S1	N_E
Training	Test	Training	Test			
1.890	0.378	0.378	0.135	2	2	48
0.881	0.330	0.236	0.126	3	3	68
0.776	0.258	0.191	0.094	4	4	90
0.627	0.221	0.167	0.086	5	5	114
0.448	0.185	0.160	0.129	6	6	140

Table 3.4: Comparison of the RMSSE for different number of nodes in hidden layer.

The results of the obtained RMSSE for the three models is summarized in table 3.5.

Model N0.	PCA		NLPCA		f	S1	N_E	n
	Training	Test	Training	Test				
1	0.336	0.109	0.202	0.060	5	5	114	125
2	0.252	0.112	0.130	0.110	5	5	114	141
3	0.420	0.196	0.115	0.083	5	5	114	255

Table 3.5: Comparison of the RMSSE for the pseudo steady state regions.

As shown, RMSSE for model number 3, i.e. the model valid for both pseudo steady state regions, is less than the others.
The results for comparison of the NLPCA and the measured data for all 14 variables are shown in appendix M.

3.3.3.9 Discussion

The essential objective of applying nonlinear PCA is to rectify the data obtained from the process which is used in process models and quality prediction models.
A linear PCA is used for assessment of the number of nodes in the input layer which is the number of factors used in the NLPCA model. Besides, the scores from PCA are used for initialization of the inputs to NLPCA. The NLPCA model is first trained by keeping the inputs constant equal to the scores from linear PCA, and then trained further by updating both the inputs and the network's parameters.
As it can be seen from the results shown in the appendix, the NLPCA model is able to reconstruct the original data. For individual variables, such as for the top product flow rate LVN, the models show some deviation for the original measured data. Model no. 3 shows generally better results, since the training data set cover a larger area and contains different steady state operation regions.
What is interesting for the future work in this field is to apply NLPCA method to trace the transient state from the data matrix. For optimization and control objectives, it is important to be able to automatically detect the transient state operation of the column.

3.4 Dynamic, Linear Methods

3.4.1 Time-series Model

System identification deals with the problem of building mathematical models of dynamical system based on observed data from the system. The characteristic of the a dynamic system is that the variables change with time, or current output value depends not only on the current external stimuli but also on their earlier values. Output of dynamical systems whose external stimuli are not observed are often called time series (Ljung, 1987, and 1994).
In this section a description of ARX (Auto Regressive with Exogenous input) method in system identification. This is basically a linear, time series regression model.
The prediction models developed in this work apply ARX model extensively. These models are described in chapter in the following chapter of this thesis.

3.4.2 Model Structure

The main concept of the modeling work is to use different methods to develop models based on input-output mapping of data by fitting the model parameters. Following the terminology and mathematical formulation presented by Ljung (Ljung, 1987), we are seeking the mapping from the data set :

$$Z^N = [u(1),\ y(1),\ .....,\ u(N), y(N)] \tag{3.34}$$

to the parameter estimate $\hat{\theta}_N$ as the following:

$$Z^N \rightarrow \hat{\theta}_N \quad \in D_M \tag{3.35}$$

in which N is a finite number denoting the dimension of data set, and D_M is a set of values over which θ ranges in a *model structure M*. A model structure is a parametrized set of models defined in equation 5.9.
In a general formulation linear time-invariant models are defined as the following:

$$y(t) = G(q, \theta)u(t) + H(q, \theta)e(t) \tag{3.36}$$

in which the G and H are functions of θ and q is the backward shift operator. Moreover, {e(t)} is a sequence of independent random variables with zero mean values and variance λ.
The extent of parameter vector θ ranges over a subset of $\boldsymbol{R}^d$ in which d is the dimension of θ. Hence, the model presented in 3.36 is no longer a model, but a set of models obtained from different values of θ.
Specification of the functions G and H will lead to a particular model. A suitable method is to choose a structure that permits the specification of G and H in terms of finite number of numerical values, for instance rational transfer functions or finite dimensional state-space descriptions.
Parametrization of G and H functions in terms of linear difference equations will lead to model structure like ARX and ARMAX.

A linear difference equation is a simple description of input-output relationship. The model expressed mathematically as the following equations.

$$\begin{aligned} &y(t) + a_1 y(t-1) + ... + a_{na} y(t-na) \; = \\ &b_1 u(t-1) + ... + b_{nb} u(t-nb) + e(t) + c_1 e(t-1) + ... + c_{nc}(t-nc) + D \end{aligned} \tag{3.37}$$

in which:
Y(t) is output measurement at time t,
U(t) is input measurement at time t,
e(t) vector of white noise sequences,
na is number of A parameters,
nb is number of B parameters,
nc is the number of C parameters,
D is constant vector.

The model formulation in 3.37 includes the moving average of white noise. This type of model is also called *equation error* model, since the white noise term is directly added in the difference equation.
The adjustable model parameter are :

$$\theta \; = \; [a_1 \; a_2 \; ... \; a_{na} \; b_1 \; b_2 \; ... \; b_{nb} \; c_1 \; c_2 \; ... \; c_{nc} \;] \tag{3.38}$$

The backward operators are defined as the following:

$$\begin{aligned} A(q) \; &= \; 1 + a_1 q^{-1} + ... + a_{na} q^{-na} \\ B(q) \; &= \; b_1 q^{-1} + ... + b_{nb} q^{-nb} \\ C(q) \; &= \; 1 + c_1 q^{-1} + ... + c_{nc} q^{-nc} \end{aligned} \tag{3.39}$$

Introducing the backward operator in the model defined in 3.37 will lead to the following formulation of the model as in 3.40 :

$$A(q)y(t) \; = \; B(q)u(t) \; + \; C(q)e(t) \tag{3.40}$$

Notice that the model in 3.37 correspond to the model defined in 3.36 by the following:

$$G(q,\theta) \; = \; \frac{B(q)}{A(q)} \qquad\qquad H(q,\theta) \; = \; \frac{1}{A(q)} \tag{3.41}$$

The ARX model is a special case of ARMAX model in which $C(q) \equiv 1$, when nc = 0.
It can be shown that the predictor for the ARX model can be defined as equation 3.41 (L. Ljung, 1987).

$$\hat{y}(t|\theta) \; = \; B(q)u(t) \; + \; [1 - A(q)] \, y(t) \tag{3.42}$$

Introducing the regression vector as the following :

$$\varphi(t) \; = \; [-y(t-1) \; ..... \; -y(t-na) \;\; u(t-1) \; ..... \; u(t-nb)]^T \tag{3.43}$$

Then equation 3.42 can be expressed as a linear regression model:

$$\hat{y}(t|\theta) = \theta^T \varphi(t) = \varphi^T(t)\,\theta \quad (3.44)$$

At time t we can evaluate how good this prediction is by calculating the prediction error:

$$\varepsilon(t,\theta) = y(t) - \hat{y}(t|\theta) \quad (3.45)$$

The model parameters are estimated by solving the following optimization problem:

$$\hat{\theta}_N = \arg \min_{\theta} \frac{1}{N} \sum_{T=1}^{N} \varepsilon^2(t,\theta) \quad (3.46)$$

This optimization problem is then solved by using a Least Squares approach. Hence, a set of optimal parameters is determined.

3.4.3 ARX model with PLS Regression

The normal procedure in estimation of parameters in ARX model is based on the least squares (LS) method minimizing the prediction error defined in equation 3.46. Another approach is to apply a PLS in parameter estimation of ARX model in order to take advantage of PLS ability to extract the useful information from collinear, noisy, input data which is relevant for modeling the output prediction.
An approach is to construct the regression vector defined in 3.43 considering number of na, and nb parameters in order to define the problem as a linear regression problem as described in equation 3.44. The regression problem can be then solved by using linear PLS regression.
The advantage of this method is that a linear time-series model can be developed by a PLS regression in which the variation and the data structure in the Y variables is directly used in PCA decomposition of the X variables.
Furthermore, in practical application in process industry there may be a lot of variables which may theoretically related to the output variable but the collected data shows no correlation due to corrupting influence of noise or effect of feed-back control. Applying PLS will use the strength of PCA in dimensional reduction of the data set and hence an effective modeling of output.

3.5 Dynamic, Nonlinear Methods

An approach for modeling a dynamic nonlinear system is based on applying nonlinear methods in time-series type of models.
As described earlier, the concept of ARX type of model can be used to define a linear regression problem as described in equation 3.44 by constructing the regression vector defined in 3.43.
A nonlinear PLS model, as described in section 3.3.1, can then be applied in order to estimate the regression parameters in the nonlinear case.
This approach is basically the same as in the case of linear ARXPLS described in section 3.4.3 in which the inner relationship in PLS is defined by a nonlinear function.
The nonlinear function can a polynomial of arbitrary order as it is expressed in equation 3.17, or alternatively using a neural network model.

3.6 Model Validation Criteria

The purpose of model validation is to test the performance of a developed model in order to assess the level of predictability of the model in the operation region of interest.
A common and natural method of validation is to simulate the model, which is developed in calibration, by using a separate data set, and compare the model predicted with the measured output.
The sperate data set called test set or validation set. It is very important that the validation data set is closely comparable to the calibration data set, with respect to sampling time, sampling condition. It is important that the validation data set is representative for the target population. The only difference between the calibration and validation should be the *sampling variance*. This sampling variance will comprise those differences between the two data sets that can *only* be explained by the two different samplings of n objects, made under identical conditions. The idea behind the model validation is to evaluate the prediction strength of the model on data with different noise than the calibration set.
There are certain criteria in the validation to be satisfied. The first criterion is the level of prediction error in validation data set, as described in the following.

3.6.1 Definition of Reference Model in Validation

One way to evaluate the performance of the model is to compare the model RMSSE defined in equation 3.22 with a reference or a pre-defined criterion.
A suitable reference which is normally used in assessment of model validation is the variance of measured output in validation data set. Comparison of the calculated standard deviation for the measured output and the RMSSE defined in 3.22 will give a measure of predictability of the obtained model. We shall illuminate the concept of this comparison further in the following.
If we use the average value of the measured output and draw an average line through all the output values, then we will have a model described by 3.47.

$$\hat{y}(t) = y_{AVG} + e(t) \qquad (3.47)$$

We shall call this model as the *average-model*. It is obvious that the purpose of the modeling is to predict the output much better than the described average model, otherwise the average

value can be used as an estimate for the future value of the output and development of a prediction model is not necessary.
This average reference model is computed by first calculating the average of all N measured output values, and then subtract the average from the output itself to calculate E_{AVG} , as the following:

$$(E_{AVG})_i = y(t_i) - y_{AVG} \quad (3.48)$$

Then, we compute a RMSS of this error by using equation (3.22), and denote it as RMSEAVG for the average-model described in 3.47. It is clear that the RMSEAVG has the same property as the standard deviation of the measured output.
We expect that the developed prediction model should predict a set of output values for a period of time which are closer to the measured output than the average value. In this sense we say that the developed model should be at least better than the average-model in order to accepted.

A second reference model can be defined as the following. Let consider a model structure of 3.44 in which the number of A-parameter is 1, i.e. na=1. Furthermore, consider that the developed prediction model find a set of B-parameters which are close to zero, and an A-parameter value close to one This is shown in the following equation:

$$\hat{y}(t) = y(t-1) + 0 \quad (3.49)$$

This means that the new prediction of y is equal to the previous y. In this case we have no effect of input variables. We shall call this as *zero-model*.
Based on this consideration, we compute a E_{ZERO} as the following in a general form:

$$(E_{ZERO})_i = y(t_i) - y(t_{i+1}) \quad (3.50)$$

Hence, a RMSS of E_{ZERO} , which is denoted by RMSEZRO will give os a reference in assessment of predictability of the obtained prediction model. Thus, the expectation is that a model with good performance characteristic should be better than the zero-model, meaning that the developed model has captured the effect of input variables.

3.7 Persistence of Excitation

One of the important issue in dynamic modeling of a physical system concern with the characteristics of the observed process data. The choice of input has a very substantial influence on how much the obtain data is representative and informative for the task of dynamic modeling. The input signal contains valuable information about the operating point and determine which part and mode of the system is excited during the period of model calibration. In the following more specific definition of informative data set and concept of persistence of excitation is presented.

3.7.1 Definition of Informative Data Set

As described in section 3.4.2, a set of linear time-invariant models can be defined, as expressed in equation 3.51, in order for input-output mapping of a set of data Z^N by fitting the model parameters θ_N, in which N is a finite number denoting the dimension of data set.

$$y(t) = G(q,\theta)u(t) + H(q,\theta)e(t) \qquad (3.51)$$

The functions G and H can be specified by rational transfer functions or finite dimensional state-space descriptions. By using linear difference equations, in order to perform a parametrization of the functions G and H, a set of model structure of ARX and ARMAX, as it is discussed earlier in section 3.4.2. Hence, the general formulation defined in 3.51 will lead to a set of model structure M obtained from different values of parameter vector θ. Number of the models that can be obtained is thus a subset of N.
The purpose of model fitting is thus to find the optimal solution of the optimization problem defined in 3.46.
If the data set Z is capable of distinguishing between these different models in the model set M, then we call the data set to be *informative* enough with respect to the model set. The assumptions here are that the data set Z is quasi-stationary and the models are linear time-invariant.
A more mathematical definition of informative data set is given by Ljung (Ljung 1987). A quasi-stationary data set is informative if the spectrum matrix :

$$Z(t) = [\ u(t)\ \ y(t)\ ]^T$$

is strictly positive definite for all ω. The spectrum matrix is defined as:

$$\Phi_z(\omega) = \begin{bmatrix} \Phi_u(\omega) & \Phi_{uy}(\omega) \\ \Phi_{yu}(\omega) & \Phi_y(\omega) \end{bmatrix} \qquad (3.52)$$

The concept of informative data is closely related to the concept of persistently exiting inputs, described in the following.

3.7.2 Concept Persistence of Excitation

One of the important aspect of choosing input variables is the second-order property of u, such as $\Phi_u(\omega)$, i.e. the spectrum of the input, and the cross spectrum $\Phi_{ue}(\omega)$ between input and the driving noise.
Let assume that the data Z is collected in an open loop experiment. Consider a quasi-stationary input signal u(t), with spectrum $\Phi_u(\omega)$, and the following filter:

$$M_n(q) = m_1 q^{-1} + ... + m_n q^{-n} \qquad (3.53)$$

The definition of *persistence of excitation* is based on the following result obtained by Ljung.

$$\left|M_n(e^{i\omega})\right|^2 \Phi_u(\omega) \equiv 0 \qquad (3.54)$$

The definition is that the input signal u is said to be *persistently exiting of order n* if for all filter $M_n(q)$ the relation 3.54 implies that $M_n(e^{i\omega}) \equiv 0$.
The direct result of this definition is that if $\Phi_u(\omega)$ is *different from zero* at least n points in the interval of $-\pi>\omega>\pi$, then the input signal u is persistency exciting of order n.
Moreover, $\left|M_n(e^{i\omega})\right|^2 \Phi_u(\omega)$is the spectrum of the signal :

$$v(t) = M_n(q)u(t)$$

Hence, the input u that is persistency exiting of order n can not be filtered to zero by a moving average filter. Consequently, there must exist a set of θ parameters that give a set of different and distinguish models due to informative characteristic of input signal. On this basis we say the data collected under open-loop control is informative if the input is persistency exciting.
It is useful to consider a more general definition as the following:
A quasi-stationary input signa u(t), with spectrum $\Phi_u(\omega)$ is said to be persistently exciting if :

$$\Phi_u(\omega) \rangle 0 \qquad \textit{for all } \omega$$

In the result and the definition above, it is assumed that input signal is collected from an open-loop experiment. However, closed-loop control is applied widely in the process industry. The impact of closed-loop control on persistence of excitation of input is discussed in the following.

3.7.3 Effect of Closed-loop Control

In practical application of input-output modeling in process industry, the input data is normally collected under output feedback. The reason for the feedback control configuration is mostly for production economy and plant safety. Most often, it is not simply allowed to manipulate the system in process industry in order to perform a set of experiments to insure an informative and excited input signal.
The information obtain from a process in a closed control can be defective for modeling the output even if the input is persistency excited. In order to illuminate this, consider the following example.
Let us consider a close-loop control configuration as the example shown in figure 3.11. Assume that we have the following first-order model structure:

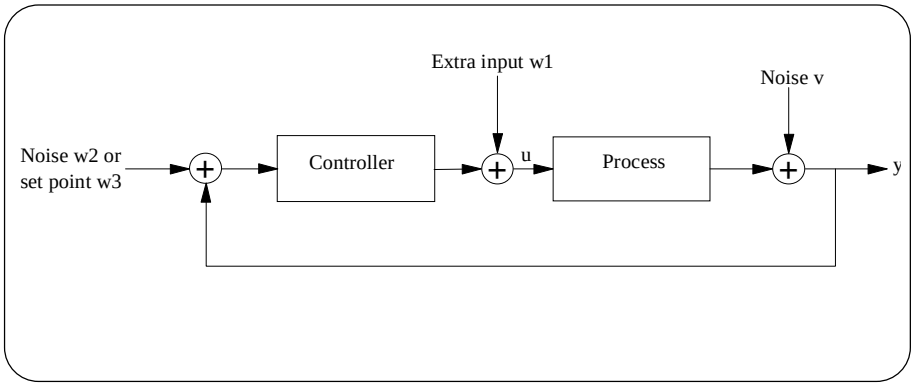

Figure 3.11: A typical closed-loop control.

$$y(t) = ay(t-1) + bu(t-1) + e(t) \tag{3.55}$$

and assume that the controller is a proportional regulator :

$$u(t) = fy(t) \tag{3.56}$$

Inserting 3.56 into 3.55 will give:

$$y(t) = (a + bf)\, y(t-1) + e(t) \tag{3.57}$$

which is the model obtained under feedback.
Now, consider the following set of $(\hat{a}, \hat{b})$ parameters, in which α is an arbitrary scalar:

$$\begin{aligned} \hat{a} &= a + \alpha f \\ \hat{b} &= b + \alpha \end{aligned} \tag{3.58}$$

It can be seen that all models that can be obtained by parameters in 3.58 will give the same description of the system as the models by parameters (a, b) in the closed-loop control. This will lead us to the conclusion that no matter the value of proportional control f, there is no way to obtain two distinguishable models from these sets of model parameters. The information obtained from this system by applying 3.55 is thus not informative enough.
Notice that this result i a valid even for an excited input u, and hence the persistency exciting of input is not a sufficient condition in closed-loop data.
Moreover, if we restrict the model 3.55 by letting parameter b to be equal to one, then the information generated by 3.55 with b=1 will be informative enough to distinguish different value of a-parameters.

However, there is chance to get informative information from a closed-loop system, if the regulator is noisy, nonlinear, time-varying or complex high-order. If it is allowed, a certain complexity can be added to a closed-loop system by adding an extra input as it is shown in figure 3.11. The input is now a sum of the output feedback plus the extra input, as equation 3.59.

$$u(t) = F_i(q)y(t) + K_i(q)w(t), \qquad i = 1, 2, ..., r \tag{3.59}$$

where $F_i(q)$ and $K_i(q)$ are linear filters.
The impact of these linear filters and the extra input is that any high frequency contribution to the signal spectrum, that is produced by changing filters F and K, can be neglected.
This can be realized as an extra input w1, noise w2 to the regulator, or set point changes w3, as it is shown in figure 3.11.

3.8 Summary

In this chapter a review of methods used in process chemometrics is presented. It is attempted to present the essence of process chemometrics in order to provide the theoretical back ground for the multivariate modeling techniques applied to develop process models in this thesis

Based on the definition of process chemometrics, the methods in model development are based on data obtained from the system, and the purpose is to develop an empirical model for estimation of one or more properties of the system.

Process chemometrics includes both linear and nonlinear approaches, and consider the static and dynamic characteristics of the system. Based on this consideration, the methods in process chemometrics are divided in four general categories according to the linear, nonlinear, static, and dynamic characteristics of the system under study.

In the class of static linear methods *Principal Component Analysis* (PCA), *Principal Component Regression* (PCR), and *Partial Least Squares Regression* (PLS) are discussed. PCA is used in data assessment, dimensional reduction through extracting the latent variables and applied mostly for process monitoring. PLS and PCR are used for developing input-output regression models.

In the class of static nonlinear approaches *Artificial Neural Networks* (ANNs) exhibit a strong ability to nonlinear functional approximation. Nonlinear PLS regression in which nonlinear function is defined for the inner relationship of the PLS is another approach in this class of chemometrics methods. Furthermore, a description of *Nonlinear Principal Component Analysis* (NLPCA) is presented. A NLPCA model is developed based on *Input Training Neural Networks* (ITNN) which is used for data rectification

The methods in the class of dynamic linear methods include knowledge based predictive modeling using linear time series regression. The linear methods include ARX and ARMAX (Auto Regressive Moving Average with Exogenous input), which are linear models based on parametric input output representations

The dynamic nonlinear, in which the time-series type of model can be integrated in a nonlinear PLS model, is discussed. Furthermore, ANNs model can be applied for estimation of inner relationship of an ARXPLS model.

A discussion about different criteria in model validation is presented in this chapter, in which two different reference model, i.e. average-model and zero-model, are presented in order to assess the predictability of the developed chemometric model

The concept of informative data set and persistence of excitation are discussed along with the issue concerning the impact of closed-loop control on persistence of excitation of input.

Chapter 4

Introduction to Model Development

4.1 Introduction

4.1.1 Purpose

The purpose of this chapter is to describe the preliminary steps in the model development work in this thesis. This introduction concerns mainly with definition of the system limit, description of the output and selected input variables, assessment of data, data scaling and sampling, and description of data treatment.
Furthermore, it is attempted to present a general scope of model development and to describe the general procedure and different steps in the chemometric approach of modeling. The described procedure in this chapter can be used as guidelines to model development.

4.1.2 Outline

In section 4.2 a general description of model development phases is presented. The review of the steps in the development procedure presented in section 4.2 can be used as guidelines for chemometric modeling. A more detailed process description and definition of system

delimitation is presented in section 4.3. This will introduce a clear view of the influence of variables in different production units on the interesting quality variables and allow the reader of this thesis to follow the description of input output variables in the following sections. In section 4.4 and 4.5 description of output and input variables are presented respectively. The name and description of the variables will be unique in this thesis. These variables are used in the developed model described in chapter 5. Section 4.6 deals with selection of suitable sample interval in which the problem concerning sample frequency is discussed. In section 4.7 a PCA analysis is presented for data obtained from catalytic reformer I, catalytic reformer II, and isomerization unit. The conclusion for this chapter is presented in section 4.8

4.2 Description of Different Steps in Model Development

Development of multivariate process model applying chemometric approach contains some essential steps that have decisive influence on the general characteristic, reliability, and performance of the obtained model.
In the following it is attempted to introduce the different stages that should be considered carefully during the model development.

4.2.1 Model Objective

The first step is to define the objective of modeling. This question is often related to which variable or quality need to be predicted and where this prediction is going to be implemented. Prediction modeling is needed when a variable, often a quality variable, is difficult, expensive, or time consuming to measure. It can also be the situation that first principles mathematical model for a complex chemical process is hardly available and hence the historical process data is used in order to develop a model to predict the future value of a variable which is used in a control or an optimization application.
Determination of the objective of modeling is important because it will determine the demands for characteristics and accuracy of the model. For example if the model is going to be applied in an optimization routine, the linear or nonlinear characteristics of the model will become important in choosing the optimization algorithm. If the model is going to be applied in a control application, the objective is to estimate the transfer function of the system and hence the stability characteristic of the obtained model will be an important issue.

4.2.2 Selection of Input Variables

In this step, it is desired to explore and determine the suitable input variables for prediction of one or several output variables. This is an essential step, because the data is the main source of information and a reasonable model performance can be expected only when necessary and sufficient information is provided. Selection of appropriate input variables is then important in order to obtain a set of data which is informative enough with respect to a model set that can be determined from the data using appropriate chemometric methods. The concept of informative data is related to persistence of excitation which is described in chapter 3.
Selection of a set of suitable input variables is in fact identifying which variable combination affect the variation of model output. This selection is naturally related to a good knowledge about the process, based on, if available, a first principles mathematical model describing the functional dependency of the model output to the input variables.
A first principles mathematical model is unfortunately not available in the complex refinery processes. However, using the basic knowledge in chemical engineering, it is possible to perform a qualitative analysis of the system in order to identify which variables are expected to affect the variation of the desired output.
Furthermore, it is possible that the collected data from a system does not reflect the expected correlation between output signal and one or several input variables. This can be often for the reasons such that data is collected under the effect of feed-back control, the choice of sampling interval is wrong, the collected input and output data are from different operation points, or the data is simply corrupted as a result of sensor fault.
A correlation analysis can be helpful to explore the functional dependency of output to input signal. Notice that the correlation analysis technique is based on the assumptions that the

system is linear time-invariant, the error or noise part of the data is normal distributed, and data is obtained from an open-loop control.
Furthermore, one should be cautious about the multivariable effect on the output signal. Applying a quick linear regression analysis, like MLR, or PCR, can give insight to the functional dependency of the variables if the relationship is linear. A possible procedure can be to start with simply all possible measured variables and then gradually exclude those variables which shows no correlation to the output.
When the appropriate calibration method is selected and the model is calibrated, then the physical meaning of the sign and magnitude of the obtained model parameters should be in agreement with the expectation and knowledge based on a first principles model or practical experiences. The predicted output will be sensitive to an input variable if a small variance is obtained in that particular. Hence, the correct choice of input variables should results in small variance for the corresponding parameters in the resulting model.

4.2.3 Data Collection and Sampling

When the input and output variables of the model is determined, the next step is to select a suitable sampling rate. Special attention should be paid in collection of data in order to detect errors, and measurement fault. It is necessary to know the level of noise and the method used for measurement of quality variables in laboratory.
Sampling rate is a significant factor which is related to the dynamic characteristic of the system defined by its bandwidth. The purpose of selecting a suitable sampling frequency is to insure that the collected data is informative enough to develop a set of models.
In many industrial application the sample time for temperature, pressure, and flow rate are around a few minutes, or even seconds. However, those variables which are measured in laboratory, often quality variables, can have a sample time of several hours due to the applied analysis methods.
Besides, if the input variables are from different production units, and eventually with different kind of chemical processes, the sample time could be different. For the modeling objective the selected sampling frequency is normally determined by a process with a faster dynamic characteristics.
It is also important to determine the time delay of the system and especially when the variables are chosen from different unit operations. An impulse or a step response of the system can provide information about time delay and time constant of the system, provided an estimate of the transfer function of the system available.
The data should be representative for the process and cover all the operation regions of interest, and thus it is important to carefully select periods of operation in order to include the desired operation modes and area.

4.2.4 Data Treatment

One of the objectives in this step is to identify the outlier, faulty, and missing data. Existence of outlier and error in measurement can completely mislead the development of the model. Missing data and error can even stop the training and development procedure in some modeling algorithm.
Most of the outliers can normally be detected just by visualizing the data in appropriate plots of respective variables. One effective way to detect error, outlier, or any abnormality in data is

scaling of the data to have zero mean and unit variance. If there is an abnormality in data it will be obvious after scaling, and can be visualized in a plot of scaled data.
Data scaling is a common procedure in chemometric modeling prior to any analysis. There are two type of scaling the data. If the data is adjusted to have a zero mean by subtracting off the original mean, then the data is called to be mean centered. This technique is useful in order to remove the effect of different dimensions in data. Autoscaling is called to the second type of data scaling and that is when the mean centered data is additionally adjusted to unit variance.
In calculation of the covariance matrix of the data in different chemometric method it is assumed that the data is mean centered. If an autoscaled data is applied then the calculated covariance matrix will give the correlation matrix of the data.
Unless other mentioned, the autoscaled data is used in the modeling work in this thesis

A PCA model can also be applied in order to uncover the abnormality in data. PCA is an effective tool used for assessment of representability of data, existence of clustering and outliers.
Another problem in data is related to the missing data and when there are periodically lack of measurement in the data set. The missing process data can be due to operation shutdown or problem in data acquisition system. Regarding the quality variables measured at laboratory, it is simply not possible to have measurement value as quick as the process variables and hence there will be lack of data for quality variables.

4.2.5 Suitable Modeling Method

The selection of suitable method is highly dependent on nonlinear and dynamic characteristic of the system under study. If there is no time dependency in relationship between output and input a static model can be a relevant choice, and further if the relationship is linear a MLR or PLS model can be applied for model development. For the case of nonlinear static relationship a ANNs model can be used to cover the nonlinearity of the system, since neural networks show great ability of nonlinear functional approximation.
The characteristic of the a dynamical system is that the variables change with time, or current output value depends not only on the current external stimuli but also on their earlier values. In this case an ARX type of model in System Identification can be applied which is basically a linear, time-series regression method. The input-output relationship in ARX model is described by a simple linear difference equation. Estimation of parameters in ARX model is based on the least squares (LS) method minimizing the prediction error. Another approach is to apply a PLS in parameter estimation of ARX model in order to take advantage of PLS ability to extract the useful information from collinear, noisy, input data which is relevant for modeling the output prediction.
When a dynamic system is also nonlinear, a nonlinear time-series method should be chosen. In a linear PLS it is assumed that the scores in output block is a linear function of the scores in inputs. A nonlinear PLS approach is based on describing the relationship between scores in input and output blocks by nonlinear functions like higher order polynomial or neural networks. Hence, for dynamic nonlinear input-output mapping a time-series ARX model structure can be applied in which nonlinear PLS approaches is chosen for its parameter estimation.

Most of the complex refinery processes are in fact nonlinear system. However, the purpose of this work is not to develop dynamic simulation models. It is desired to model the relationship

between the output quality variable and a set of input process variables in a MISO model for steady state operation. When a steady state operation is considered and the transient regions which may exist in shutdown and start-up situations are avoided, a linear time-series model can be chosen in order to obtain a linear approximation of the relationship between output quality and input process variables.
Regarding the choice of nonlinear models, special attention may be paid to the fact that selection of nonlinear approaches in quality modeling may result in a more complex optimization problem, which may cause complications in solving the problem. We may start with a linear approach, like ARX or PLS, and analysis the performance of the model and result of the validation, and then decide whether it is necessary to continue with nonlinear approaches.

4.2.6 Calibration; Estimation of Model parameter

The calibration, also called training, of the model is the main part of model development and the purpose is to estimate the model parameters, in which certain a criterion is satisfied. This criterion is basically the level of prediction error in training data set. In most of the process chemometrics methods estimation of the model parameters are based on an optimum solution for an optimization problem in which sum of squared prediction errors is minimized. The result is a set of optimum value for the model parameters. The obtained model is then applied in validation in order to assess the predictability and general characteristic of the model. In this sense, although the calibration and validation are two separate procedure, but it is often the results in validation will decide that a calibration has been performed satisfactory.

4.2.7 Model Validation

Model validation is one of the most important issues in multivariate analysis and development of process model. The purpose of model validation is to test the performance of a developed model in order to avoid overfitting or underfitting by finding the optimal number and values of model parameters. Here, we need a second data set which is called test or validation data set. It is important that the validation set is closely comparable to the calibration data set with respect to sampling frequency and condition. The validation set, similar to calibration set, must contain a set of representative data for the target population. The only difference between validation and calibration sets should be the sampling variance. The idea behind the model validation is to evaluate the prediction strength of the model on data with different noise than the calibration set.
Cross validation is a technique used in chemometric modeling, in which all the available objects are used subsequently making models on parts of the data and testing on the other parts. If we continue and make as many models as there are objects, in such a way that each time we leave one of the objects out and use that in validation, we obtain a full cross validation.
In dynamic time-series modeling, it is not appropriate to apply a full cross validation in that sense that the data are mixed over the time. It is important to secure a calibration and validation set containing time sequence of subsequent data. Hence, the calibration and validation data sets must be two different distinct sets of data.
There are several indicators and criterion that can be used in order to assess validation of the model. Naturally, the ability of prediction is the first criterion to consider. A comparison between model output and measurement indicates how well the model can predict the future

output. Another indicator is to compare the sum of square errors in prediction with two reference values. These reference values are a zero-model and an average-model as it is described in chapter 3.

The prediction error should consist the error or noise part of the output signal, and hence should be principally close to a normally distributed noise with a variance smaller than the variance of output signal itself meaning that the model is better than an average model. A histogram plot of error can be used to check the normal distribution of prediction error.

Residual Analysis is also another effective method of validation. In this method we analysis the residual, i.e. prediction error, of the model in order to determine whether there is correlation between the prediction error and the inputs. If there are such correlation, it will be an indication of there are still more system dynamics to describe than the model has already picked up. One important remark is that in residual analysis it is assumed that the input is uncorrelated with the disturbances. This means that this analysis will not work for data collected during feedback.

This is an important issue and we have to consider that the data used for calibration of the quality models in this work is collected from a real process during feedback control. Another issue to consider is that some variables used in the modes are controlled variables and hence the variation amplitude is low. This is indeed an indication of the fact that the first assumption in statistical analysis; i.e. assumption of normal distribution, is more and less violated.

4.3 System Delimitation

As it is described in chapter 2, the gasoline processing area consists of different production units. In this work, the focus will mainly be on three essential units; catalytic reformer I, catalytic reformer II, and isomerization unit. A simplified flow diagram of this part of gasoline processing area is shown in figure 4.1
The product streams of all these production units are sent to intermediate product tanks which are later used as blend components for gasoline blending, as it is shown in more detail in figure 2.1 in chapter 2.

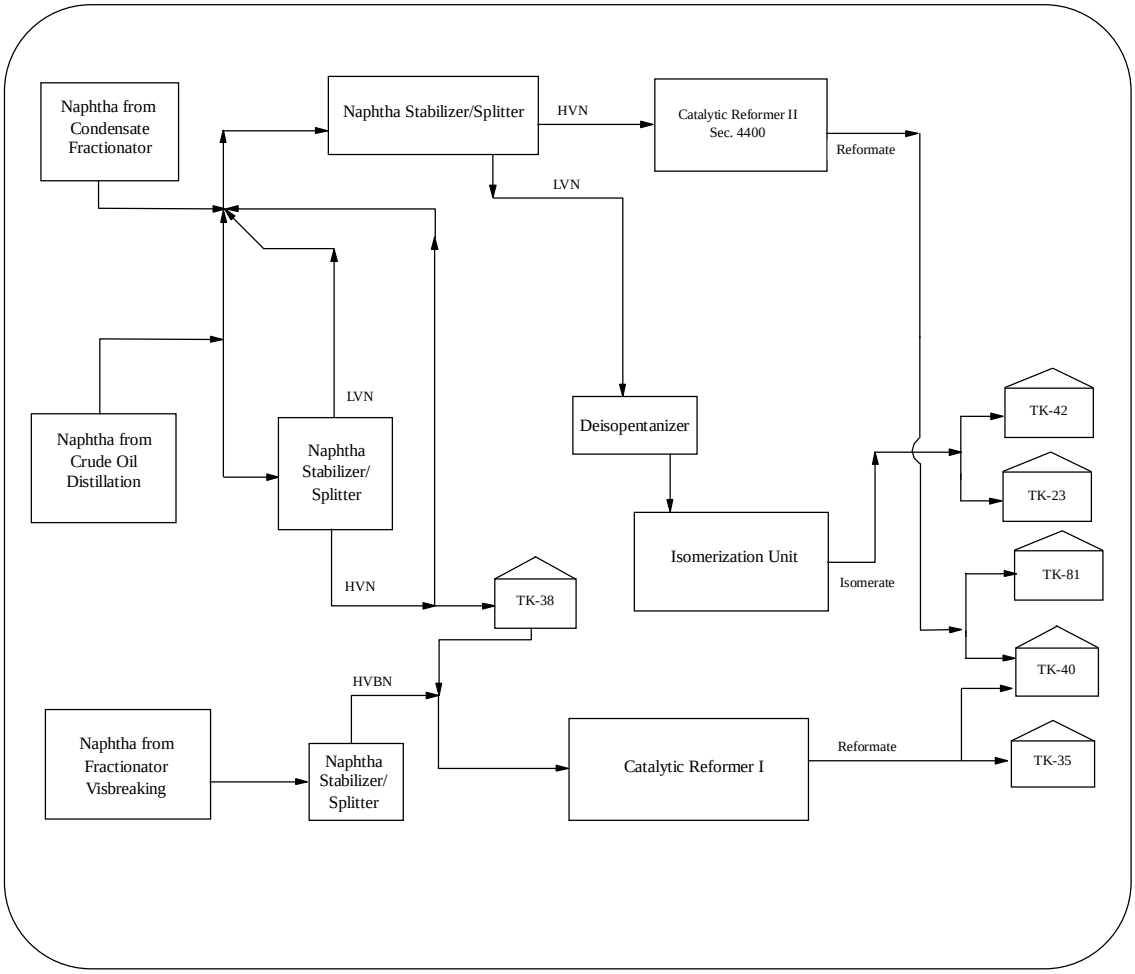

Figure 4.1 : A simplified flow diagram of gasoline processing area.

Basically, in order to be able to optimize the quality of final gasoline products, it is crucial to know the quality variables of all streams sent to the gasoline blender. Figure 2.1 in chapter 2 shows a schematic diagram of all units in gasoline processing area including the gasoline blending components.
For some blend components, it is assumed that their qualities do not change over the time and they can easily be calculated or estimated. For instance, three of the total nine blend components can be considered as almost pure components. These are oxygenate, butane and isopentane. Thus, their qualities can be reasonably estimated based on pure component property. The qualities of two other blend components LVN and LVBN; i.e. light virgin naphtha and light virgin visbroken naphtha, can also be calculated or estimated since they

contain light hydrocarbon components, mostly between C_5 to C_7 molecules, which can be identified by chromatographic analysis.

However, for the reformate products from catalytic reformer units it is not possible to calculate the qualities easily. It is mainly because these intermediate products contain numerous hydrocarbon components with different molecules structure. The isomerate product from isomerization unit contains light hydrocarbon components since the feed to this unit is LVN. Hence, the qualities of isomerate product can be calculated based on identification of the hydrocarbon components.

Besides, the variation of the qualities, especially the Research Octane Number (RON) quality, in the product streams of the catalytic reformers and isomerization units will particularly provide the possibility of producing different gasoline products with more definite octane number specification, and thus more optimization potentiality. It is essentially important to have accurate information of RON quality variable for reformate and isomerate products, since these are the high octane number products of the refinery. Furthermore, it is expensive to have on-line quality measurements for these three intermediate products in order to have the same sampling frequency as the other process variables. The only existing measurement is laboratory analyses which are available only one per day, i.e. a sample rate of 24 hours, for each quality.

As it will be discussed later, the selection of input variables for the quality prediction models includes some temperature variables in the crude oil distillation, condensate fractionator, and the main fractionator in the visbreaking units, and also flow rate variables in the respective naphtha stabilizer and splitters systems. This will extend the system limit to the beginning of the gasoline processing area, as it is shown in figure 4.1.

Consequently, the prediction of some the qualities will be affected by input variables spread over the whole processing area and special care should be paid for finding the suitable delay parameters in the obtained model.

4.4 Description of Output Variables

In this section a description of desired quality output variables is presented for the models in catalytic reformers and isomerization unit.
Among the numerous qualities of gasoline which are officially determined as specifications for the final gasoline products, we are interested in prediction of the following qualities:

1 Research Octane Number (RON)
2 Motor Octane Number (MON)
3 Reid Vapor Pressure (RVP)
4 Benzene contents of the products (BENZENE)

In the case of MON, there is an extensive lack of measurement which cause a serious complication for prediction modeling of this quality. There is not simply enough measurement for MON, neither on-line nor laboratory. However, it is possible to take advantage of strong correlation between RON and MON, which are basically both measurement of the same quality, i.e. octane number. MON can be predicted by a simple linear regression provided accurate model for RON available. For this reason, we focus on prediction of RON in this work.
It is also desired to estimate or predict the yield of reformates and isomerate products for each reformer and isomerization unit. The desired reactions in these units are dehydrogenization, cyclization and isomerization. However, hydrocracking and condensation reactions may also take place in which the first one will produce light hydrocarbons and the second will cause formation of coke. As a results, light hydrocarbon components like methane, ethane, propane, and butane will be produced which is later removed in stabilizer column. Consequently, the yield of reformate will be reduced. The change in the yield is thus correlated with RVP, and yield of reformate can be calculated as the following:

$$Yield = \frac{Reformate\ flow\ rate}{Reformer\ feed\ flow\ rate} \cdot 100 \cdot CORR \quad (4.1)$$

where:

$$CORR = C_1 - C_2\ RVP^{C_3} \quad (4.2)$$

in which C_1, C_2, and C_3 are constants. Reformate flow rate product and the feed flow rate to catalytic reformer are both measured. CORR in equation 4.2 is a correction factor used in equation 4.1 to compensate for RVP changes. Figure 4.2 shows the correction factor for a period of seven months operation in catalytic reformer I. The calculation of CORR is performed by using an on-line RVP analyzer in this unit. In table 4.1, the average, standard deviation, maximum, and minimum of CORR is shown.
As it can be seen the correlation factor has an average of 0.954 in this period with a standard deviation of 0.005. As a result, the correlation factor is a weak function of RVP. This characteristic is also observed in catalytic reformer II.
It is desired to develop a model for RVP in which the model can be used in calculation of yield. Consequently, there will be no need for a separate model for yield of production since it can be calculated using the existing measurement for feed flow rate and reformate flow rate, and then apply the predicted value of RVP model for calculation of the correction factor.
As described in chapter 2, the feed to isomerization unit is LVN, containing light hydrocarbon molecule. The benzene contents of LVN is low, and cyclization reactions is expected to take place only in a small extend in this unit. Thus, the contents of benzene in isomerate product is

expected to be small and mainly unchanged. Laboratory analyses for a period of 14 months operation has shown that the benzene content has been zero for a large period of time. There are only 54 non-zero measurements reported with an average of 0.17 wt.%. and standard deviation of 0.13. Hence, there will be no need for prediction model for benzene contents of isomerate product and i will be assumed to be constant less than 0.2 wt%.

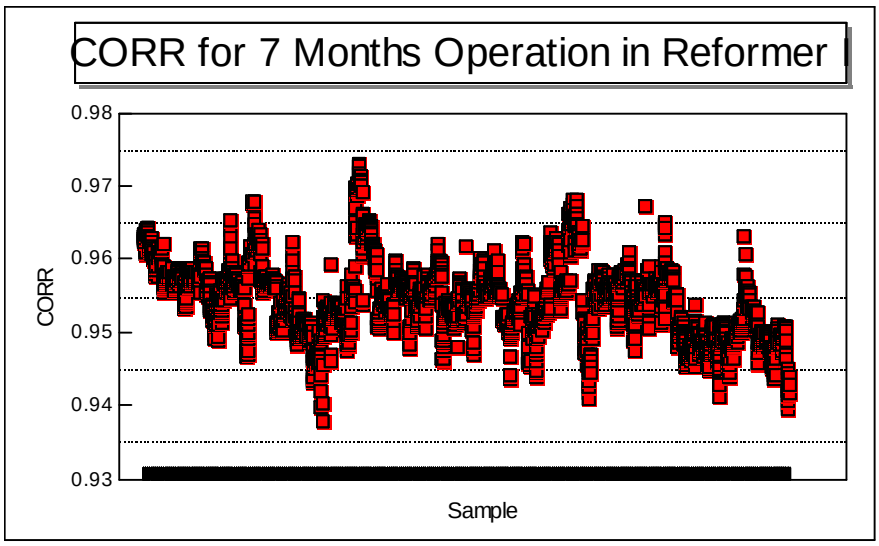

Figure 4.2: Correction factor used in equation 4.1 for calculation of yield of reformate product in catalytic reformer I.

CORR Factor	
Average	0.954
Standard Deviation	0.005
Maximum	0.973
Minimum	0.938

Table 4.1: Statistic value for correction factor.

Hence, in this work there will be focused on development of chemometric model for prediction of RON, RVP, and benzene contents for the reformate products and RON and RVP for the isomerate product..

The output variables for catalytic reformer I, II, and isomerization unit is shown in table 4.2 4.3, and 4.4 along with the calculated average, standard deviation, maximum, and minimum values. These values are for calibration and validation data sets after removal the outliers. The maximum and the minimum values for each quality can be compared with the interval of [avg.-3*std, avg+3*std] in order to assess the outliers. Equation 4.3 indicates that for a normally distributed variable the probability that a value can be out of the range $\mu \pm 3\sigma$ is less than 1%. Here, the mean value and standard deviation are denoted by μ and σ. Equation 4.3 is a results of Camp-Meidels theorem (L. Broendum and J.D. Monrad, 1987).

$$P(|X - \mu| \geq 3\sigma) = 0.27\% \quad \Leftrightarrow \quad P(|X - \mu| < 3\sigma) = 99.73\% \qquad (4.3)$$

Hence, it is expected that only one percent of data will fall out of the ranges $\mu \pm 3\sigma$ as shown in the tables 4.1, 4.2, and 4.3.
Furthermore, it can be seen that the deference between maximum and minimum values of RON is 2 for all the two reformer units and three for isomerization unit. This is in fact an indication of effective feed-back control of RON. The effect of closed-loop control is discussed in chapter 3.

Catalytic Reformer I						
	Calibration			Validation		
	RVP	RON	BENZENE	RVP	RON	BENZENE
AVG.	49.67	99.97	3.53	50.85	99.97	3.74
STD	3.02	0.37	0.43	3.74	0.45	0.36
MAX.	56.00	101.60	4.91	61.00	101.50	4.37
MIN.	39.00	98.60	2.12	44.00	98.30	2.95
AVG. + 3STD	58.72	101.09	4.82	62.05	101.33	4.83
AVG. - 3STD	40.61	98.85	2.24	39.65	98.61	2.66

Table 4.2: Output variables in Catalytic Reformer I.

Catalytic Reformer II						
	Calibration			Validation		
	RVP	RON	BENZENE	RVP	RON	BENZENE
AVG.	37.06	101.00	1.79	37.90	101.00	1.76
STD	3.73	0.25	0.45	3.46	0.20	0.43
MAX.	50.00	101.80	2.80	47.00	101.60	2.42
MIN.	23.00	100.00	0.60	32.00	100.40	1.16
AVG. + 3*STD	48.26	101.76	3.14	48.28	101.58	3.05
AVG. - 3*STD	25.86	100.25	0.44	27.51	100.41	0.47

Table 4.3: Output variables in Catalytic Reformer II.

Isomerization Unit				
	Calibration		Validation	
	RVP	RON	RVP	RON
AVG.	70.04	87.29	70.20	87.24
STD	3.54	0.47	2.98	0.32
MAX.	76.30	88.60	81.00	88.40
MIN.	58.70	85.70	65.50	86.50
AVG. + 3STD	80.65	88.70	79.13	88.21
AVG. - 3STD	59.43	85.88	61.28	86.27

Table 4.4: Output variables in isomerization unit.

4.5 Description of Input Variables

The described procedure in section 4.2.2 for selection of the appropriate input variables is followed in order to obtain a set of data which is informative enough with respect to a model set that can be determined from the data using appropriate chemometric methods.
In the following subsections the variables selected for the models will be described. These variables are going to be used for the RON, RVP, and benzene models described in chapter 5.

4.5.1 Input Variables for Catalytic Reformer I

As it is described in chapter 2, the feed to catalytic reformer I is a mix stream of heavy virgin naphtha (HVN) from splitter in crude oil distillation section and heavy virgin visbroken naphtha (HVBN) from the splitter in visbreaking section. Figure 4.3 shows a simplified flow diagram of catalytic reformer I along with the stabilizer/splitter system in crud oil distillation and in after the mail fractionator in the visbreaking section. Table 4.5, 4.6 show a list all variables that are used in developing the chemometric models described in chapter 5 for prediction of RON, RVP and benzene contents of the reformate product. The values of calculated average, standard deviation, maximum, and minimum along with the values of $\mu \pm 3\sigma$ are shown in tables 4.5 and 4.6 for respectively calibration and validation data sets.

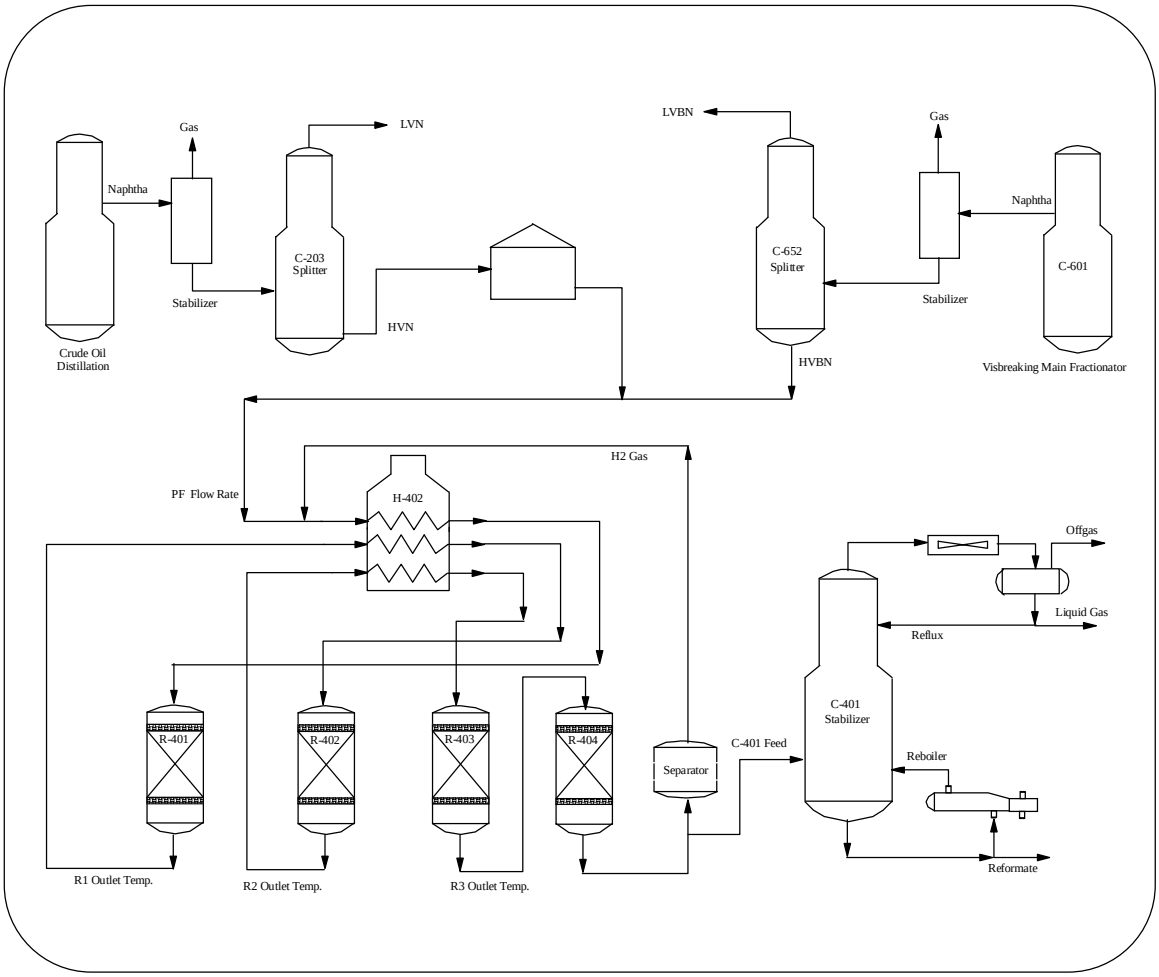

Figure 4.3 : A simplified flow diagram of catalytic reformer I in gasoline processing area.

			Calibration					
No.	Description	Unit	**AVG.**	**STD**	**MAX.**	**MIN.**	**AVG. + 3STD**	**AVG. - 3STD**
1	H / C	mol/mol	3.86	0.42	6.86	3.00	5.12	2.61
2	% H2 Gas	%	73.22	1.82	80.70	59.23	78.68	67.75
3	R1 Outlet Temp	^{0}C	430.92	4.21	442.45	421.67	443.54	418.31
4	R2 Outlet Temp	^{0}C	471.10	4.62	482.62	460.40	484.95	457.25
5	R3 Outlet Temp	^{0}C	500.33	3.54	507.32	489.89	510.95	489.71
6	Reformer Feed Flow Rate	m^3/hr	51.38	7.27	63.06	29.87	73.17	29.58
7	Reformate Flow Rate	m^3/hr	38.09	5.85	48.45	21.09	55.64	20.53
8	C401 Liquid Gas Flow Rate	m^3/hr	5.37	0.96	8.08	1.47	8.25	2.50
9	C401 Reflux Flow Rate	m^3/hr	23.06	1.05	25.33	20.99	26.22	19.90
10	C401 Feed Temp	^{0}C	165.45	2.86	174.05	158.34	174.03	156.87
11	C401 Reboiler Temp	^{0}C	254.59	2.11	261.57	239.98	260.94	248.25
12	C203 Reflux Flow Rate	m^3/hr	27.72	6.21	39.02	13.04	46.35	9.10
13	HVN Flow Rate	m^3/hr	52.79	8.28	76.57	23.69	77.63	27.94
14	LVN Flow Rate	m^3/hr	36.97	6.85	66.55	17.84	57.52	16.42
15	HVBN Flow Rate	m^3/hr	10.74	2.96	18.36	2.34	19.63	1.85
16	CP201	^{0}C.	102.33	5.43	113.72	83.23	118.63	86.04
17	CP601	^{0}C.	112.86	8.38	134.80	91.63	138.00	87.72
18	CP203B	^{0}C	105.52	2.59	114.62	91.65	113.30	97.74
19	CP652B	^{0}C	101.76	4.12	130.76	71.42	114.11	89.40

Table 4.5: Input variables used for calibration of the models in catalytic reformer I.

The first 5 variables listed in table 4.5 are expected to have a large effect on RON quality. H/C is the ratio of mole hydrogen in the recycle gas per mole hydrocarbon in the feed. Variable number 2 $\%H_2$ is the H_2 purity in %mole or %volume in the recycle gas from the product separator to the feed stream of the reformer. These two variables indicate the developed H_2 during the reactions. It is expected that a change in the feed composition will be reflected in the H/C mole ratio.

Variables number 3, 4, and 5 are the outlet temperature of the reactor number 1, 2, and 3 respectively. These temperatures indicates the type of the reactions take place in the reactors, since the hydrocracking and condensation reactions will occur at higher temperature. Variables number 6 and 7 are the feed flow rate to the reformer and the reformate product flow rate.

Variation in RVP is dependent on the light hydrocarbon contents of reformate product, and thus, it is sensitive to the operation of the stabilizer C401. Hence, variables 8, 9, 10, and 11 are chosen from stabilizer C401. The description of these variables in the table 4.5 and 4.6 are self explanatory. Variable number 12 is the reflux flow rate in the splitter column C203, which affects the changes in the boiling range of the LVN top product in C203 and hence, the split of the naphtha feed to LVN and HVN. Variables number 13, 14, 15, are flow rate of LVN, HVN, and HVBN.

			Validation					
No.	Description	Unit	**AVG.**	**STD**	**MAX.**	**MIN.**	**AVG. + 3STD**	**AVG. - 3STD**
1	H / C	mole/mole	3.96	0.47	6.17	3.26	5.36	2.56
2	% H2 Gas	%mole	72.52	1.79	79.31	63.07	77.90	67.14
3	R1 Outlet Temp	^{0}C	436.45	3.58	444.30	426.58	447.20	425.69
4	R2 Outlet Temp	^{0}C	480.89	4.24	490.08	469.46	493.61	468.17
5	R3 Outlet Temp	^{0}C	506.61	4.94	517.99	496.46	521.44	491.79
6	Reformer Feed Flow Rate	m^3/hr	51.44	6.03	62.66	36.76	69.54	33.34
7	Reformate Flow Rate	m^3/hr	39.02	4.18	47.62	27.62	51.55	26.48
8	C401 Liquid Gas Flow Rate	m^3/hr	4.57	0.92	7.09	1.04	7.32	1.82
9	C401 Reflux Flow Rate	m^3/hr	21.66	1.25	28.92	16.81	25.43	17.90
10	C401 Feed Temp	^{0}C	165.15	3.59	173.50	156.76	175.93	154.37
11	C401 Reboiler Temp	^{0}C	254.51	1.97	258.02	246.49	260.42	248.60
12	C203 Reflux Flow Rate	m^3/hr	31.05	4.66	38.89	17.90	45.04	17.06
13	HVN Flow Rate	m^3/hr	54.80	5.57	69.51	35.81	71.50	38.10
14	LVN Flow Rate	m^3/hr	32.09	7.28	53.02	14.00	53.92	10.27
15	HVBN Flow Rate	m^3/hr	55.43	6.32	67.57	43.00	74.38	36.48
16	CP201	^{0}C.	11.47	2.72	17.85	3.29	19.63	3.30
17	CP601	^{0}C.	105.24	4.81	114.15	91.93	119.66	90.82
18	CP203B	^{0}C	111.27	7.91	131.48	91.70	135.00	87.53
19	CP652B	^{0}C	106.47	2.31	115.89	100.84	113.42	99.53

Table 4.6: Input variables used for validation of the models in catalytic reformer I.

A change in the composition of the crude oil can affect the fraction of HVN and LVN in the splitter. This change is reflected by the temperature on the naphtha side stream in crude oil distillation. Variable number 16, CP201, which is also called as cut point temperature, is a

pressure corrected temperature of naphtha side stream tray in crude oil distillation, which is calculated as in equation 4.4.

$$T_{cut} = (T_i)\frac{C1 - C2Ln(P_i)}{C1 - (T_i)(Ln(P_i))} \quad (4.4)$$

where C1, and C2 are constants, T_i is the measured temperature in Kelvin, and P_i is top pressure in bar absolute.
Variables number 17, 18, and 19 are calculated by using equation 4.4 for the temperature of naphtha from main fractionator in visbreaking section, bottom temperature of splitter in crude distillation, and bottom temperature of splitter in visbreaking section respectively.

4.5.2 Input Variables for Catalytic Reformer II

There are a lot of similarity between the process in catalytic reformer I and II. Unless other mentioned, the description of the variables are the same as for the reformer I.
The feed to catalytic reformer II is HVN from splitter C-4703 as shown in figure 4.4, which shows a simplified flow diagram of catalytic reformer II along with the stabilizer/splitter system after condensate fractionator. Table 4.7, 4.8 show a list all variables that are used in developing the chemometric models for prediction of RON, RVP and benzene contents of the reformate product. Table 4.7, 4.8 shows average, standard deviation, maximum, and minimum along with the values of $\mu \pm 3\sigma$ as well for calibration and validation data sets respectively.

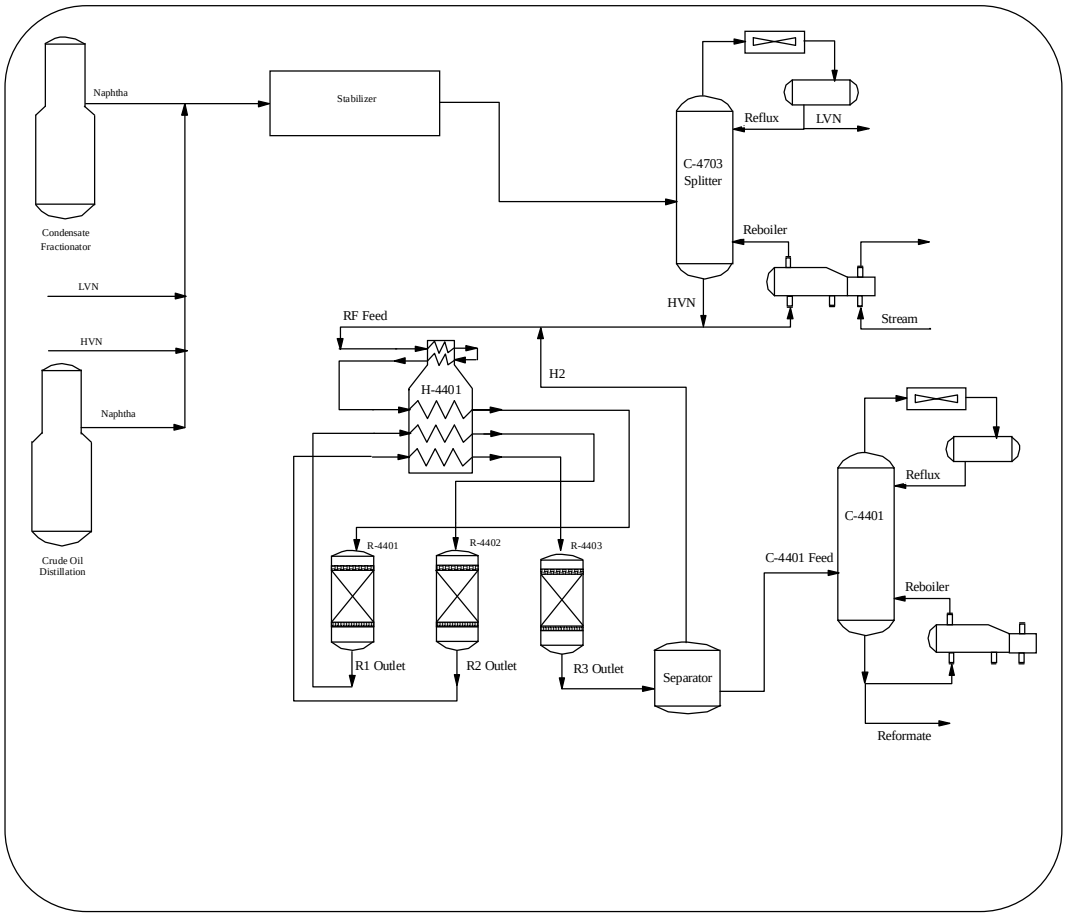

Figure 4.4 : A simplified flow diagram of catalytic reformer II in gasoline processing area.

The first 5 variables listed in table 4.7 are chosen from the reactors and the recycle gas from the reformer which are expected to have a large effect on RON quality as it is explained in reformer I.
Variation in RVP is dependent on the operation of stabilizer C4401. Hence, variables 7, 8, 9, 10, and 11 are chosen from stabilizer C4401. Variable number 12, 13 , and 14 are chosen from the splitter C4703 which affect the split of naphtha feed to LVN and HVN. Variables 15, 16, 17, are calculated by using equation 4.4. The description of these variables in the table 4.7 and 4.8 are self explanatory.

			Calibration					
No.	Description	Unit	**AVG**	**STD**	**MAX**	**MIN**	**AVG + 3STD**	**AVG - 3STD**
1	R1 Outlet Temp	^{0}C.	391.80	3.28	398.77	380.81	401.65	381.95
2	R2 Outlet Temp	^{0}C.	446.91	5.36	458.23	432.86	462.99	430.84
3	R3 Outlet Temp	^{0}C.	472.33	4.98	482.95	459.88	487.26	457.40
4	H/C	mole/mole	4.42	0.73	6.61	3.51	6.60	2.23
5	H2 Purity	%mole	82.47	1.54	87.36	72.82	87.10	77.85
6	Reformer Feed Flow Rate	m^3/hr	92.41	10.66	107.98	64.99	124.38	60.44
7	C4401 Feed Temp	^{0}C.	187.47	3.42	194.63	158.52	197.73	177.20
8	C4401 Reboiler Temp	^{0}C.	248.82	1.52	252.85	203.60	253.39	244.26
9	Reformate Product Flow Rate	m^3/hr	76.29	8.85	89.21	54.16	102.85	49.73
10	C4401 Reflux Flow Rate	m^3/hr	4.98	1.76	11.53	0.61	10.26	-0.29
11	C-4401 Feed Flow Rate	m^3/hr	77.81	9.18	91.36	54.00	105.36	50.27
12	C4703 Reflux Flow Rate	m^3/hr	119.33	9.26	146.00	68.03	147.10	91.56
13	C4703 LVN Flow Rate	m^3/hr	76.77	3.29	89.10	21.41	86.64	66.91
14	C4703 Reboiler Steam Flow Rate	ton/hr	17.68	1.19	20.00	12.48	21.26	14.11
15	CP201	^{0}C.	103.62	5.45	113.90	83.23	119.98	87.26
16	CP4201	^{0}C.	101.26	3.80	123.64	71.64	112.65	89.88
17	CP4703B	^{0}C.	110.10	1.31	113.48	78.29	114.02	106.18

Table 4.7: Input variables used for calibration of the models in catalytic reformer II.

			Validation					
No.	Description	Unit	**AVG**	**STD**	**MAX**	**MIN**	**AVG + 3STD**	**AVG - 3STD**
1	R1 Outlet Temp	^{0}C.	397.61	2.71	403.82	391.92	405.74	389.49
2	R2 Outlet Temp	^{0}C.	456.99	4.62	465.29	448.60	470.84	443.14
3	R3 Outlet Temp	^{0}C.	482.32	6.25	493.09	471.82	501.09	463.56
4	H/C	mole/mole	4.48	0.43	5.66	3.73	5.77	3.20
5	H2 Purity	%mole	80.53	1.55	84.46	69.82	85.19	75.88
6	Reformer Feed Flow Rate	m^3/hr	90.86	9.76	103.16	75.45	120.15	61.58
7	C4401 Feed Temp	^{0}C.	182.97	4.05	188.64	171.90	195.11	170.82
8	C4401 Reboiler Temp	^{0}C.	247.93	1.54	250.41	243.58	252.54	243.33
9	Reformate Product Flow Rate	m^3/hr	74.89	7.50	87.46	63.10	97.37	52.40
10	C4401 Reflux Flow Rate	m^3/hr	5.91	2.33	11.25	1.86	12.91	-1.09
11	C-4401 Feed Flow Rate	m^3/hr	76.50	8.07	87.61	63.91	100.71	52.29
12	C4703 Reflux Flow Rate	m^3/hr	117.74	4.12	127.91	89.99	130.11	105.38
13	C4703 LVN Flow Rate	m^3/hr	76.55	3.03	86.46	60.76	85.65	67.44
14	C4703 Reboiler Steam Flow Rate	ton/hr	17.90	0.73	19.28	13.17	20.08	15.72
15	CP201	^{0}C.	103.50	5.21	114.15	91.93	119.14	87.85
16	CP4201	^{0}C.	100.61	1.94	106.21	95.84	106.44	94.78
17	CP4703B	^{0}C.	110.75	0.94	113.56	108.13	113.57	107.93

Table 4.8: Input variables used for validation of the models in catalytic reformer II.

4.5.3 Input Variables for Isomerization Unit

The feed to isomerization unit is light virgin naphtha (LVN) after removal of isopentane (IC5) in deisopentanizer (DIP) as it is described in chapter 2. LVN is the top product of the splitter C4703 which is sent to DIP, as it is shown figure 4.5. A typical LVN contains mostly of hydrocarbon molecules between C_5 to C_7 with a True Boiling Point (TBP) range of 32-88 ^{0}C. Hence, it would be possible to identify the hydrocarbon components in the feed to this unit by chromatographic analysis.
The low octane LVN is converted to high octane number isomerate product by catalytic isomerization process. The reactions in this process are mainly exothermic.
Table 4.9, and 4.10 show a list all variables used for model development in this unit. These model are developed for prediction of RON, and RVP. The benzene contents of the isomerate product is assumed to be contestant. The values of calculated average, standard deviation, maximum, and minimum along with the values of $\mu \pm 3\sigma$ are shown in tables 4.9 and 4.10 for respectively calibration and validation data sets.

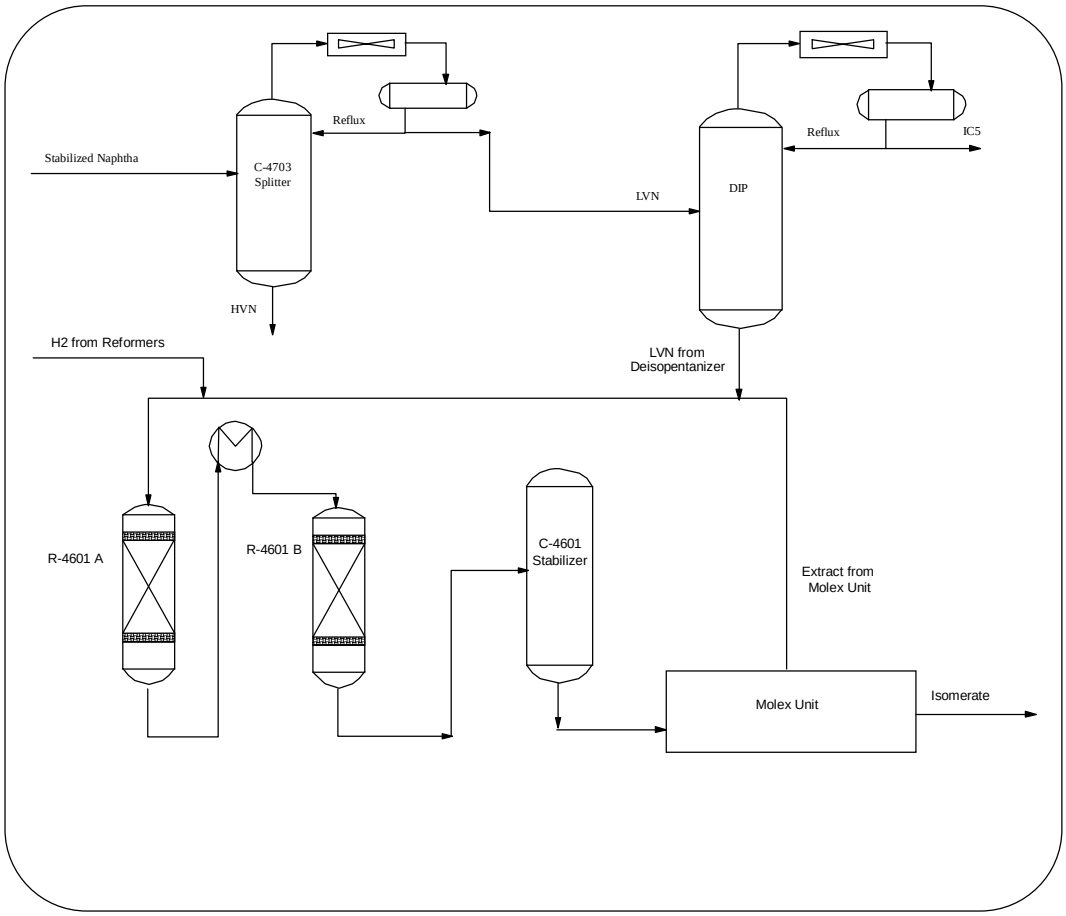

Figure 4.5 : A simplified flow diagram of isomerization unit in gasoline processing area.

Variable number one is the inlet temperature of the feed to the reactor A. The outlet stream of the reactor A is cooled down to about the same temperature as the feed to the first reactor and sent to reactor B. Variables 2, and 3 are outlet temperatures of the reactors. Variable number 4 is Liquid Hourly Space Velocity (LHSV) which is calculated as the total inlet flow rate of the feed to the reactors divided by the catalyst volume. Variable number 5 is the hydrogen consumption in this unit, since the isomerization reactions consume H_2. Variables 6. 7. 8. and

9 contain information about the operation in DIP in which the IC5 is removed, and thus these variables will reflect de degree of efficiency in DIP.

			Calibration					
No.	Description	Unit	**AVG**	**STD**	**MAX**	**MIN**	**AVG + 3STD**	**AVG - 3STD**
1	Reactor Inlet Temp	^{0}C	142.96	2.21	147.30	100.27	149.60	136.32
2	Reactor A Outlet Temp	^{0}C	189.10	2.28	193.81	132.52	195.95	182.25
3	Reactor B Outlet Temp	^{0}C	159.80	2.52	164.20	112.18	167.35	152.26
4	LHSV	1/hr	2.15	0.14	2.29	1.50	2.58	1.72
5	H2 Consumption	Sm^3/hr	2553.09	245.48	3136.48	1563.97	3289.53	1816.65
6	DIP Tray 8 Temp	^{0}C	84.88	1.28	92.53	81.59	88.72	81.03
7	DIP Bottom Flow Rate	m^3/hr	55.63	3.40	60.59	41.03	65.83	45.44
8	DIP Feed Flow Rate	m^3/hr	75.54	2.91	83.05	60.44	84.27	66.81
9	DIP Reflux Flow Rate	m^3/hr	98.18	8.50	152.48	65.96	123.68	72.68
10	CP4703T	^{0}C	52.46	1.79	57.56	37.68	57.83	47.09
11	C-4703 Reflux Flow rate	m^3/hr	124.24	3.38	146.00	87.57	134.39	114.09
12	C-4703 Feed Flow Rate	m^3/hr	172.69	12.09	193.77	112.81	208.97	136.42

Table 4.9: Input variables used for calibration of the models in isomerization unit.

Variables number 10 is the LVN cut point in the splitter C4703 calculated by using equation 4.4.
Variables 10, 11, and 12 together with variable number 7 which is in fact the flow rate of LVN contain information about the mass balance and split efficiency in the splitter C4703.

			Validation					
No.	Description	Unit	**AVG.**	**STD**	**MAX.**	**MIN.**	**AVG. + 3STD**	**AVG. - 3STD**
1	Reactor Inlet Temp	^{0}C	143.40	1.09	147.72	131.12	146.67	140.13
2	Reactor A Outlet Temp	^{0}C	189.35	1.54	193.11	175.91	193.98	184.72
3	Reactor B Outlet Temp	^{0}C	159.74	1.04	162.50	147.22	162.86	156.62
4	LHSV	1/hr	2.18	0.07	2.27	1.44	2.40	1.95
5	H2 Consumption	Sm^3/hr	2787.02	176.93	3183.27	1706.59	3317.79	2256.24
6	DIP Tray 8 Temp	^{0}C	85.33	1.49	89.30	76.55	89.79	80.87
7	DIP Bottom Flow Rate	m^3/hr	55.81	3.33	59.20	10.89	65.81	45.82
8	DIP Feed Flow Rate	m^3/hr	75.39	2.82	84.19	53.58	83.85	66.93
9	DIP Reflux Flow Rate	m^3/hr	96.98	14.73	144.71	24.63	141.18	52.78
10	CP4703T	^{0}C	53.94	1.33	57.66	50.96	57.93	49.95
11	C-4703 Reflux Flow rate	m^3/hr	112.28	7.60	127.91	89.99	135.08	89.47
12	C-4703 Feed Flow Rate	m^3/hr	172.75	11.04	189.34	129.98	205.87	139.64

Table 4.10: Input variables used for validation of the models in isomerization unit.

4.6 Selection the Sample Interval

In most of the modern industry process control system today, one or another type of commercial Distributed Control System (DCS) is used in which the conventional process variables such as temperature, pressure and flow rate are measured in a sample rate of a few seconds. This high frequency sampling rate is due to important consideration of process control applications, and occur at the lowest level of conventional control instrumentation which is then connected normally to the Control Processor of the respective control loop. The data acquisition system will then collect the data and possibly perform a kind of suitable data treatment later. This data treatment can be such as a moderate filtering in order to prevent aliasing in process control, or eliminate obvious sensor fault.

Right from the beginning of every input/output modeling work or any process control application the question about sampling frequency will arise. How fast should a sampling rate be in order to data would then be informative enough to meet the demands for a robust prediction model?.
This question is closely related to the dynamic characteristic of the system under study. Furthermore, it is not always possible to measure the desired quality variable considering the possibility and limitation of measurement devices and what existing hardware and software system may offer in the data sampling field. Hence, the question above is effectively related to a second question about the possibility of whether we can actually get a set of measurements in a suitable sample frequency which we desire.
It is attempted to answer these questions in the following subsections.

4.6.1 Suitable Sample Frequency

An appropriate sampling rate should be relative to the time constants of the system. A good choice can be a sample frequency of ten times of the bandwidth of the system (Ljung, L. 1987). In the frequency domain, the dynamic behavior of a system is characterized by its bandwidth. There is a reciprocal relationship between bandwidth and dynamic response time. For a first order dynamic system the bandwidth is equal to natural frequency of the system, this means that the product of bandwidth and dynamic response time is exactly one. Friedland (Friedland, B. 1987) has shown that this product is approximately one for a properly damped second order system. The same relation is also valid for higher system order as well.
On the other hand, very fast sampling is undesirable since the developed model fits in high frequency bands, and hence produce bias in model prediction. Besides, fast sampling leads to numerical difficulties in model parameter estimation.

4.6.2 Sample Frequency for Input Variables

The input signals are mostly temperature, pressure and flow rate and also some calculated variables based on these measurements, as it is described in chapter 4.
The data acquisition system of the refinery offer 4 sampling rates which are average of 3 and 6 minutes data, average of one hour, and average of one day data.
As it is described in chapter 4, we are seeking the relevant effect of some process variables on the qualities of intermediate gasoline products after reforming and isomerization units. Thus, our system-limit covers the whole gasoline processing area from the naphtha side-stream of

crude oil distillation column in the beginning of this processing area to the intermediate storage tank in gasoline blending at the end.
The question is, what is a good estimate for bandwidth and dynamic response time of this system? or more practically which of those existing 4 sample frequencies should be selected in order to be around 4-10 times higher than the bandwidth of this system?
Notice that, for example, in catalytic reformer I, there are 4 reactors, 1 distillation column, 1 heater, and 1 gas-liquid separator. Consequently, we may think of several hours of response times.
Based on the above considerations, the sample rate of average of one hour is chosen for the input variables in order to satisfy the demand for a sampling rate of 4-10 times faster than the response of the system. This choice rely on some previous practical experience in modeling in this project, and also in previous control application at the refinery. These experiences suggest that day-average sampling rate is too slow and six-minutes sampling rate is too fast.

4.6.3 Sample Frequency for Output Variable

The qualities of reformates and isomerate products, which make the output signal of the models, are measured at laboratory once every 24 hours. This low sampling frequency for model output has given rise to a challenging problem in this work. It may appear, at the first place, that in some modeling cases, it would not simply be possible to obtain a reliable and robust model by this low sampling frequency.
The proposed solution for output low sample frequency is that we simply consider the sampling rate of the system to be one hour, in which we have a lot of missing data in output measurement. And then, depending on the variation in output signal from one day to another, we can choose one of the following way to overcome this issue.
If the output variation is slow moving over one day to the next day, this is as the case of RON, we can perform an interpolation in output and perform model calibration but avoid interpolation in model validation.
In the opposite case, in which there is a considerable variation in the output signal, indicating a possible faster dynamic response, such as the case of RVP, it would not be a good idea to replace the missing output by interpolation. The proposed solution here is we choose a suitable structure for ARX model in which we take hourly sampled input together with the last existing output signal in order to model the prediction of output at time t. Since this solution is inherently integrated in the ARX structure, it will be described in the following section.

4.7 PCA Analysis

PCA model is developed in order for assessment of representability of data used for the developed models. It is important to analysis the data to discover any abnormality in the data which can for instance be a departure from the normal operation point. It is assumed that outliers are already removed and thus the abnormality can be due to an operation point which is not normal or is shutdown or startup period. Another aspect of a PCA model is to examine the existence of distinct clusters of data which suggest totally different operation of the plant. In this case it should be considered to develop prediction models for each area separately. A third aspect is to discover any collinearity in the selected input.
The PCA analysis is performed for the input data which is used for the development of the chemometric models for catalytic reformer I. In the next subsection a description of the PCA model is presented.

4.7.1 PCA Model for Catalytic Reformer I

The input to the PCA model is called X matrix. Number of selected input variables for the prediction models for this unit are 19 as it is described in section 4.5. Hence, the data X contains 19 columns. Furthermore, the data X contains both the calibration and the validation sets used in the prediction model development, and hence number of rows in X matrix, which is denoted by n in equation 4.5, corresponds to the total number of data after removing the faulty, and outliers. Number of n is 10012 samples. The input X data is then autoscaled. The covariance matrix of X is calculated by equation 4.5, since the data is autoscaled this equation will give the correlation matrix.

$$cov(X) = \frac{X^T X}{n-1} \tag{4.5}$$

As it is described in chapter 3, PCA model relies on an eigenvector decomposition of the correlation matrix of the data X. The eigenvectors are called Principal Components (PC), and the associated eigenvalues of the correlation matrix are a measure of the captured variance for each pair of score and loading vector. Principal Components are also called Latent Variables (LV).
In table 4.11, the percent variance captured by each PC is shown for the PCA analysis. Figure 4.6 shows the eigenvalues of the correlation matrix versus principal component number. These information is used in order to decide how many PCs should be included to the model. It appears that a number of 10 PCs can be a suitable choice. This choice rely on an assessment of the level of the noise in the data, and it is expected that most of the samples in input data reflect the interesting operation points, as shown in the tables 4.5 and 4.6.

A graphical approach is the best way to represent the results of a principal component analysis which makes the interpretation of the results much easier. By examining scores and loading plots, we can explore the quality of substantial information in the data. We can examine the scores and loadings plot one by one or plot the score or loading vectors for PC number 1 versus PC number 2.
Figure 4.7 shows the scores on PC number 1 along with 95 % limit of confidence interval. The scores are the effect of observations on the PC. Figure 4.8 shows a scatter plot of the scores for PC number 1 versus the scores for PC number 2 along with the 95% and 99% confidence

interval limits. These two figures shows no significant outliers, and the major part of the data is indeed inside the 95% limit.

Percent Variance Captured by PCA Model			
Principal Component Number	Eigenvalue of Correlation (X)	%Variance Captured by this PC	%Variance Captured Total
1	4.84E+00	25.49	25.49
2	3.41E+00	17.96	43.44
3	2.73E+00	14.35	57.79
4	2.21E+00	11.65	69.44
5	1.28E+00	6.73	76.18
6	1.07E+00	5.61	81.79
7	7.51E-01	3.95	85.74
8	6.33E-01	3.33	89.07
9	4.79E-01	2.52	91.59
10	3.60E-01	1.90	93.49
11	3.22E-01	1.70	95.18
12	2.93E-01	1.54	96.73
13	2.37E-01	1.25	97.97
14	1.32E-01	0.70	98.67
15	9.26E-02	0.49	99.15
16	8.51E-02	0.45	99.60
17	5.61E-02	0.30	99.90
18	1.42E-02	0.07	99.97
19	5.20E-03	0.03	100.00

Table 4.11 : Percent Variance Captured by PCA Model

.

What we observe in the figures is that there is apparently a systematic variation in data, and there are at least two major areas of operation. It is an indication of two different operation modes. It is important for the task of the modeling that it is necessary separate these two operation modes and develop model for each mode alone.

The interesting observation is that the periods of these two operation modes are not the same, meaning that the operation points are independent of season changes. This indication suggests that the operation modes are mainly related to the desired quality of RON and benzene contents rather than RVP. The specification for RVP quality is different for summer and winter periods, but there is no season change for specification of neither RON nor benzene content. There are two major type of reformate products regarding the level of benzene content characterized by low and high aromatic; i. e. mostly benzene, contents. Examining the figures shows that the change from one to another variant is around a few weeks.

Consequently, it is not necessary to distinguish between these two regions and it should be investigated that one model for both operation modes will be appropriate.

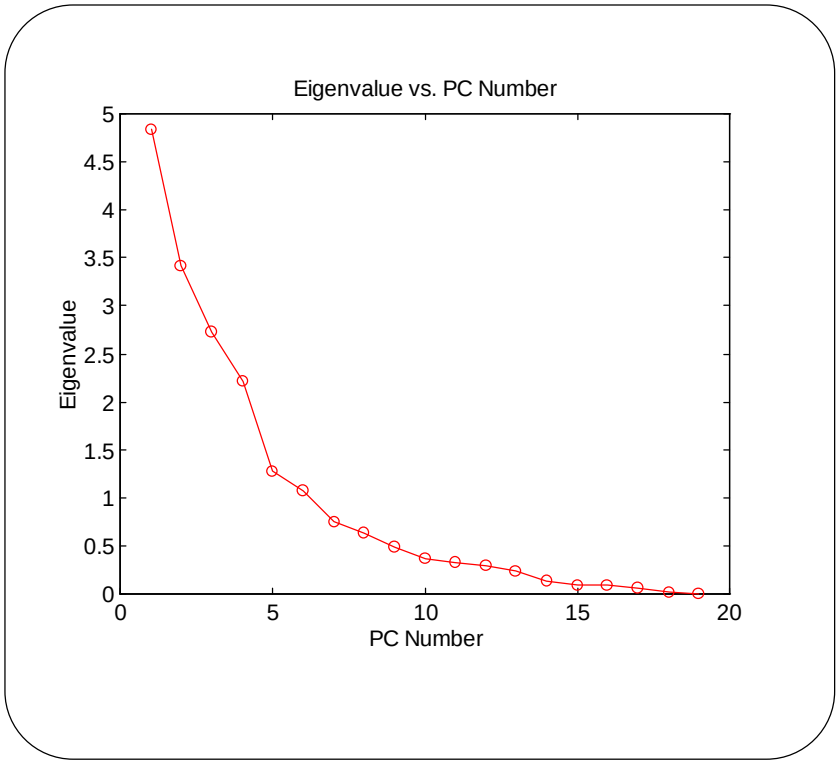

Figure 4.6 : Percent variance Captured by PCA model .

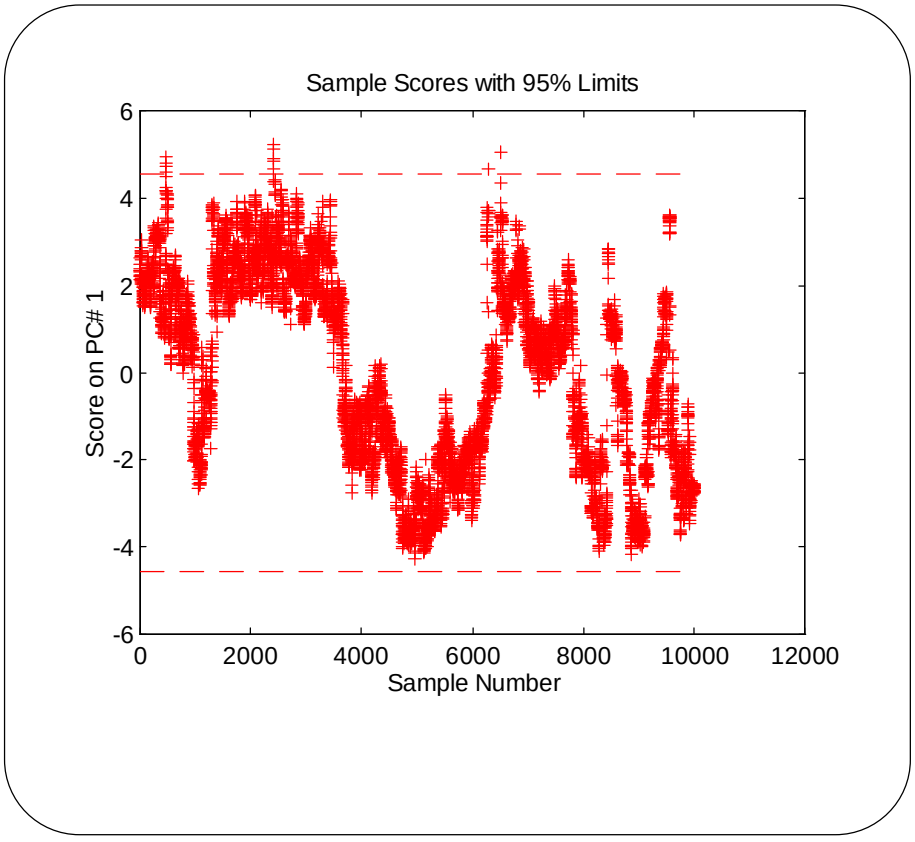

Figure 4.7 : Scores on PC number 1.

Figure 4.9 shows the loadings for PC number 1 versus number 2. The loading plot exhibit the effect of variables on the PC. Examining the loading plot is interesting to discover which variables have the most effect on the PC. It can be seen that for example variables 6 and 7, which are flow rate of the feed to the reformer and the reformate product, have the same effect on both PCs, meaning that they are correlated, and thus one of them is enough to be included in a model to represent the corresponding information.

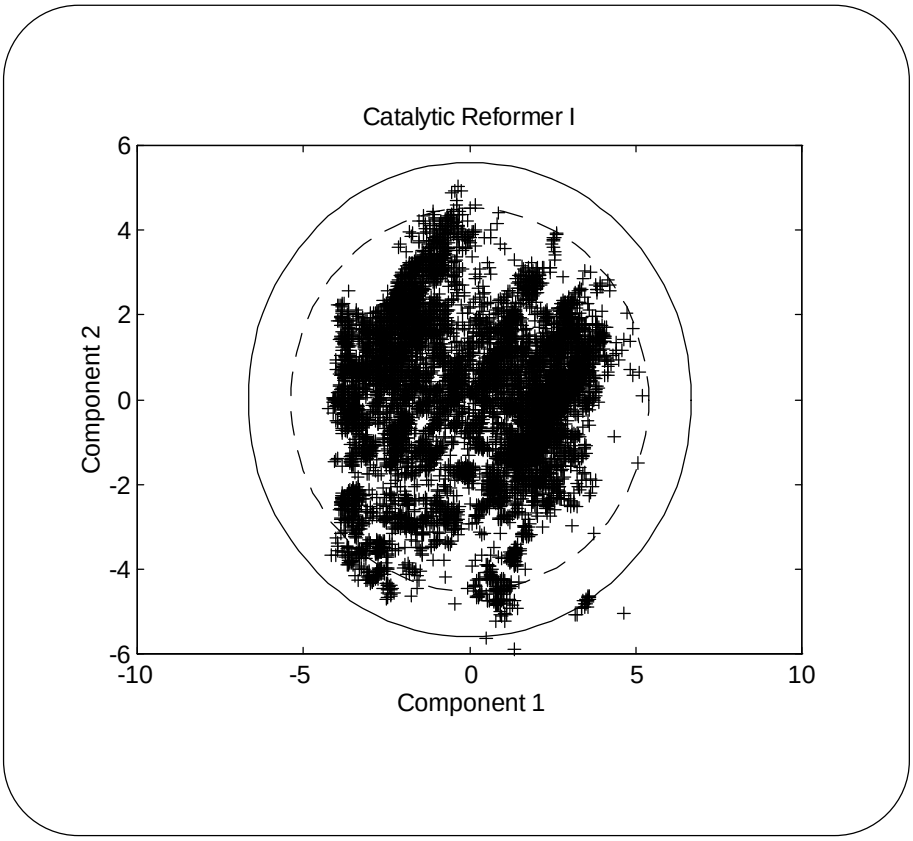

Figure 4.8 : Scatter plot of PC # 1 vs. PC # 2 showing the 95% and 99% limits .

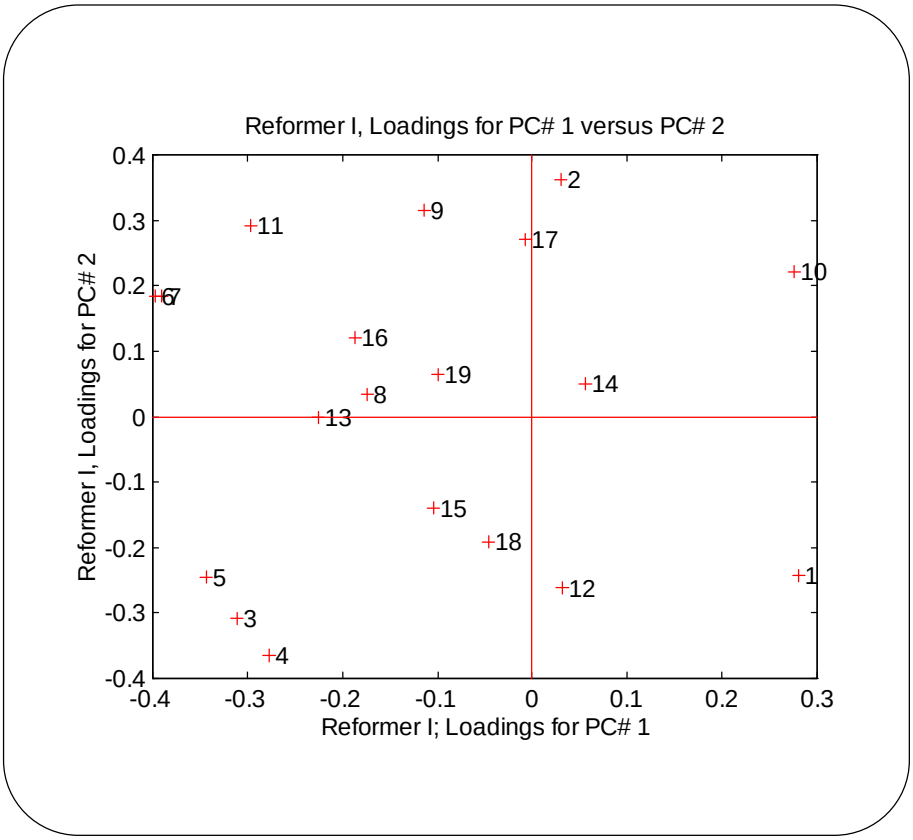

Figure 4.9 : Loading for PC number 1 and 2.

4.7.2 PCA Model for Catalytic Reformer II

The input to the PCA model is X matrix containing 17 columns, representing 17 variables used in models for reformer II, and 9703 rows for the total number of existing objects. The input X data is autoscaled, and contains both the calibration and the validation sets used in the prediction model development.

Table 4.12 shows the percent variance captured by each PC in the PCA analysis. Figure 4.10 shows the eigenvalues of the correlation matrix versus principal component number. It this case, it can be seen that the choice of PC is in the region of the 4-8. To avoid loss of significant information 7 PCs is chosen to be included in the model.
Figure 4.7 shows the scores on PC number 1 along with 95 % limit of confidence interval. The scatter plot of the scores for PC number 1 versus the scores for PC number 2 along with the 95% and 99% confidence interval limits is shown in figure 4.8. These two figures shows no significant outliers, and the major part of the data is indeed inside the 95% limit
Examining these figures shows a systematic variation in data, indicating two different operation modes. The period of these two operation modes are much less than the period in catalytic reformer I and is around 10 days. This suggests again that the operation modes are related mainly to the desired quality of RON and benzene contents rather than RVP. The conclusion is again that it is not appropriate to separate this two regions.

Percent Variance Captured by PCA Model			
Principal Component Number	Eigenvalue of Correlation (X)	%Variance Captured by this PC	%Variance Captured Total
1	7.21E+00	44.02	44.02
2	3.02E+00	18.42	62.44
3	2.50E+00	15.28	77.72
4	9.22E-01	5.63	83.35
5	7.62E-01	4.65	88.01
6	5.99E-01	3.66	91.66
7	3.53E-01	2.16	93.82
8	2.70E-01	1.65	95.47
9	2.22E-01	1.36	96.83
10	1.86E-01	1.13	97.96
11	1.19E-01	0.73	98.69
12	1.04E-01	0.63	99.32
13	6.16E-02	0.38	99.70
14	4.12E-02	0.25	99.95
15	5.63E-03	0.03	99.98
16	2.13E-03	0.01	100.00
17	6.73E-04	0.00	100.00

Table 4.12 : Percent Variance Captured by PCA Model

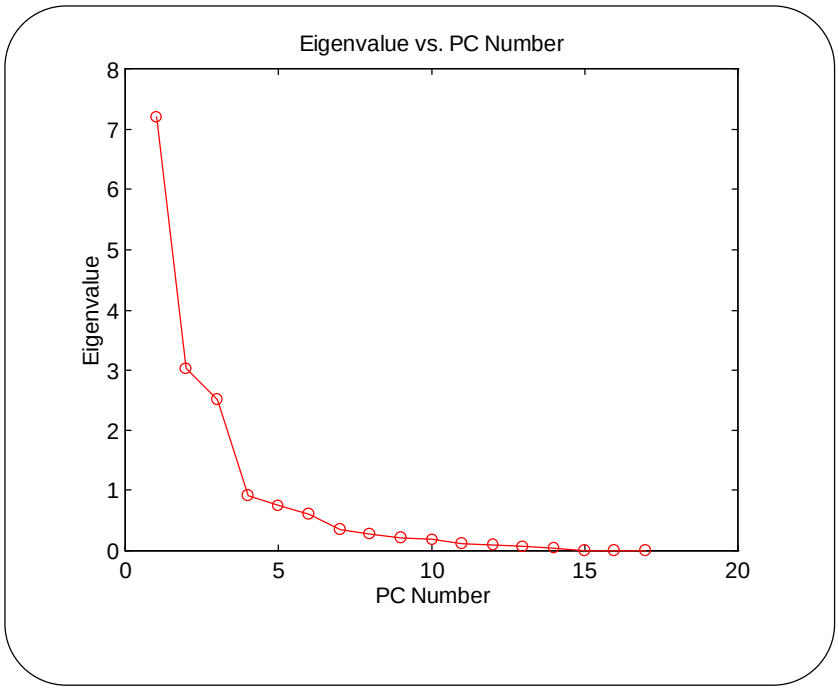

Figure 4.10 : Eigenvalue versus PC number.

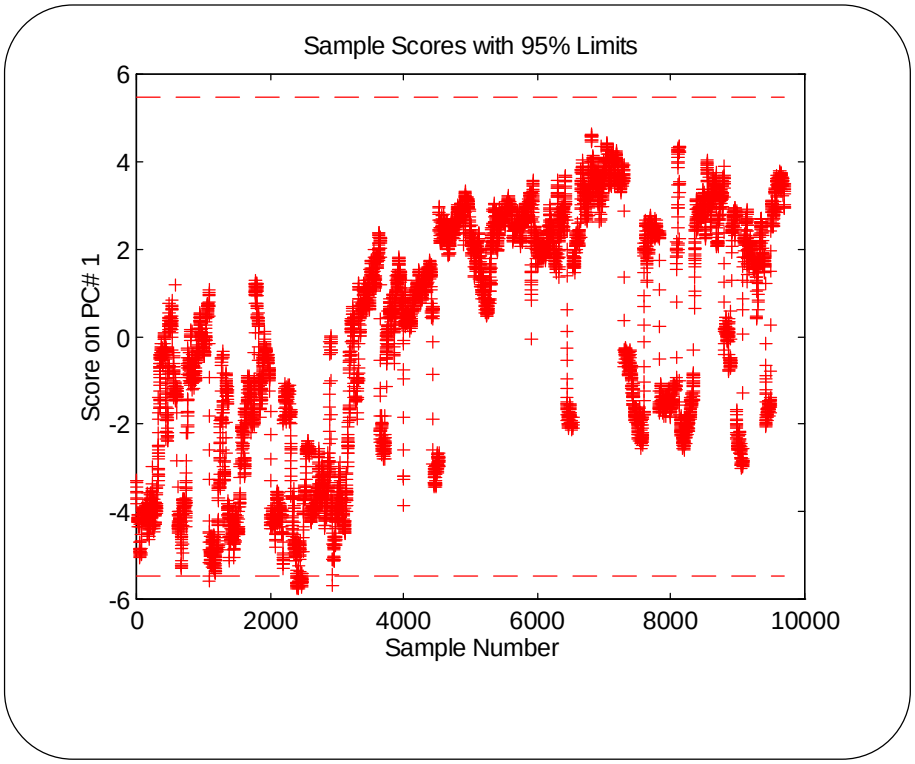

Figure 4.11 : Scores on PC number 1.

The loadings for PC number 1 versus number 2 is shown in figure 4.13, indicating the effect of variables on the two PCs. It can be seen that for example variables 6 , 9 and 11, which are flow rate of the feed to the reformer, feed to the stabilizer, and the reformate product, have the same effect on both PCs. The correlation means that one of them is enough to be included in this PCA model.

Notice that these variables are used in different prediction models, which can have different representation of the dynamic information.

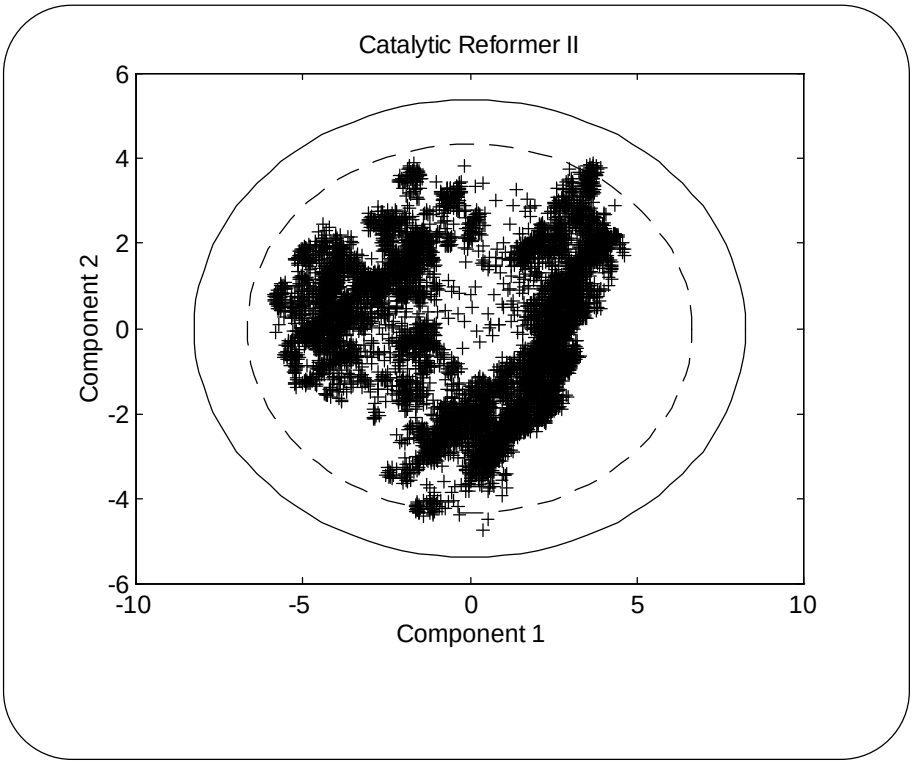

Figure 4.12 : Scatter plot of PC # 1 vs. PC # 2 showing the 95% and 99% limits .

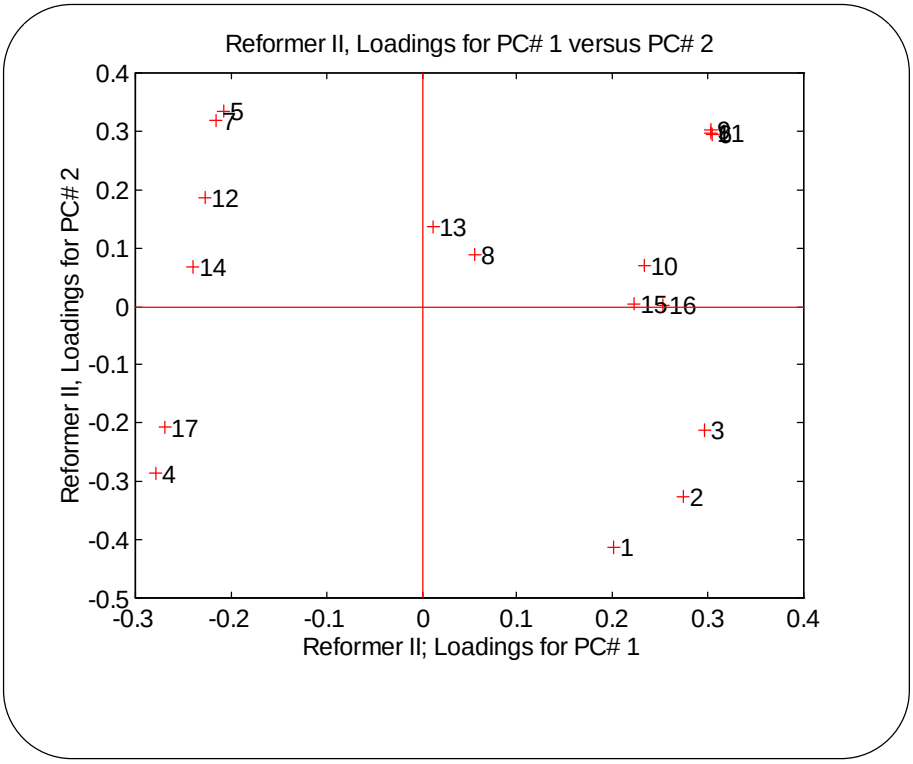

Figure 4.13 : Loading for PC number 1 and 2.

4.7.3 PCA Model for Isomerization Unit

The input to the PCA model is X matrix containing 12 columns for the 12 variables used in prediction models for isomerization unit, and 9467 rows for the total number of objects. The input X data is autoscaled.

Percent Variance Captured by PCA Model			
Principal Component Number	Eigenvalue of Correlation (X)	%Variance Captured by this PC	%Variance Captured Total
1	5.36E+00	44.65	44.65
2	2.20E+00	18.32	62.97
3	1.50E+00	12.47	75.44
4	9.93E-01	8.27	83.71
5	6.77E-01	5.64	89.35
6	4.00E-01	3.34	92.69
7	3.14E-01	2.62	95.31
8	2.06E-01	1.71	97.02
9	1.50E-01	1.25	98.27
10	1.09E-01	0.91	99.18
11	7.29E-02	0.61	99.79
12	2.56E-02	0.21	100

Table 4.13 : Percent Variance Captured by PCA Model

Figure 4.14 and table 4.13 show the eigenvalues of the correlation matrix versus principal component number and percent variance captures by each PC. Based on these information. a choice of 6 PCs is suitable for the PCA model in this case.

The plot of scores on PC number one, which is shown in figure 4.15, indicate that there are some objects that are out of the range of 95% limit of confidence interval. These objects are not outlier and identified to be from an operation point in which the temperature of the inlet stream to the first reactor and the temperature of both outlet streams are lower than the rest of the objects. The amount of data in this region is 12% of total number of data including both calibration and validation which consist of 9467 hours operation (about 13 months). This abnormality has obviously occurred in several periods mostly in calibration data as it can be seen in figure 4.15. It is more clear in figure 4.17 which shows the scatter plot of the scores for PC number 1 versus the scores for PC number 2 along with the 95% and 99% confidence interval limits.

Figures 4.16, and 4.18 shows the corresponding plot for scores on PC number 2, and a scatter plot scores on PC number 2 versus PC number 3 respectively. A comparison between figure 4.15 and figure 4.16 indicate that the abnormality is captured mostly by the first PC. However, figure 4.16 suggest that there is clear systematic variation in data indicating existence of different operation modes as in the reformer units.

Whatever reason for this abnormality may be, it will not affect the development of the prediction models since that portion of data should be excluded from the calibration set and hence the obtained model will be valid only for the normal operation point.

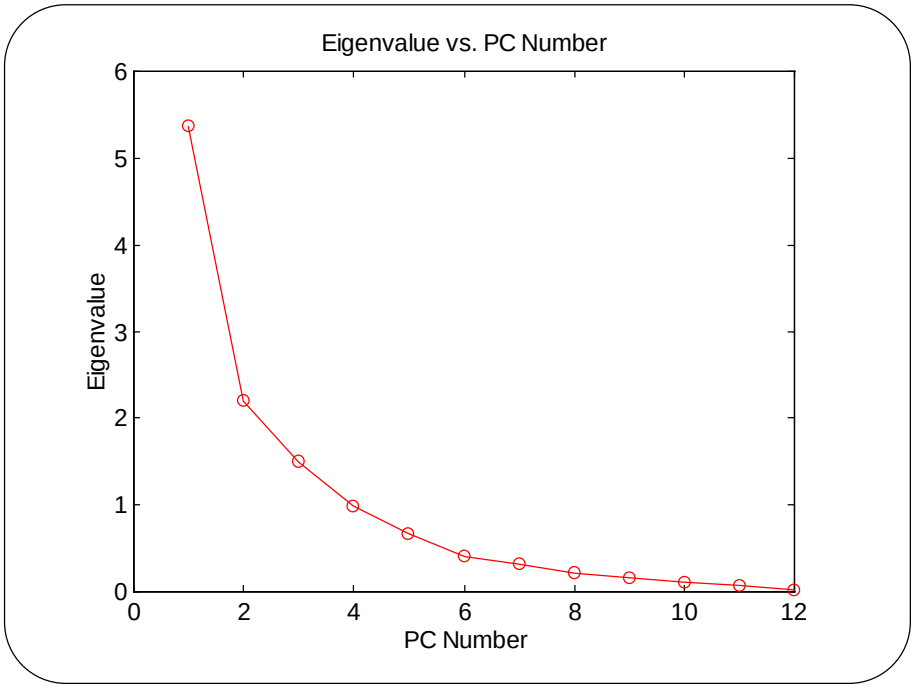

Figure 4.14 : Eigenvalue versus PC number.

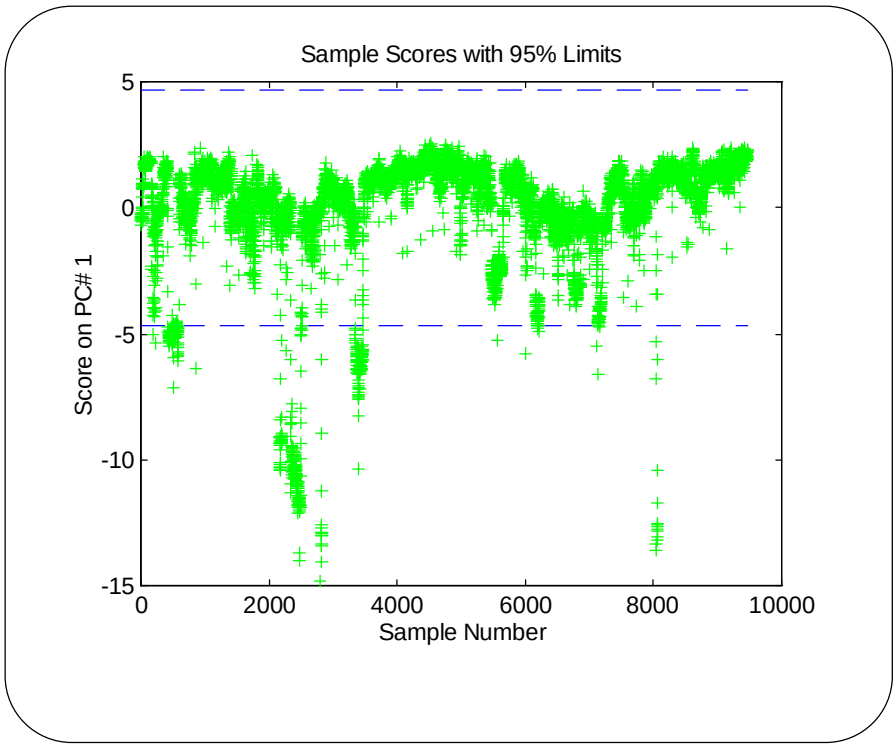

Figure 4.15 : Scores on PC number 1.

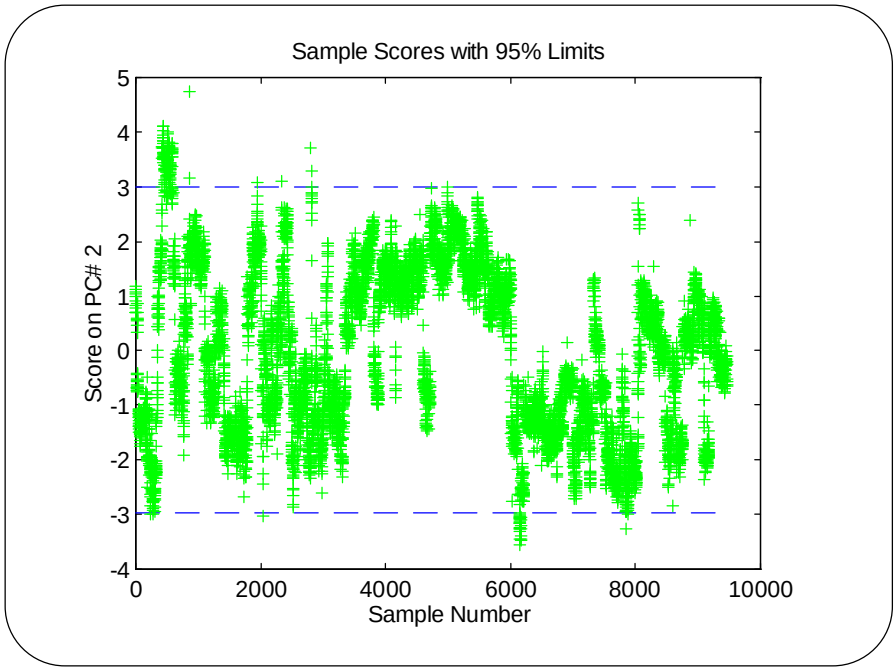

Figure 4.16 : Scores on PC number 2.

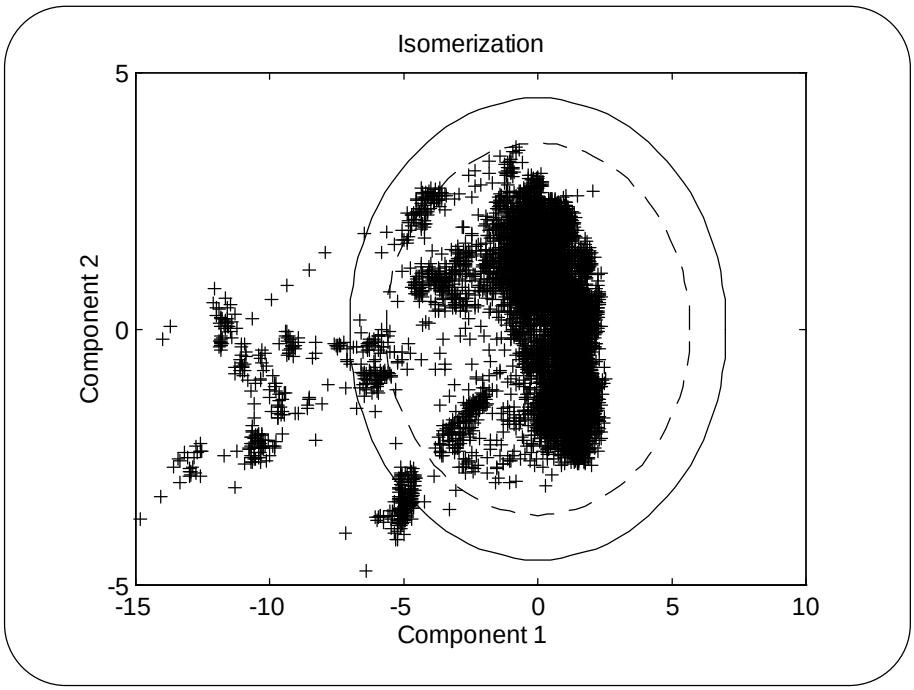

Figure 4.17 : Scatter plot of PC # 1 vs. PC # 2 showing the 95% and 99% limits .

The loadings for PC number 1 versus number 2 is shown in figure 4.19 indicating the effect of variables on the two PCs. It can be seen that variables 1, 2, 3, 4, 5, 7, 8, and 12 have the strongest effect on PC number one, and it is thus expected that these variables have the largest effect on the abnormality described before and observed in figures 4.17 and 4.15. A description of these variables can be seen in table 4.9.

Variables 1 through 5 represent the operation in the reactors. It is interesting to discover that variables 7, 8, and 12 which are the flow rate of DIP bottom, DIP feed, and splitter feed, has also an effect on the abnormality. Examination of the data in the period of abnormality reveal that the average of these flow rate were also lower than the average of the rest in the data set.

Hence, a possible explanation for this abnormality is problem with low LVN product indicating a possible cooling capacity limitation in the splitter distillation column.
This PCA analysis is in fact an example of uncovering the source of hidden information in the data, which can be use as guide for discovering the source of a potential problem in the process of a large plant.

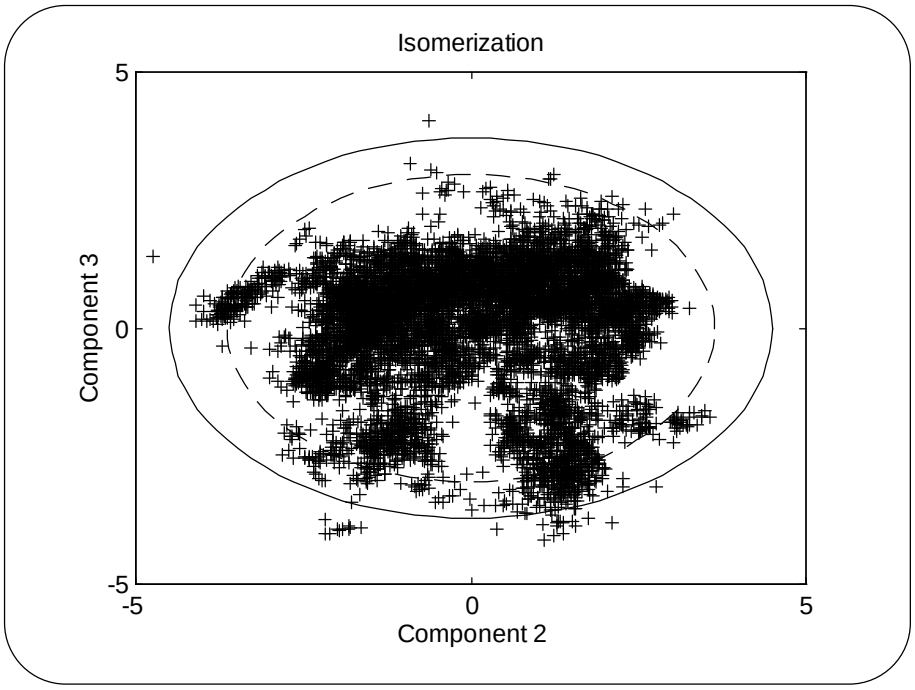

Figure 4.18 : Scatter plot of PC # 2 vs. PC # 3 showing the 95% and 99% limits .

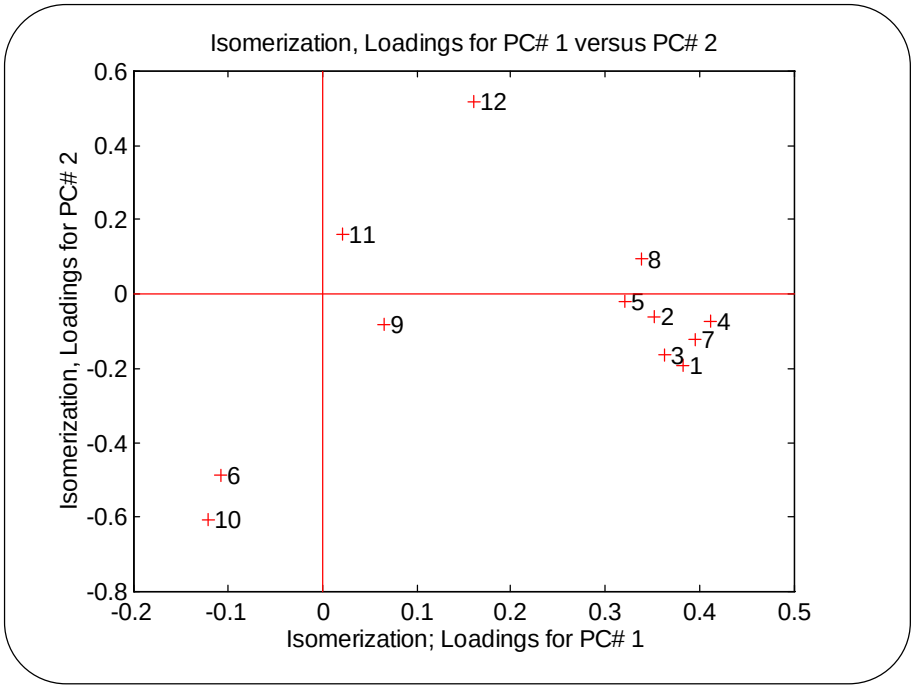

Figure 4.19 : Loading for PC number 1 versus 2.

4.8 Conclusion

In this chapter, the essential steps in development of multivariate process model applying chemometric approach is presented. The purpose is to describe how every single step is performed in this work in order to provide a strong background for description of the models presented in chapter 5.

A general scope for chemometric model development and a procedure with its different steps is presented and discussed in section 4.2. This procedure includes discussions of the objective of modeling, selection of variables, data collection and sampling, data treatment and scaling, selection of suitable method, calibration and validation of the obtained chemometric model.
A more specific system definition and description of the process with different production units is presented in order to clarify the background and the motivation for the selected input variables to be used for the prediction models.
In section 4.4, and 4.5 the description of output and input variables are presented. The calibration and validation data sets along with sample mean, variance, maximum and minimum values are presented for each variable in order to assess the data.
The problem regarding sampling frequency is discussed in section 4.6. This discussion will serve as a background for the proposed solution discussed in chapter 5.
A PCA model is developed for each catalytic reformer and isomerization unit. The PCA analysis is performed in order to assess and classify the type of behaviours represented in the data used for the developed models. The PCA analysis shows that there is a systematic variation in data, indicating the existence of at least two different operation modes. The period of operation in each mode is around a few days up to a few weeks. The results of PCA suggests that the operation modes are related mainly to the desired quality of RON and benzene contents rather than RVP. It is concluded that it is not appropriate to separate these two regions, and a common model should be developed for the whole calibration data set containing 10 months operation data.
The PCA analysis has also shown an abnormality in isomerization unit, indicating that a portion of the obtained data corresponding to 12% of total 13 months operation lies outside the 95% confidence interval. The analysis has shown that the operating temperatures in the reactors, along with the feed flow rate to isomerization and deisopentanizer unit were low in those period of abnormality, suggesting a possible obstacle in operation of the splitter. The data for the period of abnormality is excluded from the calibration set and hence the obtained model will be valid only for the normal operation.

Chapter 5

Models for Reformate and Isomerate Products

5.1 Introduction

5.1.1 Purpose

The purpose of this chapter is to describe structure, calibration, validation, and performance of the chemometric models developed for quality prediction of reformate and isomerate products in gasoline processing area. The objective is to describe the basic principles in model development, and to demonstrate how the problems are solved. In this chapter the developed models for prediction of RON, RVP, and aromate (benzene) contents of the product in catalytic reformer I are presented and discussed. The corresponding models for catalytic reformer II and isomerization unit can be found in appendix A and B since the basic steps and the results of these model identifications are quite similar to those described in this chapter.

5.1.2 Background

As it is described in chapter 2 and discussed further in chapter 4, there are a total of 9 different blend components used in the gasoline blending unit. It is important to know the quality variables of all streams sent to the gasoline blender in order to optimize the quality of final gasoline products. For some blend components, it is easy to estimate or calculate the qualities. For instance, three of the total nine blend components can be considered as almost pure components. These are oxygenate, butane and isopentane. Their qualities can be estimated based on pure component property. Regarding LVN and LVBN; i.e. light virgin naphtha and light virgin visbroken naphtha, the qualities can be calculated or estimated since they contain light hydrocarbon components which can be identified by chromatographic analysis.
However, for the reformate and isomerate products from catalytic reformers and the isomerization unit it is not possible to estimate or calculate the qualities easily due to the complex process of reforming and isomerization, and predictive models are needed. Furthermore, it is expensive to have on-line quality measurements for these three intermediate products in order to have the same sampling frequency as the other process variables. The only existing quality measurement for the reformate and isomerate streams are laboratory analyses, which are available only once per day, i.e. a sample interval of 24 hours, for each quality. Input variables in these units are measured on-line, and a sample interval of one hour is chosen for these variables in order to satisfy the demand for a sampling rate 4-10 times faster than the response of the system, as described in chapter 4.
The output from the models are Research Octane Number (RON), Reid Vapor Pressure (RVP), and aromatic contents of the products (Benzene). The inputs are a set of selected process variables. In order to fulfill the assumption of an informative input set as described in chapter 3, a set of inputs is selected based on general chemical engineering principles and knowledge of the process, in which the selected inputs are expected to have significant influence on the output variable. The inputs and output variables were presented and discussed in chapter 4.

The reformate and isomerate products from catalytic reformers and the isomerization unit are especially interesting since they have high octane number and low aromatic characteristics. The gasoline blending process can be considered as a batch process in which the blend components are used in order to produce the final gasoline products. The quality and volume of the final products are fixed by the refinery production schedule. If there is no feed introduced to the blendstock tanks during the period of blending, i.e. so called standing tanks, then the measured or predicted qualities of the blend components remain constant during the blending, and hence a Linear Programming (LP) approach in optimization of blending process would be successful (Singh et al. 2000).
The normal practice for the gasoline blending process in the refinery is that the qualitiy of the material in the blendstock tanks are measured once a day and used to estimate the first recipe for the blending process by using a LP optimization approach. The process is controlled by performing feed-back control using on-line measurements of the qualities in the outlet stream of the gasoline blending.
The blendstock tanks are supplied by continuous feed streams during the blending process, i.e. so called running tank. In this situation the assumption of constant quality of the blend components is no longer valid and applying the LP formulation will not handle such time-varying feedstock qualities adequately. A solution for the blending problem is obtained since feed-back control is used based on on-line quality measurements of the product stream.

However, this solution will not be an optimal solution due the time-varying qualities of the blend stocks.
The RON quality of the reformate streams is predicted using a multivariate regression methods and then applying bias updating. The bias updating is based on the daily measurements of RON reported by the laboratory and involves comparing the measured RON with those predicted by the model and then the difference is added as a constant in order to update the model with the new measurement. In this approach it is assumed that the RON quality remain unchanged during the period of bias updating, which is not an appropriate assumption.
The solution strategy proposed in this thesis is based on applying a moving horizon optimization for the blending problem in which the quality variables are predicted based on the variation of the upstream process and then provide the blending optimization problem with the predicted previous, present and future qualities of blendstocks (Nikalou 1998, and Singh et al. 2000).

In this chapter, the prediction models for the qualities of reformate and isomerate products sent to the blendstock tanks are presented and discussed.

5.1.3 Outline

In section 5.2 a discussion about selection of the methods applied for model development is presented. In this section the proposed technique for handling the missing values of output variables are described.
The models for prediction of the quality of the reformate product stream of catalytic reformer I, along with model development procedure, are presented and discussed in section 5.3.
The conclusions for the quality prediction models are summarized in section 5.4

5.2 Selection of the Method

In chapter 3, the different methods in process chemometrics are discussed. These are static and dynamic modeling methods in which both of them may use linear or nonlinear approaches in parameter estimation. All these type of modeling methods has been examined for prediction of the quality of reformate products. A great del of time has been spent on developing neural network models as a static nonlinear method, PLS as a static linear, and ARX as a time-series linear method.

It has been found that the output signal exhibit correlation to the past values of input and output signals. As it is motivated by the discussion in chapter 4 regarding sample rate of output signal, the best result has been found using the ARX identification method.

Parameter estimation in an ARX model is conventionally performed by a Least Squares (LS) regression method, as it is suggested by Ljung (Ljung, 1987). Better result in this work is obtained by ARX model in which the parameters θ are estimated by a PLS model in analogy with the described relationship between a PCA and ARX model by Wise and Gallagher 1996. Applying a PLS method in parameter estimation of ARX model has increased the strength of the predictability by taking advantage of the ability of PLS method to extract the useful information from collinear, noisy, input data which is relevant for modeling the output .

The low sampling frequency for model output has given rise to a challenging problem in this work. A solution to this problem is proposed which is related to one of the following two situations.

If the output variation is slow moving from one day to the next day, as the case of RON, a linear interpolation of the output is performed in order to estimate the missing output values in calibration data set. However, applying interpolation is avoided in validation data set. During the validation the model predicted output at time t is used to compute the output at time t+1.

If there is a considerable variation in the output signal, indicating a possible faster dynamic during a day, such as the case of prediction of RVP and aromatic contents, interpolation is not an appropriate approach in order to estimate the missing output. The solution here is that the information of the pervious outputs is imposed in the suitable structure of the ARX model. The structure of the ARX model is based on a form of regression vector in which the hourly sampled input variables are used together with the previous existing output measurement, normally measured at time t-24 hour, in order to model the output at time t. This solution is integrated in the regression matrix of the ARX structure, in which the delay time for output is inherently 24 hours. In the following a review of the ARX model structure is presented:

$$A(q)y(t) = B(q)u(t) + e(t) \tag{5.1}$$

$$\begin{aligned} A(q) &= 1 + a_1 q^{-1} + \ldots + a_{na} q^{-na} \\ B(q) &= b_1 q^{-1} + \ldots + b_{nb} q^{-nb} \end{aligned} \tag{5.2}$$

$$\theta = [a_1\ a_2\ \ldots\ a_{na}\ b_1\ b_2\ \ldots\ b_{nb}\ ] \tag{5.3}$$

$$\varphi(t) = [y(t-1)\ldots.y(t-na)\quad u(t-1)\ldots.\ u(t-nb)]^T \tag{5.4}$$

$$\hat{y}(t|\theta) = \varphi^T(t)\,\theta \tag{5.5}$$

where:

y(t) is output measurement at time t,

u(t) is input measurement at time t,

e(t) vector of white noise sequences,
q is the backward shift operator q^{-1} in which q^{-1} u(t) = u(t-1)
na is number of A parameters,
nb is number of B parameters,
θ is the model parameters which include a and b parameters,
ϕ(t) is the regression vector,
$\hat{y}(t|\theta)$ is prediction of new y at time t as a function of model parameters.

Following the terminology in ARX model, na and nb are defined as the number of A and B parameters for previous output y and inputs u, respectively. The order of the ARX model is defined to be the number of past input and output variables considered in the model.
Furthermore, there may well be a delay from output to each of the input variables. These delays are defined as a vector K of scalar value k_i corresponding to delay for each input and output. Thus, nk would be the number of k delay parameters which is equal to number of input variables.
Thus the standard ARX model may be written as equation 5.1, where inputs and outputs are sampled with the same sample interval. However for many quality variables the standard sampling procedure involves a n_q =24-hours sampling interval, whereas the input variables are assumed known every hour. In order to use the available data for model development a suitable model representation must be developed. This is here accomplished based upon the ARX model in equation 5.1. Since however the outputs between ....t-n_q, t, t+nq.... are unknown, a model predict these values would be convenient for developing a predictor for the quality variables at their sampling times. The development of this predictor is here based on a simple first order predictor based upon the sampling interval for the input vector u(t):

$$\hat{y}(t+1) \quad = \quad ay(t) \; + \; bU(t-k) \tag{5.6}$$

where k is the vector of input delays.
The above model is used to predict the quality variable y(t+n_q) as follows:

$$\hat{y}(t+2) = a\hat{y}(t+1) + bU(t-k+1) = a^2 y(t) + abU(t-k) + bU(t-k+1)$$
$$\hat{y}(t+3) = a\hat{y}(t+2) + bU(t-k+2) = a^3 y(t) + a^2 bU(t-k) + abU(t-k+1) + bU(t-k+2)$$
$$\vdots$$
$$\hat{y}(t+n_q) = a^{n_q}\, \hat{y}(t) + \sum_{i=1}^{n_q} a^{i-1}\; bU(t-k+n_q-i)$$

Thus defining new parameters as follows:
$a_1 \quad = \quad a^{n_q}$
and
$b_i \quad = \quad a^{i-1}b \qquad for \qquad i = 1, ..., n_q$ (5.7)

the predictor may be written as:

$$\hat{y}(t+n_q) = a_1\, \hat{y}(t) + \sum_{i=1}^{n_q} b_i U(t-k+n_q-i)$$
$$= a_1\, y(t) + b_1 U(t-k+n_q-1) + b_2 U(t-k+n_q-2) + ... + b_{n_q} U(t-k) \tag{5.8}$$

In this predictor the parameter values may be determined from plant data. It must be noted however that according to equation 5.7 there are only n_u+1 unknown parameters, where n_u is the number of inputs in U(t-k). Since however the parameters in equation 5.8 are nonlinearly interdependent, it will be attempted to estimate all $n_u * n_q+1$ parameters in equation 5.8 using a linear parameter estimation method.
To summarize, the above predictor is based upon a first order model in the input sample time. In the following presentation and discussion, the model order will be labeled one with respect to the output and the number of inputs included in the predictor should be n_q according to equation 5.8. When applying equation 5.8 for modeling, it will be attempted to use n_q number of previous inputs. However if it becomes difficult to estimate all $n_u * n_q$ input parameters one should apply a nonlinear parameter estimation method to determine the a and b parameters directly.
The derivation of the first order predictor in equation 5.8 was based on a first order model in the input sample time. A higher order model might also be used, which would lead to usage of older quality variable measurements, which means that y(t-24) and y(t) would be used to predict $\hat{y}(t+24)$. Such a model would lead to an even higher number of parameters to be estimated using a linear estimation method. For higher order models initialization also becomes an issue.

It is noteworthy to mention here that we start with a number of nb equal to 24 in order to assure that all existing variation in input variables is included. It is important for the predictability and quality of the model that the existing dynamic variation in input which is most relevant for modeling of output is covered. Thus, the choice of nb=24 seems to be most appropriated. However, choosing number of nb equal to 24 has a disadvantage that the number of model parameters will become high.
These issues will be discussed further in calibration and validation of the RVP model in section 5.3.2. Hence, the parameter nb, i.e. the number of input vector, is assumed unknown for the time being and must be determined during the model development. In this work when we talk about the number of input vector, we actually mean only the number of past input variables to consider, i.e. number of nb, since we have a fixed number of na equal to one.
The regression vector in equation 5.4 is used in the objective function in the LS method, in which it is minimized with respect to θ, in order to find the best fit, as it is discussed in chapter 3. The regression vector can be expressed as follows:

$$\begin{aligned}
\varphi(24) &= [y_0\ U_{23}\ U_{22}\ U_{21}\ .....U_{24-nb}] \\
\varphi(48) &= [y24\ U_{47}\ U_{46}\ U_{45}\ .....U_{48-nb}] \\
&\ldots \\
&\ldots \\
\varphi(N) &= [y_{N-24}\ U_{N-1}\ U_{N-2}\ U_{N-3}\ .....U_{N-24-nb}]
\end{aligned} \qquad (5.9)$$

where N refer to number of existing output variable.
It is assumed that the delay parameter k is equal to one in equation 5.9. All the regression vectors in equation 5.9 are used to form a regression matrix, which is used in a PLS regression model in order to estimate the model parameters θ defined in equation 5.3.
An example of how the regression vector is built for this application is shown in table 5.1. The three columns in table 5.1 represent respectively time in hours, output y and input U. Let assume that the starting time is t0, and then y0 and U0 is the corresponding output and inputs at time 0. The first predicted output in this formulation would be y24.

Thus, the regression vector used for prediction of y24 contains the previous output, i.e. y0, and also all the previous values of U inputs, from U23 and backward corresponding to the determined model order, i.e. nb, and the delay k. The same procedure is used to form the regression vector corresponding to prediction of y48.
Determination of the nb parameter will lead to the number of the previous inputs considered in the model, which is shown by gray area in table 5.1. It is assumed that the delay parameter k is equal to one in table 5.1.

T	**Y**	**U**
0	y0	U0
1		U1
2		U2
..		..
..		..
..		..
21		U21
22		U22
23		U23
24	y24	U24
25		U25
26		U26
..		..
..		..
45		U45
46		U46
47		U47
48	y48	U48
49		U49
50		U50
..		..
..		..

Table 5.1 : The structure of regression vector with delay k=1.

The software used in this work is MATLAB for windows, The MathWorks, Inc., version 4.2c.1, 1994. For ARX and PLS model development the Identification Toolbox of Matlab, and a university version of the PLS-Toolbox version 1.5.1, 1995 by B.M. Wise, as well as the routine developed by the author of this thesis are used.

5.3 Models for Catalytic Reformer I

5.3.1 Introduction

In this section the models for prediction of RON, RVP, and benzene contents of reformate production from catalytic reformer I will be presented and discussed.
The input variables used for the models are described in chapter 4, along with a detailed Principal Component Analysis (PCA), and a description of data treatment.
In the following sub-sections, there will be more focus on model structure, calibration, validation, and performance of the models.

5.3.2 RVP Model

5.3.2.1 Inputs and Output

The selection of input variables relies basically on general knowledge of the process and an assessment of which variable have the largest effect on the output. The principles in selection of input variables in order to obtain a set of informative input data are discussed in chapter 3.
If a model is sensitive to one or more input variables meaning that those variables have a large influence on the predicted output, then the corresponding parameter values will be large compared to the other parameters.
This concept is used in model development by using all candidate input variables in the beginning and then after validation the non sensitive variables are excluded from the inputs. The advantage of this procedure is to prevent exclusion of those variables that can be influential on the output prediction due to a multivariable effect or existence of an unknown phenomena.
In this section an example of this procedure is presented. The following input variables are used in the RVP model as the initial selected variables.

1 Mole H2/ Mole C in recycle gas
2 % H2 purity in recycle gas
3 Reactor 1 outlet temperature
4 Reactor 2 outlet temperature
5 Reactor 3 outlet temperature
6 Reformer Feed flow rate
7 C-401 Reformate flow rate
8 C-401 Liquid gas flow rate
9 C-401 Reflux flow rate
10 C-401 Feed temperature
11 C-401 Reboiler temperature
12 C-203 Reflux flow rate
13 C-201 Naphtha side stream temperature (Pressure Corrected)
14 C-601 Naphtha side stream temperature (Pressure Corrected)
15 C-203 Bottom temperature (Pressure Corrected)
16 C-652 Bottom temperature (Pressure Corrected)

It will be shown that for the final model variables number 7 through 12 will have the largest effect among all variables, as it is expected.
The output is RVP measured by laboratory. Thus, the mode will be a Multi Input Single Output (MISO) case.

The calibration data set is chosen from a period of approximately 9 months operation starting from October 1. 1996 to June 13. 1997. The validation data cover approximately the rest of 1997, i.e. June 13. 1997 to December 30. 1997. There are some days, both i calibration and validation, where both input and output data are missing. These missing datas are mainly due to operation shutdowns. Besides, there are also some missing data for a few hours because of problems in data acquisition system or sensor faults.

5.3.2.2 Model Structure

As mentioned earlier, the model structure is based on an ARX model in which the parameters θ are estimated by a PLS model. Based on the discussion in section 5.2, we will take only the effect of y(t-24) for output, as in equation 5.8, and hence we will have only one A parameter. Regarding the B parameters we are seeking for as much effect from the inputs, and hence the B parameters will be as many as necessary to get an acceptable low prediction error and a satisfactory model performance, as it is discussed in the following. We are specially interested to examin the case of nb=24, as it is discussed in the following section.
In parameter estimation we use a PLS model in which we need to determine a suitable number of principal components PC. The number of PC is also called number of Latent Variables (LV).
Thus, there are two parameters in model development, i.e. nb and the number of LV, which have to be found.
This task is handled by developing a recursive routine in Matlab, in which the Root Mean Sum Squares Error in Validation (RMSSEV) is used as the criterion for optimization of nb and number of PC. RMSSEV is defined as the following.

$$RMSSEV = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (5.10)$$

where $\hat{y}_i$ is the model predicted output and y_i is the measurement for all data over time t. Notice that we have only one output and n is the total number of y. The results for these simulation are described in the next subsection.

5.3.2.3 Identification

As mentioned before, the purpose of calibration is to estimate the optimum values for the model parameters θ, along with nb, and the delay parameter k for each input variables. Notice that na is one in our case. The purpose of the validation is, however, to evaluate the model obtained in the calibration. Since the model has time-series dynamic characteristic, it is important to secure a calibration and validation set containing time sequence of subsequent data. For that reason, it is not desirable to mix the data and select a random test set data for validation. Furthermore, based on process knowledge it is known that the operation mode is different in summer and winter seasons. Thus, the validation is performed applying a

completely distinct set of data, and it is attempted to cover both winter and summer operation mode both in the validation and calibration data sets.
Calibration and validation of the obtained models are inherently related, and the models are evaluated based on some criteria concering both calibration and validation phases, as we shall see in the following sections.

5.3.2.3.1 ARX Model With All Inputs

Referring to the developed model structure defined in equation 5.8, it is especially interesting to study the case of nb=24, in which the effect of inputs is covered all the way back to y(t-24). This case is discussed in this section.
Nevertheless, as we shall see in the next section, it will be shown that special cases exist in which the number of model parameters can be reduced with no significant loss of prediction ability.
Number of delay parameter in this case is k=1 for all input variables, since it is desired to take the effect of all previous input values on the prediction of output, even if the effect is small.
One parameter remaines to be estimated, and that is number of latent variables LV. Figure 5.1 shows the result of a series of recursive simulations, in which RMSSEV is calculated as a function of LV.

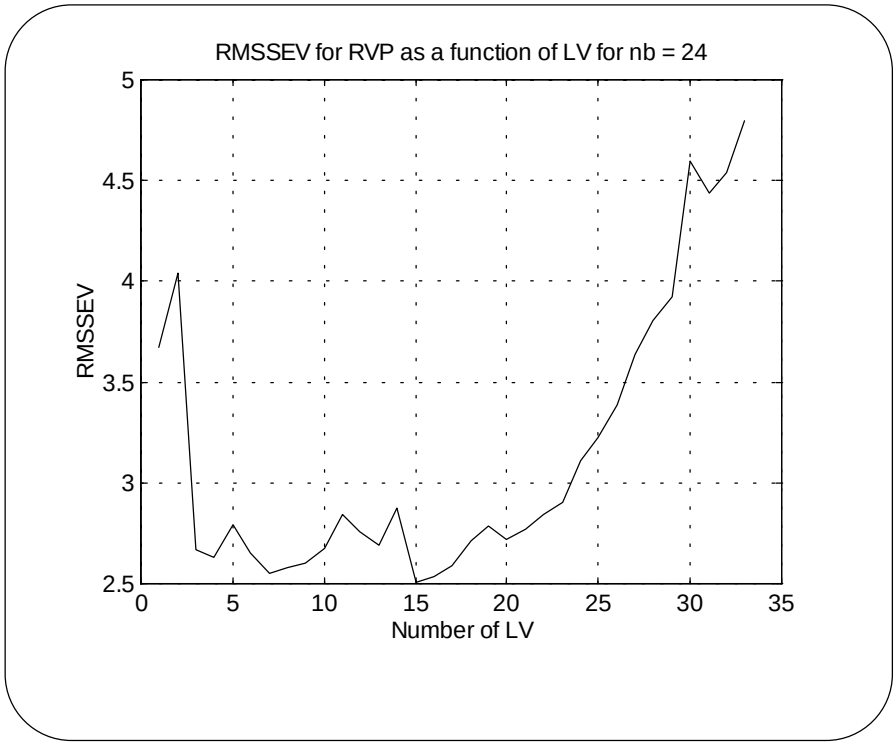

Figure 5.1 : RVP Model, RMSSEV as a function of LV for nb =24, and delay K=1.

It can be seen in figure 5.1, there are three local minima around LV equal to 4, 7, and 15. Choosing more LV will result in increasing total number of model parameter, which will cause over fitting, as discussed in chapter 3.
As we shall see in the next section, where we discuss the optimum number of LV, we will choose a number of LV = 6. Notice that with LV=6 a RMSSEV=2.65 is obtained, which is not far from the local minimum RMSSEV=2.5 at LV=15.

Another reason for choosing LV=6 is that the results of the case nb=24 presented in this section, is desired to be comparable with the results that will be presented in the next section.

Consequently, calibration of the model is performed, and the model parameters θ are estimated by choosing LV=6, delay parameter for all inputs equal to one, na =1, and nb=24. Notice that by having 16 inputs, we will get a total number of 385 parameters in the θ vector, according to equation 5.3.
Table 5.2 shows the percent variance captured by PLS model. As it can be seen, the captured variance in X-block, i.e. inputs, and Y-block, i.e. output, are respectively 75.99% and 50.35%.

Percent Variance Captured by PLS Model				
	X-Block		Y-Block	
LV #	This LV	Total	This LV	Total
1	30.17	30.17	18.08	18.08
2	21.92	52.09	8.52	26.60
3	3.34	55.43	16.85	43.45
4	5.56	60.99	3.37	46.82
5	4.96	65.95	2.77	49.58
6	10.04	75.99	0.76	50.35

Table 5.2 : RVP model, percent variance captured by PLS model.

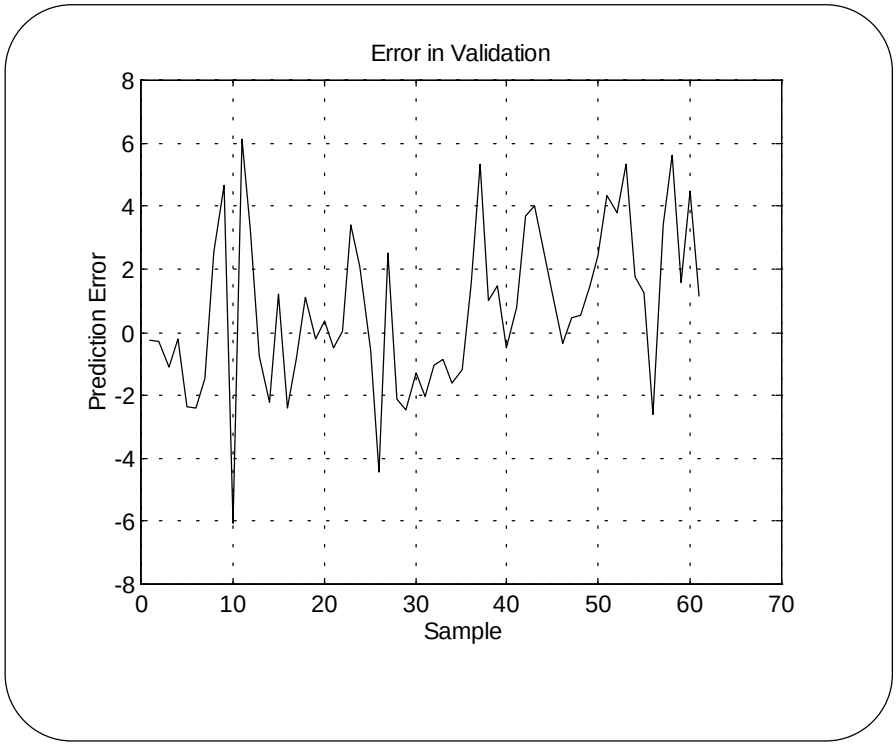

Figure 5.2 : RVP Model, prediction error in validation, nb=24, delay K=1.

Figure 5.2 shows the prediction error in validation. Notice that the error is the difference between the model predicted and the measured RVP and it has the pressure unit. i.e. kPa. Recal the measureed RVP data for catalytic reformer I in table 4.2 in chapter 4, in which the

average RVP is 49.67 kPa, and the standard deviation is 3.02 in calibration data. We shall later compare these data with the results presented in table 5.3. The corresponding prediction error in calibration is shown in figure 5.4.

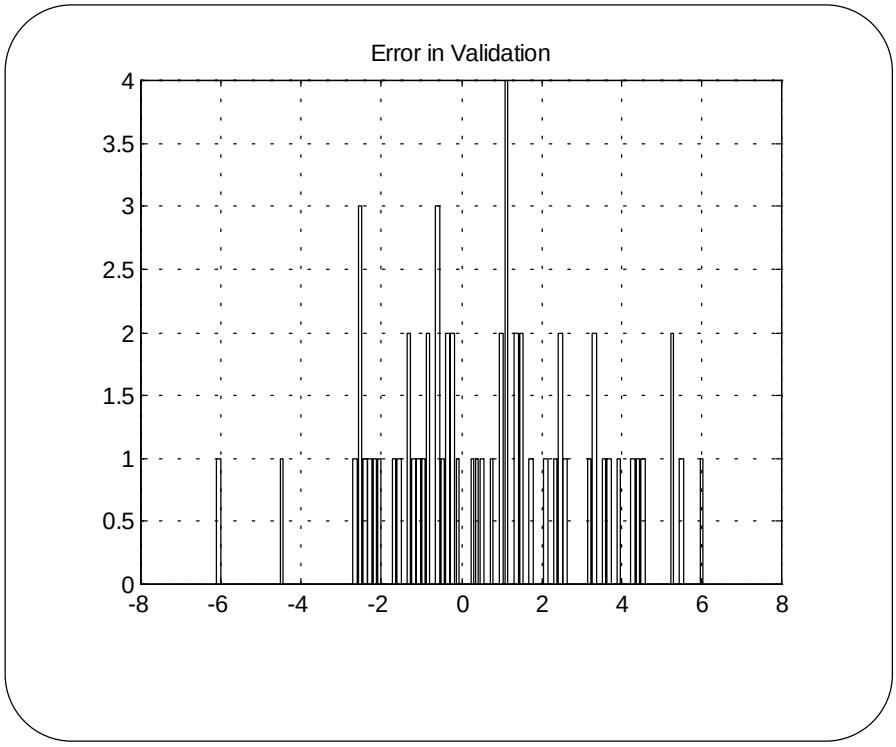

Figure 5.3 : RVP Model, histogram plot of prediction error in validation.

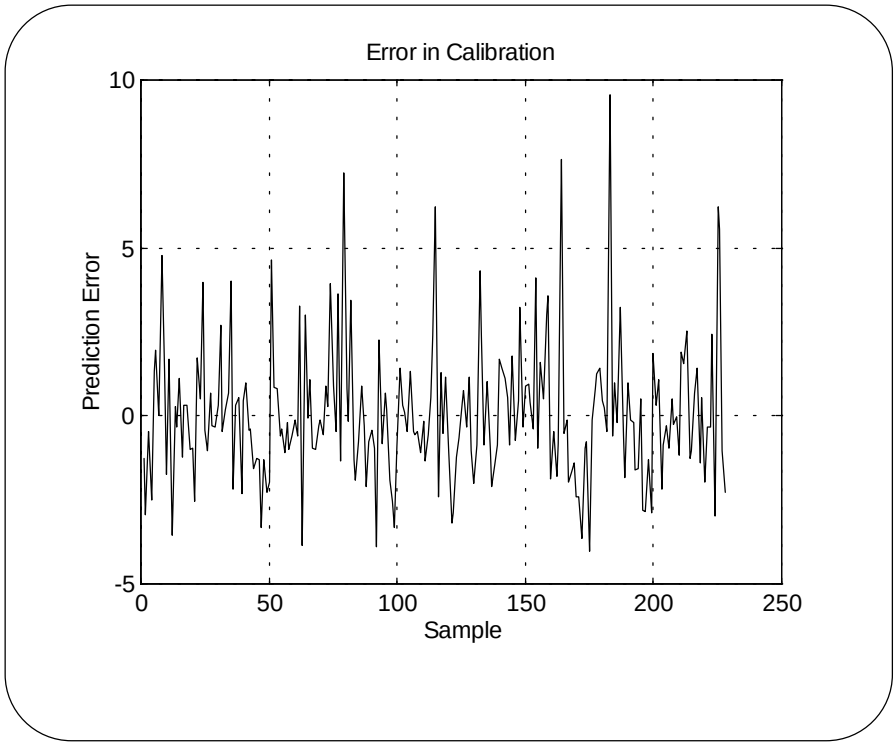

Figure 5.4 : RVP model, prediction error in calibration, nb=24, delay K=1.

Figure 5.3 shows a histogram plot of prediction error in validation. This plot will give us an impression of how close the error signal is to a normal distributed zero mean noise. The corresponding histogram polt for calibration is shown in figure 5.5.

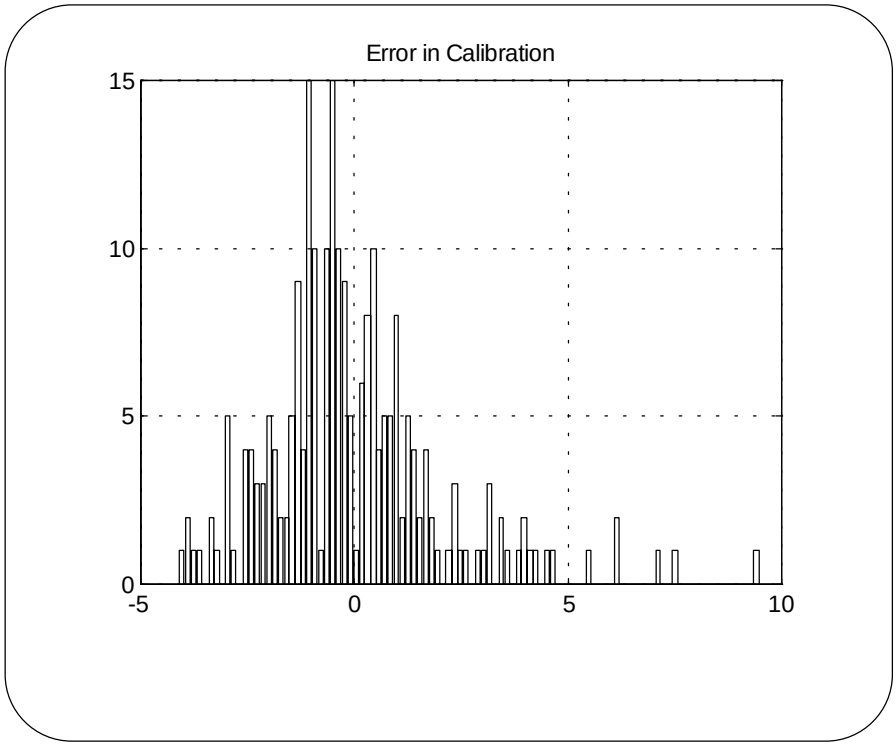

Figure 5.5 : RVP model, histogram plot of prediction error in calibration.

As it is described in chapter 3, in order to evaluate the performance of the model the RMSSE is compared with a reference obtained from either a zero-model or an average-model.
The average reference model is computed by first calculating the average of all N measured output values, and then subtract the average from the output itself to calculate E_{AVG} , as follows:

$$(E_{AVG})_i = y(t_i) - y_{AVG} \qquad (5.11)$$

Then, we compute a RMSS of this error by using equation 5.10, and denote it as RMSEAVG. It is clear that the RMSEAVG has the same property as the standard deviation of the measured output. We expect that the developed prediction model should predict a set of output values for a period of time which is closer to the measured output than the average value. In this sense we say that the developed model should be at least better than the average value model in order to be accepted.
A second reference model is defined based on the following consideration. Consider a model structure as given in equation 5.8 in which the number of A-parameter is 1, i.e. na=1. Furthermore, consider that the developed prediction model find a set of B-parameters which is close to zero, and an A-parameter value close to one This is shown in the following equation:

$$\hat{y}(t) = y(t-24) + 0 \qquad (5.12)$$

This means that the new prediction of y is equal to the previous measured y. In this case we have no effect of input variables and the model just predicts the next output equal to the

previous measured one. Based on this consideration, we compute a E_{ZERO} in a general form as follows:

$$E_{ZERO} = y(t_i) - y(t_{i+1}) \quad (5.13)$$

Hence, a RMSS of E_{ZERO} , which is denoted by RMSEZRO will give us a reference in assessment of predictability of the obtained prediction model. Thus, the expectation is that a model with good performance characteristics should be better than the zero-model, meaning that the developed model has captured the effect of input variables.
Table 5.3 shows RMSSE in both calibration and validation along with the average-model and zero-model in this RVP model.

Validation		Calibration	
RMSSEV	2.65	RMSSEC	2.05
RMSEAVGV	3.28	RMSEAVGC	2.91
RMSEZROV	4.10	RMSEZROC	3.05

Table 5.3 : RVP model, RMSSE, average-model, and zero-model in validation and calibration.

It can be seen that the RMSSEV is less than average- and zero-model, indicating that the model has captured essential variation both in input and output.

Another way to evaluate the model performance is a so called open-loop simulation of the model. In this simulation the predicted output at time t is applied instead of measured output in order to predict the next output value at time t+1. This model simulation is performed after model calibration. We shall call this simulation as open-loop simulation in which the new predicted value is used instead of the measured output for prediction of the next output value.
Open-loop simulation will show the predictability of the model during a period of operation without having the actual output measurement.
It is interesting to see the open-loop simulation in both calibration and validation for RVP model. These are shown in figure 5.6 and 5.7 respectively

In Figure 5.8 a plot of predicted versus measured RVP in calibration is shown. These plot shows how successful the calibration is performed. However, it is more interesting to study this plot in the validation case. The corresponding plot for the validation can be seen in figure 5.9. As it can be seen the model has captured the essential variation of RVP with a RMSSEV of 2.65.

The results obtained in this section will be discussed later in the next section, where possibility of reducing the number of model parameters will be discusssed.

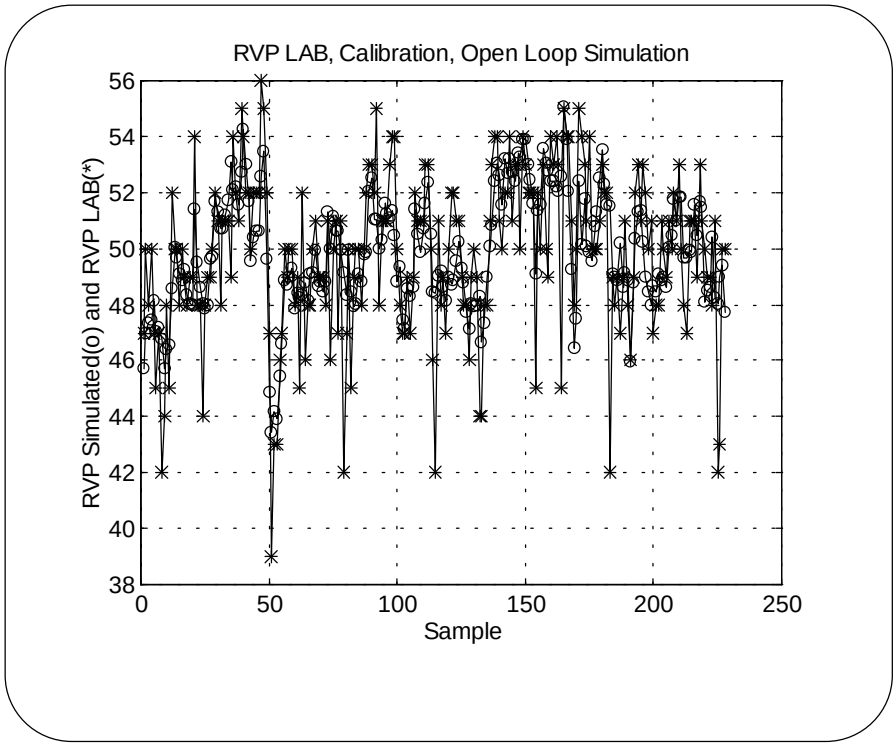

Figure 5.6 : RVP Model, Open Loop Simulation in calibration

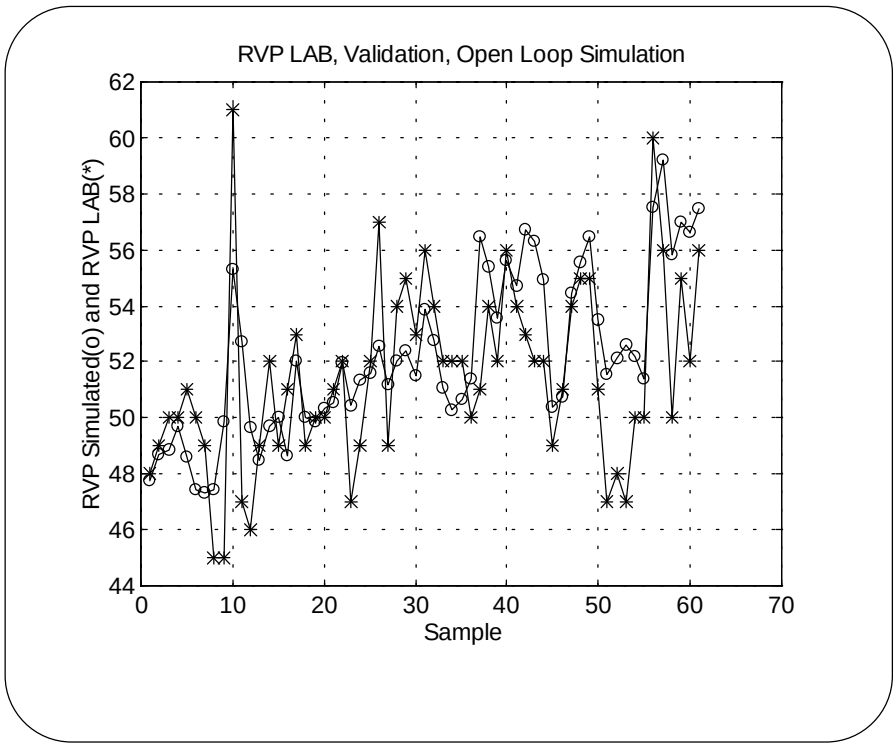

Figure 5.7 : RVP Model, open loop simulation in validation

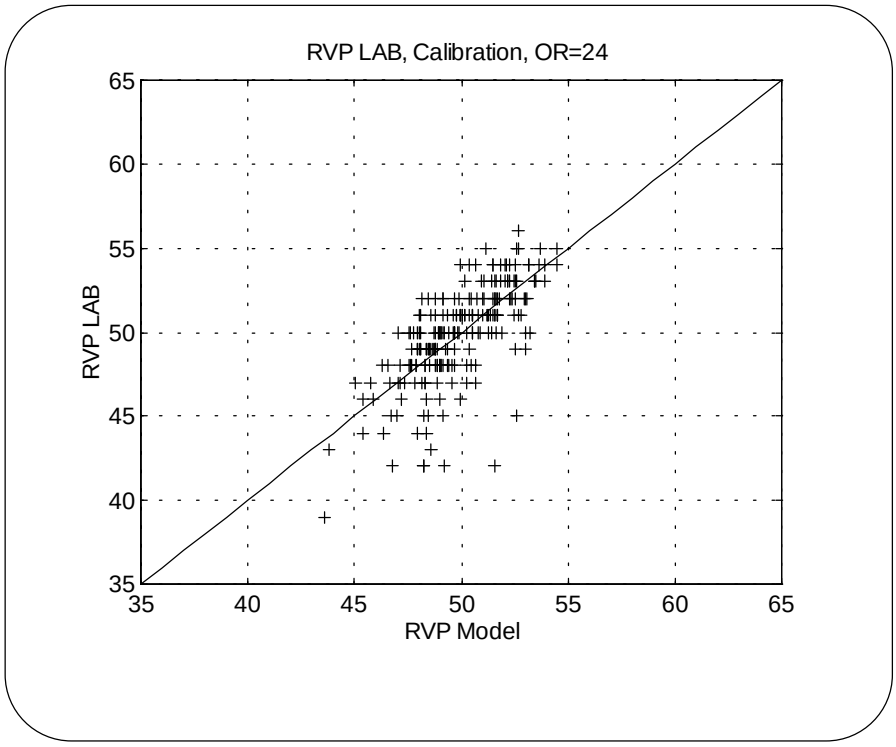

Figure 5.8 : RVP model predicted versus RVP measured in calibration.

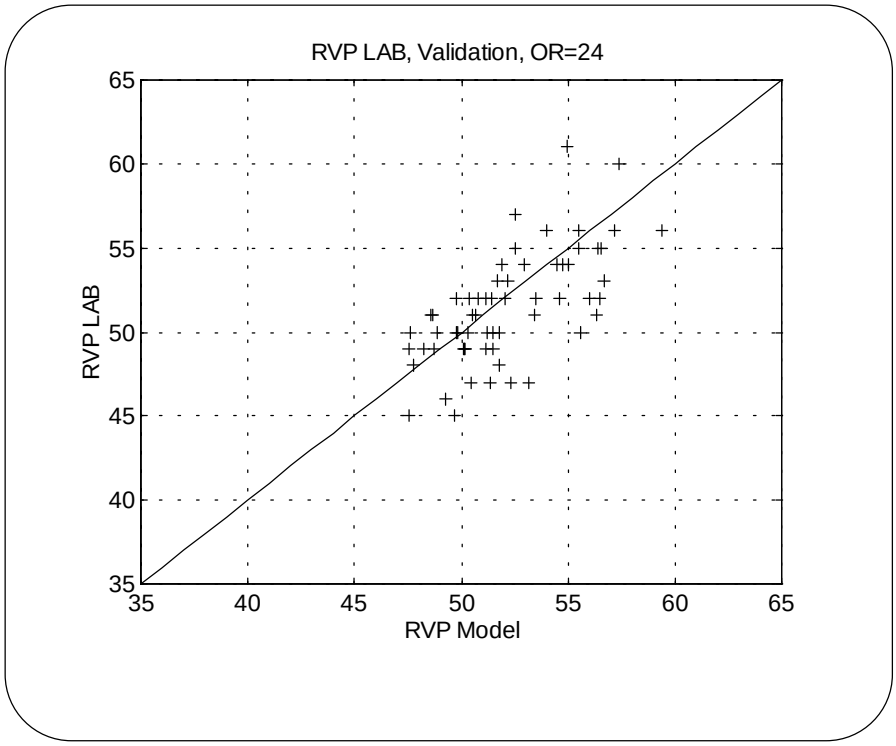

Figure 5.9 : RVP model predicted versus RVP measured in validation

5.3.2.3.2 Reducing the Number of Model Parameters

As it is mentioned in the previous section, a number of nb equal to 24 would be logical in order to cover the variation of input, since the predictor in eqution 5.8 contains y(t-24). The disadvantage of choosing nb=24 is the high number of model parameters, which may cause an over fitting problem. However the validation results in table 5.3 and figue 5.9 do not indicate problems with overfitting.
In this section, it is attempted to reduce and find the optimum number of model parameters, with no significant loss of predictability.

Apart from number of input vectors nb and number of LV, one more parameter has to be determined, which is the delay for each variables as expressed in equation 5.8.
One way to find the number of delay parameters is to perform calculation of residence time in tanks, vessels, units and pipeline. Moreover, if a first principles mathematical model was available, the effect of variables in energy balance, such as reactor outlet temperature could be investigated.
Another way is to let the model find the delay parameters. This can be done by a series of recursive simulations in which the best set of delay parameters are found giving the minimum RMSSEV defined in 5.10.
In this work, these two approaches are combined in which the search for delay parameters is limited by some qualified estimate according to process knowledge and physical restrictions, such as the length of the pipeline, the volume of tanks and etc, and then let the model find the best delay parameters.

The search for the optimal ARX model order for input variables, i. e. nb, and number of LV is carried out by a series of simulations, in which nb and LV are changed from 1 to 25 for nb, and from 1 to 33 for the number of latent variables (LV). The choice for maximum number of LV is based on the following considerations.
Number of LV is a function of number of variables, and ARX order, as shown in equation 5.14.

$$Max.\,LV = na \cdot ny + nb \cdot nu \qquad (5.14)$$

where ny is the number of output, and nu is the number inputs variables.

Since we have na = 1 and ny = 1, the product of na and ny is equal to one. Notice that we have 16 input variables, and hence the maximum number of LV will be 17 and 33 for respectively nb = 1 and nb = 2:

Max. LV = 17 for nb = 1
Max. LV = 33 for nb = 2

Selecting more LV has two disadvantages. First, choosing more LV means adding more noise to the structure part, and second, total number of model parameters will increase and it will cause overfitting, as it is discussed in chapter 3.
For that reason, a maximum of 33 number of LV has been chosen in this investigation for number of nb larger than 2.

Figure 5.10 shows the calculated RMSSEV as a function of nb and LV, for nb from 2 to 25 and LV from 1 to 33. The plot for nb=1 is shown separately in figure 5.13 since maximum of LV is 17 according to equation 5.14.
It can be seen that there is a region of nb less than 5-7 that RMSSE has its minimum. Moreover, RMSSEV increases in the region of LV larger than 15-17 due to additional noise in the structure part.
In table 5.4, the minimum RMSSEV is shown for each nb, along with number of LV at the minimum. The percent variance captured by PLS model is also shown both for input (X-Block) and for output (Y-block). As we can see in table 5.4, the minimum RMSSEV is found for nb = 2 and LV = 23 at a value of 1.89. Besides, as it is shown in figure 5.10 another local minimum appear to be around nb=7.
The next job is now to study the progress of RMSSEV for some nb parameters in more detail, and eventually obtain a model with fewer parameters, with no significant loss of prediction ability.

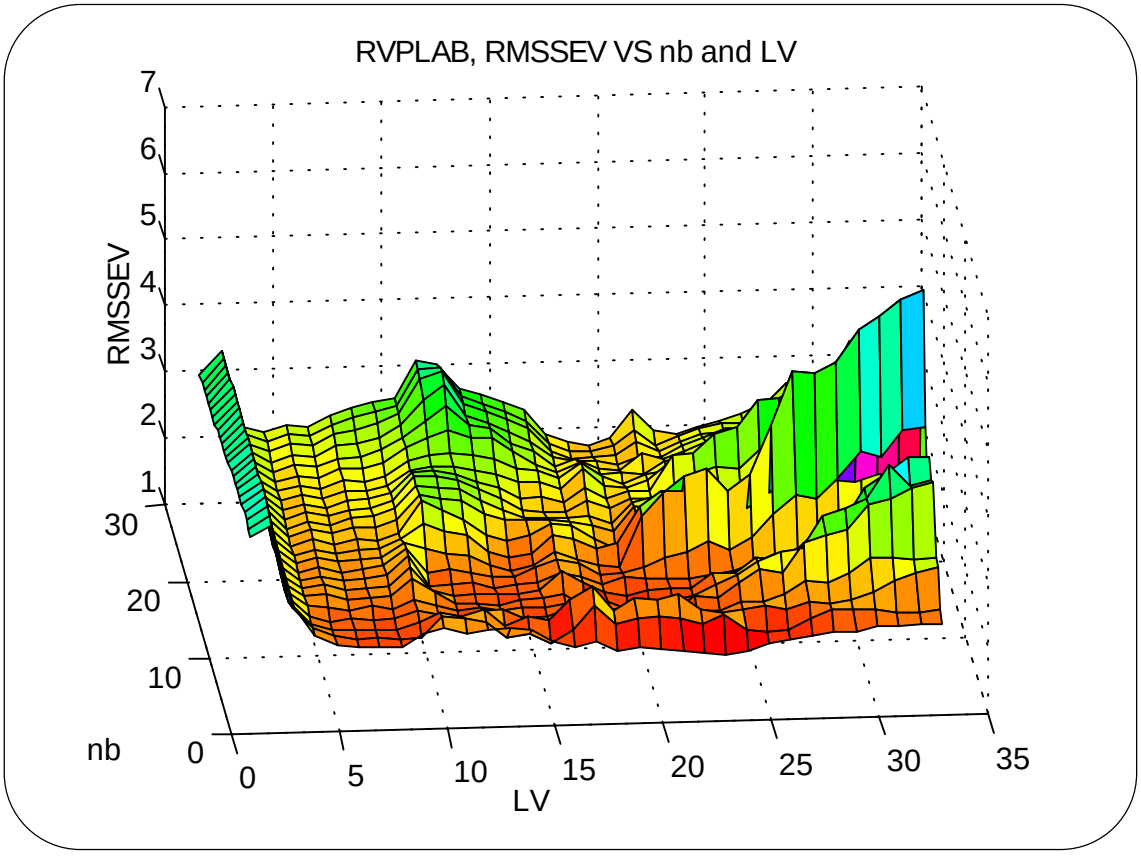

Figure 5.10 : RVP Model, RMSSEV as a function of nb and LV.

Figure 5.11 shows the RMSSEV for nb=1, which we could not see in figure 5.10. Figure 5.12 shows the RMSSEV for nb=2. As it can be seen in these two diagrams a local minimum appear at LV around 5-6 and then another minimum RMSSEV at 17 and 23 number of LV respectively.
Furthermore, it can be seen that the value of RMSSEV is around 2.2 for nb=2 and LV=6, which is actually the first local minimum. It seems that this case is more preferable rather than the case with nb=2 and LV=23 since the total number of model parameters is smaller and the difference between the two RMSSEV is not too large.

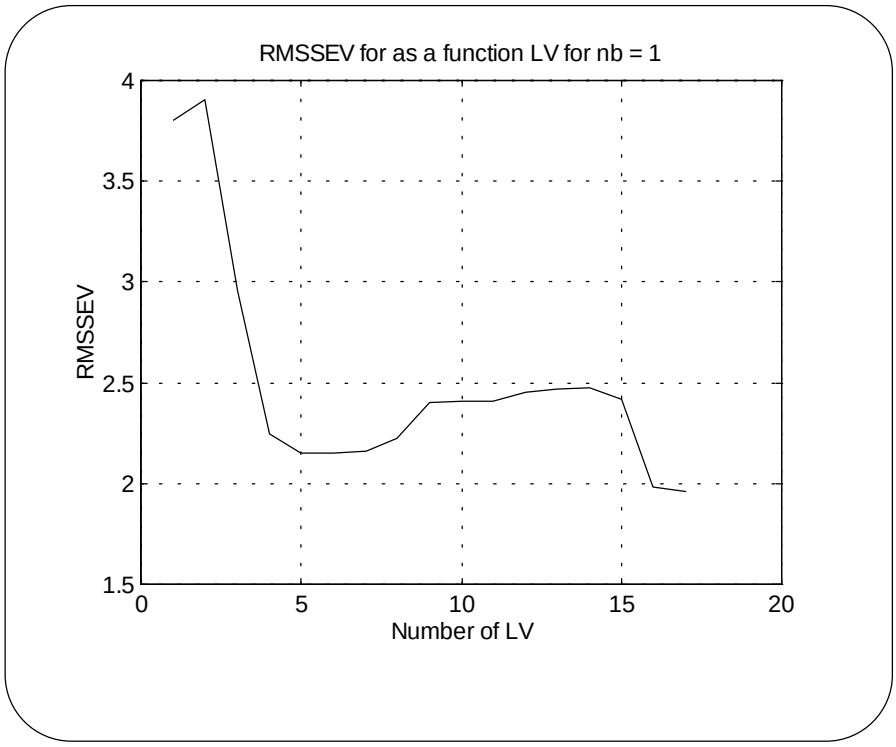

Figure 5.11 : RVP Model, RMSSEV as a function of LV for nb =1.

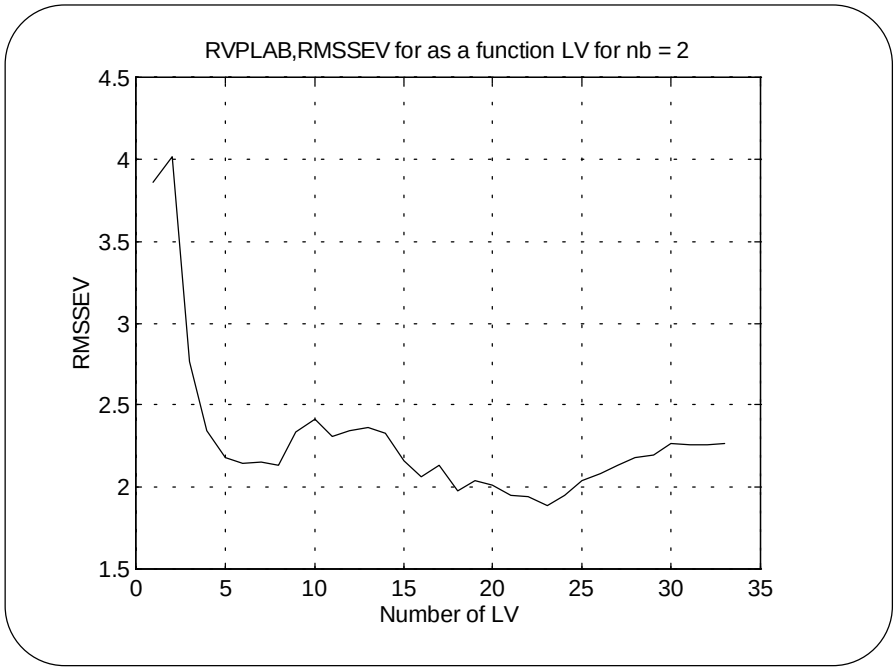

Figure 5.12 : RVP Model, RMSSEV as a function of LV for nb =2.

nb	Min. RMSSEV	X-Block	Y-Block	LV
1	1.96	99.99	43.86	17
2	1.89	99.94	47.69	23
3	2.02	97.88	48.39	15
4	2.16	83.57	44.00	7
5	2.13	93.60	48.41	11
6	2.13	98.55	55.05	21
7	2.03	97.56	55.84	19
8	2.13	97.25	57.05	19
9	2.18	94.13	53.46	13
10	2.22	94.45	55.08	14
11	2.18	94.23	55.91	14
12	2.27	93.98	57.24	14
13	2.26	96.35	63.40	19
14	2.41	96.08	66.05	19
15	2.41	95.91	66.13	19
16	2.46	94.91	64.64	17
17	2.49	62.18	44.55	4
18	2.51	61.92	44.52	4
19	2.54	95.16	70.38	20
20	2.54	95.21	72.66	21
21	2.51	94.44	72.03	20
22	2.47	94.44	71.92	20
23	2.46	93.52	69.27	18
24	2.44	93.78	70.68	19
25	2.43	93.54	70.92	19

Table 5.4 : RVP model, Minimum RMSSEV for different nb, and LV.

As mentioned earlier, another local minimum appear to be around nb=7. Figure 5.13 shows the RMSSEV for nb= 7. A comparison of between this diagram and figure 5.12 shows that the obtained RMSSEV in the case of nb=2 and LV=6 is still preferable, since both the value of RMSSEV and the model order is smaller in the latter case.

Selecting the best nb and LV is of course based on performance of the obtained model in validation. The important issue is to capture the maximum effect of input variables in prediction of output, and obtain a model which has an acceptable general characteristic. As mentioned previously, it is important to selcect one of the best models with fewer model parameters among a set of model candidate .

In all phases of model development procedure from determination of optimal delay parameters to calibration of the model along with determining optimum number of LV and nb, validation is an essential part of the development work.

In the following the result in calibration and validation of the selected model with nb=2 and LV = 6 will be presented.

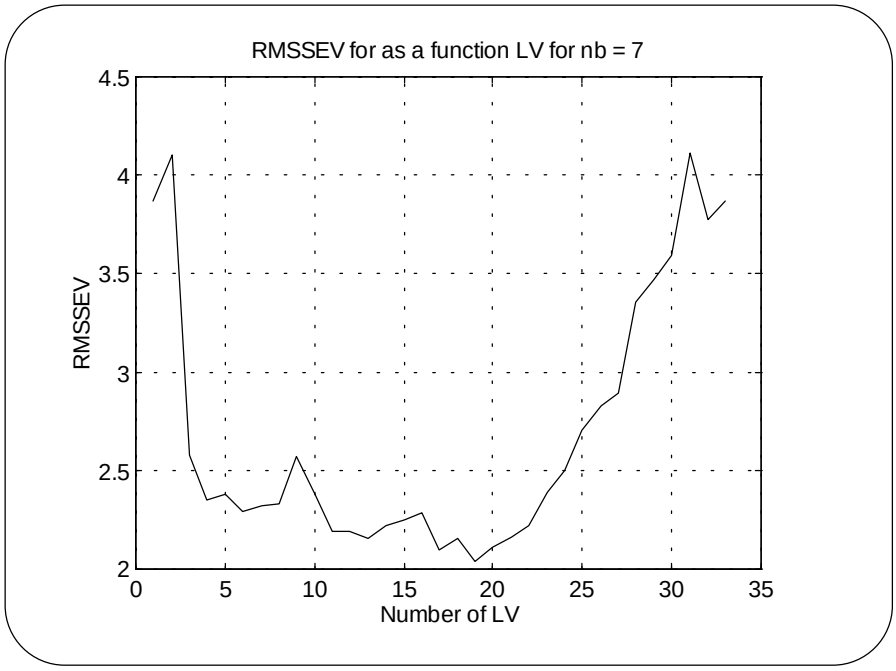

Figure 5.13 : RVP Model, RMSSEV as a function of LV for nb =7.

The order of the ARX model is thus as follows: na = 1, and nb = 2.
The following delay parameters has been found for each input variables, using a series of recursive simulation:

K = [2 2 3 3 3 2 1 4 2 2 1 4 17 17 5 10]

It is interesting to see the progress of PLS regression. Table 5.5 shows the percent variance captured by PLS model. As we can see the captured variance in X-block, i.e. inputs, and Y-block, i.e. output, are respectively 77.34% and 43.63%.

Percent Variance Captured by PLS Model				
	X-Block		Y-Block	
LV #	This LV	Total	This LV	Total
1	26.02	26.02	24.58	24.58
2	27.98	54.01	6.09	30.67
3	4.34	58.34	9.00	39.67
4	4.65	62.99	2.17	41.84
5	8.35	71.34	0.97	42.81
6	6.00	77.34	0.82	43.63

Table 5.5 : RVP model, percent variance captured by PLS model.

The first thing we are interested in to examine is the level of prediction error in both calibration and validation. Figure 5.14 shows the prediction error for validation. Notice that the error here is the difference between model output and RVP measured in the laboratory, and the error value is not calculated based on autoscaled data, but it has the real unit, i.e. kPa. If we could obtain a perfect model, then we would expect that the error signal would have approximately the same characteristic as white noise. Thus, a histogram plot of the prediction error will give an impression of how close the error signal is to a normal distributed zero mean noise. The histogram plot of error signal in validation is shown in figure 5.15.

Examining the same plots in calibration, would give os an impression of how well the calibration is performed. If the prediction error in calibration is very small and much closer to zero than in the validation, it could be a sign of an overfitted model or perhaps the validation and calibration data are different and possibly from two different regions of operation. Figure 5.16 and 5.17 show the respective plots of prediction error and histogram of error in calibration.

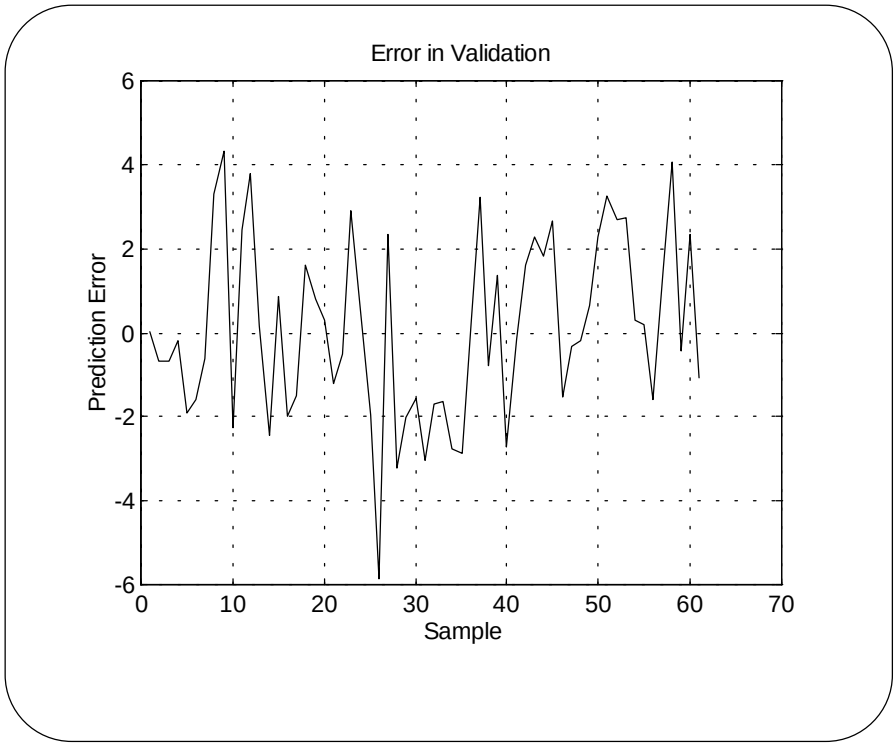

Figure 5.14 : RVP Model, prediction error in validation.

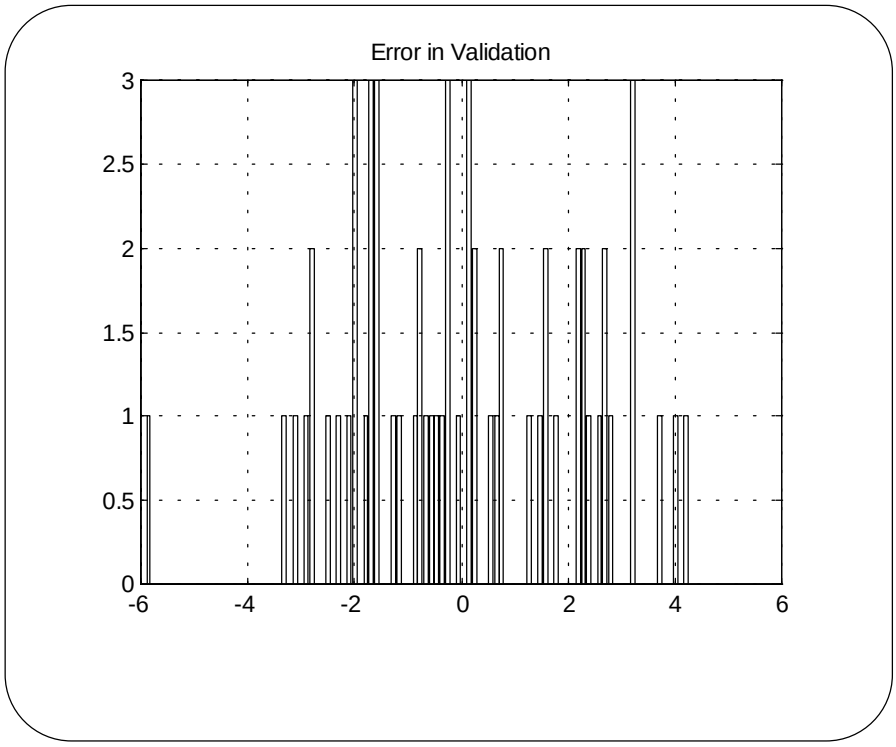

Figure 5.15 : RVP model, histogram plot of prediction error in validation.

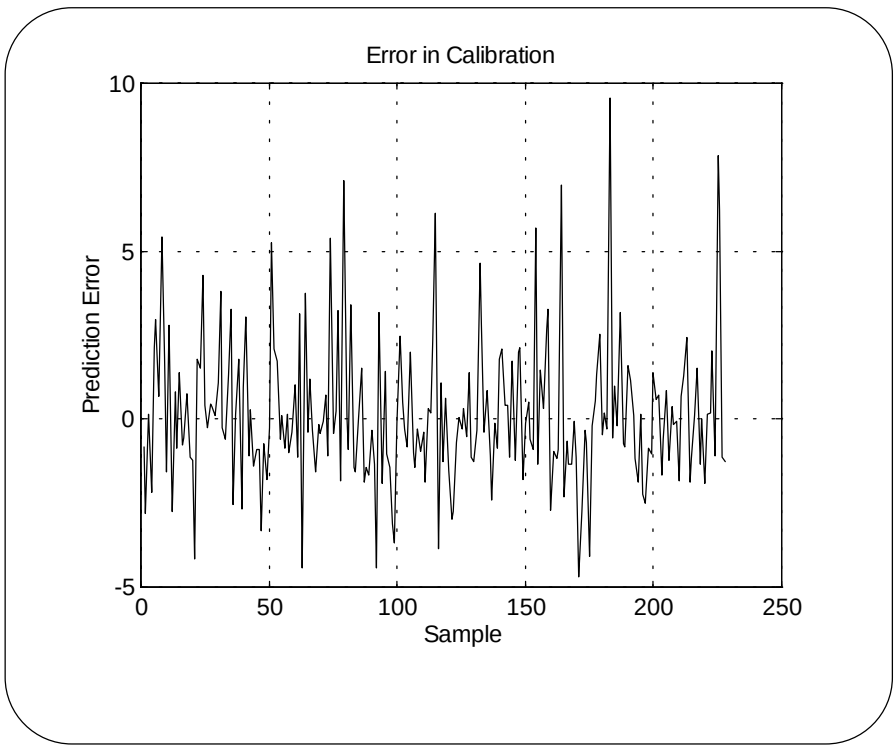

Figure 5.16 : RVP model, prediction error in calibration.

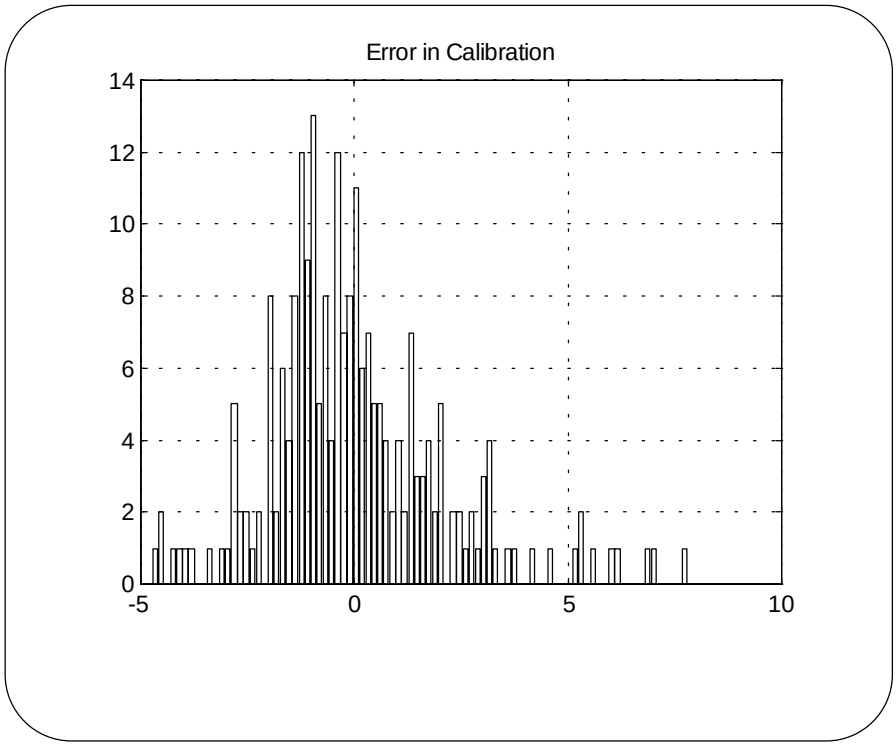

Figure 5.17 : RVP model, histogram plot of prediction error in calibration.

Table 5.6 shows RMSSE in both calibration and validation along with the average-model and zero-model in RVP model. The development of zero-model and average-model are discussed in the previous section.

Validation		Calibration	
RMSSEV	2.14	RMSSEC	2.19
RMSEAVGV	3.28	RMSEAVGC	2.91
RMSEZROV	4.10	RMSEZROC	3.05

Table 5.6 : RVP model, RMSSE, average-model, and zero-model in validation and calibration.

It can be seen that the RMSSEV is less than average- and zero-model, indicating that the model has captured the essential variation both in input and output.

The open-loop simulation in both calibration and validation for RVP model are shown in figure 5.18 and 5.19 respectively.
In open-loop simulation predicted output at time t is used instead of measured output in order to predict the output value at time t+1. Open-loop simulation will show the predictability of the model during a period of operation without having the actual output measurement.

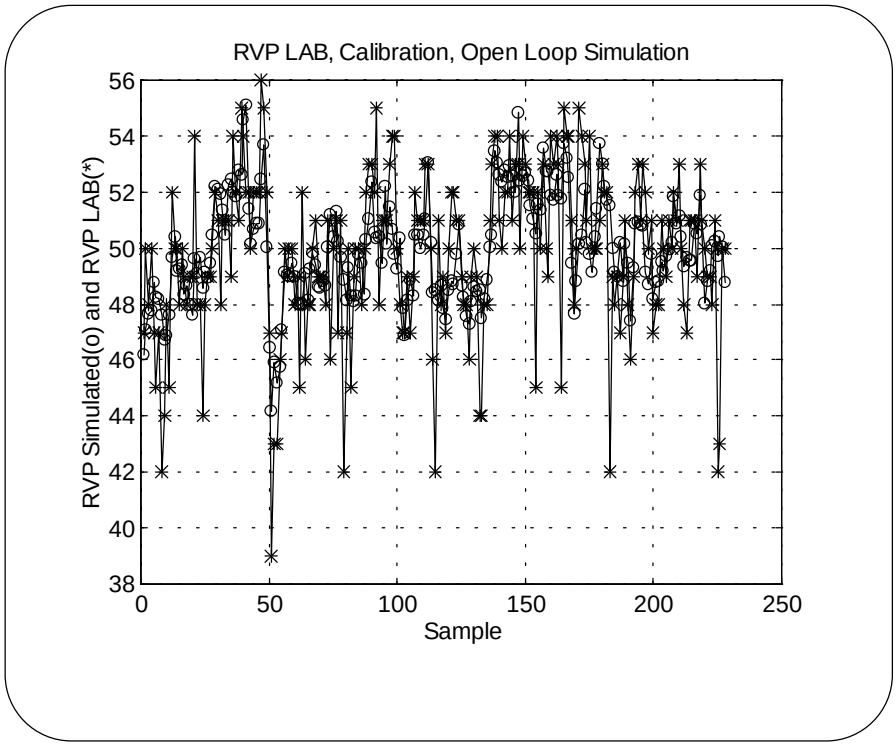

Figure 5.18 : RVP model, open loop simulation in calibration

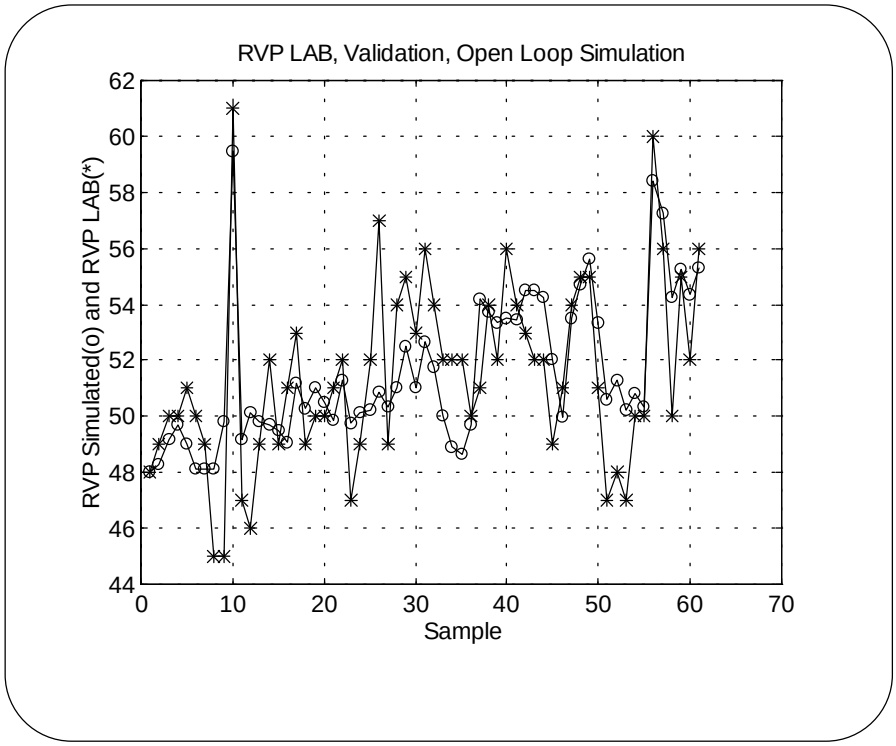

Figure 5.19 : RVP model, open loop simulation in validation

As we can see the model has captured the essential variation and follow the variation of RVP op and down, indicating a good performance.

In the following the results the model simulation is presented when the measured output is used in the model for prediction of the next output. This model simulation is performed using both calibration and validation data set. Simulation of the model applying the same data set which is used for calibration seems superfluous. However, it will give an impression of how well the calibration is performed. The expectation is that the model is capable to reproduce the calibration satisfactory.

Figure 5.20 shows the result of prediction of RVP in calibration in which the output measurement is plotted versus model predicted RVP. It can be seen how well prediction follows the measured output. Notice that RVPLAB is RVP measured at laboratory, and RVPMODEL is the predicted output.
This result demonstrates that the model has captured the essential variation but there are some points that the model has difficulty to fit, mostly low RVP measurements.

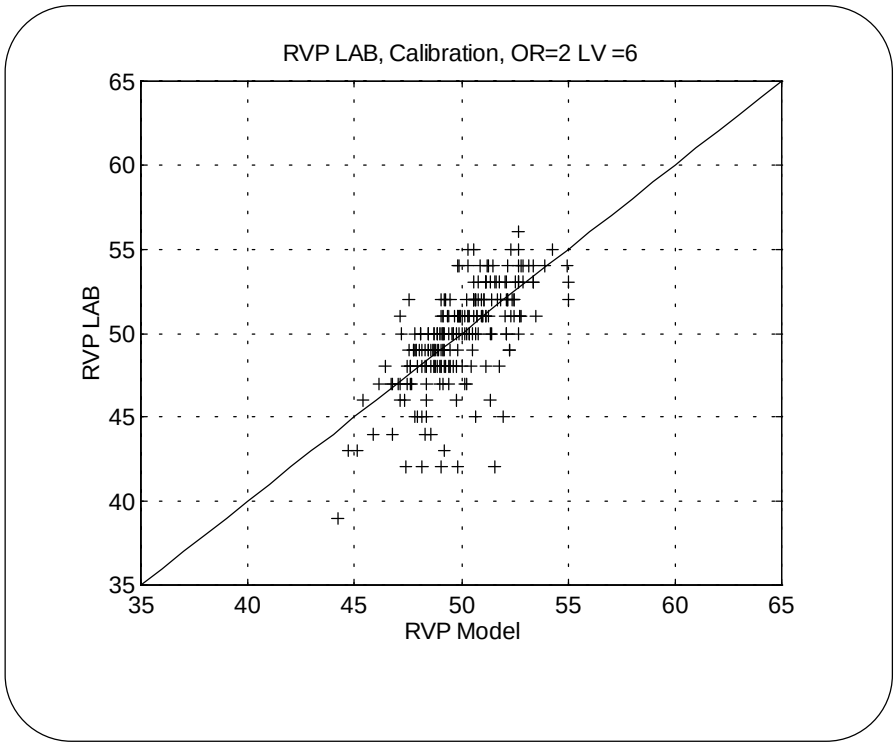

Figure 5.20 : RVP model predicted versus RVP measured in calibration

As we can see in figure 5.20, most of the points are lying around the diagonal line indicating that there is virtually no bias in the model except for some points that are slightly away from the diagonal line.

Figure 5.21 and presents the same simulation using validation data set. It shows RVP model predicted versus RVP measured in validation. It can be seen clearly that the predicted values follow the variation of the measured outputs and demonstrate a good predictability characteristic.

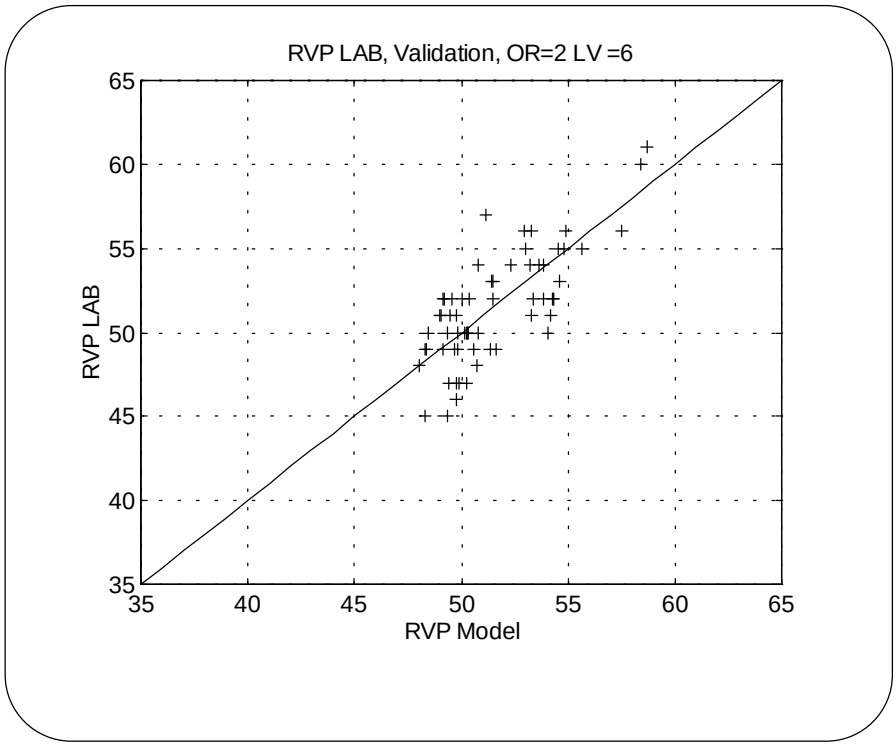

Figure 5.21 : RVP predicted versus RVP measured in validation

As it is discussed in chapter 3, in linear PLS it is assumed that the scores in the output Y-block is a linear function of scores in input X-block, as it is expressed in equation 5.15.

$$u = b\,t + h \qquad (5.15)$$

b is called the inner relationship, or internal regression coefficients.
A plot of score u versus score t can be useful in order to visualize and examine the functionality of u=f(t).
For the RVP model this plot can be seen in figure 5.22 for the first u vs. the first t. As it can be seen from figure 5.22, there is no obvious nonlinear relationship between t and u, and thus this justify the use of linear PLS.
In fact, nonlinear PLS has been investigate by the author. A number of simulations has been carried out using both neural network and different degree of nonlinear polynomials. The results show no significant improvement by using nonlinear PLS in the RVP model.

The last step in assessment of model validation is to examine the model parameters and evaluate the sign and quantity of parameters in order to interpret the physical sense of the parameters.

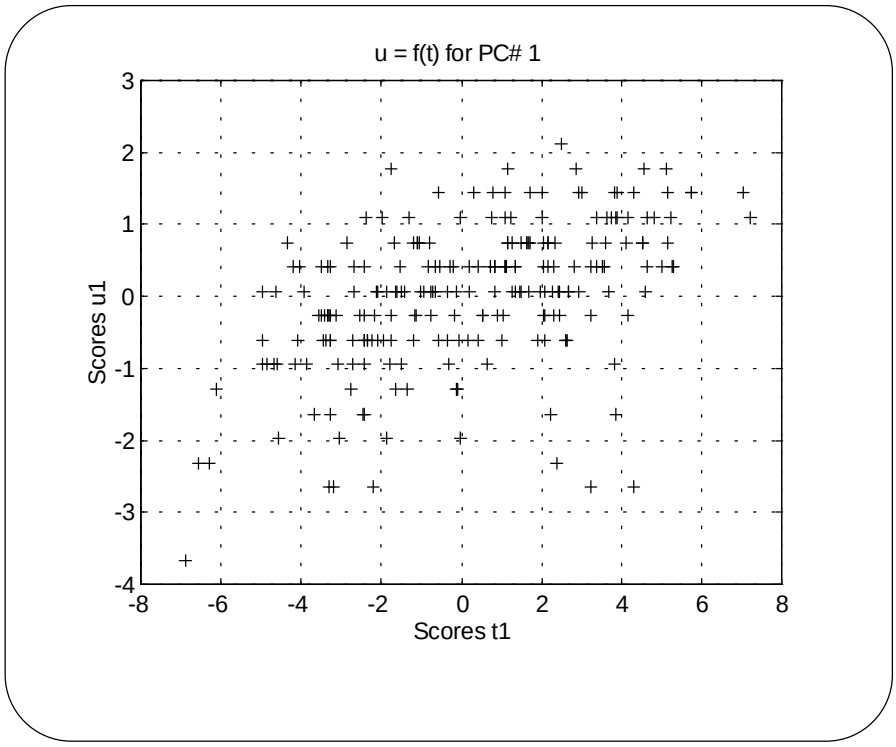

Figure 5.22 : RVP model, scores u1 as a function of scores t1.

As it is mentioned in the beginning of this section, the objective is to give an example of the general procedure in model development, in which the performance of the developed model is evaluated and the model is accepted if the results indicate satisfactory prediction ability.
Examining the model parameters at this point has shown that the developed model is not sensitive to some variables in which the respective parameter values are small.
These variables are excluded from the inputs, and hence the whole procedure is repeated. It has been found that the following input variables have the largest effect on RVP output.

1 Reformer Feed flow rate
2 C-401 Reformate flow rate
3 C-401 Liquid gas flow rate
4 C-401 Reflux flow rate
5 C-401 Feed temperature
6 C-401 Reboiler temperature
7 C-203 Reflux flow rate

The obtained B-parameters are shown in table 5.7. There is only one A-parameter which is:
a = 0.160.

The largest effect stem from variables number 2, 3, and 6, which are the variables chosen from the stabilizer column in catalytic reformer I shown in figure 4.3 in chapter 4.

The negative effect of variables 3 and 6 are correct since an increment in both reboiler temperature and liquid gas flow rate, which is the top distillate flow rate, will decrease RVP as a result of removing more light hydrocarbon components from the reformate product.

Variable number 2 is the feed flow rate of reformate product itself. It has positive effect because more bottom product in the stabilizer means less top distillate, and hence more light component in the reformate.
Variable number 4 is reflux flow rate in the stabilizer column. An increment in reflux flow rate means less distillate product and more liquid down-stream at the top of the column. It has positive effect, meaning more light hydrocarbon components will be sent down through the stabilizer column, to prevent flooding at the top, which eventually end in the reformate product.

Variables	Coef.for u(t-1)	Coef. for u(t-2)	Sign	Description
1	0.101	0.103	+	Reformer Feed flow rate
2	0.234	0.230	+	C-401 Reformate flow rate
3	-0.233	-0.216	-	C-401 Liquid gas flow rate
4	0.101	0.084	+	C-401 Reflux flow rate
5	-0.055	-0.141	-	C-401 Feed temperature
6	-0.261	-0.276	-	C-401 Reboiler temp.
7	-0.030	-0.053	-	C-203 Reflux flow rate

Table 5.7 : RVP model, B- parameters in ARX model.

Variable number 5 is the temperature of the feed to the stabilizer column. This temperature represent the magnitude of the enthalpy introduced to the column, and hence has the same effect as reboiler temperature, i.e. sending more light hydrocarbon components upward and thus decreasing RVP.
Variable number 1 is feed flow rate to catalytic reformer. In the first place, it is expected that the sign of parameter relating to this variable should be negative for the reason that the higher flow rate will increase the heavy components, since the feed to the catalytic reformer is Heavy Virgin Naphtha (HVN). It is difficult to say anything in more detail about the sign of this parameter because of the complex reactions take place in the reactors. The positive sign can be just an indication of promotion of cracking reactions and formation of more light hydrocarbon components.
Variable number 7, which has the smallest effect, is the reflux flow rate in the naphtha splitter. The main objective of the splitter is to split naphtha into LVN and HVN. Increment in this reflux flow rate means more light components toward LVN and more heavy component to HVN, and thereby that a negative effect on RVP.

5.3.2.4 Discussion of Full ARX Model versus Reduced Parameters

In the previous two sections the results of two RVP models are presented. One model with nb=24 and delay parameters K=1, and another with reduced nb and optimum number of delay parameters K.
Based on the developed structure of the ARX model presented in equation 5.8, in which we take the previous existing output y, it is important that the model cover all existing dynamic variation of input which is most relevant for modeling of output. Thus, the choice of nb=24 seems to be most appropriated.
On the other hand, there will be a risk of over fitting problem by applying nb=24, since it results in a high number of model parameters. By the results presented in the previous sections, it is shown that the number of nb can be reduced with less significant loss of predictability by applying a set of optimum delay parameters K and LV. It has been shown that nb can be reduced to 2, and at the same time keep almost the same level of predictability. However, the developed model with the reduced model order is a special case of the model with nb=24 and depends much on the condition that data has been obtained from.
Dynamic variation of the input variables has a significant effect on the general predictability of the obtained model. Figure 5.23 and 5.24 show examples of variation of two important inputs used in modeling of RVP during one week of operation. Examining the variation of the inputs along the whole period of the calibration and validation shows that there are significant low frequency changes in the characteristics of the variation in the different periods, presumably based on the changes in the operation points related to the different seasons. Thus, applying nb=2 may not be optimal, and there will be risk of loosing the general predictability characteristic.
The conclusion is the model with nb=24 is the basic recommendable model, and the reduced model is a special case of the basic model, which only can predict low frequency changes in inputs.

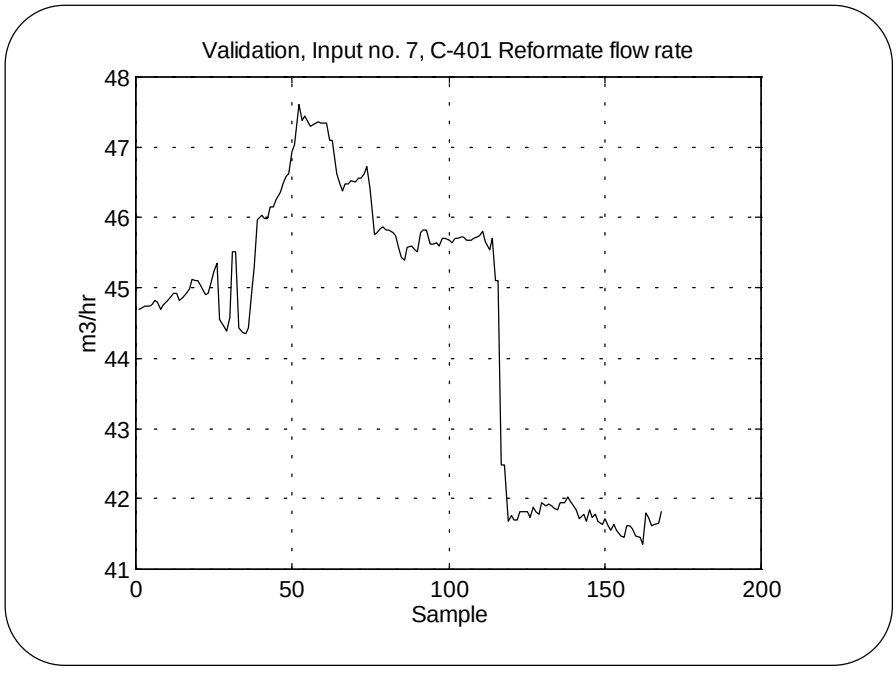

Figure 5.23 : RVP model, one week data for reformate flow rate.

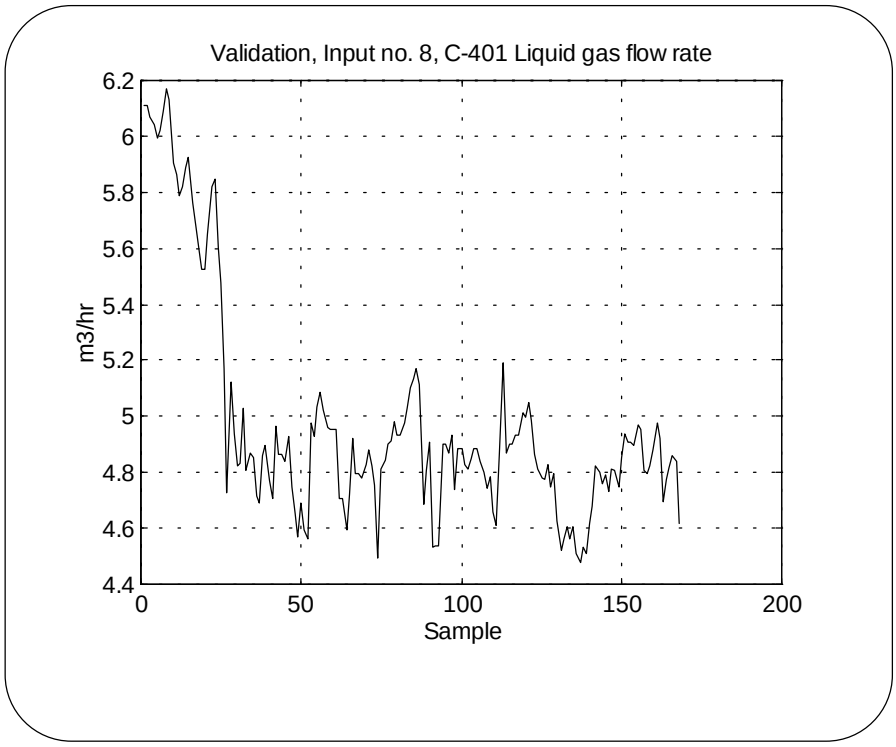

Figure 5.24 : RVP model, one week data for liquid gas flow rate.

5.3.3 RON Model

Research Octane Number RON is an important quality variable and has a tremendous profit effect on the economy of the refinery. It is quite possible to use MTBE in order to compensate for low octane quality in some gasoline products. However, this solution is expensive due to high price of the MTBE. Moreover, there is a maximum limit on the oxygenate contents in gasoline products due to environmental regulation and restrictions in different countries in Europe. These are two main reasons that make using oxygenate not a feasible solution for compensating RON quality.
The purpose of large investment on catalytic reformer and isomerization units is justified by achieving high octane number of gasoline products. By this it is possible to meet the market demands on octane quality and at the same time produce more environment friendly products.
This is the strong motivation for control of RON quality in catalytic reformer in order to meet the demands for production specifications and economy. However, the close-loop control of RON has caused a difficulty in input-output modeling and prediction of RON quality, as it is discussed in chapter 3.

5.3.3.1 Inputs and Output

The following variables are chosen as the input variables in the RON model. These are the variables expected to be most influential on the output RON.

1 Mole H2/ Mole C in recycle gas
2 % H2 purity in recycle gas
3 Reactor 1 outlet temperature
4 Reactor 2 outlet temperature
5 Reactor 3 outlet temperature
6 Reformer Feed flow rate
7 Reformate flow rate

The total amount of input data is 5600, and 4334 in the calibration and validation data sets respectively. These data sets correspond to 9 months of operation data in calibration and 6 months operation data in validation. Notice that the data corresponding to the periods of process shutdown and outliers has been omitted. Regarding the output RON, there are only 232 and 164 laboratory measurement available in respectively for calibration and validation periods.

5.3.3.2 Model Structure

The control strategy of the reformer unit is based on close-loop control of RON quality due to the great importance of octane number on production economy. The control action has effectively resulted in small variation in output RON, as it is shown in table 5.8. The average, standard deviation, maximum, and minimum values for RON and the reactors outlet temperature are shown in table 5.8. It can be seen that the variation of RON and the temperatures are small due to the small values of the standard deviations. Notice that the data in table 5.8 are calculated based on the calibration data set that covers 9 months of operation.

It has been found that the type of model structure used in the case of RVP model is not suitable and effective for prediction of RON. As it is shown in chapter 4, there is more variation in RVP than in RON.

Calibration		Reactor Outlet Temperarture		
	RON	R1	R2	R3
Average	99.97	430.92	471.10	500.33
Std. Deviation	0.37	4.21	4.62	3.54
Maximum	101.60	442.45	482.62	507.32
Minimum	98.60	421.67	460.40	489.89

Table 5.8 : Calibration data set, RON and reactor outlet temperatures

To overcome the problem of missing output value, linear interpolation is preferred between the existing RON values, which is consequently based on the assumption that the variation of RON from one day to another is small enough to permit a rough estimation of RON between two subsequent existing RON measurement by interpolation. Since interpolation is applied in order to estimate the missing RON output, an equal number of observations in both input and output data set is thus obtained.

The model structure is based on an ARX model of the general type:

$$A(q)y(t) = B(q)u(t) + e(t)$$

in which the parameters are estimated by a PLS regression. The number of A-parameter, and B-parameters, are determined by model order, and we will have the same number of na, and nb.
It is very important to emphasize that the interpolation is performed only in calibration data set. In the validation, we let the model apply its own predicted output in order to predict the next output.

5.3.3.3 Calibration

As it is described in the case of RVP model, a set of suitable delay parameters, optimal model order and number of LV parameter need to be determined. These parameters has been found by numerous recursive simulations.
The following delay parameters has been found for the input variables:

$$K = [16 \;\; 16 \;\; 18 \;\; 18 \;\; 18 \;\; 4 \;\; 7 \;]$$

There is no delay for output RON.

Table 5.9 shows the values of RMSSV obtained for the different ARX orders and LV, in which the model order is changed from 1 to 20 in order to search for all possible effect of variables up to t-24, i.e. the previous measured RON. The maximum number of LV is a function of model order.

ARX Order	Min. RMSSEV	X-Block	Y-Block	LV
1	0.896	12.78	83.46	1
2	0.111	97.26	99.29	4
3	0.104	97.91	99.48	4
4	0.116	99.11	99.58	4
5	0.042	99.02	99.60	14
6	0.040	99.28	99.61	14
7	0.053	97.39	99.42	13
8	0.053	98.07	99.47	16
9	0.052	97.83	99.45	16
10	0.055	97.59	99.44	16
11	0.056	95.78	99.11	12
12	0.057	98.12	99.54	21
13	0.050	97.98	99.53	21
14	0.048	97.87	99.51	21
15	0.050	97.92	99.52	22
16	0.054	97.83	99.51	22
17	0.054	97.71	99.51	22
18	0.055	97.72	99.54	23
19	0.050	97.61	99.53	23
20	0.050	97.50	99.52	23

Table 5.9 : RON model, min. RMSSEV for different value of ARX order and LV.

Validation		Calibration	
RMSSEV	0.104	RMSSEC	0.054
RMSEAVGV	0.453	RMSEAVGC	0.365
RMSEZROV	0.683	RMSEZROC	0.553

Table 5.10 : RMSSE, average-model, and zero-model in validation and calibration

Table 5.10 shows the obtained RMSSE in calibration and validation along with the values of RMSSE for the average-model and the zero-model respectively. It can be seen in table 5.9 that already by a third order ARX model, the value of RMSSEV reaches a first local minimum at is 0.104 and a comparison with the reference models in table 5.10 shows this value can be accepted.

It is important to notice again that these RMSSEV values are calculated based on that the model apply its own predicted output in order to predict the next output.

Based on the these results, it is concluded that a third order model with LV=4 can be an appropriate candidate for the accepted model considering the discussion about fewer model parameters in order to avoid overfitting.

Hence, the model structure of ARX order=3 , and LV = 4 is chosen. That means we will get 3 a-parameters, and 21 b-parameters corresponding to 7 input variables.

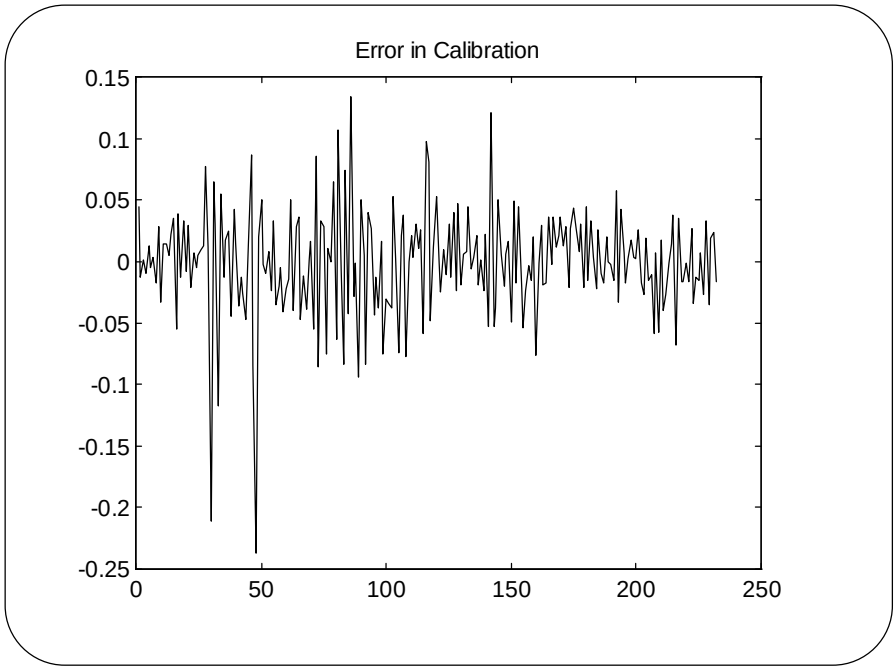

Figure 5.25: RON model, prediction error in calibration.

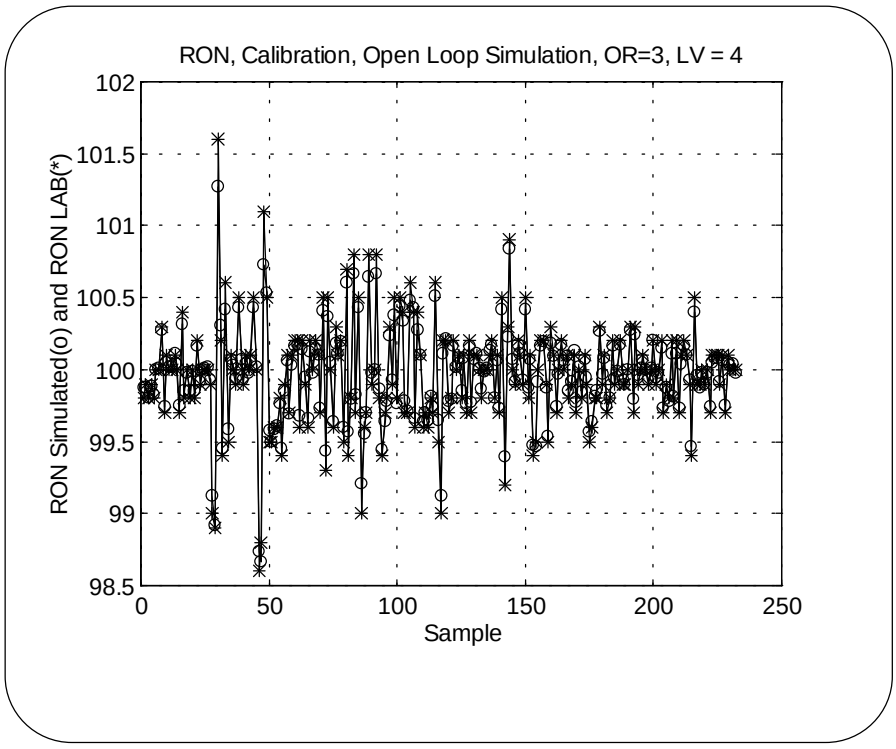

Figure 5.26: RON model, open loop simulation in calibration.

The prediction error in calibration is shown in figure 5.25. Notice that the obtained RMSSE in calibration is 0.054. The calibration data consists of 5600 input-output data, in which only 232 of them are measured RON by laboratory and the rest is estimated by linear interpolation. In

figure 5.25, only the error corresponding to existing measured RON is shown, and the error corresponding to interpolated data is omitted.
Figure 5.26 shows the so called open-loop simulation of the model in calibration, in which the new predicted value is used instead of the measurement for prediction of the next output value. The open loop simulation indicate that the calibration is satisfactory.
Figure 5.27 shows a histogram plot of error in calibration, which exhibits an approximate zero mean error.

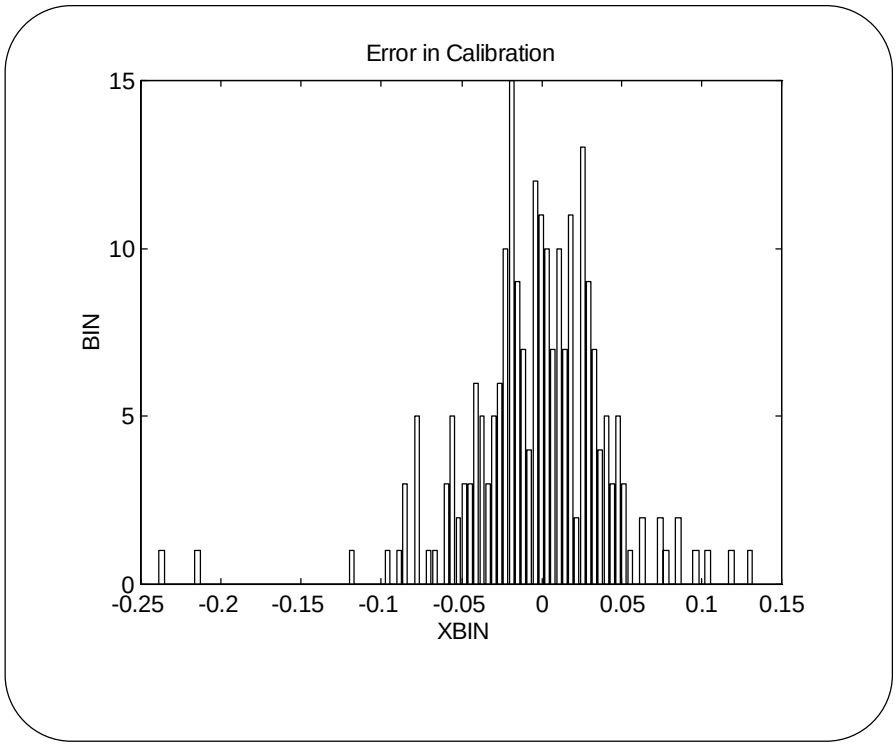

Figure 5.27: RON model, histogram plot for prediction error in calibration.

5.3.3.4 Validation

The validation is performed applying a completely distinct set of data. As mentioned before the input data in validation set consists of 4334 number of data covering 6 months of operation. In this period, after omitting the outliers, there are only 164 laboratory measurements of output RON remained. Omitting the outlier has been discussed in chapter 4 and the missing data in input variables are due to operation shutdown.
Recall the discussion in RVP model, we have defined two different reference models in order to assess the value of prediction error. These are defined as average-model and zero-model.
Table 5.10 shows the values of these two reference models along with RMSSEV in both calibration and validation. The model is thus accepted since the RMSSEV in validation is less than the two reference models.
Figure 5.28 shows the prediction error in validation. This figure, along with the corresponding histogram plot in figure 5.29 are used for the assessment of the obtained prediction error. Figure 5.29 shows a histogram plot of prediction error which exhibit a zero mean error.

It is interesting to study the model performance in validation by applying the new predicted RON value for calculation of the next output. As known, this simulation is called as open-loop

simulation, in which it shows the predictability of the model during a period of operation without having the actual output measurement. Then, we can compare the model-predicted output with the existing RON measurement. This is shown in figure 5.30. Notice that we have no interpolation in validation output. It can be seen that the model is capable of capturing the essential part of the output variation.

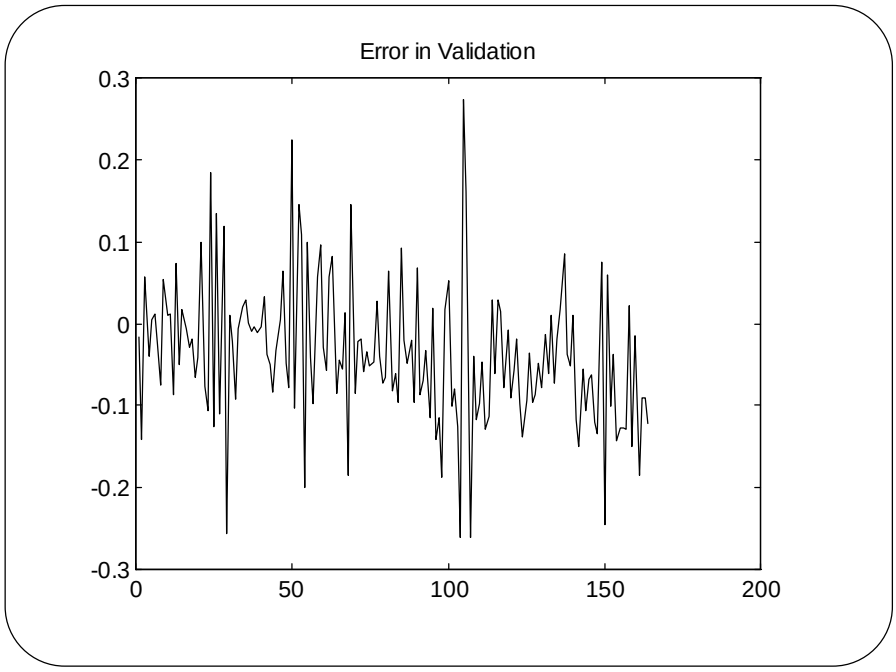

Figure 5.28: RON model, prediction error in validation.

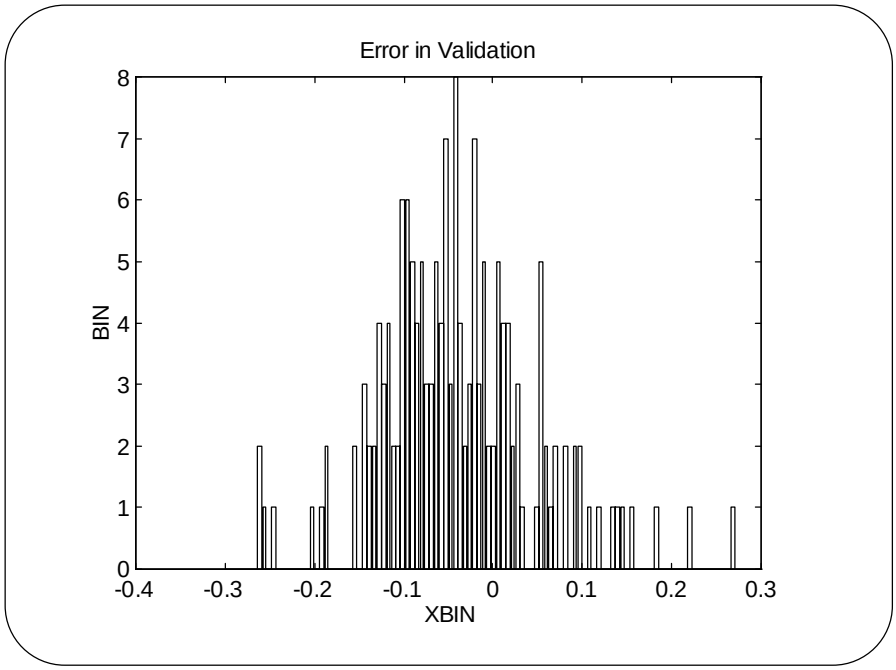

Figure 5.29: RON model, histogram plot for prediction error in validation.

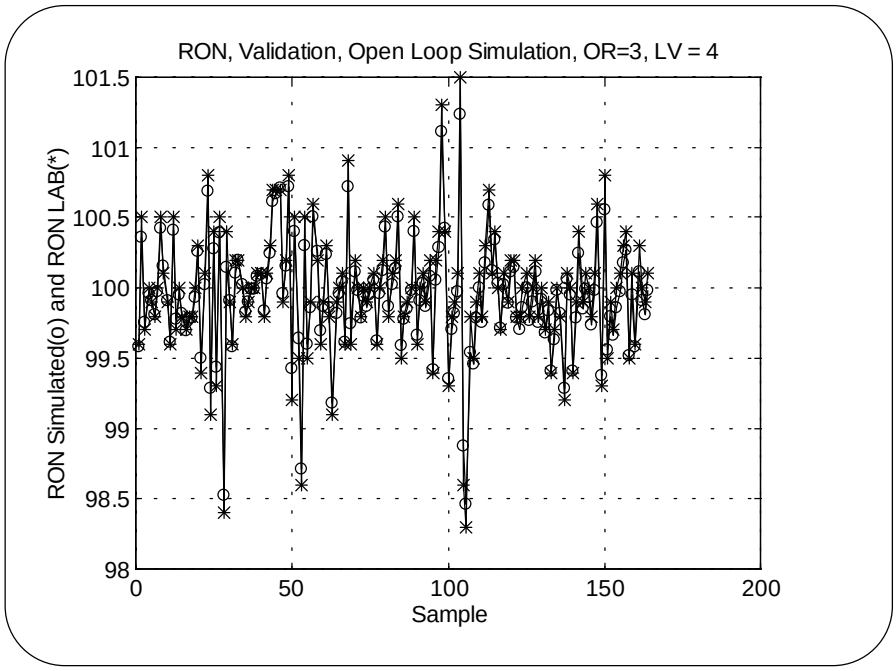

Figure 5.30: RON model, Open Loop Simulation in validation.

It is more easier to show the agreement between predicted RON by the model and the measured RON in figure 5.31.

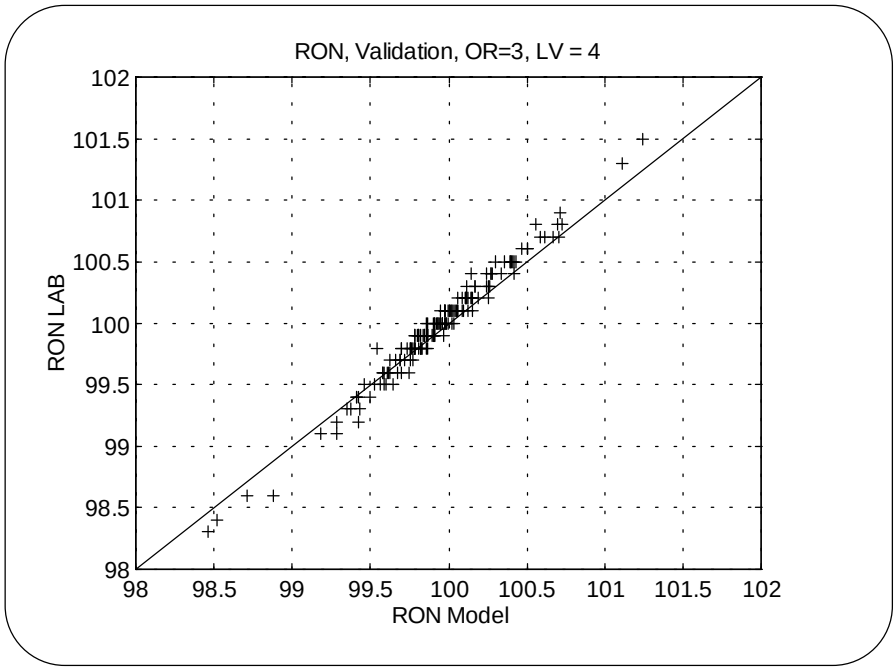

Figure 5.31: RON predicted versus RON measured in validation.

5.3.4 Benzene Model

The ARX model for prediction of aromatic (benzene) contents of the reformate product in catalytic reformer is presented in the following sections.

5.3.4.1 Inputs and Output

The following variables are chosen as the input variables in the benzene model. Selection of the variables has been discussed in chapter 4.

1 Mole H2/ Mole C in recycle gas
2 % H2 purity in recycle gas
3 Reactor 1 outlet temperature
4 Reactor 2 outlet temperature
5 Reactor 3 outlet temperature
6 Reformate Feed flow rate
7 C-401 Reformate flow rate
8 C-401 Liquid gas flow rate
9 C-401 Reflux flow rate
10 C-401 Reboiler temp.
11 C-203 Reflux flow rate
12 C-203 HVN flow rate
13 C-203 LVN flow rate
14 C-652 HVBN flow rate
15 C-201 Naphtha side stream temperature (Pressure Corrected)
16 C-601 Naphtha side stream temperature (Pressure Corrected)
17 C-203 Bottom temperature (Pressure Corrected)
18 C-652 Bottom temperature (Pressure Corrected)

The total amount of input data is 5627, and 4240 in the calibration and validation data sets respectively, which correspond to 9 months of operation data in calibration and 6 months operation data in validation. Notice that the data corresponding to the periods of process shutdown and outliers has been omitted. Regarding the output variable, there are only 144 and 33 laboratory measurement available in respectively for calibration and validation periods.

5.3.4.2 Model Structure

The model structure in this case is similar to the structure in the case of RVP model.
Based on the discussion in section 5.2 and equation 5.8, effect of y(t-24) for output is taken in the model along with the inputs. Hence, number of A-parameter na will be one and number of B-parameters nb are determined along with number of LV by a series of recursive simulation of the model. The criterion in determination of nb and LV is RMSSEV as described earlier in the case of RVP model.
This procedure is performed in the calibration of the model and described in the following sub-section.

5.3.4.3 Calibration

As it is described in the case of both RON and RVP models, a set of suitable delay parameters, optimal model order and number of LV parameter need to be determined. These parameters has been found by numerous recursive simulations.
The following delay parameters has been found for the input variables:

K = [2 2 3 3 3 2 1 4 2 1 4 7 7 7 17 17 6 10]

Number of suitable nb and LV are found by a separate series of model simulation examining nb from 1 to 25 and LV from 1 to an arbitrary value 33. The maximum limit of LV can be chosen according to the discussion in section 5.3.2.3.2 of this chapter.
The result is shown in figure 5.32 in which the obtained RMSSEV is shown as a function of nb and LV. Table 5.11 shows RMSSEV for the first 20 nb. As it can be seen, a local minimum for the RMSSEV is obtained at nb=2, and LV=10, and another local minimum exist at nb=19, and LV=12. Recall the discussion about approving a model with fewer parameters, the model with nb=2, and LV=10 would be a good candidate. Figure 5.33 shows RMSSEV as a function of LV for nb=2.

ARX Order	Min. RMSSEV	X-Block	Y-Block	LV
1	0.2289	93.19	92.32	11
2	0.2287	87.60	92.27	10
3	0.2320	87.34	92.24	10
4	0.2337	88.19	92.25	10
5	0.2341	88.52	92.32	10
6	0.2336	87.36	92.46	10
7	0.2289	86.12	92.58	10
8	0.2286	88.11	92.49	10
9	0.2298	87.99	92.45	10
10	0.2320	90.12	92.76	11
11	0.2308	93.69	93.67	14
12	0.2304	93.42	93.75	14
13	0.2328	93.10	93.85	14
14	0.2320	89.08	92.80	11
15	0.2362	91.16	93.60	13
16	0.2311	89.95	93.06	12
17	0.2244	89.58	93.07	12
18	0.2188	89.21	93.06	12
19	0.2171	88.82	93.16	12
20	0.2219	88.41	93.34	12

Table 5.11 : Minimum RMSSEV for different value of ARX order and LV.

Validation		Calibration	
RMSSEV	0.229	RMSSEC	0.091
RMSEAVGV	0.358	RMSEAVGC	0.327
RMSEZROV	0.288	RMSEZROC	0.137

Table 5.12 : RMSSE, average-model, and zero-model in validation and calibration

Table 5.12 shows the values of the RMSSE obtained for the average and zero models. The obtained RMSSEV for nb=2 and LV=10 is compared with the reference models and it can be seen that the obtained model can be acceptable.

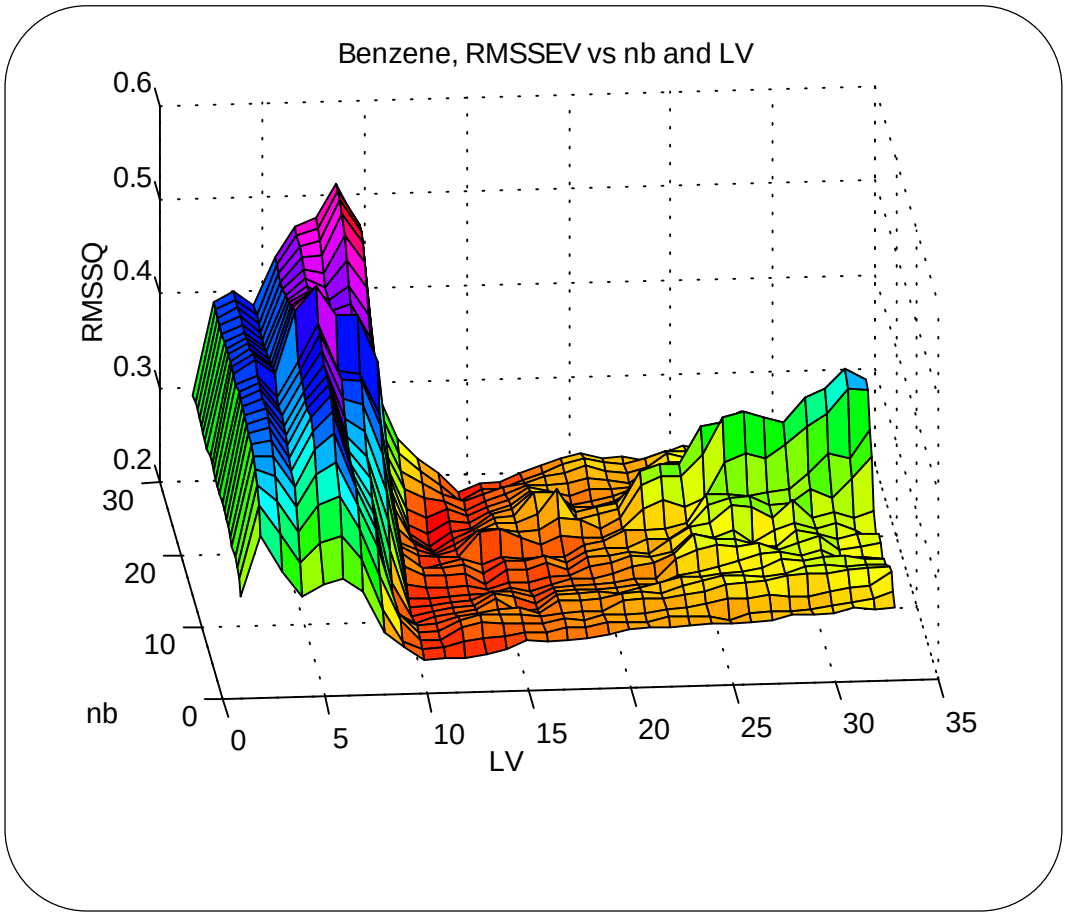

Figure 5.32: RMSSEV as a function of nb and LV.

Hence, we choose LV=10 and nb=2, which means that the number of B-parameters would be 36 in this model. Number of A-parameters would be one according the structure of the model described previously.
In order to assess how well the calibration is performed, we begin with examining the plot of the prediction error in calibration shown in figure 5.34, which indicate small error.

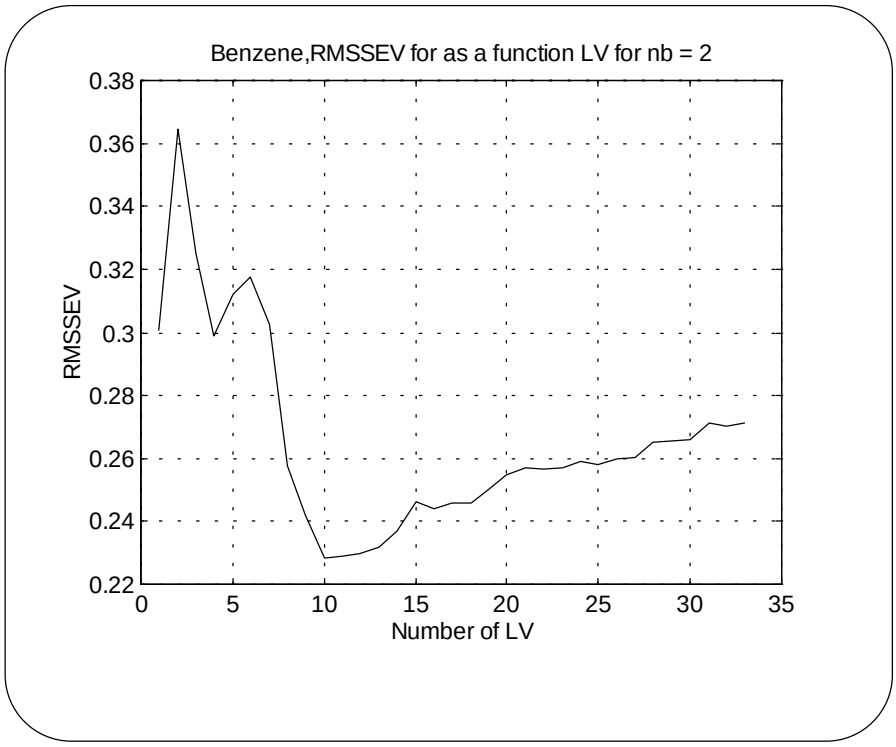

Figure 5.33: RMSSEV as a function of LV for nb=2.

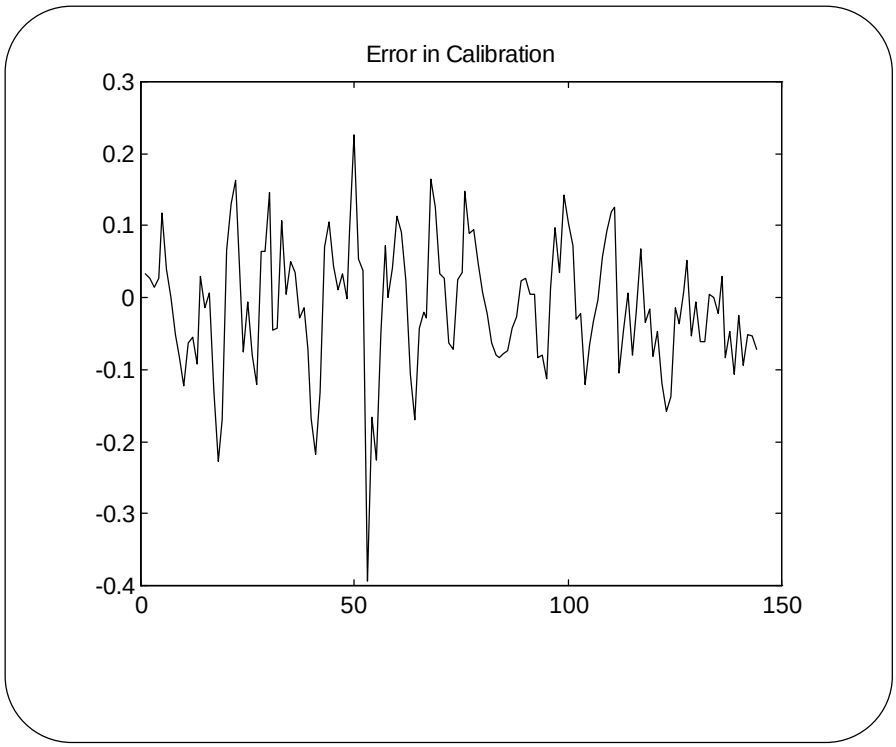

Figure 5.34: Prediction error in calibration.

Figure 5.35 shows the histogram plot of prediction error in calibration indicating that the residuals can be considered approximately normal distributed with an approximate zero mean error.

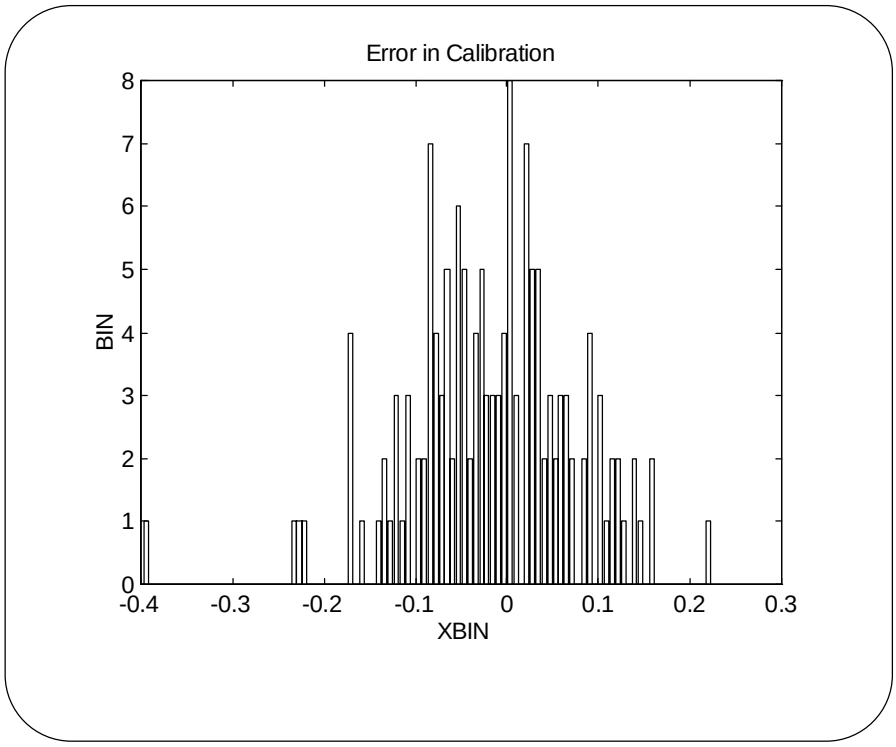

Figure 5.35: Histogram plot of prediction error in calibration.

The obtain model is simulated using the calibration data set in which the model predicted output is used in the ARX model, so called open-loop simulation, which is shown in figure 5.36. It can be seen that the open loop simulation is satisfactory.

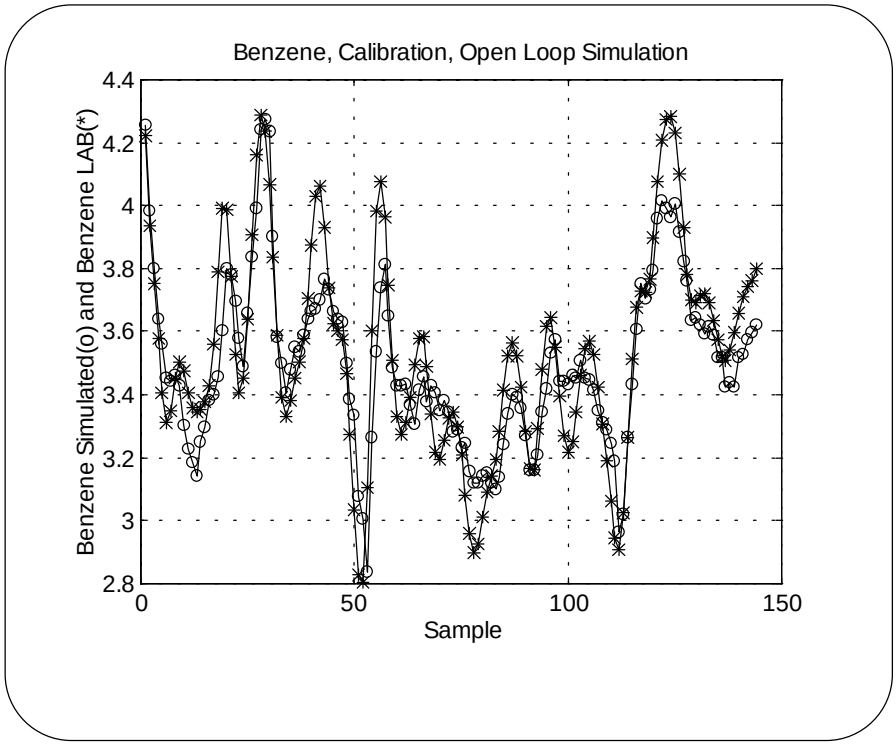

Figure 5.36: Open- loop model simulation in calibration.

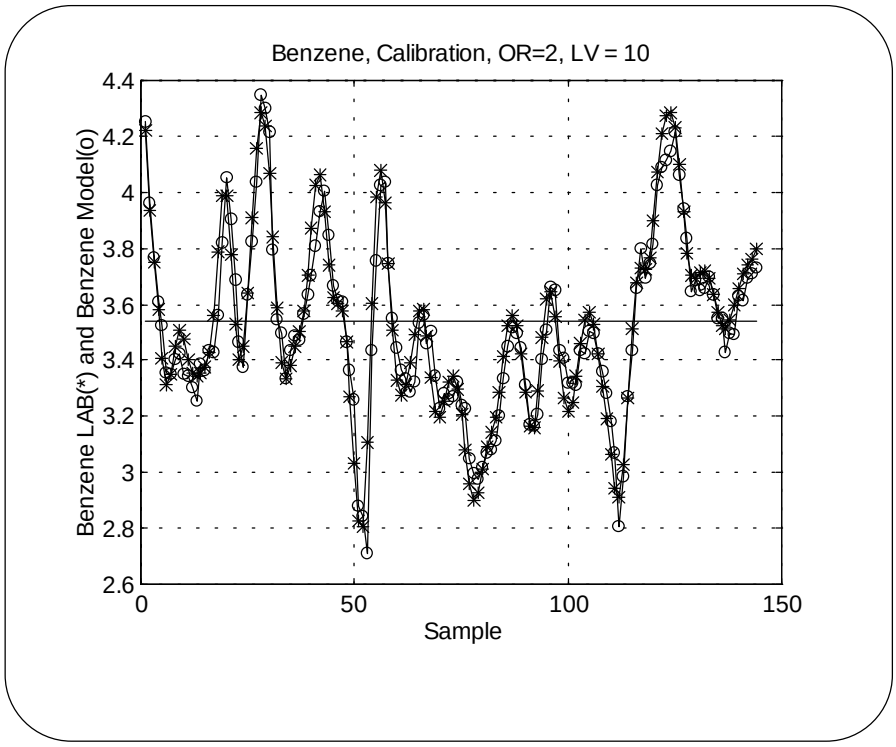

Figure 5.37: Prediction ability in calibration.

Figure 5.37 shows the result for simulation of the model in which the actual output measurement is used for prediction of the next output. It is expected that the developed model is capable to reproduce the calibration satisfactory. As it can be seen from figure 5.36 and 5.37, the model has captured the essential variation in the calibration.

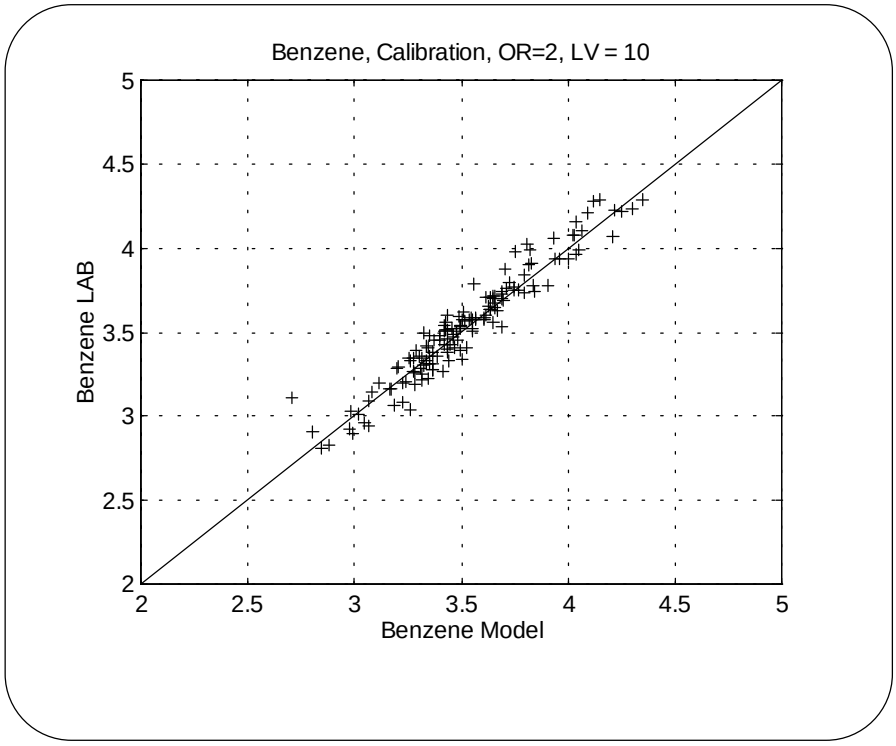

Figure 5.38: Benzene contents predicted versus measured in calibration.

The result in figure 5.37. can be better seen in figure 5.38, which shows a plot of measured output versus model predicted output.

5.3.4.4 Validation

The validation is performed applying a completely distinct set of data. The input data in validation set consists of 4240 input-output covering 6 months operation. In this case, after omitting the outliers and missing data, only 33 laboratory measurements of output remaines.
The obtained RMSSEV is 0.229 which shown in table 5.12 along with the values of the two reference models in both calibration and validation. The model is thus accepted since the RMSSEV is less than the two reference models.

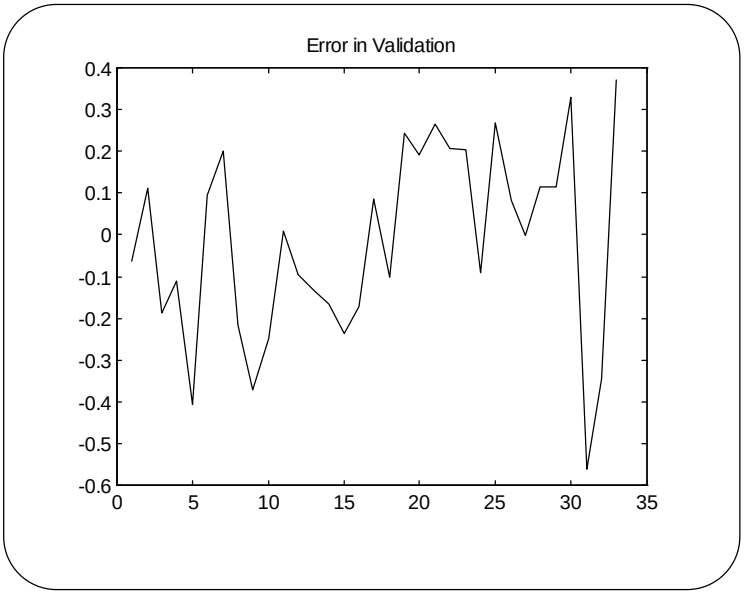

Figure 5.39: Prediction error in validation.

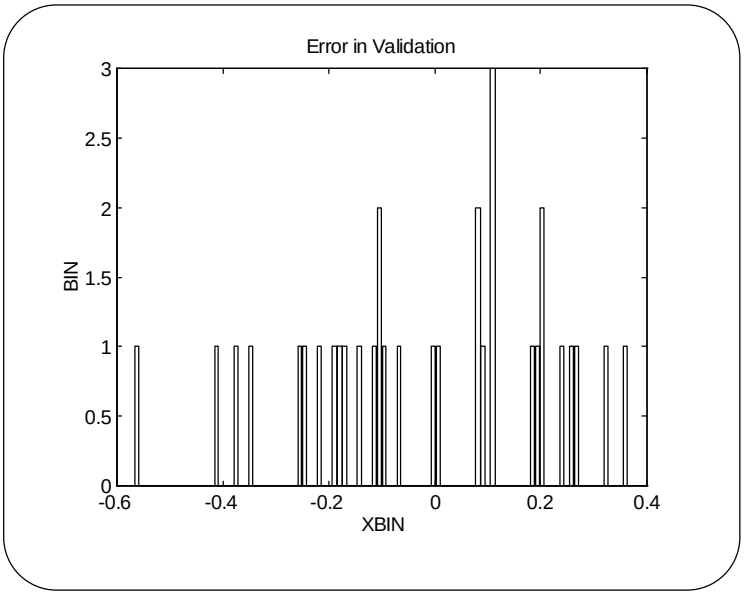

Figure 5.40: Histogram plot of prediction error in validation

Figure 5.39 shows the prediction error in validation. It can be seen that there is one point that produce a large error. The histogram plot of prediction error is shown in figure 5.40. Notice that there are extremely few output measurements available in this case.

The plot for the open-loop simulation of the model is shown in figure 5.41. As known, the open-loop simulation shows the predictability of the model during a period of operation without having the actual output measurement.
As it can be seen, the prediction by open-loop simulation produce a bias in the middle range of validation set. However, examining the simulation of the model, when the actual measurements are used, which can be seen in figure 5.42 and 5.43, indicate that model can be acceptable.

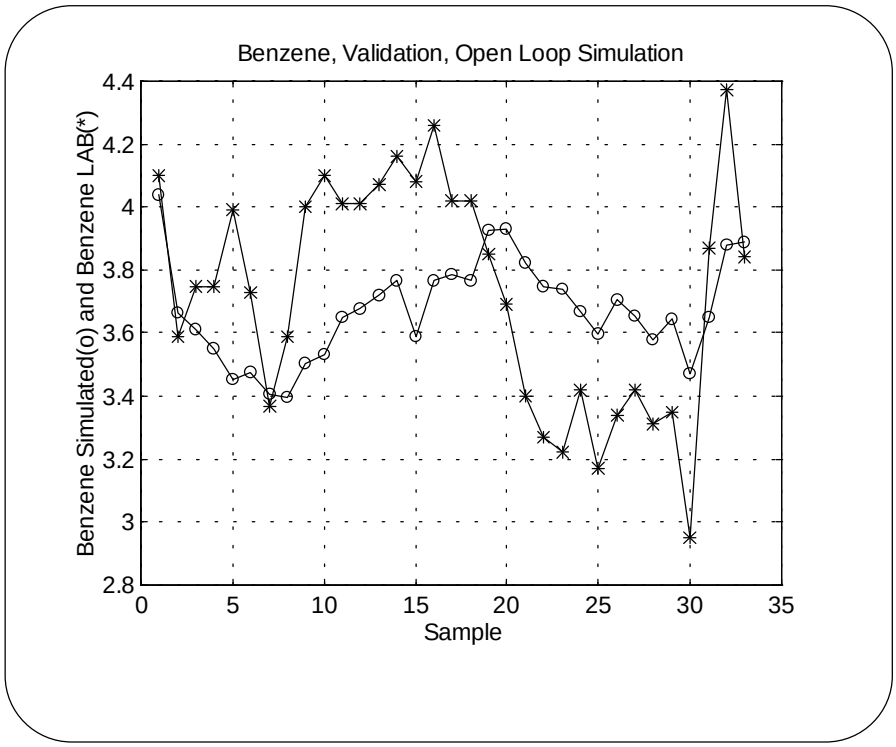

Figure 5.41: Open- loop model simulation in validation.

One possible explanation for the poor performance of the model in open-loop simulation is that there are only 144 output measurements available for calibration of the model for a period of 9 months. Number of existing measurement for the period of nine months is expected to be around 270 if the output is measured only once a day.
The other explanation could be that the choice of nb and LV is not perfectly suitable. Figure 5.32 and table 5.11 show another local minimum for RMSSEV at nb=19 and LV=12. This choice would not be appropriate since the total model parameters would become 240 which is more than the total input-output of 114. Hence, the option of nb=19 and LV=12 is rejected due to the risk for overfitting. Better result in for this modeling can be investigated only when more data is available.
Figure 5.42 shows the result of the simulation when the actual measurements are used in prediction of the next output. Figure 5.43 shows the predicted versus the measured benzene contents in the validation.

From figure 5.42 and 5.43 can be seen that the developed model has captured the essential variation in the data. However, more data is needed in order to improve the predictability of the model.

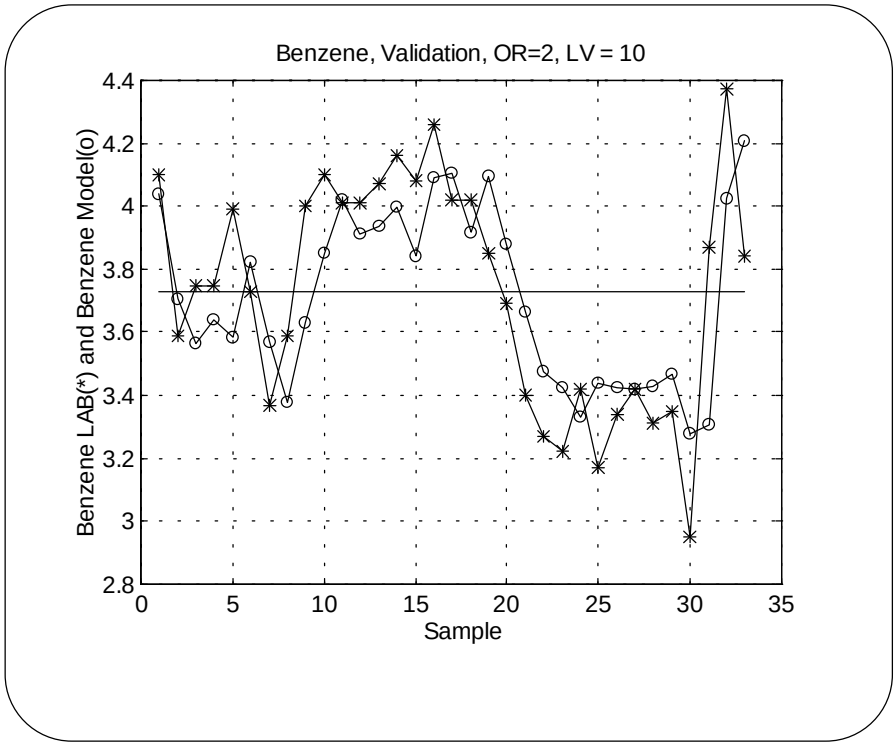

Figure 5.42: Prediction ability in validation.

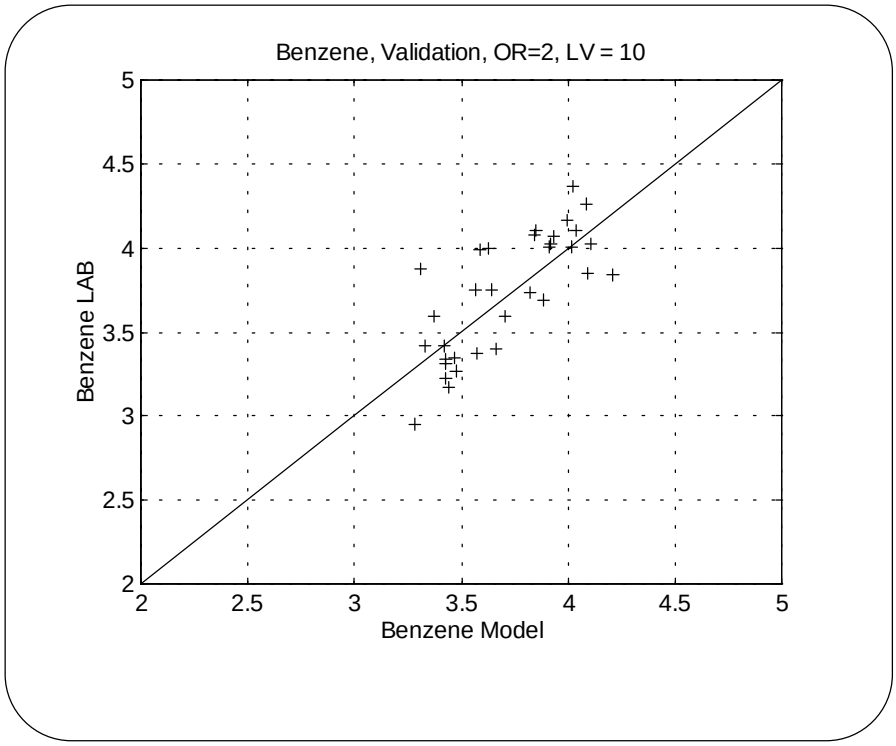

Figure 5.43: Benzene contents predicted versus measured in validation.

5.4 Conclusion

In this chapter the structure, calibration, and validation of the multivariate predictive models developed for quality prediction of reformate product from catalytic reformer I are presented. The corresponding models for prediction of the qualities of reformate and isomerate products of catalytic reformer II and isomerization unit can be found in appendix A and B respectively. The multivariate models are developed for prediction of RON, RVP, and benzene contents of the products.

It is observed that the quality variables are dependent on the earlier values of them selves and the inputs. This means that an input-output type dynamic modeling approach is a suitable choice. ARX model is chosen as the model type used in model calibration, in which the parameters are estimated by a PLS model. Applying PLS approach in parameter estimation of ARX model has been useful in which the ability of prediction has increased.

A solution to the problem of low sampling frequency for model output is proposed as follows.

In the case of RVP, and benzene models, a suitable structure of the ARX model is developed in which the information of the pervious available outputs is imposed in the regression vector of the ARX model. In RVP and benzene models, the data set are informative enough for prediction of these qualities since they are not the target of feed-back control.

The case of applying a full model, i.e. nb=24, for the ARX model has been investigated in RVP modeling, since it is expected that the nb=24 will cover all the variation of the input, and consequently will result in improving model performance. It is shown that the number of nb can be reduced without significant loss of predictability by applying a set of optimal delay parameters K and latent variables LV. However, the model with nb=24 is the basic model, and the reduced model is a special case of the basic model, which only enable modeling of low frequency variation in inputs. This conclusion will be thus valid also for other models developed later in this work.

In the case of RON, a linear interpolation in output is performed to recover the output in calibration data set, while output interpolation is avoided in validation. It has been found that the input-output data set is little informative with respect to the prediction of the output RON due to the effect of closed-loop control and the effect of little variability of the RON set point.

Consequently, better results are obtained for prediction of RVP and benzene contents of the products.

Validation of the benzene model for catalytic reformer I show a poor performance in the simulation in which the predicted output is used in the model for calculation of the next output (open-loop simulation) while the normal simulation using measured output indicate satisfactory performance. One explanation for the poor performance of the model in open-loop simulation is that there are few output measurements available. There are only 144 output measurements available during a period of 9 months for model calibration, while number of existing measurements is expected to be around 270 if the output is measured only once a day. Better results in prediction modeling can be investigated only when more quality measurements are available.

Chapter 6

Optimization

6.1 Introduction

In this chapter a multi-period optimization model for optimization of gasoline blending is presented. The optimization model assumes that the prediction models for the streams sent to gasoline blending are available. These models are discussed in chapter 5. The objective is to minimize the cost of operation for gasoline production such that the quality and quantity demands are satisfied. The objective function is a cost function which represent the cost of operation for production of blending components plus the inventory cost. This objective function is minimized subject to a set of constraints which represent the demands for quality and quantity of final gasoline products. The optimum solution will yield in quality and quantity needed for blend components and with that the optimum value for decision variables. These are also called *Targets*, which will be sent to the advanced control level for implementation.
The optimization model assumes that the qualities of final gasoline product is a linear function of the qualities of the streams sent to the blending unit.
A case study is considered as a scenario in production scheduling and the optimum solution for this case is discussed.

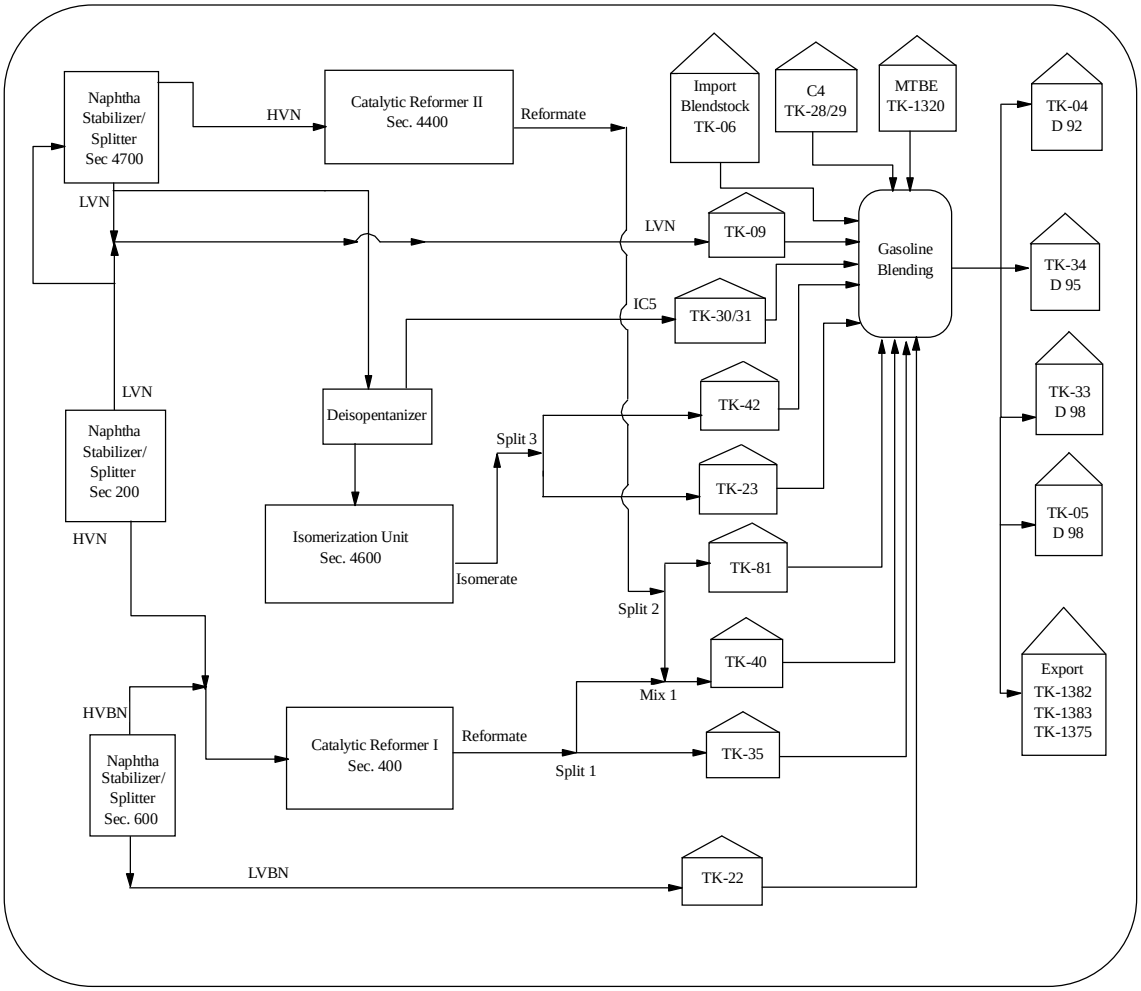

Figure 6.1: A schematic diagram of the gasoline blending unit and inventory tanks.

6.2 Optimization Model

In this section the optimization model is presented. The model integrates the gasoline blending and short term production planning for the gasoline blending. Short term implies that the scheduling horizon will be approximately 7-10 days.

6.2.1 Nomenclature

6.2.1.1 Index Sets

The plant consists of a set of objects. These objects are defined by the flow diagram shown in figure 6.1 . Abstractly, these objects are described by the sets:

- $i \in I$ where I is the set of inventory tanks in the plant.
- $j \in J$ where J is the set of outlet streams of inventory tanks.
- $n \in N$ where N is the set of inlet streams to inventory tanks.
- $o \in O$ where O is the set of all other streams including input and output streams of mixing point and splitting points.
- $k \in K$ where K is the set of quality characteristics considered.
- $l \in L$ where L is the set of gasoline products produced.
- $m \in M$ where M is the set of mixing points in the plant.
- $s \in S$ where S is the set of splitting points in the plant.
- $t \in T$ where T is the set of time points considered.
- $u \in U$ where U is the set of catalytic reformers, and isomerization units in the plant..

The structure of the plant is defined by the connection of the objects defined by the above sets. The interconnections are defined by the following subsets:

- II(i) is the set of inlet streams to tank i , $i \in I$ and $II(i) \in N$.
- OI(i) is the set of outlet streams from tank i, $i \in I$ and $OI(i) \in J$.
- IM(m) is the set of inlet streams to mixing point m, $m \in M$ and $IM(m) \in O$.
- OM(m) is the outlet streams from mixing point m, $m \in M$ and $OM(m) \in O$.
- IS(s) is the inlet streams to split point s, $s \in S$ and $IS(s) \in O$.
- OS(s) is the set of outlet streams to split point s, $s \in S$ and $OS(s) \in O$.
- IU(u) is the set of inlet streams to production unit u, $u \in U$ and $IU(u) \in O$.
- OU(u) is the set of outlet streams from production unit u, $u \in U$ *and* $OU(u) \in O$.

6.2.1.2 Variables

The variables in the model are:

- F_{jt} flow rate (m^3/hr) in stream j at time point t.
- f_{jt} volume (m^3) in stream j during the period starting at time point t.
- G_{jt} flow rate (m^3/hr) in stream N at time point t.
- g_{jt} volume (m^3) in stream n during the period starting at time point t.

- H_{ot} flow rate (m^3/hr) in stream *O* at time point *t*.
- V_{it} volume (m^3) in tank *i* at time point *t*.
- Q_{ikt} measure of quality *k* in tank *i* at time point *t*.
- q_{jkt} measure of quality *k* in stream *j* during the period starting at time point *t*.
- p_{nkt} measure of quality *k* in stream *n* during the period starting at time point *t*.
- r_{okt} measure of quality *k* in stream *o* during the period starting at time point *t*
- W_{lkt} measure of quality *k* in product *l* at time point *t*.
- *VC* is the variable cost of products over a given time horizon.

6.2.1.3 Parameters

The parameters in the model are:

- D_{lt} is the demand (m^3) for product *l* in the period starting at time point *t*.
- c_{it} is the cost ($/$m^3$), i.e. market price, of the content in tank *i* in the period starting at time point *t*.
- b_{it} is the price ($/$m^3$) of blend stock storage in tank *i* form time point *t* to time point *t +1*. Basically this parameter should be a discount factor accounting for the working capital tied up in inventory.

6.2.2 Tank Models

The models for tanks are based on total volume balance and quality characteristic balances, in which the qualities are assumed to blend linearly. The dynamic equations in the optimization model are discretized using Euler discretization.
Furthermore, the tanks are assumed to be well stirred such that the quality of the effluent stream from the tank is equal to the quality inside the tank. An upper and a lower bound for each variable is also included in the model.
The assumptions mentioned above is listed and summarized as follows:

1 Quality characteristics blend linearly.
2 The tanks are well mixed.
3 The quality characteristics of an effluent stream from a tank is equal to the quality characteristics of the material in the tank.
4 Each tank has an upper and a lower volume capacity limit.

6.2.2.1 Balance Equations

The total volume balance around tank *i* is:

$$\frac{dV_i(t)}{dt} = G_n(t) - F_j(t) \qquad \in II(i),\ j \in OI(i) \qquad (6.1)$$

in which it is assumed that there is only one input, and one output stream for each tank. The corresponding Euler approximation can be written as:

$$\frac{dV_i(t)}{dt} \approx \frac{V_i(t+\Delta t) - V_i(t)}{\Delta t} = G_n(t) - F_j(t) \qquad n \in II(i),\ j \in OI(i) \qquad (6.2)$$

The discrete time model for the total volume balance of tank i is consequently:

$$V_{i,t+1} = V_{it} + g_{nt} - f_{jt} \qquad n \in II(i),\ j \in OI(i) \qquad (6.3)$$

Where:

$$V_{it} = V_i(t) \qquad (6.4)$$

$$g_{nt} = \Delta t\, G_n(t) \qquad (6.5)$$

$$f_{jt} = \Delta t\, F_j(t) \qquad (6.6)$$

Note that $F_j(t)$ is a flow rate (m^3/h) while f_{jt} is a volume (m^3).
Similarly, the quality balance equation around each tank:

$$\frac{d}{dt}(Q_{ik}(t)V_i(t)) = p_{nk}(t)G_n(t) - q_{jk}(t)F_j(t) \qquad n \in II(i),\ j \in OI(i) \qquad (6.7)$$

is discretized as follows:

$$Q_{i,k,t+1}V_{i,t+1} = Q_{ikt}V_{it} + p_{nkt}g_{nt} - q_{jkt}f_{jt} \qquad n \in II(i),\ j \in OI(i) \qquad (6.8)$$

where:

$$Q_{ikt} = Q_{ik}(t) \qquad (6.9)$$

$$q_{jkt} = q_{jk}(t) \qquad (6.10)$$

$$p_{nkt} = p_{nk}(t) \qquad (6.11)$$

6.2.2.2 Well stirred tank assumption

The assumption of well stirred tank means that the quality of the tank effluent is identical to the quality of the material in the tank:

$$q_{jkt} = Q_{ikt} \qquad j \in OI(i) \qquad (6.12)$$

6.2.3 Mixing Point

The mixing point are modeled by a static total volume balance :

$$H_{\alpha t} = \sum_{\beta \in IM(m)} H_{\beta t} \qquad \alpha \in OM(m) \qquad (6.13)$$

and static quality balances in which the qualities are assumed to blend linearly:

$$r_{\alpha kt}H_{\alpha t} = \sum_{\beta \in IM(m)} r_{\beta kt}H_{\beta t} \qquad \alpha \in OM(m) \qquad (6.14)$$

6.2.4 Splitting Points

A total volume balance around each split point is :

$$H_{\alpha t} = \sum_{\beta \in OS(S)} H_{\beta t} \qquad \alpha \in IS(s) \qquad (6.15)$$

The quality of each effluent stream from the splitter is identical to the quality of the inlet to the splitter stream:

$$r_{\alpha kt} = r_{\beta kt} \qquad \alpha \in IS(s),\ \beta \in OS(s) \qquad (6.16)$$

6.2.5 Qualities of Isomerate, and Reformate Streams

The qualities of the products from catalytic reformers and isomerization units are calculated by the chemometric models described in chapter 5. In optimization model they are expressed in a general form of function Φ, as follows:

$$r_{\alpha kt} = \Phi_{kt}(\bullet) \qquad \alpha \in OU(u) \qquad (6.17)$$

6.2.6 Blending Model

The qualities of the outlet stream of the gasoline blending unit, which is the final gasoline product, are assumed to blend linearly as function of the qualities of the intermediate streams sent to the gasoline blending. It is expressed mathematically as follows:

$$W_{lkt} = \frac{\sum_{j \in J} q_{jkt} f_{jt}}{\sum_{j \in J} f_{jt}} \qquad (6.18)$$

The demand for each product is :

$$\sum_{j \in J} f_{jt} = D_{lt} \qquad (6.19)$$

Combination of equation 6.12, 6.18, and 6.19 gives :

$$W_{lkt} D_{lt} = \sum_{j \in J} Q_{ikt} f_{jt} \qquad (6.20)$$

6.2.7 Restrictions

The quality restrictions for each quality *k* of each product *l* is given by upper and lower bounds:

$$W_{lk}^{L} \le W_{lkt} \le W_{lk}^{U} \qquad (6.21)$$

Combination of equation 6.12, 6.18, 6.19, and 6.21 gives :

$$W_{lk}^{L} D_{lt} \le \sum_{j \in J} Q_{ikt} f_{jt} \le W_{lk}^{U} D_{lt} \qquad (6.22)$$

It is further assumed that there are upper and lower restrictions on the volume of a given blend component used in the final gasoline product.

$$f_{jt}^{L} \leq \frac{f_{jt}}{\sum_{j \in J} f_{jt}} \leq f_{jt}^{U} \tag{6.23}$$

which can be rearranged to the following equation:

$$D_{lt} f_{jt}^{L} \leq f_{jt} \leq D_{lt} f_{jt}^{U} \tag{6.24}$$

6.2.8 Bounds on Variables

A lower and an upper bound on the quality, and material in each tank is added to the model by the bounds:

$$V_i^{L} \leq V_{it} \leq V_i^{U} \tag{6.25}$$

$$Q_i^{L} \leq Q_{it} \leq Q_i^{U} \tag{6.26}$$

$$W_i^{L} \leq W_{it} \leq W_i^{U} \tag{6.27}$$

Lower and Upper bounds on the streams is as follows:

$$G_j^{L} \leq G_{jt} \leq G_j^{U} \qquad j \in N \tag{6.28}$$

$$H_j^{L} \leq H_{jt} \leq H_j^{U} \qquad j \in O \tag{6.29}$$

To avoid too drastic changes in operation conditions it could be relevant to put limits on the change of the flows from one period to the next:

$$\Delta F_j^{L} \leq F_{j,t+1} - F_{jt} \leq \Delta F_j^{U} \tag{6.30}$$

$$\Delta G_n^{L} \leq G_{n,t+1} - G_{nt} \leq \Delta G_n^{U} \tag{6.31}$$

$$\Delta H_o^{L} \leq H_{o,t+1} - H_{ot} \leq \Delta H_o^{U} \tag{6.32}$$

6.2.9 Objective Function

It is assumed that the corporate strategy is to run the refinery at full capacity. Therefore the objective is to produce minimum cost product in such quantities that the demand is satisfied. Consequently the objective function can be stated as follows:

$$VC = \sum_{t\in T}\sum_{i\in I}\sum_{j\in J} c_i f_{jt} + \sum_{t\in T}\sum_{i\in I} b_{it}V_{it} \qquad (6.33)$$

The first term in the objective function accounts for the value of the blend components put into to the final gasoline products. This term implicitly accounts for the processing cost and price of raw material used to produce the blend components. In this formulation it is assumed that the processing cost is independent of the processing conditions and only depends on the throughput. Independence of production conditions and cost is an assumption which is open for discussion. In reality there will obviously be some relation between the processing conditions and the costs of the blend components. This relation can be incorporated by considering a discrete set of processing and relating the cost coefficients at these discrete processing conditions only. If the cost coefficients, i.e. c_i are regarded functions of the processing conditions the decomposition of the problem presented in the next subsection of this chapter will not be possible.

The second term accounts for the working capital tied up in carrying an inventory. Basically it represent the interest rate paid to finance working capital used for carrying the inventory. It is assumed the coefficients in this term, i.e. b_i is constant during the period of optimization, since the time horizon of the optimization problem in this formulation is 7-10 days applied for short time planning and scheduling.

6.2.10 Total Optimization Model

The optimization model is expressed as follows:

$$\min\left\{ \sum_{t\in T}\sum_{i\in I}\sum_{j\in J} c_i f_{jt} + \sum_{t\in T}\sum_{i\in I} b_{it} V_{it} \right\} \tag{6.34}$$

$$s.t. \quad \sum_{j\in J} f_{jt} = D_{lt} \tag{6.35}$$

$$W_{lkt} D_{lt} - \sum_{j\in J} Q_{ikt} f_{jt} = 0 \tag{6.36}$$

$$V_{i,t+1} - V_{it} - g_{nt} + f_{jt} = 0 \qquad n \in II(i),\ j \in OI(i) \tag{6.37}$$

$$Q_{i,k,t+1} V_{i,t+1} - Q_{ikt} V_{it} - p_{nkt} g_{nt} + Q_{ikt} f_{jt} = 0 \qquad n \in II(i),\ j \in OI(i) \tag{6.38}$$

$$H_{\alpha t} - \sum_{\beta\in IM(m)} H_{\beta t} = 0 \qquad \alpha \in OM(m) \tag{6.39}$$

$$r_{\alpha kt} H_{\alpha t} - \sum_{\beta\in IM(m)} r_{\beta kt} H_{\beta t} = 0 \qquad \alpha \in OM(m) \tag{6.40}$$

$$H_{\alpha t} - \sum_{\beta\in OS(S)} H_{\beta t} = 0 \qquad \alpha \in IS(s) \tag{6.41}$$

$$r_{\alpha kt} - r_{\beta kt} = 0 \qquad \alpha \in IS(s),\ \beta \in OS(s) \tag{6.42}$$

$$r_{\alpha kt} = \Phi_{kt}(\bullet) \qquad \alpha \in OU(u) \tag{6.43}$$

$$W_{lk}^{L} D_{lt} \le \sum_{j\in J} Q_{ikt} f_{jt} \le W_{lk}^{U} D_{lt} \tag{6.44}$$

$$D_{lt} f_{jt}^{L} \le f_{jt} \le D_{lt} f_{jt}^{U} \tag{6.45}$$

$$V_i^{L} \le V_{it} \le V_i^{U} \tag{6.46}$$

$$Q_i^{L} \le Q_{it} \le Q_i^{U} \tag{6.47}$$

$$W_i^{L} \le W_{it} \le W_i^{U} \tag{6.48}$$

$$p_i^{L} \le p_{it} \le p_i^{U} \tag{6.49}$$

$$G_j^{L} \le G_{jt} \le G_j^{U} \qquad j \in N \tag{6.50}$$

$$H_j^{L} \le H_{jt} \le H_j^{U} \qquad j \in O \tag{6.51}$$

$$\Delta f_j^{L} \le f_{j,t+1} - f_{jt} \le \Delta f_j^{U} \tag{6.52}$$

$$\Delta G_n^{L} \le G_{n,t+1} - G_{nt} \le \Delta G_n^{U} \tag{6.53}$$

$$\Delta H_o^L \leq H_{o,t+1} - H_{ot} \leq \Delta H_o^U \quad (6.54)$$

This problem is a nonlinear dynamic optimization problem. The multiperiod gasoline blending problem includes gasoline blending unit, the blend component tanks, flow rate and qualities of the streams to the blending tanks. The optimal solution to this problem specifies the flow rate and qualities of streams sent to the final gasoline product tank during each time period, the qualities of blend components inside each tank, the qualities of the intermediate products sent to blend component tanks.
The optimum solution for the qualities and flow rate of the streams sent to each blend component tank will be the targets for the advanced control level.

Another possibility to facilitate the mathematical tractability of the optimization problem would be to relax the NLP by linearization of the quality balances. The linearization is based upon the new variables as presented in the following equations:

$$v_{ikt} = Q_{ikt} V_{it} \quad (6.55)$$

$$y_{nkt} = p_{nkt}\, g_{nt} \quad (6.56)$$

$$x_{ikt} = Q_{ikt} f_{it} \quad (6.57)$$

$$z_{\beta kt} = r_{\beta kt}\, H_{\beta t} \quad (6.58)$$

which are to be used in equations 6.36, 6.38, 6.40, 6.41, and 6.44.

6.3 Decomposition

The assumption that the blend component cost is independent of the processing conditions implies that the optimization model can be decomposed into sub-problems. One sub-problem is a gasoline blending problem and the other sub-problem is a production planning problem.
The gasoline blending problem includes gasoline blending unit including the component and final product tanks, and the objective is to produce final products assuming that the quality and the amount of the blend components are known. The quality variables of the blend components are calculated by using the developed process chemometrics models in this work. The optimal solution to this problem specifies the volume of each component used for the final product, and the qualities of the final gasoline product during each time period.
The production planning problem includes calculation of the qualities and volume in tanks, the qualities and flow rate of the streams in the remaining part of the plant. For production planning it is assumed that the optimal amount of consumed volume of the blend component tanks is known and provided by the gasoline blending optimization. Hence, the production planning problem solves the quality and material balances in the plant to obtain the targets for the advanced control level.
The basis for this decomposition is that the inlet flows to the component tanks are continuos stream of the products from splitters, catalytic reformers, and isomerization units. However, the contents of component tanks are used only when the gasoline blending is running, and hence the outlet flow of the blend tanks are zero between the batches of the blending.

In summary the production planning problem solves the material balances of the plant and provides the optimal inlet flow, volume and qualities of the blend component tanks. The gasoline blending optimization determine the optimal volume of the different blend component used for production of the desired final product.
The decomposition presented here gives a global optimal solution provided that the optimal solution to the gasoline blending problem makes the production planning problem feasible

6.3.1 Gasoline Blending

The gasoline blending part of the problem covers the plant form the blend component tanks to the final products, i.e. the downstream section of the gasoline plant. In this formulation it is assumed that volume and the quality variables of the component tanks are known. The qualities are calculated by the chemometrics models during the last period of filling up the tanks. The flow rate of the streams to the component tanks are measured and hence the volumes are easily calculated.

$$\min \left\{ \sum_{t\in T} \sum_{i\in I} \sum_{j\in J} c_i f_{jt} + \sum_{t\in T} \sum_{i\in I} b_{it} V_{it} \right\} \tag{6.59}$$

$$s.t. \quad \sum_{j\in J} f_{jt} = D_{lt} \tag{6.60}$$

$$V_{i,t+1} - V_{it} - g_{nt} + f_{jt} = 0 \qquad n \in II(i),\ j \in OI(i) \tag{6.61}$$

$$W_{lk}^{L} D_{lt} \le \sum_{j\in J} Q_{ikt} f_{jt} \le W_{lk}^{U} D_{lt} \tag{6.62}$$

$$Q_{i,k,t+1}V_{i,t+1} - Q_{ikt}V_{it} - p_{nkt}g_{nt} + Q_{ikt}f_{jt} = 0 \qquad n \in II(i),\ j \in OI(i) \qquad (6.63)$$

$$D_{lt}\, f_{jt}^{L} \le f_{jt} \le D_{lt}\, f_{jt}^{U} \qquad (6.64)$$

$$V_i^L \le V_{it} \le V_i^U \qquad (6.65)$$

$$Q_i^L \le Q_{it} \le Q_i^U \qquad (6.66)$$

$$W_i^L \le W_{it} \le W_i^U \qquad (6.67)$$

$$_i^L \le p_{it} \le p_i^U \qquad (6.68)$$

$$g_j^L \le g_{jt} \le g_j^U \qquad j \in N \qquad (6.69)$$

$$\Delta f_j^L \le f_{j,t+1} - f_{jt} \le \Delta f_j^U \qquad (6.70)$$

Consequently, the gasoline blending problem is a dynamic optimization problem, since equation 6.61, and 6.63 account for the dynamic term in this formulation. It is further assumed that the corporate strategy is to run the refinery at full capacity, and hence an estimate of the inlet streams G_{it} will be provided at each time period.

6.3.2 Intermediate Production Planning

The production planning problem consists of the remaining equations i.e. the tanks which are not used in gasoline blending, isomerization unit, catalytic reformers, splitters, mixing and splitting points.
This problem is considerably smaller than the original problem. And should be solvable at least locally.
Another and perhaps more realistic model for the production planning would be to include cost of different operation points in the catalytic reformers and isomerization unit as well as the splitters. This seems necessary as the main objective of the model is to prevent the give-away which means the quality of the products are better than the desired specifications for that product.

6.3.3 Discussion

The weakness of the model formulation above is that, it is assumed that the prices of the blend components are independent of the processing conditions. We should have the prices for reformate and isomerate as a function of octane characteristic or perhaps other qualities. Logically there should be higher prices for higher qualities. This is not the case with the current model. It is not easy to define and determine the relationship between price and quality. However, this restriction can be partly removed by partitioning the characteristics in certain discrete intervals and introducing binary variables indicating which cost region applies. In practice we define different type of products based on certain qualities and determine the prices based on their qualities.

6.4 Scheduling

In order to test the optimization model for gasoline blending, a case study is considered described in the following. The name of the production units, products, and tanks refers to the description of the process in chapter 2. The flow diagram is shown in figure 6.1.
The assumptions are described in the next subsection and a schedule for gasoline production is considered by the scenario described in the following. It should be emphasized that the information of tank capacity, flow rate, and qualities are fictitious and they just serve as an example for testing the model.

6.4.1 Assumptions

6.4.1.1 Term

In this test a period of 9 days, i.e. 216 hours, will be considered, which starts from day number one at 00:00 o'clock and end with day number 10 at 00:00 o'clock.
The discretization time interval is 4 hours. This means we get the optimum values every 4 hours.

6.4.1.2 Tank Capacity

Let's just assume that we are working with the following capacities. Table 6.1 and 6.2 show the maximum capacity of the blending component tanks and final gasoline product tanks respectively.

Blend Component	Tank no.	Volume (m^3)
MTBE	TK-20	1400
Butane	TK-28+29	800
Import Naphtha	TK-06	16000
LVN	TK-09	5000
IC5	TK-30+31	1600
Isomerate	TK-23	5000
Isomerate	TK-42	5500
Reformate II	TK-81	15000
Reformate I+II	TK-40	6500
Reformate I	TK-35	5000
LVBN	TK-22	1500
Total		63300

Table 6.1 : Capacity of Gasoline Blending Component Tanks

It is assumed that a volume of about 5% of maximum tank capacity is the minimum limit for the inventory tanks. However, the butane gas tanks; i.e. TK 28 and 29, are exception.

This will give 60135 m^3 maximum volume of all blending components which can be used for blending. It is also assumed that there are orders for seven different type of products as suggested in table 6.2, and the total capacity of this final product tanks is 76200 m^3.

Tank no.	Volume (m^3)	Final Gasoline Products
TK-04	2600	D92
TK-34	7500	D95
TK-05	2600	D98
TK-82	15000	S95
TK-33	7500	S98
TK-83	15000	G91
TK-75	26000	G95
Total	76200	

Table 6.2 : Capacity of Final Gasoline Product Tanks

6.4.1.3 Gasoline Blending Input and Output Flow Rate

It is further assumed that the output flow rate of gasoline blending system is 600 m^3/hr. Table 6.3 shows the assumed upper and lower limits for feed flow rate of different components to the blending component tanks.

	Flow Rate m^3/hr	
Blend Component	Minimum	Maximum
LVN	30	100
IC5	4	10
Isomerate	40	70
Reformate 4400	60	85
Reformate 400	30	50
LVBN	4	10
Total	168	325

Table 6.3 : Upper and lower limit for feed to blend component tanks

6.4.1.4 Price Index

The component prices used in calculation of the objective function is shown in table 6.4. These prices are taken from different issue of Oil & Gas Journal.

Blend Component	$/ m^3
MTBE	255.10
Butane	108.5
Import Naphtha	168.1
LVN	143.4
IC5	171.4
Isomerate	114.3
Isomerate	114.3
Reformate II	140.2
Reformate I+II	140.2
Reformate I	133.9
LVBN	153.5

Table 6.4: Blend component prices

6.4.2 Gasoline Blending Production Plan

It is further assumed that the blend component tanks are about 50% full at the beginning of the blending period in this scenario. It is thus assumed to be 30000 m^3 total volume of all blending components available at the start of blending period. A minimum total feed flow rate of 168 m^3/hr to the blending tanks will give 36288 m^3 for the whole period of 216 hours. Hence, the total volume of all blend stock at the end of the time period would be 66288 m^3.
Consequently, taking the capacity of the final product tanks under the consideration, it would be possible to plan for a total volume of 64800 m^3 of seven final products as suggested in table 6.5. The quality specification of the seven products is assumed to be like the suggested values in table 6.6.
A production plan for the gasoline blending is suggested as shown in table 6.7.

Final Gasoline Product	m^3/hr	Tank no.
D92	4800	TK-04
D95	15000	TK-34
D98	9600	TK-05
S95	10800	TK-82
S98	6000	TK-33
G91	10800	TK-83
G95	7800	TK-75
Total	64800	

Table 6.5 : Assumed capacity of final product tanks

	RON min.	RVP kPa	BENZENE max. vol-%	MTBE max. vol-%
D92	92	60-95	2.00	10.00
D95	95			
D98	98			
S95	95	65-95	3.00	11.00
S98	98	65-95		
G91	91	60-90	5.00	15.00
G95	95			

Table 6.6 : Assumed quality specifications for final products

Order No.	Product	Date	Time Start	Time Stop	Product Tank No.	Volume (m^3)
1	D92	Day 1	00:00	04:00	4	2400
2	D95	Day 1	08:00	16:00	34	4800
3	D98	Day 2	04:00	08:00	5	2400
4	S95	Day 3	00:00	09:00	82	5400
5	S98	Day 3	16:00	21:00	33	3000
6	G91	Day 4	00:00	09:00	83	5400
7	D92	Day 4	20:00	24:00	4	2400
8	G95	Day 5	00:00	05:00	75	3000
9	D98	Day 5	16:00	20:00	5	2400
10	D95	Day 6	08:00	16:00	34	4800
11	G91	Day 7	00:00	09:00	83	5400
12	S98	Day 7	16:00	21:00	33	3000
13	S95	Day 8	00:00	09:00	82	5400
14	D98	Day 8	19:00	23:00	5	2400
15	D95	Day 9	00:00	09:00	34	5400
16	D98	Day 9	16:00	20:00	5	2400
17	G95	Day 10	15:00	23:00	75	4800
TOTAL						64800

Table 6.7 : Orders and production plan

6.4.3 Results

The value of the objective function is M$ 8.4 for production of 64800 m^3 different gasoline products. This will give an average cost of $ 0.13 per litter produced gasoline. The multiple blending results for the 17 orders are presented in tables in appendix C. An example of the result is presented in table 6.8 which shows the results of gasoline blending for order number 9.

The first two rows in the table is the volume and the qualities of the final product D98. These are in good agreement with the demands. The rest of the table 6.8 shows the blending components volume used in the blend and their actual qualities. Notice that for MTBE, Butane, Import Naphtha, and IC5, the qualities are almost constant. Furthermore, the qualities of LVN, LVBN, reformates and isomerate are calculated value, based on the tank model.

Final Product		Volume m^3	RON	Benzene vol%	RVP kpa
D98		2400	98	2	95
Component	Time Period	Volume m^3			
MTBE	28	0	0	0	0
Butane	28	290,65	93	0	460
Import	28	0	0	0	0
LVN	28	0	0	0	0
IC5	28	0	0	0	0
Isomerate (42)	28	332,63	89	1	70
Isomerate (23)	28	0	0	0	0
Reformate II	28	893,47	101	5	35
Reformate I+II	28	0	0	0	0
Reformate I	28	883,25	100	0	45
LVBN	28	0	0	0	0

Table 6.8 :Qualities and the volume of the final product and the blend components

The results presented in appendix C exhibit generally good agreement with the demands, and is an indication for a feasible and local optimum solution.

6.5 Discussion

An optimization model for operation of the gasoline processing area of the refinery has been developed. The model concerns production of the blend components and gasoline blending over multiple periods. The model consist basically of material balances, quality requirements, and upper and lower bounds on the variables.
The model is decomposed into two sub-problems, one covering the production of the blend components and the other covering the final gasoline product.

The main assumption is that the gasoline qualities blend linearly. This assumption is based on the results obtained in this work in chapter 3 in which a neural network model is developed for prediction of the qualities of the final gasoline products. The result for this modeling indicates that linear approaches can be applied with a reasonable accuracy. Furthermore it is assumed that the reliable models for prediction of the qualities of isomerate and reformate products are available. It is further assumed that the processing cost is independent of the processing conditions and only depends on the throughput.

Although the data used in this scenario is fictitious, but the general evaluation of the multiple blending model is that the solution is a feasible, local optimum solution, and there is good agreement with the specifications and demands of the products. The value of the objective function in this scenario is M$ 8.4 for production of 64800 m^3 different gasoline products which gives an average value for the variable cost of $ 0.13 per litter produced gasoline, in which a comparison with 1992 prices (Gary, 1994) shows 15% revenue per litter gasoline.
Furthermore, the obtained optimum values of the flow rate and qualities of the reformate and isomerate products at the end of the blending period are used as the suggesting target for operation of these production units.

The main weakness of the model is that the prices of the blend components are independent of the processing conditions. Further development of the optimization model should include determination of the price of the blend components as a function of qualities. This is also a challenging job, which make the model more complex.

Chapter 7

Conclusions

7.1 Introduction

The process of gasoline blending is based on in-line blending of blendstocks, i.e. while continuous feed to the component blend tanks are introduced. In this situation applying the bias-updated regression model for on-line prediction of blend component qualities would not be adequate since the qualities of the blend stocks will change due to the upstream process variation. The existing LP plus bias-updating formulation may not handle such time-varying feedstock qualities in order to find the optimal solution for the blending problem. Thus, improved and reliable prediction of the qualities in the blendstock tanks are needed based on the variation of the upstream process as an important basis for optimization of the gasoline blending process.

The main purpose of this work has been to develop data-based dynamic models in order to predict the qualities of the blend components and supply the optimization system with the previous, present and predicted future values of the qualities. The developed models are then used in a multiperiod nonlinear optimization problem for the gasoline blending.

The models are mainly developed for prediction of Research Octane Number (RON), Reid Vapor Pressure (RVP), concentration of aromatic compounds, e.g. benzene, in the blendstocks.

The optimization is concentrated around the gasoline blending unit of the refinery, and the objective is to determine the targets for the advanced control and conventional process control system by minimizing a cost function subject to a set of process and quality constraints in such a way that the needed quantities of the different final gasoline products can be produced on-time, with the desired specifications. The objective function represent the cost of operation for production of blending components plus the inventory cost. The objective function is minimized subject to a set of constraints which represent the demands for quality and quantity of final gasoline products, provided the prediction of the qualities of the blend components are available.

In the following the conclusions for the modeling and optimization of the gasoline blending process is presented.

7.2 Modeling

7.2.1 Conclusion

Artificial Neural Networks (ANNs) models are developed for prediction of qualities of final gasoline products using the data from intermediate gasoline blend component tanks. These models are developed in order to explore nonlinear effects in the blending process.
The results for the nonlinear approach for prediction of the qualities for the final gasoline products indicate that linear approaches can be applied with a reasonable accuracy.
ANNs models exhibit good performance of prediction ability in the case of static nonlinear modeling. However, when the system exhibit dynamic behavior, static ANN models will not work. The solution to be investigated in this work is to use a dynamic or time series models.

Principal Component Analysis (PCA) is performed in order to assess the representability of the data, discover any collinearity in the selected inputs, detection of distinct clusters of data due to plant operation.
The results from PCA analysis have shown that there are systematic variations in data, and hence existence of different operation points in catalytic reformer and isomerization units. The interesting observation is that the systematic variations are related mainly to the desired quality of RON and benzene contents rather than RVP. The RVP specification for final gasoline product is different for summer and winter period, but there is no season change for specification of neither RON nor for benzene content. Consequently, for prediction of RON and benzene contents of the isomerate and reformate products it is not necessary to separate the modeling into two regions of summer and winter operation. Regarding RVP quality of isomerate and reformate, there are reasons to believe that one model will be appropriate to cover variation of RVP in both summer and winter period since the operation of reforming and isomerization processes are mainly to maintain the desired RON and benzene content of reformate and isomerate respectively.
A set of suitable input variables is grouped in each modeling case in order to secure a feasible model structure. The main excitation is benzene content which is a set point. Hence this excitation signal will be sufficient to ensure identifiability of the benzene loop. Further analysis is necessary for the whole plant section investigated.

Multivariate predictive models, applying methods in process chemometrics, are developed for quality prediction of isomerate and reformate products in gasoline processing area.
It has been observed that the quality variables are dependent on the previous value of themselves and the input variables. This means that a dynamic, time-series modeling approach is a suitable choice in this application.
The applied model in this work is ARX (Auto Regressive with Exogenous input) type, in which Partial Least Squares Regression (PLS) method is used for parameter estimation. Applying PLS parameter estimation of ARX model has increased the strength of the predictability by taking advantage of the ability of PLS to extract the useful information from collinear, noisy, input data which is relevant for modeling the output prediction.
Nonlinear PLS approaches has also been examined in order to explore and model the nonlinear relationship between input and output. The approaches include Neural Net PLS (NNPLS) and Polynomial PLS. In this work no significant nonlinear relationship has been observed for relationship between scores in inputs and output.

For the output variables, there are few hourly samples available. The qualities are measured only once per day, due to economical consideration and time consuming laboratory analyses. This low sampling frequency for model output has given rise to a challenging problem in this work. The solution to this problem is based on the one of the following two situations.
If the output variation is slowly moving over one day to the next day, as it is in the case of RON, a linear interpolation of the output is performed to recover the model output in calibration. However, it is avoided to perform output interpolation in validation.
In the opposite case, in which there is a considerable variation in the output signal, indicating a possible faster dynamic response, such as the case of prediction of RVP and benzene contents, it would not be a good solution to replace the missing output by interpolation. The proposed solution here is that a suitable structure for the ARX model is chosen in which the hourly sampled input variables are used together with the previous existing output measurement, normally measured every 24 hours, in order to model the prediction of output at time t. This solution is integrated in the regression matrix of the ARX structure, in which the regression for output is over 24 hours whereas the inputs are available every hour.

Another problem is concerned with the cross spectrum between input and the driving noise realized by output feedback. The control strategy of the reformer and isomerization units is based on feed-back control of RON quality due to the great importance of octane number on production economy. This has effectively caused that the obtained input-output data set is only little informative with respect to prediction of the output RON due to the effect of an apparently well tuned closed-loop control.

The models are validated using cross validation. Since the model has time-series dynamic characteristics, it is important to secure a calibration and validation set containing time sequence of subsequent data. Thus, the validation is performed applying a completely distinct set of data, and it is attempted to cover both winter and summer operation both in the validation and calibration data sets.
Even if the selected data in calibration cover almost 9-10 months operation and the validation period is about 4 months, it must be emphasized that the developed models have a moderate general characteristics in which the models are valid only for the operation regions that they are calibrated for, and hence, implementation of the models for other operations points will need further calibration. The limited characteristics of the models are due to the effect of closed-loop control and low sample frequency of the output. However, under the existing

circumstances, the results exhibit acceptable prediction ability and performance of the ARX models in time-series regression.

7.2.2 Future Work

By applying the techniques developed in this work attacking the problem of low sampling frequency, there will be no need for providing a massive amount of quality measurements. Having in mind that a massive amount of quality measurement would not be economically feasible. However, an investment in a reasonable higher sampling rate is recommended for a limited period of time. This can be for instance a sampling rate of two or more laboratory analysis per day for a period of few weeks covering both winter and summer operation modes.
Regarding elimination of the effect of feedback, a carefully planed experimental design should be performed. There are several methods that can be applied in a closed-loop control in order to get informative (excited) input-output data set (Nikolaou 1998). An idea in this field can be introducing an extra input to the regulator which is responsible for adding a controlled extra disturbance to the system.
Another direction in predictive quality modeling should be to combine the models for one production unit, and develop Multi Input Multi Output (MIMO) models. This will improve the strength of the models effectively since there is cross correlation between the inputs and the outputs used in all models in each production unit.

7.3 Optimization

7.3.1 Conclusion

An optimization model for operation of the gasoline processing area of the refinery has been developed. The model concerns production of the blend components mainly in catalytic reformers and isomerization unit, and gasoline blending over multiple periods.
It is assumed that the corporate strategy is to run the refinery at full capacity. Therefore the objective is to produce minimum cost product in such quantities that the demands are satisfied.

The objective function includes the value of the blend components used for the final gasoline products, which implicitly accounts for the processing cost and price of raw material used to produce the blend components. It is thus assumed that the processing cost is independent of the processing conditions and only depends on the throughput. A second term in objective function accounts for the working capital tied up in carrying an inventory. It is assumed that the interest rate is constant during the period of optimization, since the time horizon of the optimization problem in this formulation is 7-10 days applied for short time planning and scheduling.

A heuristic decomposition of the model has been performed and two sub-problems are obtained. One covering the production of the blend components and the other covering gasoline blending.

The multiple period gasoline blending model covers the gasoline blending unit including the blendstocks and final product tanks. It is assumed that the gasoline qualities blend linearly and the models for prediction of the qualities for isomerate and reformate blendstocks are available. The assumption of gasoline linear blending is based on the results of an analysis performed in this work by a nonlinear approach for prediction of the qualities for the final gasoline products indicating that linear approaches can be applied with a reasonable accuracy.

The optimization problem is based on available accurate prediction of the qualities in the streams sent to blendstock tanks reflecting the variation of mainly the reforming and isomerization process.

A case study is considered and production of different types of final gasoline is scheduled in a complete scenario for a multiple period of blending for a period of 10 days. A feasible local optimum solution is obtained and the resulting values of optimum variables in this case study indicate that there is good agreement with the specifications and demands of the final gasoline products.
A value of the objective function is obtain for this scenario which gives an average value for the variable cost of $ 0.13 per liter produced gasoline, in which a comparison with 1992 prices (Gary, 1994) shows 15% revenue per liter gasoline.

The obtained optimum values of the flow rate and qualities of the reformate and isomerate products at the end of the blending period are used for suggesting an operation target of these production units.

7.3.2 Future Work

The multiple period optimization model in this work is developed for the refinery blending process, which is used for the short term planning and scheduling. The future development of this model should be in the direction of including a longer time horizon for an intermediate-range planning and scheduling for the gasoline processing area, in which the targets for flow rate and qualities of the reformate and isomerate products can be determined. This is important for reduction of give-away in which the optimum solution can be found for a larger time horizon and targets can be used for more appropriate planning for the operation of the respective units.
Further development of the optimization model should also include determination of the price of the blend components as a function of qualities. This is also a challenging job, which may add more nonlinear characteristics to the model, and with that making the model more complex.

Real-time application the multiple period gasoline blending optimization model should be one of the steps in future work, since in-line blending and prediction of the blendstock qualities has the potential to provide a competitive benefit for the refinery.

References

1 Agrawal, S.S., (1995) *"Integrated blending control, optimization and planning"*, Hydrcarbon Processing, August 1995, 129-139.

2 Biegler, L. T., Grossmann, I.E., and Westerberg, A.W.(1997). *"Systematic Methods of Chemical Process Design"*, Prentiv Hall.

3 Brown, D. Steven (1998). *"Information and data handling in chemistry and chemical enginerring: The state of the field from the perspective of chemometrics"*, Computer and Chemical Engineering 23, 1998, 203-216.

4 Brabrand, Henrik, (1991), "*Dynamics. Identification, and control of a fixed bed reactor with reactant recycle*", Ph.D. Thesis, Department of Chemical Engineering, Technical University of Denmark

5 Dayal, Bhupinder S., and MacGregor John F. (1996). *"Identification of Finite Impulse Response Model: Methods and Robustness Issues"*, Ind. Eng. Chem. Res. Vol. 35, No. 11, 1996, P. 4078-4090.

6 Deming, Stanley N. and Morgan Stephen L. (1987), *"Experimental Design: A Chemometric Approach"* , "Data Handling in Science and Technology -Volume 3", Elsevier Science Publisher B.V.

7 Dong, D. and Thomas J. McAvoy (1994), *"Nonlinear Principal Component Analysis-Based on Principal Curves and Neural Networks"* , American Control Conference, Baltimore, Maryland, June 1994.

8 Edgar, T.F. and Himmelblau, D.M. (1989). "*Optimization of Chemical Processes*", McGraw-Hill International Edition.

9 Esbensen, K., Schönkopf, S., and Midtgaard T. (1994). "*Mutivariate Analysis in Practice*", A training Package by Computer-Aided Modeling AS, CAMO.

10 Friedland, Bernard (1987). *"Control System Design, An Introduction to Stat Space Methods "*, McGraw-Hill.

11 Gary, James H., Handwork, Glenn E. (1994). *"Petroleum Refining Technology and Economics"*, Third Edition, Marcel Dekker, New York.

12 Gill, Philip E., Walter Murray, and Margaret H. Wright (1981), *"Practical Optimization"* , Academic Press, inc.

13 Hallager, Louis, (1984), "*Multivariable Self-Tuning Control of a Fixed-bed Chemical Reactor Using Structured, Linear Models*", Ph.D. Thesis, Department of Chemical Engineering, Technical University of Denmark

14 Haykin, Simon (1994), "*Neural Networks, A Comprehensive Foundation"* .

15 Hime, David M., Robert H. Storer, and Christos Georgakis (1994), *"Determination of the Number of Principal Components for Disturbance Detection and Isolation"*, American Control Conference, Baltimore, Maryland, June 1994.

16 Herts, J., Krogh, A., and Palmer, R. G. (1991). "*Introduction to the Theory of Neural Computing*", Lecture Notes Volume I, Santa Fe Institute Studies in the Sciences of Complexity.

17 Jørgensen, Sten B., and Hangos, Katalin M., (1995). *"Gray Box Modelling for Control: Qualitative Models as a Unifying Framework"*, International Journal of Adaptive Control and Signal Processing, vol. 9 pp 547-562.

18 Kourti, Theodora, Nomikos Paul, and MacGregor, John F. (1995), *"Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway PLS"* , Journal of Process Control vol. 5, No. 4, pp. 277-284.

19 Kourti, Theodora, and MacGregor, John F. (1995), *"Tutorial: Process analysis, monitoring and diagnosis, using multivariate projection methods"* , Chemometrics and Intelligent Laboratory Systems 28, 3-21.

20 Kramer, Mark A. (1991), *"Nonlinear Principal Component Analysis Using Autoassociative Neural Network"* , AICHE Journal, February 1991, vol. 3, No. 2, P. 233-243.

21 Ljung, Lennart. (1987). *"System Identification-Theory for the user"*. Prentice Hall, Englewood Cliffs, N.J.

22 Ljung, L., and Glad, T. (1994). "*Modeling of Dynamic System*". Prentice Hall, Englewood Cliffs, N.J.

23 Martens, Harald and Næs, Tormod (1989), *"Multivariate Calibration"*.

24 Moeler, Martin (1993), *"Efficient Training of Feed-Forward Neural Network"*, Ph.D. thesis, Computer Science Department, AArhus University, Denmark, December 1993.

25 Munck, L., Nørgaard, L., Engelsen, S.B., Bro, R., and Andersson, C.A. (1998), *"Chemometrics in food science-A demonstration of the feasibility of a*

highly exploratory, inductive evaluation strategy of fundamental scientific significance" , Chemometrics and Intelligent Laboratory Systems 44, 31-60.

26 Nikolau, Michael (1998). *"NSF/NIST Workshop, Process measurement and control: Industry Needs"*, Workshop on Identification and Adaptive Control. Computer and Chemical Engineering 23, 1998, 217-227.

27 Palmer, F.H., Smith A.M. (1985), *" The performance and specification of gasolines"*, In : E.G. Handbook (Ed.) Technology of Gasoline, Blackwell Scientific London.

28 Russin, M.H., Chung H.S., Marshall, J.F. (1981), *" A transformation method for calculating the research and motor octane numbers of gasoline blends"*, Ind. Eng. Chem. Fund. 20, 195-204

29 Russin, M.H., (1975), *" The structure of nonlinear models"*, Chem. Eng, Sci. 30, 935-988.

30 Shi, Ruijie, and MacGregor, John F. (2000), *" Modeling of dynamic systems using latent variable and subspace methods"*, Journal of Chemometrics, Volume 14. no. 5-6, 423-439

31 Singh, A., Forbes, J.F., Vermeer, P.J. and Woo, S.S. (2000) *"Model-based real-time optimization of automotive gasoline blending operations"* , Journal of process Control 10, 43-58.

32 Sullivan, T.L., (1990) *"Refinery-wide blend control and optimization"*, Hydrcarbon Processing, May 1990, 93-96

33 Shinskey, F. Greg (1988). *"Process Control Systems, Application, Design, and Tuning"*, Third Edition, McGraw-Hill.

34 Tan, Shufeng, and Michael L. Mavrovouniotis (1995), *"Reducing Data Dimensionality through Optimizing Neural Network Inputs"* , AIChE Journal, June 1995, vol. 41, No. 6, P. 1471-1479.

35 Visweswaren, V. and Floudas, C.A. (1996). "*Computational result for an efficient implementation of the gop algorithm and its variants*", in I.E. Grossmann (ed.), *"Global Optimization in Engineering Design"*, kluwer.

36 Williams, H.P. (1993). *"Model Building in Mathematical Programming"*, 3rd edition, wiley.

37 Wise, Barry M., (1991), *"Adopting Mutivariate Analysis for Monitoring and Modeling of Dynamic Systems"* , Ph.D. Thesis, Department of Chemical Engineering, University of Washington, USA

38 Wise, Barry M., and Gallaher Neal B. (1996), *"The process chemometrics approach to process monitoring and fault detection"* , Journal of Process Control, Vol. 6, No. 6, pp. 329-348.

39 Wold, S., Kettaneh-wold, N., and Skagerberg, B., (1989), *"Nonlinear PLS Modeling"* , Chemometrics and Intelligent Laboratory Systems, 7, pp. 53-65.

Appendix A

Models for Catalytic Reformer II

1 Introduction

The developed models for prediction of RON, RVP, and benzene contents of reformate product from catalytic reformer II are presented in this appendix. The different steps in model development are essentially similar to the procedure applied for the models described in chapter 5.

It has been found that the output signal exhibit great correlation to the past values of input and output signals. The method used for development of the models is Auto-Regressive with Exogenous input (ARX) in which Partial Least Squares Regression (PLS) method is used for its parameter estimation.

A description of the plant can be found in chapter 2. The input variables used for the models are described in chapter 4, along with a Principal Component Analysis (PCA), and a description of data treatment.

It is expensive to have on-line quality measurements for the reformate product in order to have the same sampling frequency as the other process variables. The only existing measurement is laboratory analyses which are available only once per day for each quality, i.e. sample rate of 24 hours.

In the following sections, there will be more focus on model structure, calibration, validation, and performance of the models.

2 RVP Model

In this section the model for prediction of Reid Vapor Pressure (RVP) for reformate product from catalytic reformer II will be presented.

2.1 Inputs and Output

The following input variables are used in the RVP model.

1 Reactor 1 outlet temperature
2 Reactor 2 outlet temperature
3 Reactor 3 outlet temperature
4 Mole H2/ Mole C in recycle gas
5 % H2 purity in recycle gas
6 Reformer Feed flow rate
7 C-4401 Feed temperature
8 C-4401 Reboiler temperature
9 C-4401 Reformate product flow rate
10 C-4401 Reflux flow rate
11 C-4401 Feed flow rate
12 C-4703 Reflux flow rate
13 C-4703 LVN flow rate
14 C-4703 Reboiler Steam flow rate
15 C-201 Naphtha side stream temperature (Pressure Corrected)
16 C-4201 Naphtha side stream temperature (Pressure Corrected)
17 C-4703 Bottom temperature (Pressure Corrected)

The output is RVP measured by laboratory, and thus the model will be a Multi Input Single Output (MISO) model.
The calibration data set is chosen from a period of approximately 9 months operation. The total number of input data in calibration is 7516. The validation data cover approximately 6 months operation in which the total number of observation of input data is 2532. As described in chapter 4, the data corresponding to the periods of process shutdown and outliers has been omitted. Regarding the output RVP, there are only 305 and 93 laboratory measurements available for calibration and validation periods respectively.

2.2 Model Structure

The model structure is based on an ARX model in which the parameters θ are estimated by a PLS model. The structure of the ARX model is based on a form of regression vector in which the hourly sampled input variables are used together with the previous existing output, which is normally measured at time t-24 hour, in order to model the prediction of output at time t. As described in chapter 5, this solution is integrated in the regression matrix of the ARX structure, in which the delay time for output is inherently 24 hours. Thus, prediction of the next output can be calculated using equation A.1, which is derived based on equation 5.8 in chapter 5.

$$\hat{y}(t) = a_1 y(t-24) + B_1 U(t-K-1) + ... + B_{nb} U(t-K-nb) \qquad (A.1)$$

in which $\hat{y}(t)$ is the predicted output, U is a vector of input variables, and K is a vector of delay parameters for inputs.
Hence, there will be only one A-parameter, i.e. na=1, and number of B-parameters nb will be as many as it is necessary to get an acceptable low prediction error, compared to the defined reference models described later.
In parameter estimation a PLS model is used. A suitable number of principal components or Latent Variable LV need to be found. Thus, two sets of parameters in model development, i.e. nb and number of LV, has to be determined. This task is handled by developing a recursive routine in Matlab, in which the Root Mean Sum Squares Error in Validation (RMSSEV) is used as the criterion for optimum number of nb and LV. RMSSEV is defined as the following.

$$RMSSEV = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (A.2)$$

where $\hat{y}_i$ is the model predicted output, y_i is the output measurement, which is only RVP in this case, and n is the total number of y.
The results for these simulations are described in the next subsection.

2.3 Calibration

As discussed in chapter 5, and expressed in equation A.1, three parameters have to be determined. These are optimum number of ARX order i.e. nb, number of LV, and delay parameter for each variable.
In this work, the delay parameters are determined by recursive simulations, in which the search for delay parameters is limited by some qualified estimate according to process knowledge and physical restrictions, and then let the model find the best delay parameters found for minimum RMSSEV.
The following delay parameters has been found for the input variables in this model:

K = [3 4 4 2 2 1 3 1 1 1 3 11 5 11 1 18 6]

The search for optimum number of ARX order for input variables, i.e. nb, and number of LV is carried out by a series of separate recursive simulations, in which number of nb and LV are changed from 1 to 25 for nb, and from 1 to 35 for number of latent variables (LV). The choice for maximum number of LV is based on the following considerations.
Number of LV is a function of number of variables, and ARX order, as shown in equation A.3.

$$Max.\ LV = na \cdot ny + nb \cdot nu \qquad (A.3)$$

where ny is the number of output, and nu is the number inputs variables.
In this case na = 1, ny = 1, and nu =17. For nb=2, there will be 35 maximum number for LV.
Selecting more LV is disadvantageous, in which it will add more noise to the structure part. Moreover, total number of model parameters will increase by choosing more LV, which is not desirable, due to the risk of overfitting. These issues are discussed in chapter 3
For these reasons, maximum number of 35 LV has been chosen in this case for all nb larger than 2.

Figure A.1 shows the calculated RMSSEV as a function of nb and LV, for nb from 2 to 25 and LV from 1 to 35.
It can be seen that minimum RMSSEV can be found for LV= 4. In table A.2, the minimum RMSSEV and number of LV at the minimum are shown for nb from 1 to 20. The results for nb larger than 20 are skipped since the value of RMSSEV increase for the rest of the nb. It can be seen that the minimum RMSSEV is found for nb=2 with LV = 4. Furthermore, there is a region of nb=15 and nb=16 that RMSSEV has another local minimum, which is not a good model candidate due to the large model parameters.
Based on evaluation of the model performance in validation the case with nb=2 and LV=4 is chosen as the best model in this case.

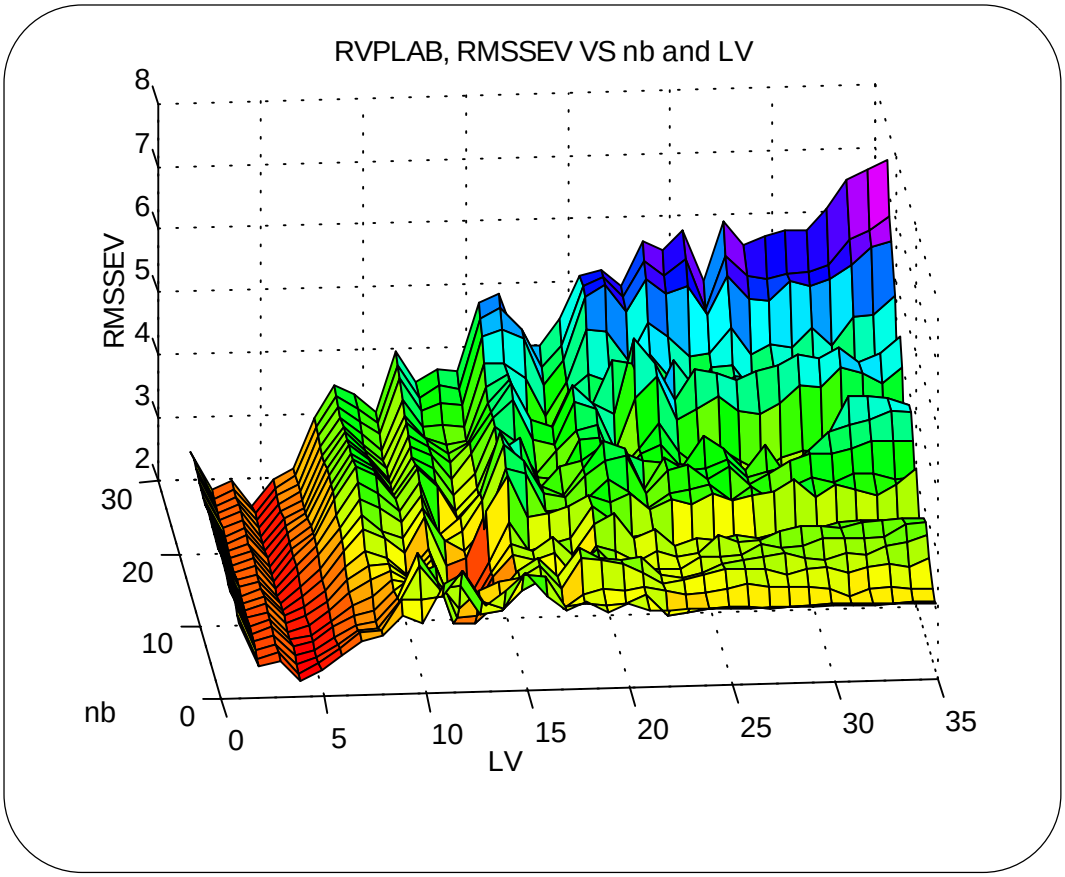

Figure A.1 : RVP Model, RMSSEV as a function of nb and LV.

The obtained RMSSEV for nb=2 is compared with the reference models, i.e. the average-model RMSEAVGV and the zero-model RMSEZROV, as it is shown in table A.1. The reference models are described in chapter 3, section 3.6 and has been used in description of the models for catalytic reformer I in chapter 5.

Validation		Calibration	
RMSSEV	2.12	RMSSEC	2.66
RMSEAVGV	3.46	RMSEAVGC	3.75
RMSEZROV	3.11	RMSEZROC	3.84

Table A.1 : RVP model, RMSSE, average-model, and zero-model in validation and calibration.

As it can be seen in table A.1 the value of RMSSEV and RMSSEC, which are the obtained Root Mean Sum Squares Error in model validation and calibration respectively, are less than RMSSE for average-model and zero-model both in validation and calibration.

nb	Min. RMSSEV	X-Block	Y-Block	LV
1	2.188	76.69	48.31	4
2	2.125	78.01	49.71	4
3	2.137	78.22	49.36	4
4	2.165	78.28	49.37	4
5	2.187	78.03	49.46	4
6	2.211	77.95	49.10	4
7	2.229	77.81	48.77	4
8	2.231	77.56	48.69	4
9	2.241	77.32	48.76	4
10	2.240	77.07	48.62	4
11	2.231	76.85	48.46	4
12	2.225	75.62	48.51	4
13	2.214	75.08	48.52	4
14	2.220	74.99	48.53	4
15	2.217	74.02	48.55	4
16	2.216	74.07	48.47	4
17	2.220	74.04	48.39	4
18	2.229	74.07	48.37	4
19	2.234	74.13	48.18	4
20	2.244	74.19	48.04	4

Table A.2 : RVP model, Minimum RMSSEV for different nb.

The plot for RMSSEV as a function of LV for nb=2 is shown in figure A.2, which shows clearly that the local minimum appears at LV=4.

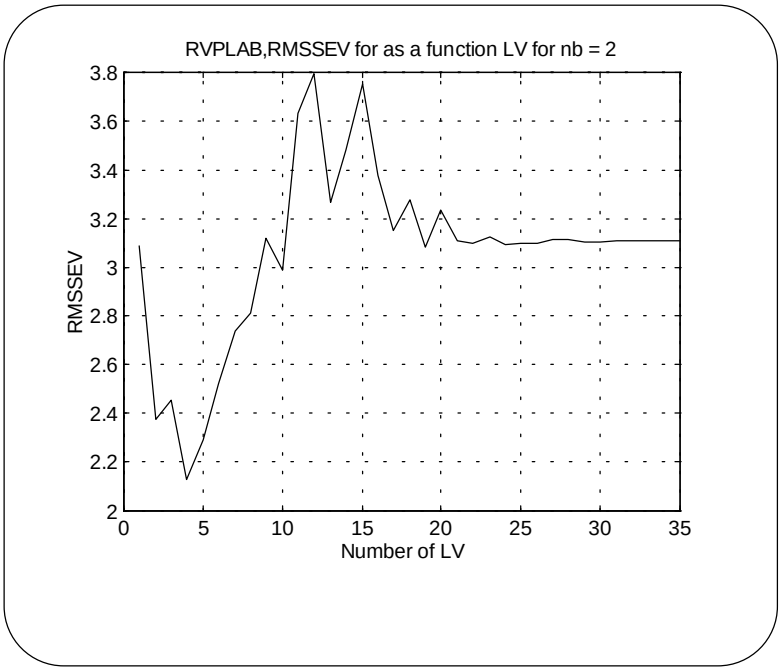

Figure A.2 : RVP Model, RMSSEV as a function of LV for nb =2.

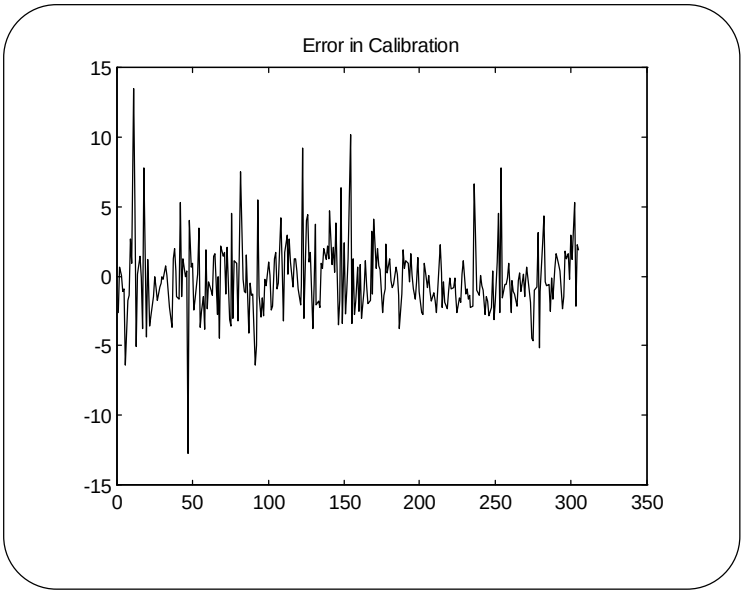

Figure A.3 : RVP Model, prediction error in calibration.

Following the procedure described in chapter 5, the next step is model calibration. The prediction error in calibration for this model is shown in figure A.3. Notice that the obtained RMSSE in calibration is 2.66 kP RVP. Figure A.4 is also used for the assessment of calibration. It can be seen that the histogram of error in calibration shown in figure A.4 exhibit an approximate zero mean error.

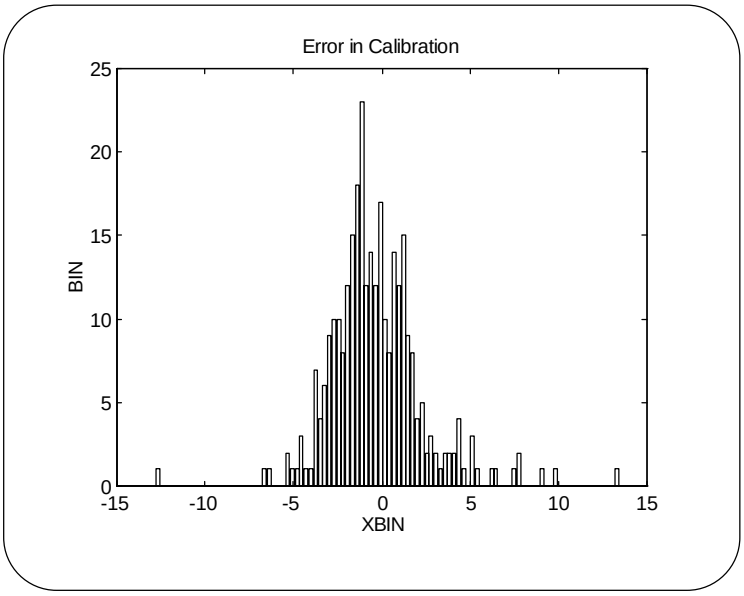

Figure A.4 : RVP Model, histogram plot of prediction error in calibration

Performance of the model in calibration can be evaluated by simulation of the model using calibration data. The simulation is performed using the obtained parameters. As described in chapter 5, an open-loop simulation can be used for assessment of model calibration. In this simulation the model predicted output at time t is used in the model instead of output measurement in order to predict the output value at time t+1. The result for this simulation is shown in figure A.5. The open loop simulation will show the predictability of the model during a period of operation without having the actual output measurement.

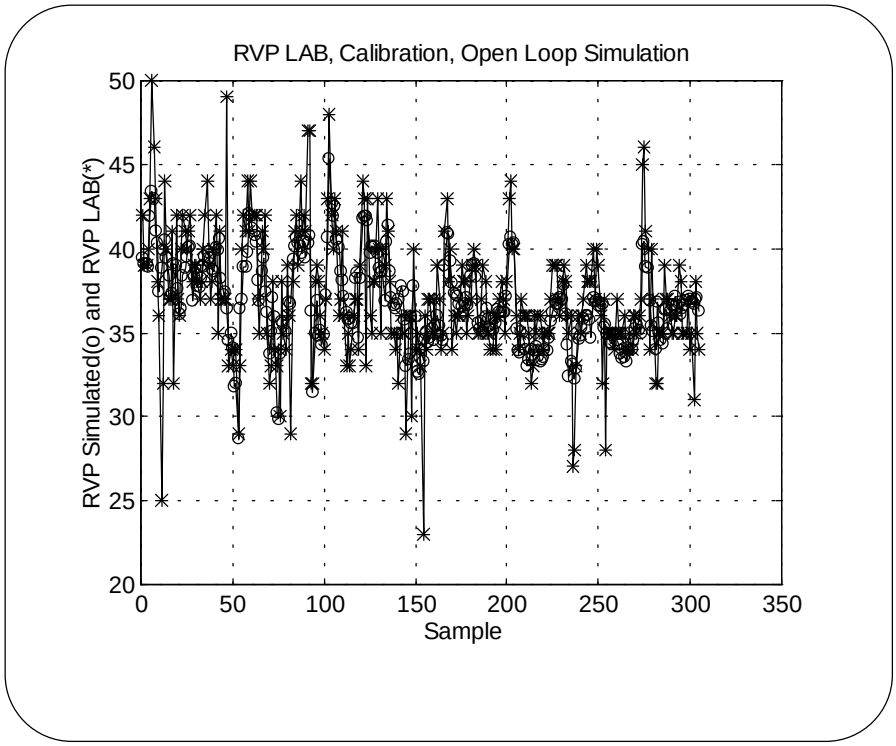

Figure A.5: RVP Model, Open Loop Simulation in calibration

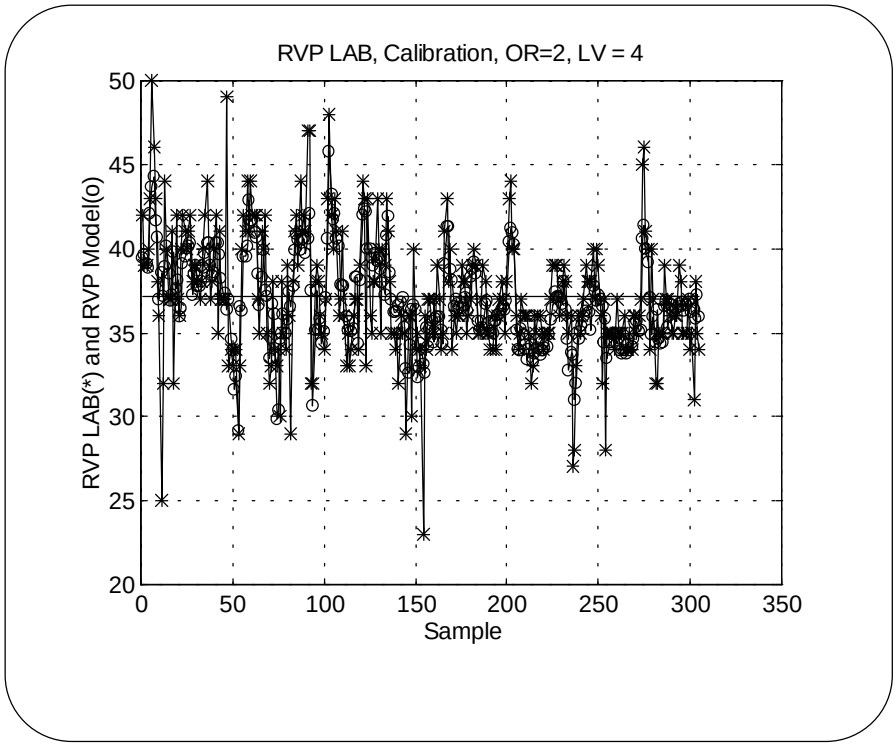

Figure A.6: RVP Model, prediction ability in calibration

Figure A.6 shows the result for simulation of the model in which the actual output measurement is used for prediction of the next output. This simulation is performed in order to assess the calibration of the model. It is expected that the developed model is capable to reproduce the calibration satisfactory.

As it can be seen from figure A.6 and A.5, the model has captured the essential variation of RVP. The result in figure A.6 can be better expressed in figure A.7. Figure A.7 shows measured RVP at laboratory versus model predicted RVP in calibration. It can be seen that although the model has captured the essential variation but it has difficulty to capture the high frequency variation of RVP.

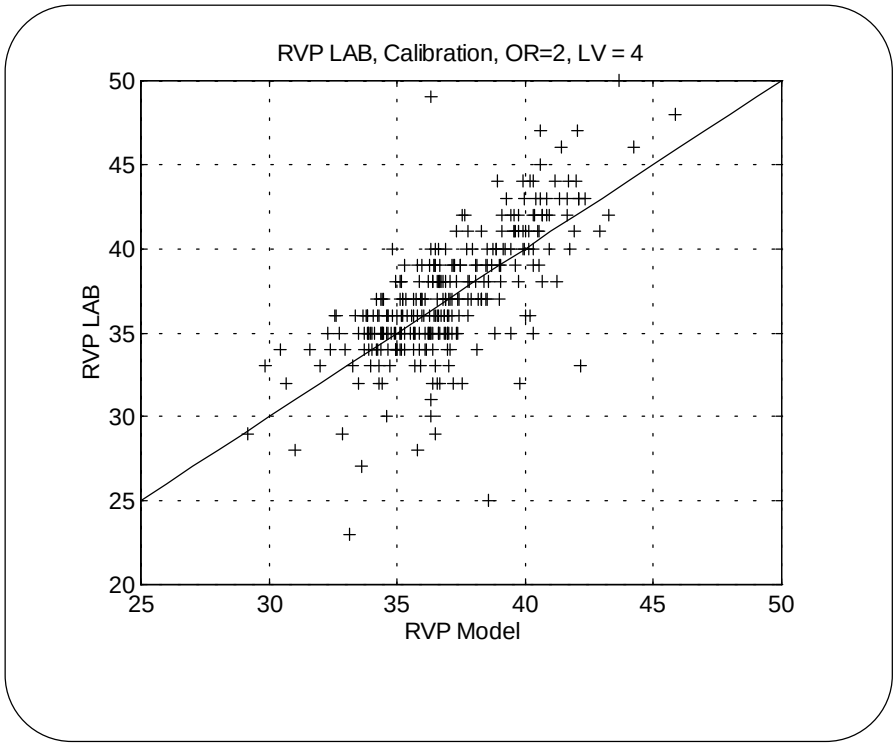

Figure A.7 : Measured RVP vs. model predicted RVP in calibration.

Selection the best nb and LV is based on performance of the obtained model in validation. The important issue is to capture the maximum effect of input variables on prediction of output and obtain a model with minimum prediction error. The issues in validation of the selected model is discussed in the following subsection.

2.4 Validation

As mentioned earlier in section 2.1, the validation data set is chosen from a period of approximately 6 months operation in which the total number of observation of input data is 2532. The data corresponding to the periods of process shutdown and outliers has been omitted, and thereby there are only 93 laboratory measurements of output RVP available for the validation period.
As discussed in the calibration section, a suitable model with fewer model parameters is chosen among a set of model candidate. The selected model has the following parameters.
The order of the ARX model is: na = 1, and nb = 2. Number of latent variable in PLS regression LV is equal to 4. The following delay parameters has been found for each input variables:

K = [3 4 4 2 2 1 3 1 1 1 3 11 5 11 1 18 6]

The RMSSE in validation, the average-model RMSEAVGV and the zero-model RMSEZROV are shown in table A.1. It can be seen that the RMSSEV is less than average- and zero-model, indicating that the model has captured the essential variation both in input and output.

The prediction error and a histogram plot of error in the validation are shown in figure A.8 and A.9. As it can be seen the error is less than in calibration, however a small bias exists.

Notice that the error here is the difference between model output and RVP measured at the laboratory, and the error value is not calculated based on autoscaled data, but it has the real unit, i.e. kPa.

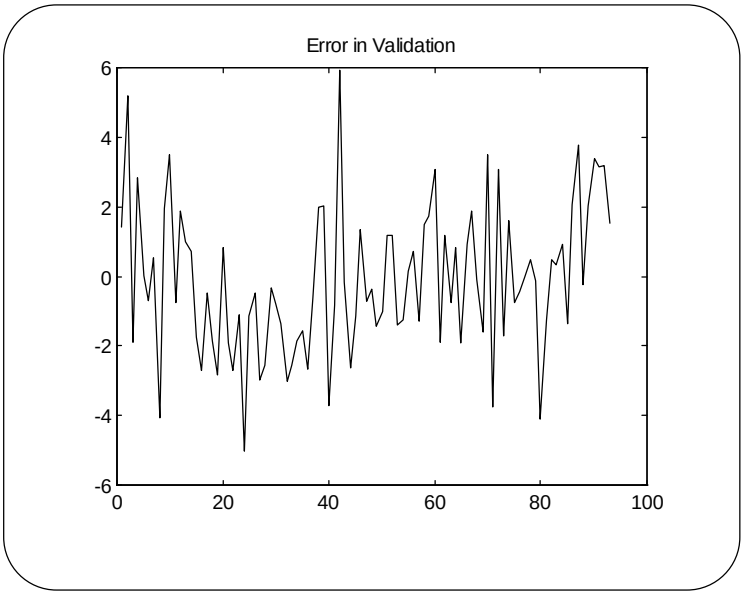

Figure A.8: RVP Model, prediction error in validation.

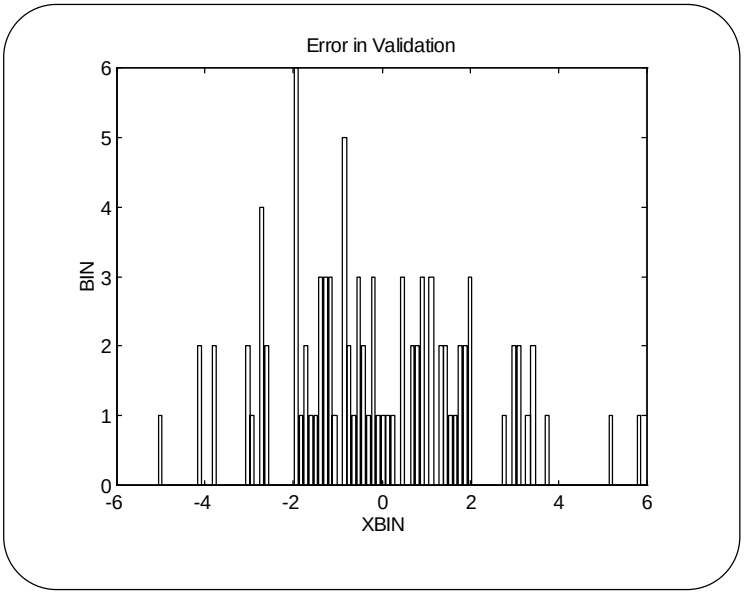

Figure A.9: RVP Model, histogram plot of prediction error in validation.

Another way to evaluate the model performance is an open-loop simulation of the model. Open-loop simulation is performed by letting the new predicted value of the output be used instead of measurement for prediction of the next output value. Open-loop simulation will tell us how well the model will predict the output values during a period of operation without having the actual output measurement.
It is interesting to see the open-loop simulation of the obtained RVP model using validation data set, which is shown in figure A.10.

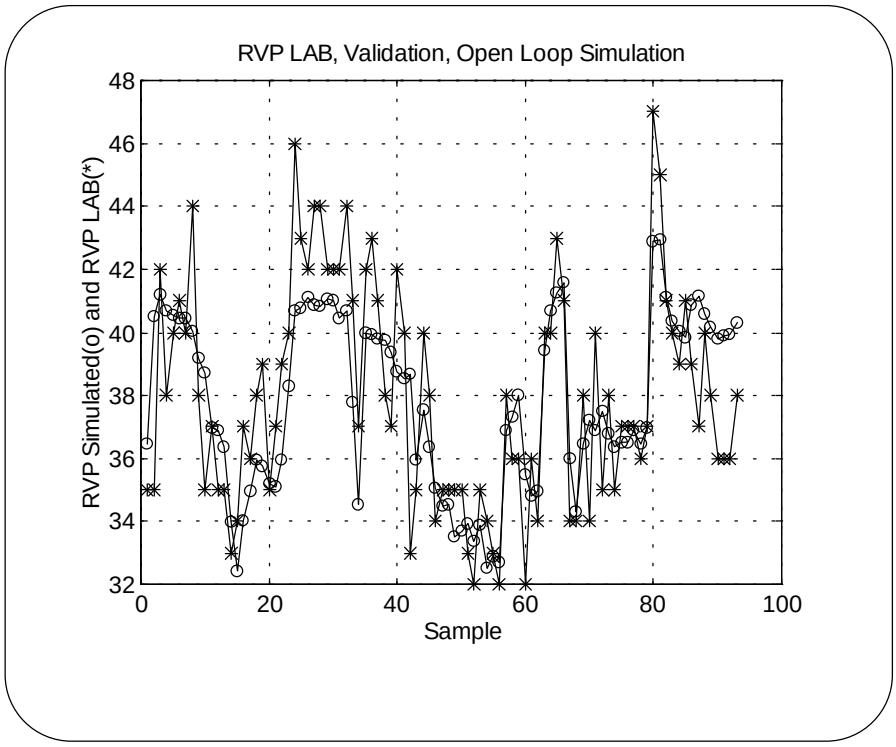

Figure A.10: RVP Model, Open Loop Simulation in validation

As we can see the model has actually captured the essential variation and follow the variation of RVP op and dawn, indicating acceptable performance.

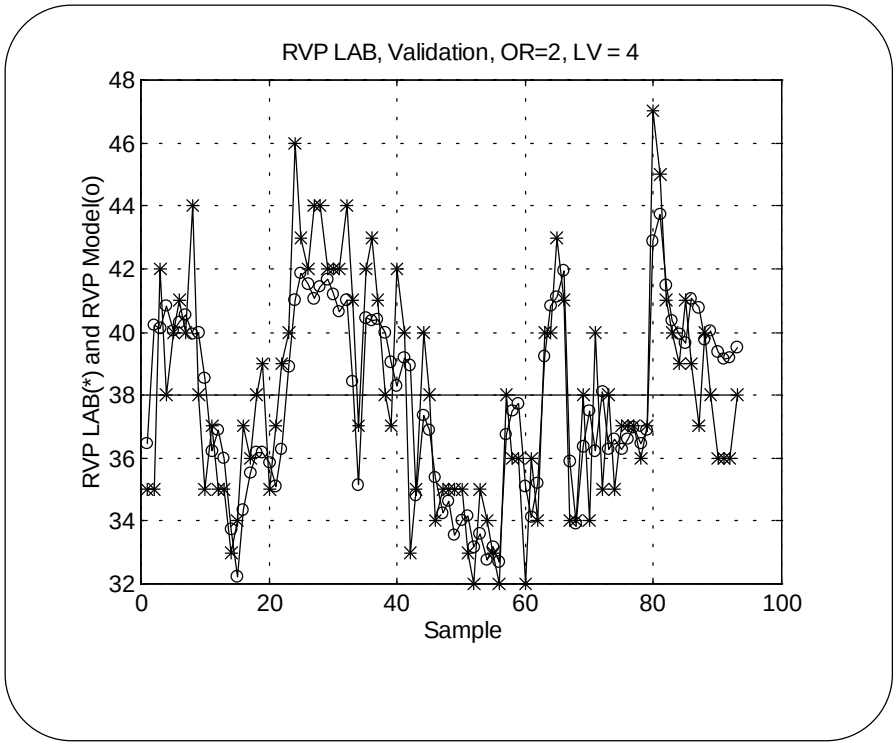

Figure A.11 : RVP Model, prediction ability in validation

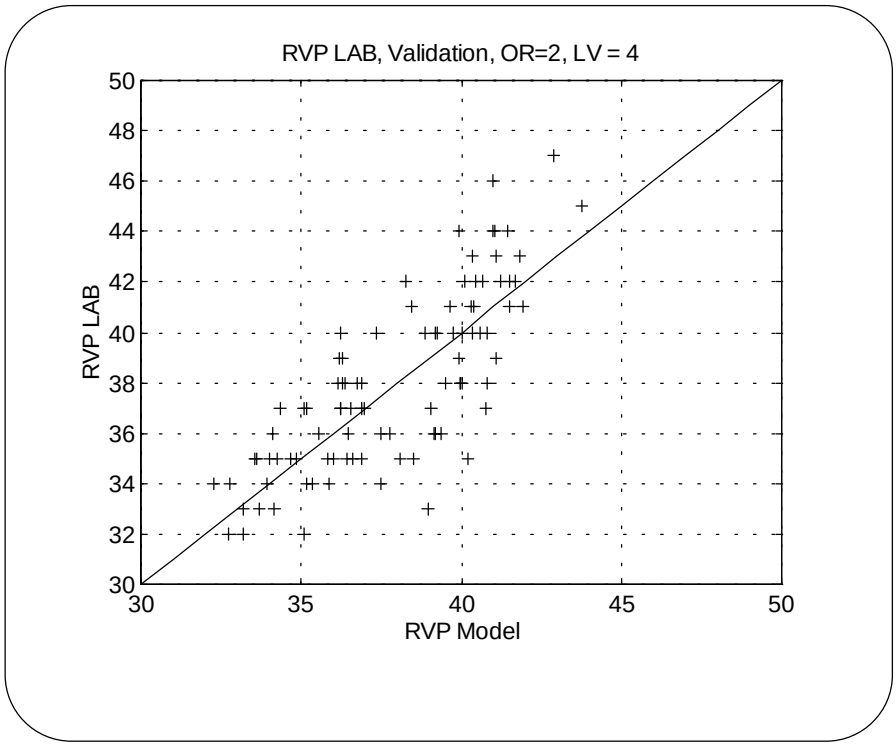

Figure A.12: RVP model predicted versus RVP measured in validation

Figure A.11 shows the result of the simulation when the actual measurements are used. It is obvious from figure A.11 and A.12 that these two simulations are similar. This similarity is an indication that the calibration has captured the essential variation in the input data. Figure A.12 shows RVP predicted versus RVP measured in validation.

Parameter	Coef.for (t-1)	Coef. for (t-2)	Sign	Description
A	0.1961	0.0000	+	Previous RVP
B8	-0.1546	-0.1717	-	C-4401 Reboiler temp.
B10	-0.1789	-0.2200	-	C-4401 Reflux flow rate
B12	-0.0219	-0.0165	-	C-4703 Reflux flow rate

Table A.3 : The largest parameters in RVP model.

Table A.3 shows a list of the parameter values of those variables that have the largest effect on the RVP. The same discussion as in chapter 5 is valid here for interpretation of the effect of different input variables.

There is only one A-parameter, which is the effect of the last measured RVP. The largest effects of input variables come from variables number B8 (stabilizer C-4401 reboiler temperature) and B10 (stabilizer C-4401 reflux flow rate) both with negative effects. This negative effects are correct since an increment in both reboiler temperature and reflux flow rate will decrease RVP as a result of removing more light hydrocarbon components from the bottom product (the reformate product).

Variable number B12 is the reflux flow rate in the naphtha splitter, in which its main objective is to split naphtha into LVN and HVN. Increment in the reflux flow rate of the splitter means more light components toward LVN and more heavy component to HVN, and by that a negative effect on RVP.
The rest of the variables is considered to have less effect on RVP.

3 RON Model

3.1 Introduction

In this section the model for prediction of Research Octane Number RON for reformate product from catalytic reformer II will be presented.

3.2 Inputs and Output

It has been found that the following variables have the most effect on output RON.

1 Reactor 1 outlet temperature
2 Reactor 2 outlet temperature
3 Reactor 3 outlet temperature
4 Mole H2/mole C in recycle gas
5 % H2 purity in recycle gas
6 Reformer feed flow rate

The total number of input data is 7515, and 2527 in calibration and validation data set respectively. These number of data set corresponds to 11 months operation data in calibration and 4 months operation data in validation. Notice that the data corresponding to the periods of process shutdown and outliers has been omitted. Regarding the output RON, there are only 331 and 100 laboratory measurements available in the calibration and validation periods respectively.

3.3 Model Structure

The RON quality is the main control variable in the catalytic reformer. Effective feedback control, based on manipulating the temperature of the inlet streams to the reactors, has caused small variation in the RON quality, as it is shown in table A.4. Notice that this is the calibration data set that cover 11 months of operation.
It has been found that the type of model structure used in the case of RVP model is not suitable for prediction of RON. As it is shown in chapter 4, there is more variation in the RVP. This is particularly due to the control strategy in the reformer unit, which is based on effective control of RON quality.

Calibration		Reactor Outlet Temperature		
	RON	R1	R2	R3
Average	101.00	391.80	446.91	472.33
Std. Deviation	0.25	3.28	5.36	4.98
Maximum	101.80	398.77	458.23	482.95
Minimum	100.00	380.81	432.86	459.88

Table A.4 : Calibration data set, RON and reactor outlet temperatures

As it is discussed in chapter 5, linear interpolation is performed in order to estimate the missing RON values. This is consequently based on the assumption that the variation of RON from one day to another is small enough to permit a rough estimation of RON between two subsequent existing RON measurement.
Since we are applying interpolation in order to estimate the missing RON output, we will have an equal number of observations in both input and output data set. The number of A-parameters, and B-parameters are then determined by model order, and we will have the same number of na, and nb. The model structure is based on an ARX model, in which PLS is used for parameter estimation.
It is important to emphasize that the interpolation is performed only in calibration data set. In the validation, we let the model apply its own predicted output, in order to predict the next output.

3.4 Calibration

As it is described earlier, we need to find a set of suitable delay parameters, optimum number of model order and LV parameter.
These parameters has been found by numerous recursive simulations. The following delay parameters has been found for the input variables:

$$k = [20 \quad 20 \quad 19 \quad 6 \quad 6 \quad 5]$$

There is no delay for output RON.

Table A.5. shows the values of RMSSV obtained for the different ARX orders and LV, in which the model order is changed from 1 to 25 in order to search for all possible effect of variables up to t-24, i.e. the previous measured RON.
It can be seen that a local minimum appear already by a second order ARX model and LV=3, the value of RMSSEV is 0.045. Furthermore, the value of RMSSEV can not be much less than 0.03 for all possible ARX orders and all LVs, and another local minimum appear at nb=10, LV= 6, which gives RMSSEV= 0.032.
The progress of RMSSEV for different LV is shown in figures A.13 for the first ARX order, and in figure A.14 for the 10th ARX order. Notice that in these figures only that part of the diagram is shown that include the minimum. The rest of the plot is skipped because including more LV produce large RMSSEV.

As discussed in chapter 5, it is preferable to choose a model structure with fewer parameters. Hence, the model structure with nb=2, and LV=3 is chosen, since the difference between RMSSEV in this case and the next local minimum is small.

It is important to notice again that these RMSSEV values is calculated based on that the model apply its own predicted output in order to predict the next output. In other word these are the values of RMSSEV in open-loop simulation of the model.

ARX Order	Min. RMSSEV	X-Block	Y-Block	LV
1	0.1950	99.28	98.68	5
2	0.0446	86.72	97.88	3
3	0.0444	86.62	96.97	3
4	0.0476	86.51	95.89	3
5	0.0434	98.76	98.78	7
6	0.0374	98.39	98.60	6
7	0.0318	96.28	98.42	5
8	0.0328	96.12	98.25	5
9	0.0324	97.39	98.14	6
10	0.0316	97.42	97.90	6
11	0.0318	97.27	97.66	6
12	0.0324	97.09	97.43	6
13	0.0335	96.88	97.18	6
14	0.0349	97.94	98.56	9
15	0.0355	97.84	98.49	9
16	0.0360	97.74	98.42	9
17	0.0362	97.63	98.35	9
18	0.0364	97.52	98.27	9
19	0.0371	97.40	98.20	9
20	0.0375	98.43	98.77	13
21	0.0375	98.34	98.76	13
22	0.0369	98.00	98.67	12
23	0.0363	97.90	98.66	12
24	0.0358	97.80	98.65	12
25	0.0354	97.70	98.64	12

Table A.5 : Minimum RMSSEV for different value of ARX order and LV.

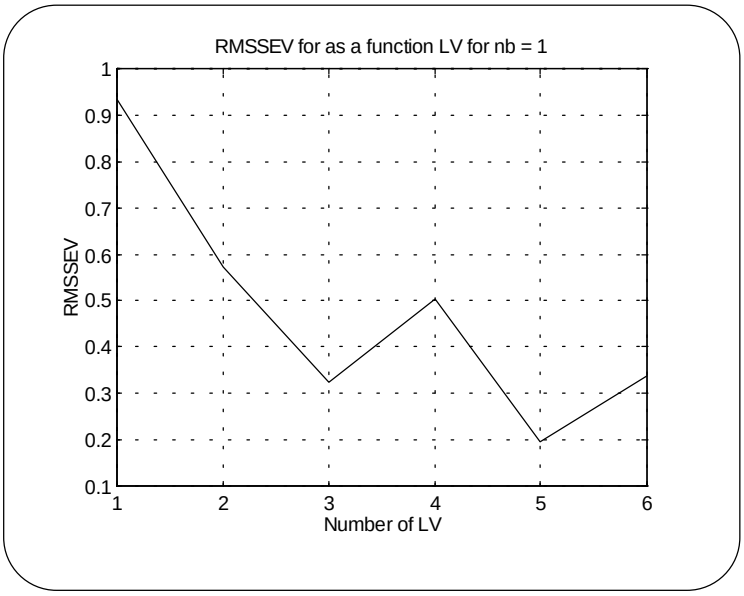

Figure A.13: RON model, RMSSEV as a function of LV for nb=1.

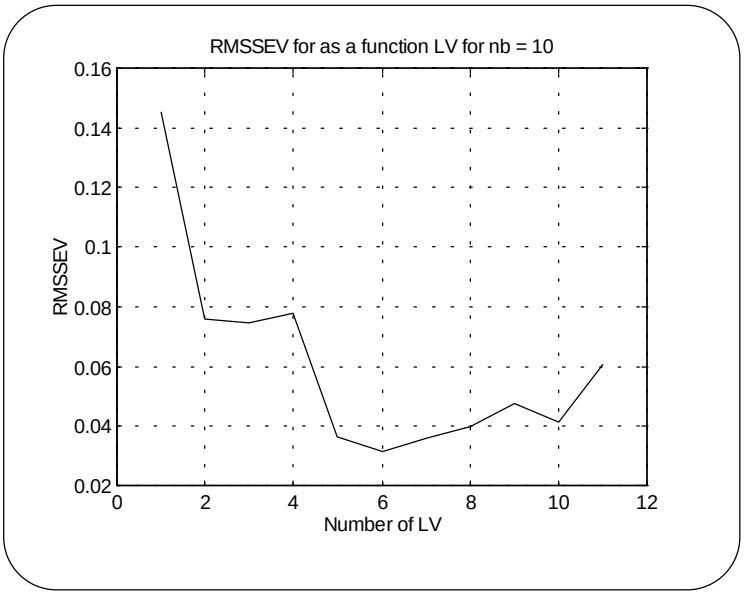

Figure A.14: RON model, RMSSEV as a function of LV for nb =10.

The prediction error for this model is shown in figure A.15. Notice that the size of the data set in calibration is 7515 observations. However, number of actual measured RON by laboratory is only 331. In figure A.15, only the error corresponding to existing measured RON is shown, and the error corresponding to interpolated data is omitted. It can be seen that the prediction error is small.

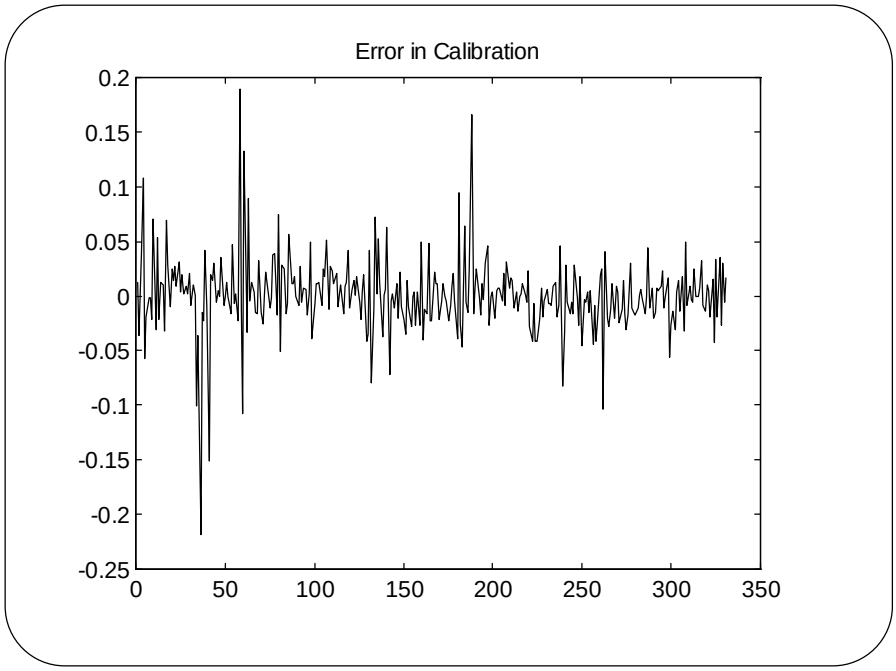

Figure A.15: RON model, prediction error in calibration.

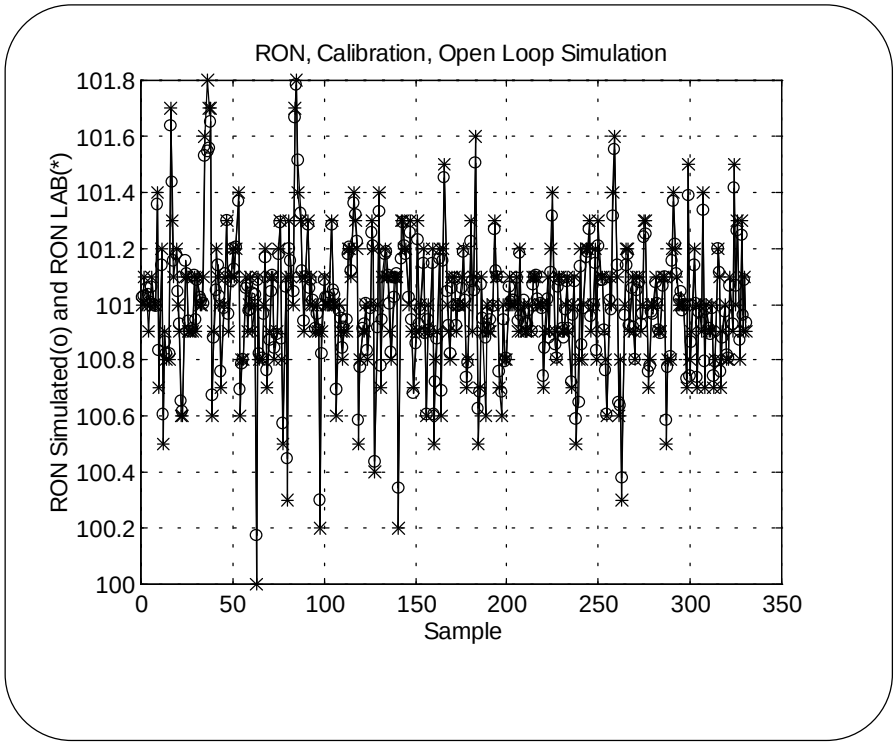

Figure A.16: RON model, open loop simulation in calibration.

Figure A.16 shows the open-loop simulation of the model by using calibration data set. Open-loop simulation is performed by letting the new predicted value of the output be used instead of measurement for prediction of the next output value. It can be seen that the prediction ability is satisfactory.

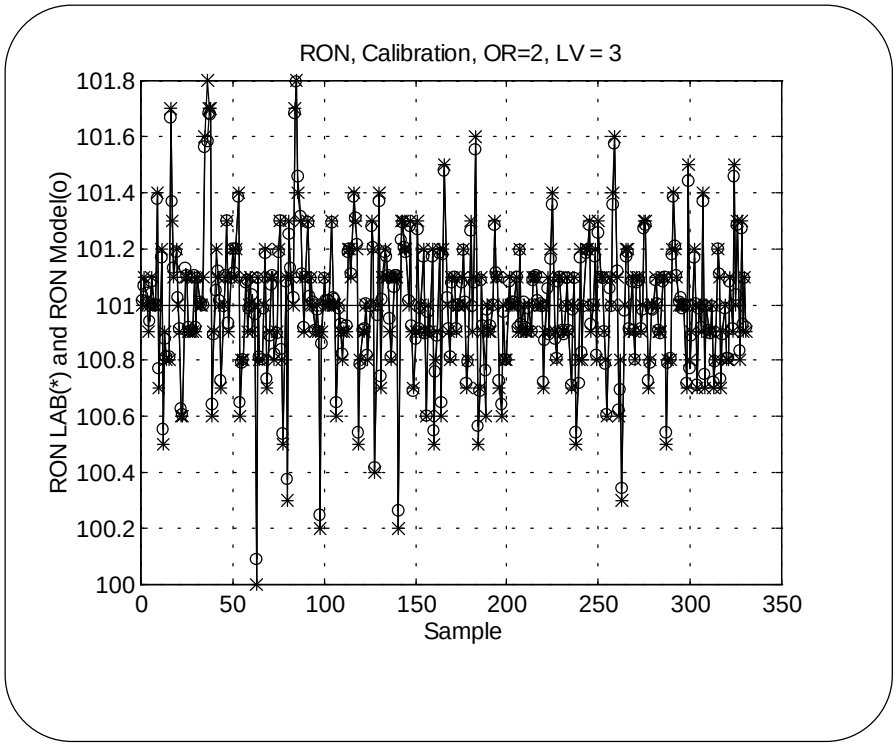

Figure A.17: RON model, prediction in calibration.

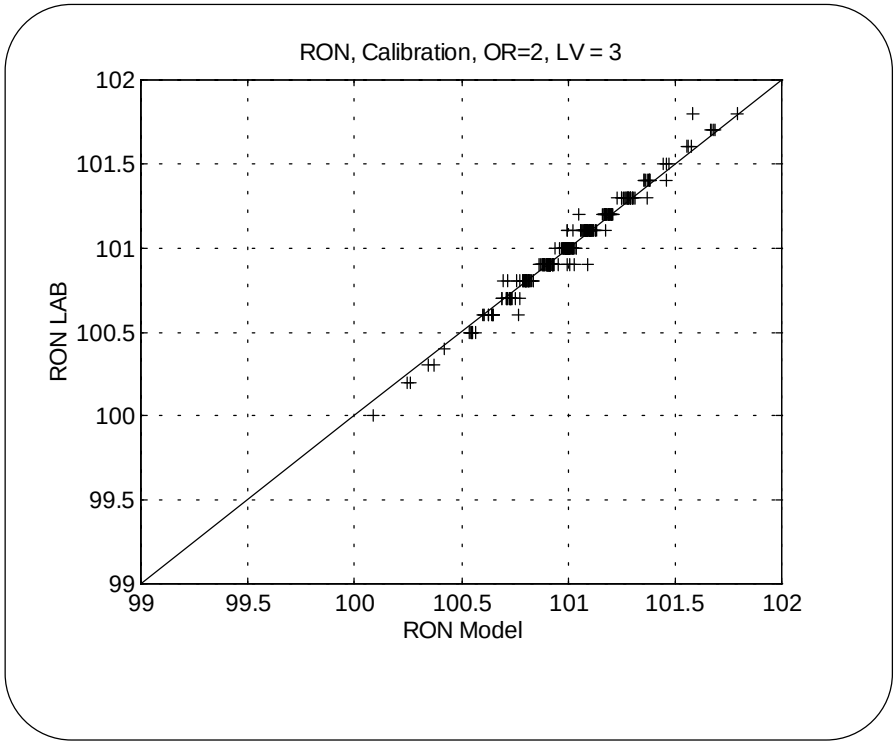

Figure A.18: Measured RON vs. model predicted RON in calibration.

Figure A.17 shows the result for simulation of the model in which the actual output measurement is used for prediction of the next output. This simulation is performed in order to assess the calibration of the model. It is expected that the developed model is capable to reproduce the calibration satisfactory.

As it can be seen from figure A.16 and A.17, the model has captured the essential variation of RON. The result in figure A.17 can be better expressed in figure A.18, which shows measured RON at laboratory versus model predicted RON in calibration.

Figure A.19 shows the histogram plot of prediction error in calibration, which exhibits an approximate zero mean error.

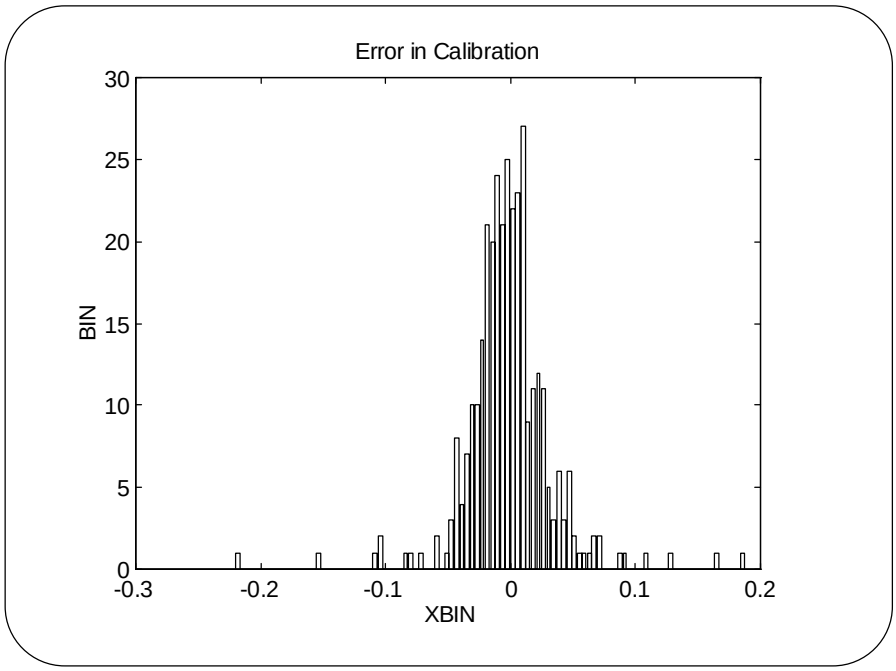

Figure A.19: RON model, histogram plot for prediction error in calibration.

3.5 Validation

The validation is performed applying a completely distinct set of data. As mentioned before the input data in validation set consists of 2527 data set covering 4 months operation. In this period, after omitting the outliers, there are only 100 laboratory measurements of output RON is remained.

The RMSSE in validation, the average-model RMSEAVGV and the zero-model RMSEZROV are shown in table A.6. It can be seen that the RMSSEV is less than average- and zero-model, indicating that the model has captured the essential variation both in input and output.

Validation		Calibration	
RMSSEV	0.0446	RMSSEC	0.0352
RMSEAVGV	0.1957	RMSEAVGC	0.2502
RMSEZROV	0.2834	RMSEZROC	0.3308

Table A.6: RMSSE, average-model, and zero-model in validation and calibration.

The prediction error in the validation is shown in figure A.20. Notice again that in figure A.20, only the error corresponding to existing 100 measured RON is shown. It can be seen that the prediction error is small.

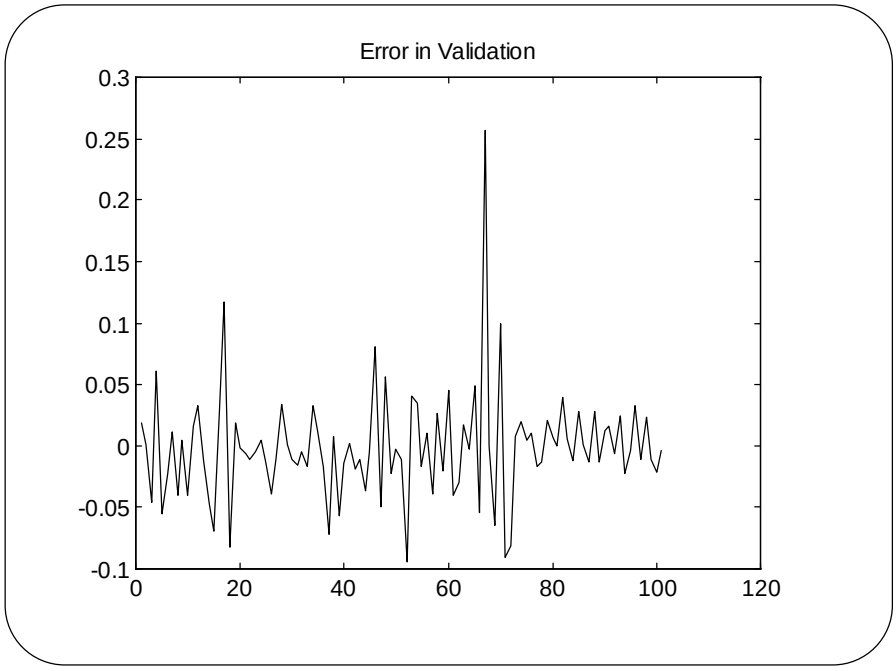

Figure A.20: RON model, prediction error in validation.

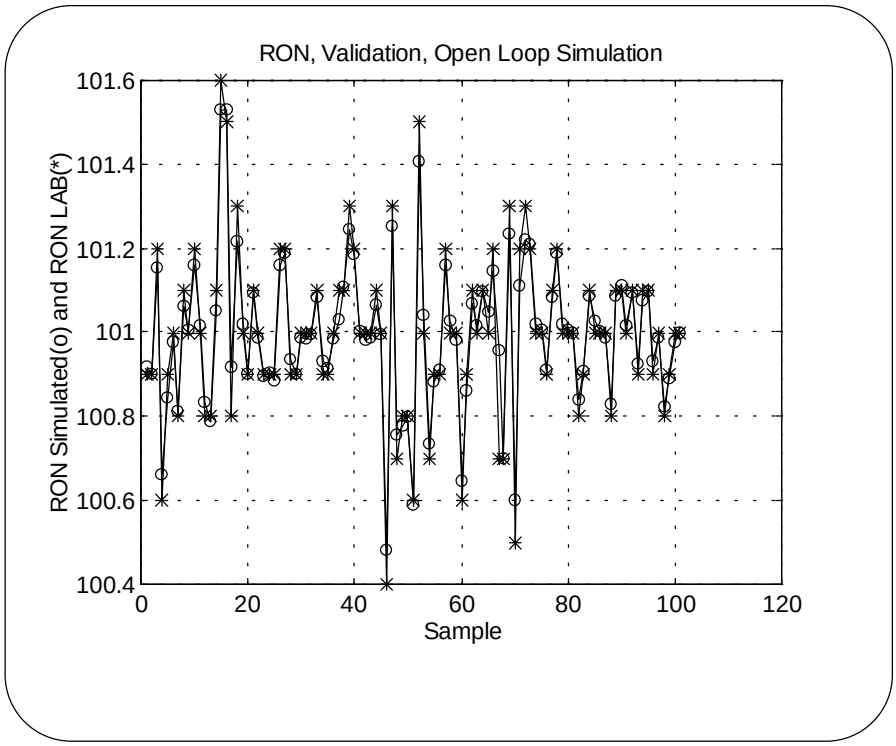

Figure A.21: RON model, open loop simulation in validation.

Figure A.21 shows the open-loop simulation in the validation, in which the predicted value of the output is used to predict the next output value. It can be seen that the prediction ability is satisfactory.

The result in figure A.21 can be better expressed in figure A.22, which shows measured RON at laboratory versus model predicted RON in validation.
Figure A.23 shows the histogram plot of prediction error in validation, which exhibits an approximate zero mean error.

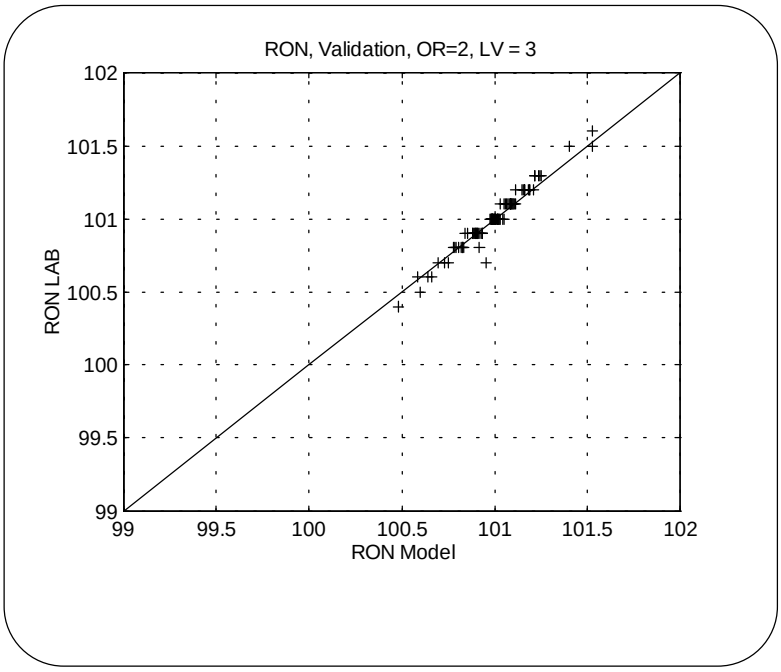

Figure A.22: Measured RON vs. model predicted RON in validation.

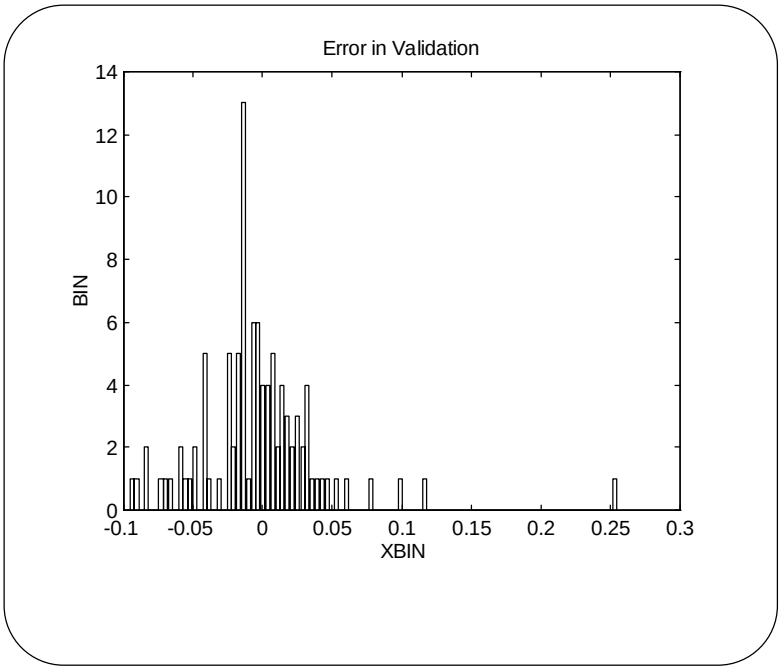

Figure A.23: RON model, histogram plot for prediction error in validation.

4 Benzene Model

In this section the model for prediction of benzene (aromatics) contents of reformate product from catalytic reformer II will be presented.

4.1 Inputs and Output

The following input variables are used in the benzene model.

1 Reactor 1 outlet temperature
2 Reactor 2 outlet temperature
3 Reactor 3 outlet temperature
4 Mole H2/ Mole C in recycle gas
5 % H2 purity in recycle gas
6 Reformer Feed flow rate
7 C-4401 Feed temperature
8 C-4401 Reboiler temperature
9 C-4401 Reformate product flow rate
10 C-4401 Reflux flow rate
11 C-4401 Feed flow rate
12 C-4703 Reflux flow rate
13 C-4703 LVN flow rate
14 C-4703 Reboiler Steam flow rate
15 C-201 Naphtha side stream temperature (Pressure Corrected)
16 C-4201 Naphtha side stream temperature (Pressure Corrected)
17 C-4703 Bottom temperature (Pressure Corrected)

The output is benzene contents (wt%) measured by laboratory. The total number of input data are 7516, and 2532 in calibration and validation data set respectively. These number of data set corresponds to 11 months operation data in calibration and 4 months operation data in validation. Notice that the data corresponding to the periods of process shutdown and outliers has been omitted. Regarding the output benzene, there are only 328 and 103 laboratory measurements available in the calibration and validation periods respectively.

4.2 Model Structure

The model structure in benzene model is similar to the structure of the model in RVP case. It is based on an ARX model in which the parameters are estimated by the PLS regression model. As it is discussed in the RVP model, and shown in equation A.1, the effect of previous output y(t-24) is taken along with the hourly sampled input variables. As described in chapter 5, this solution is integrated in the regression matrix of the ARX structure, in which the delay time for output is inherently 24 hours.
Hence, there will be only one A-parameter, i.e. na=1, and number of B-parameters nb will be as many as it is necessary to get an acceptable low prediction error. In parameter estimation a PLS model is used. A suitable number of principal components or Latent Variable LV need to be found. Number of LV, and nb are determined by a series of recursive simulations, in which the Root Mean Sum Squares Error in Validation (RMSSEV) is used as the criterion for optimum number of nb and LV.

4.3 Calibration

The following delay parameters has been found for the input variables:

K = [3 4 4 2 2 1 3 1 1 1 3 11 5 11 1 18 6]

Table A.7. shows the values of RMSSEV obtained for the different ARX orders, in which the model order is changed from 1 to 25 in order to search for all possible effect of variables up to time t-24. Figure A.24 shows a plot of the obtained RMSSEV for nb from 2 to 25, and LV from 1 to 35. Figure A.25 shows the RMSSEV for nb=1.

ARX Order	Min. RMSSEV	X-Block	Y-Block	LV
1	0.127	67.21	86.96	2
2	0.132	67.08	86.14	2
3	0.135	66.98	85.69	2
4	0.138	66.86	85.37	2
5	0.139	66.72	85.08	2
6	0.141	66.58	84.79	2
7	0.127	97.70	92.33	13
8	0.109	96.16	91.86	11
9	0.104	95.89	91.96	11
10	0.111	95.69	92.06	11
11	0.134	95.42	92.04	11
12	0.147	65.01	83.10	2
13	0.146	95.22	92.49	12
14	0.139	94.93	92.54	12
15	0.129	94.50	92.60	12
16	0.131	94.10	92.66	12
17	0.145	93.87	92.74	12
18	0.154	64.06	81.55	2
19	0.156	64.03	81.30	2
20	0.157	64.01	81.07	2
21	0.159	63.99	80.85	2
22	0.162	63.98	80.63	2
23	0.164	63.96	80.41	2
24	0.166	63.94	80.20	2
25	0.168	63.92	79.98	2

Table A.7 : Minimum RMSSEV for different value of ARX order and LV.

It can be seen that a local minimum appear already by first and second order ARX model and LV=2. Furthermore, another local minimum appear at nb=9, LV= 11, which is also shown separately in figure A.26.
As discussed in chapter 5, it is preferable to choose a model structure with fewer parameters. Hence, the model structure with nb=2, and LV=2 is chosen, since the difference between RMSSEV in this case and the next local minimum is small.

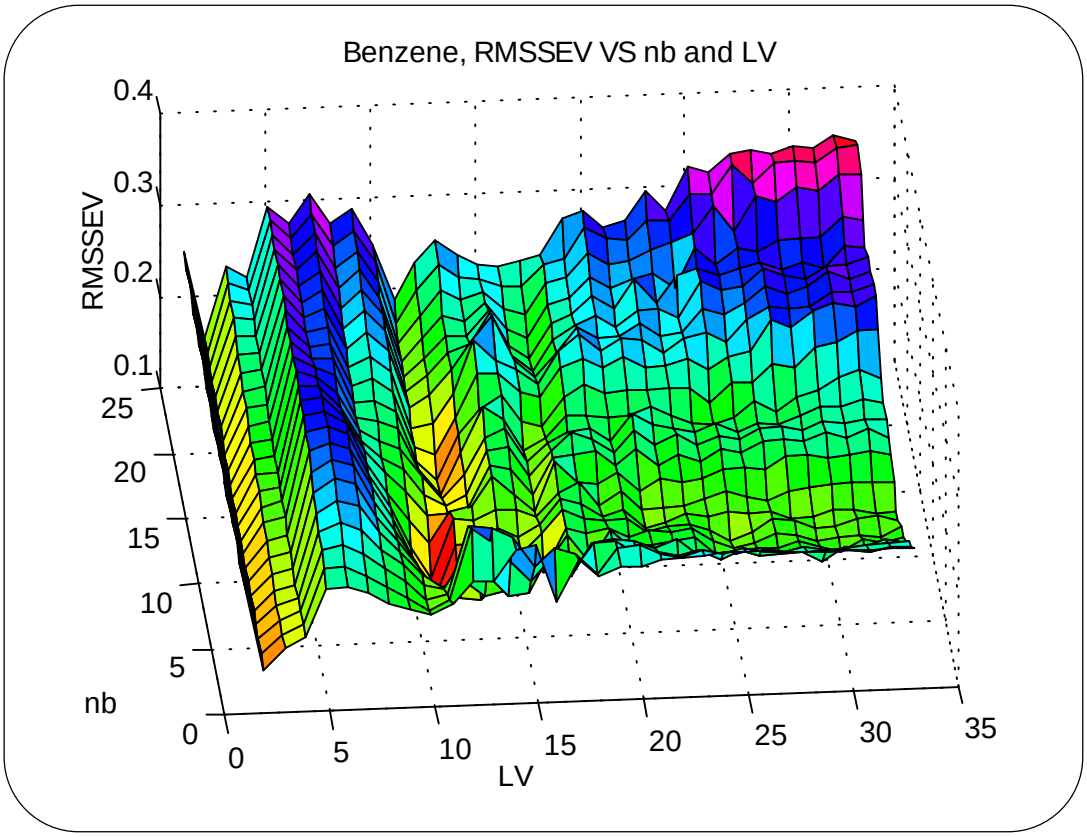

Figure A.24: RMSSEV as a function of nb and LV.

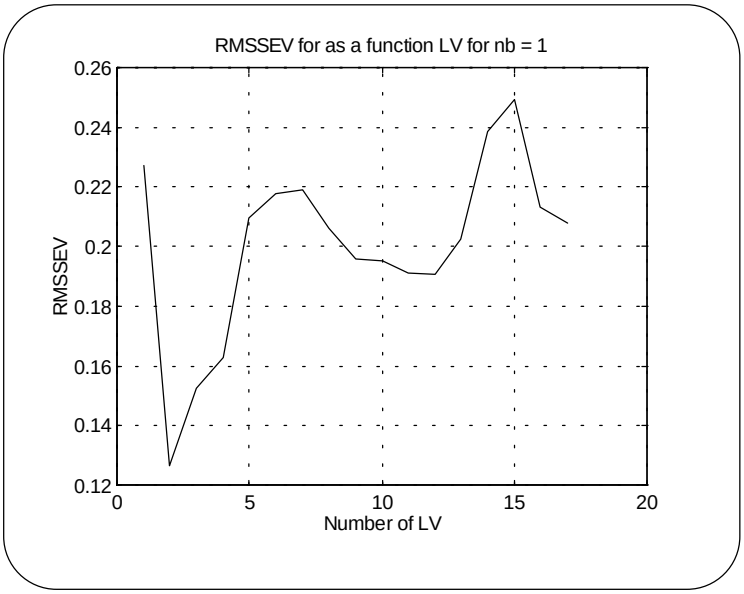

Figure A.25: RMSSEV as a function of LV for nb=1.

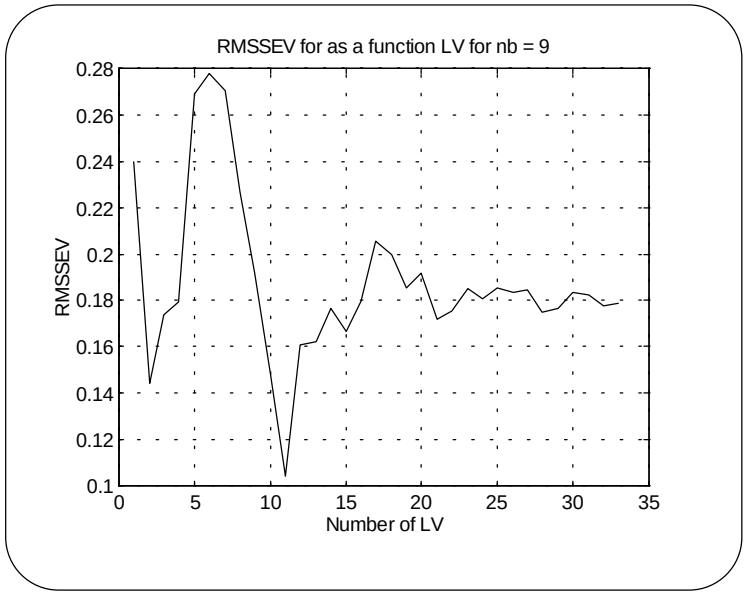

Figure A.26: RMSSEV as a function of LV for nb=9.

The prediction error for this model is shown in figure A.27. Notice that the actual number of measured benzene contents by laboratory is only 328. In figure A.27 only the error corresponding to existing measured output is shown. It can be seen that the prediction error is small. Figure A.28 shows the histogram plot of prediction error in validation, which exhibits an approximate zero mean error.
Figure A.29 shows the open-loop simulation in the calibration, in which the predicted value of the output is used to predict the next output value. It can be seen that the prediction ability is satisfactory.

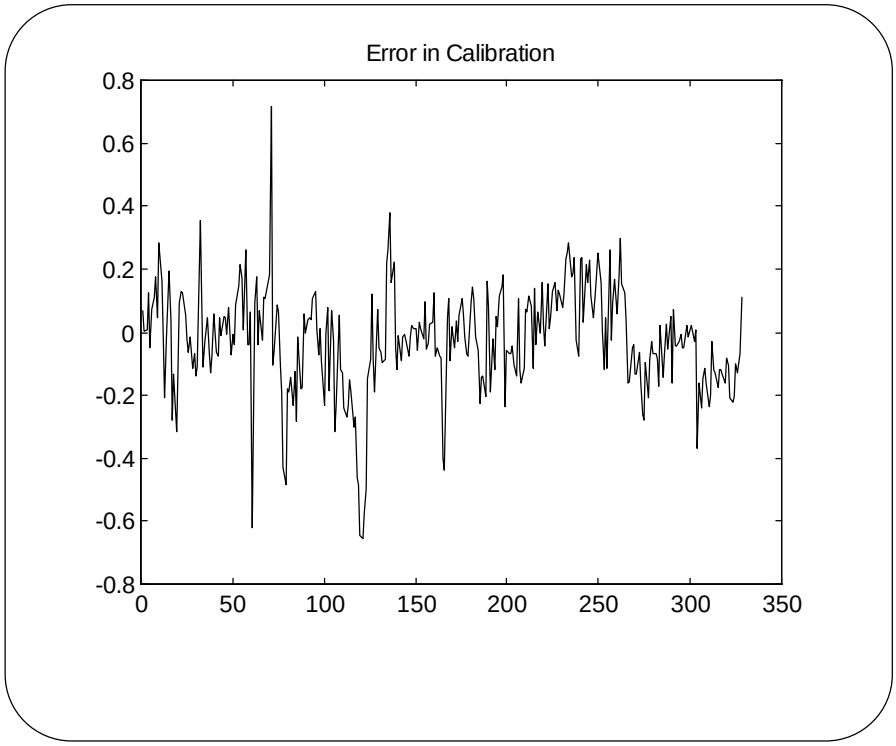

Figure A.27: Prediction error in calibration.

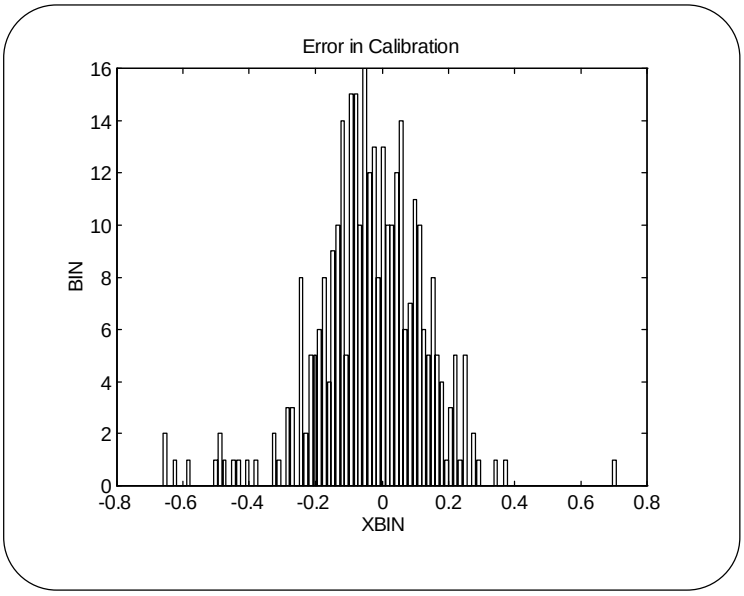

Figure A.28: Histogram plot for prediction error in calibration.

Figure A.30 shows the result for simulation of the model in which the actual output measurement is used for prediction of the next output. It is expected that the developed model is capable to reproduce the calibration satisfactory. As it can be seen from figure A.29 and A.30, the model has captured the essential variation of the output.
Figure A.30 shows the result for simulation of the model in which the actual output measurement is used for prediction of the next output. It is expected that the developed model is capable to reproduce the calibration satisfactory. As it can be seen from figure A.29 and A.30, the model has captured the essential variation of the output.

Figure A.30 shows the result for simulation of the model in which the actual output measurement is used for prediction of the next output. It is expected that the developed model is capable to reproduce the calibration satisfactory. As it can be seen from figure A.29 and A.30, the model has captured the essential variation of the output.

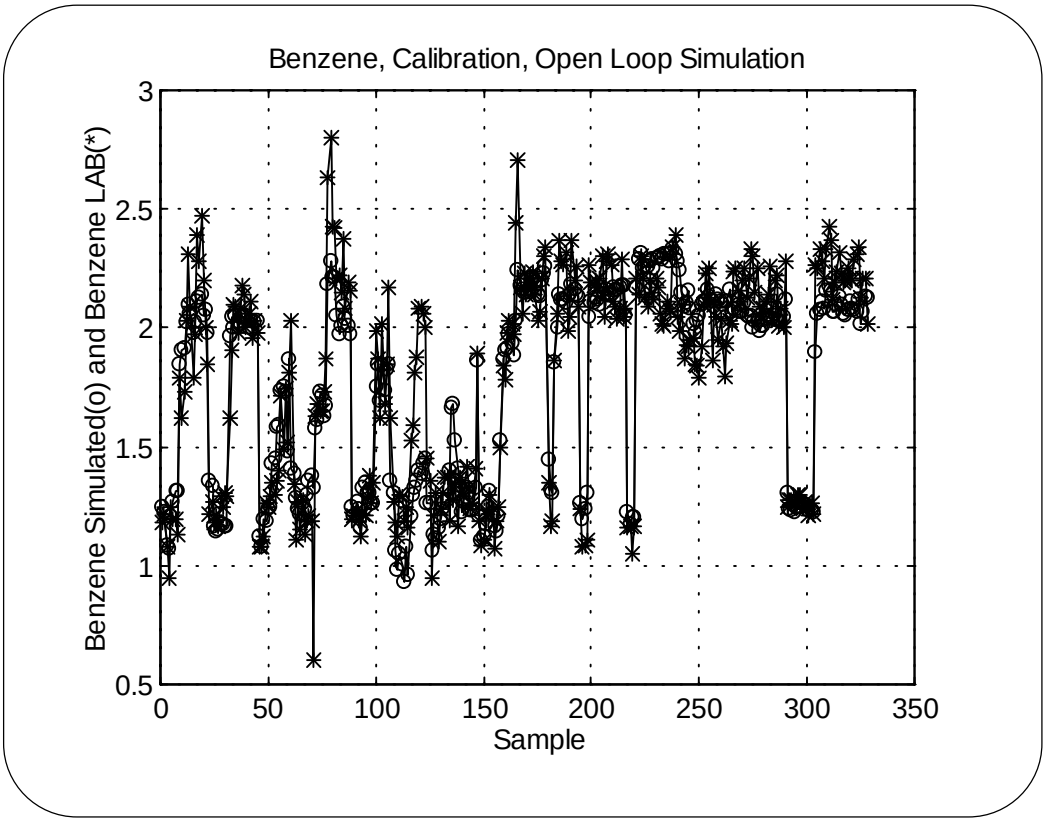

Figure A.29: Open loop simulation in calibration.

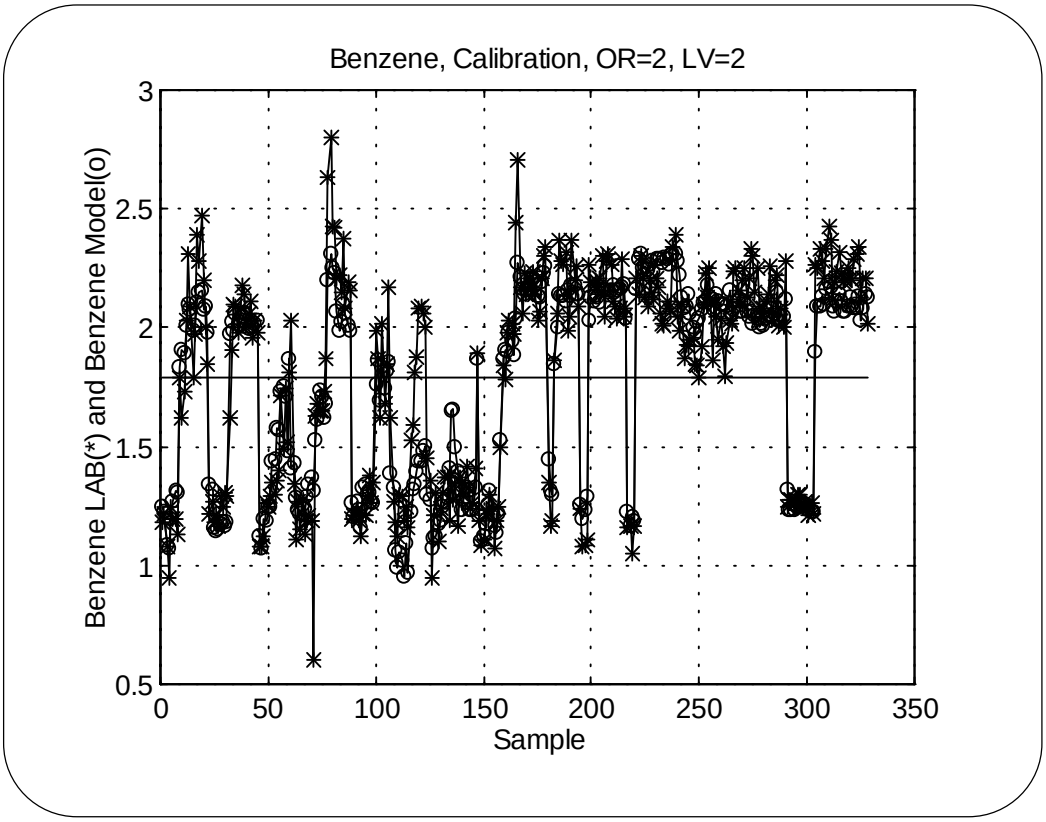

Figure A.30: Prediction in calibration.

The result in figure A.30 can be better expressed in figure A.31, which shows measured versus model predicted benzene contents in calibration.

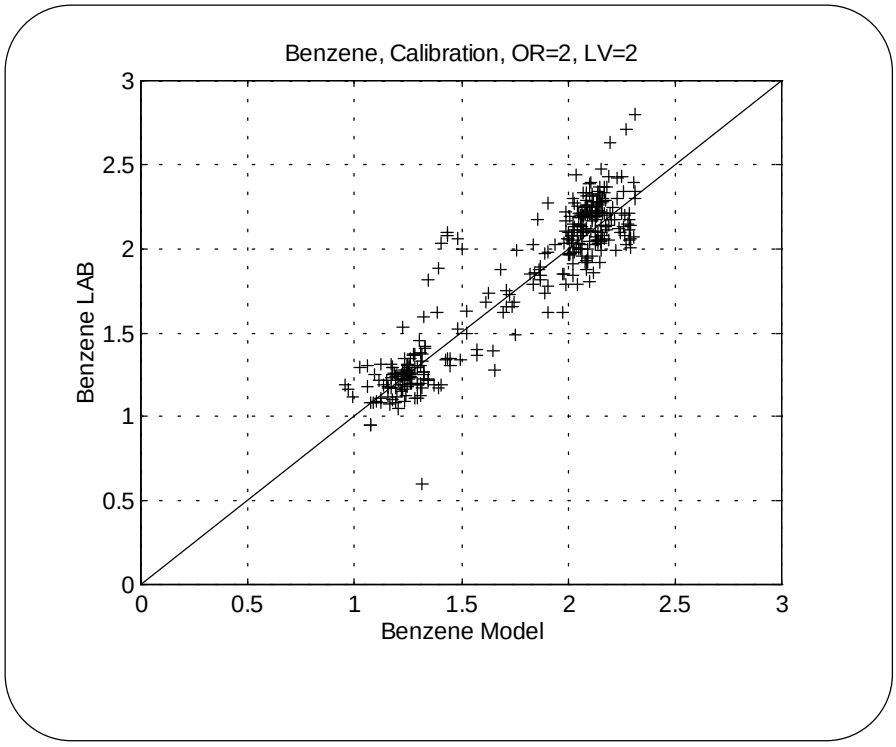

Figure A.31: Measured vs. model predicted benzene contents in calibration.

4.4 Validation

The validation is performed applying a completely distinct set of data. As mentioned before the input data in validation set consists of 2532 data set covering 4 months operation. In this period, after omitting the outliers, there are only 103 laboratory measurements of output is remained.
The RMSSE in validation, the average-model RMSEAVGV and the zero-model RMSEZROV are shown in table A.8. It can be seen that the RMSSEV is less than average- and zero-model, indicating that the model has captured the essential variation both in input and output.

Validation		Calibration	
RMSSEV	0.132	RMSSEC	0.168
RMSEAVGV	0.431	RMSEAVGC	0.451
RMSEZROV	0.266	RMSEZROC	0.255

Table A.8: RMSSE, average-model, and zero-model in validation and calibration.

The prediction error in the validation is shown in figure A.32. Notice again that in figure A.32, only the error corresponding to existing 103 measured output is shown. It can be seen that the prediction error is small.

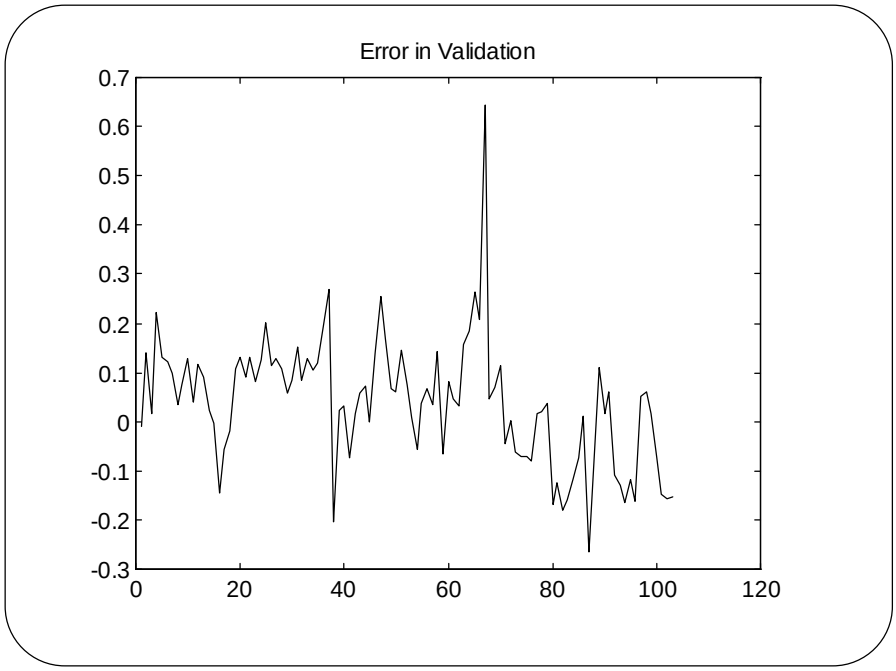

Figure A.32: Prediction error in validation.

Figure A.33 shows the histogram plot of prediction error in validation, which exhibits an approximate zero mean error.
Figure A.34 shows the open-loop simulation in the validation, in which the predicted value of the output is used to predict the next output value. It can be seen that the prediction ability is satisfactory.

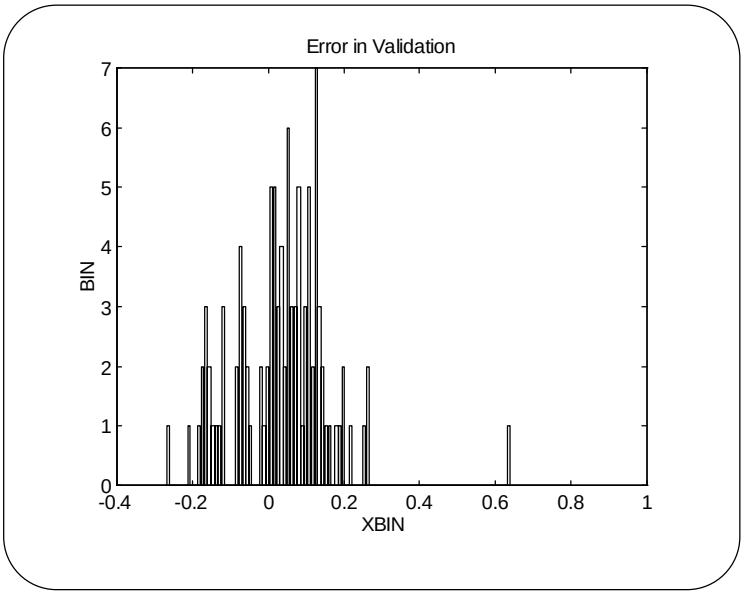

Figure A.33: Histogram plot for prediction error in validation.

Figure A.35 shows the result of the simulation when the actual measurements are used. It can be seen from figure A.34 and A.35 that the calibration has captured the essential variation in the input. It can also be seen that there are two distinct region in the output values; one around 1.4% and another around 2.1% benzene. The developed model is capable to cover both region at the same time.

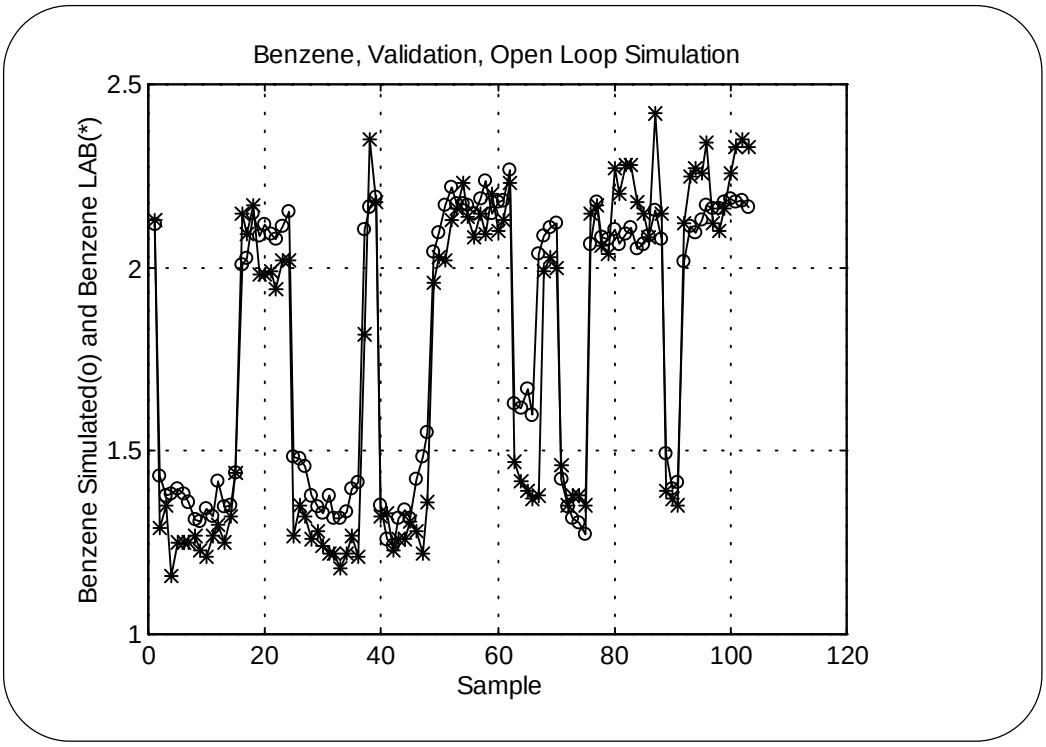

Figure A.34: Open loop simulation in validation.

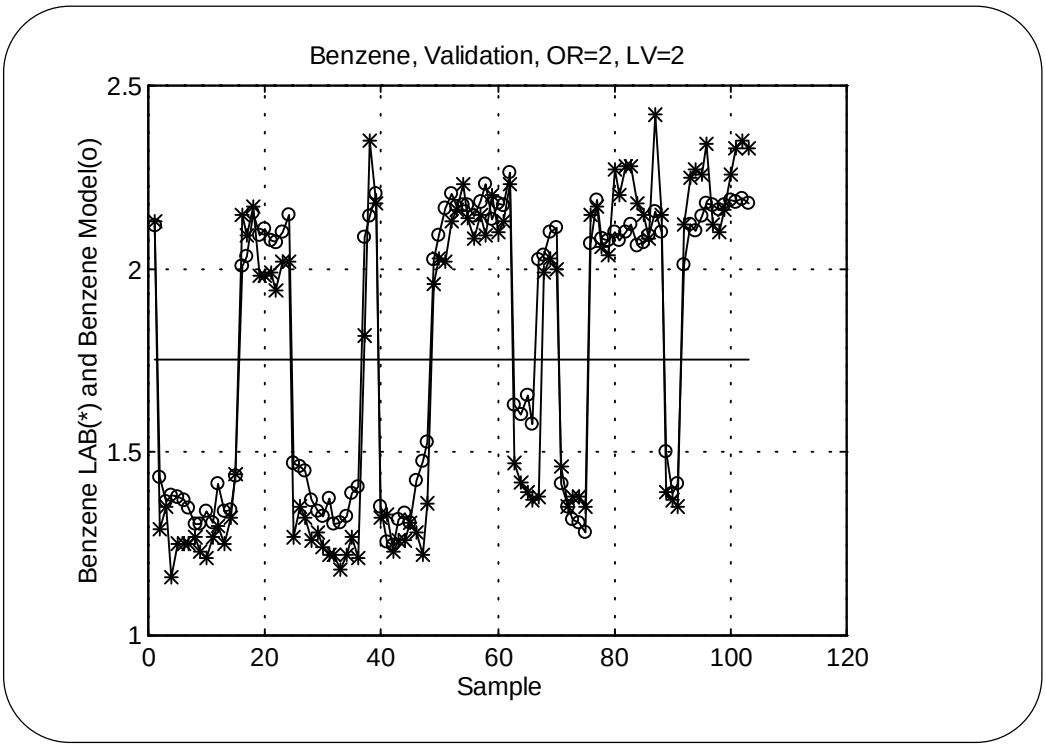

Figure A.35: Prediction in validation.

The result in figure A.35 can be better expressed in figure A.36, which shows measured versus model predicted benzene contents in validation.

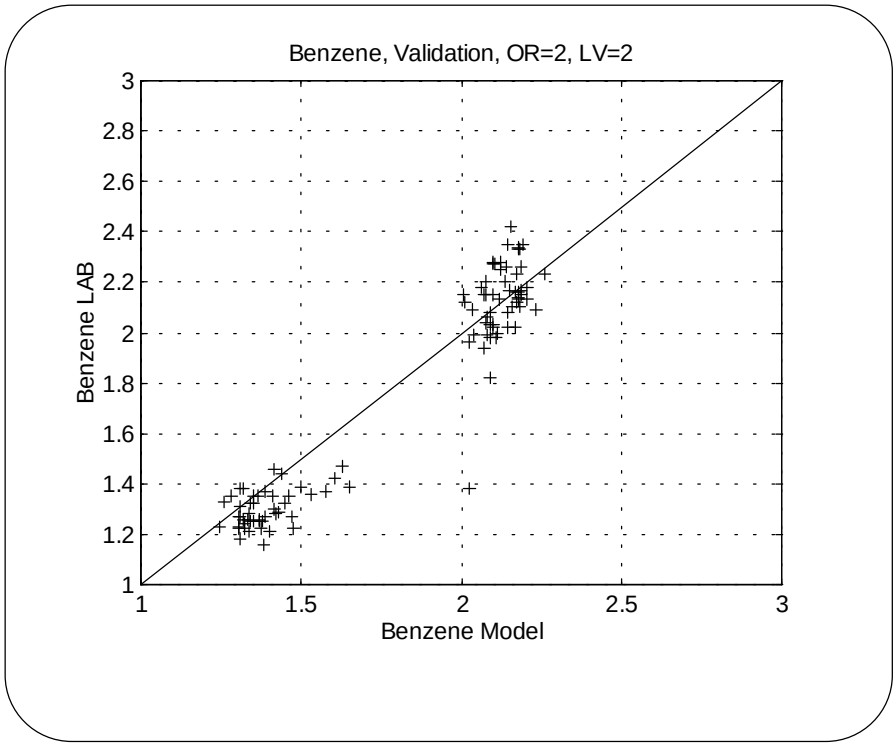

Figure A.36: Measured benzene vs. model predicted benzene in validation.

Appendix B

Models for Isomerization Unit

1 Introduction

The developed models for prediction of RON, and RVP for isomerate product from isomerization unit are presented in this appendix.

The different steps in model development are essentially similar to the procedure applied for the models described in chapter 5, and appendix A.

A description of the plant can be found in chapter 2. The input variables used for the models are described in chapter 4, along with a Principal Component Analysis (PCA), and a description of data treatment.

In this appendix, there will be more focus on model structure, calibration, validation, and performance of the models. The reader is encouraged to see chapter 5 for more detail.

2 RVP Model

In this section the model for prediction of Reid Vapor Pressure (RVP) for isomerate product from isomerization unit will be presented.

2.1 Inputs and Output

The following input variables are used in the RVP model.

1 Reactor Inlet temperature ^{0}C
2 Reactor A outlet temperature ^{0}C
3 Reactor B outlet temperature ^{0}C
4 Liquid Hourly Space Velocity (LHSV) 1/hr
5 H2 Consumption Sm^3/hr
6 Deisopentanizer (DIP) tray 8 temperature ^{0}C
7 DIP Bottom Flow Rate m^3/hr
8 DIP Feed Flow Rate
9 DIP Reflux Flow Rate
10 CP4703 top temperature (Pressure Corrected)
11 C-4703 Reflux Flow rate
12 C-4703 Feed Flow Rate

The output is RVP measured by laboratory. The data is chosen from a period of approximately 9 and 6 months operation for calibration and validation respectively. The total number of input data in calibration and validation are 5697, and 3184 respectively. The data corresponding to the periods of process shutdown and outliers has been omitted. Consequently, there are only 233 and 53 laboratory measurements of RVP available for calibration and validation periods respectively.

2.2 Model Structure

The structure of the model is based on an ARX model in which the hourly sampled input variables are used together with the previous existing output at time t-24 in order to predict the output at time t. As described in chapter 5, this solution is integrated in the regression matrix of the ARX structure, in which the delay time for output is inherently 24 hours.
Hence, there will be only one A-parameter, i.e. na=1, and number of B-parameters nb will be as many as it is necessary to get an acceptable low prediction error, compared to the defined reference models described in chapter 5.

In parameter estimation a PLS model is used. A suitable number of principal components or Latent Variable LV need to be found. Number of B-parameters nb and number latent variable LV are determined by a series of recursive simulation of the ARX model, in which minimum of the Root Mean Sum Squares Error in Validation (RMSSEV) is used as the criterion for optimum number of nb and LV.

2.3 Calibration

Another parameter needs to be determined. That is the delay parameters involved with each input variables. The delay parameters are also determined by numerous recursive simulations. The following has been found for the input variables.

K = [6 5 10 7 7 5 5 11 15 9 9 16]

Table B.1 shows the values of RMSSV obtained for the different nb and LV, in which the model order is changed from 1 to 20 in order to search for all possible effect of variables up to t-24, i.e. the previous measured RVP. The maximum number of LV is chosen to be 25 in this case.
It can be seen that there is only one local minimum that appear already by a second order ARX model and LV=4. It can also be seen in figure B.1, which shows a plot of RMSSEV versus both LV and nb, and in figure B.2, which shows RMSSEV as a function of LV for nb=2. Hence, there is only one solution, and the model structure with nb=2, and LV=4 is chosen.

nb	Min. RMSSEV	X-Block	Y-Block	LV
1	1.043	78.56	82.97	4
2	1.036	78.07	82.87	4
3	1.085	78.42	82.84	4
4	1.086	78.68	82.45	4
5	1.099	78.76	82.10	4
6	1.140	78.71	81.87	4
7	1.190	83.88	82.65	5
8	1.201	82.11	82.90	5
9	1.211	81.22	82.96	5
10	1.222	80.95	83.07	5
11	1.225	80.66	82.95	5
12	1.211	79.73	82.79	5
13	1.207	78.50	82.68	5
14	1.215	78.54	82.74	5
15	1.221	78.64	82.71	5
16	1.209	78.70	82.64	5
17	1.200	78.73	82.63	5
18	1.198	78.83	82.51	5
19	1.199	78.83	82.54	5
20	1.218	78.8395	82.606	5

Table B.1 : RVP model, Minimum RMSSEV for different nb, and LV.

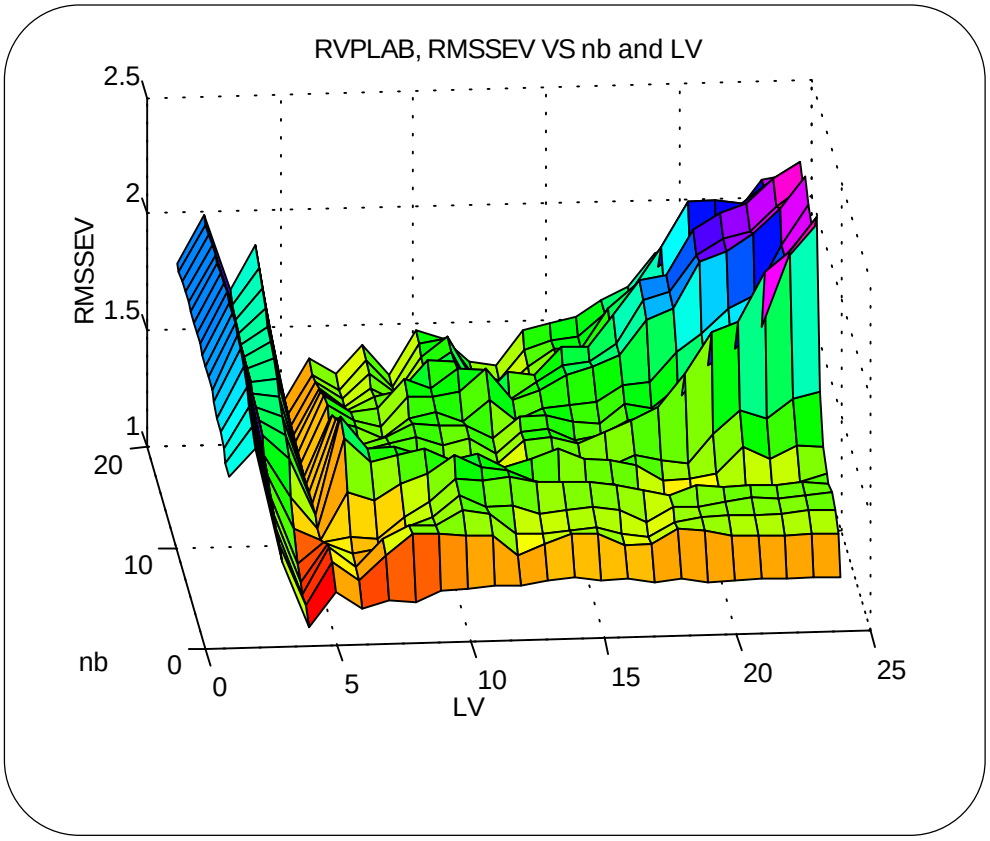

Figure B.1 : RVP Model, RMSSEV as a function of nb and LV.

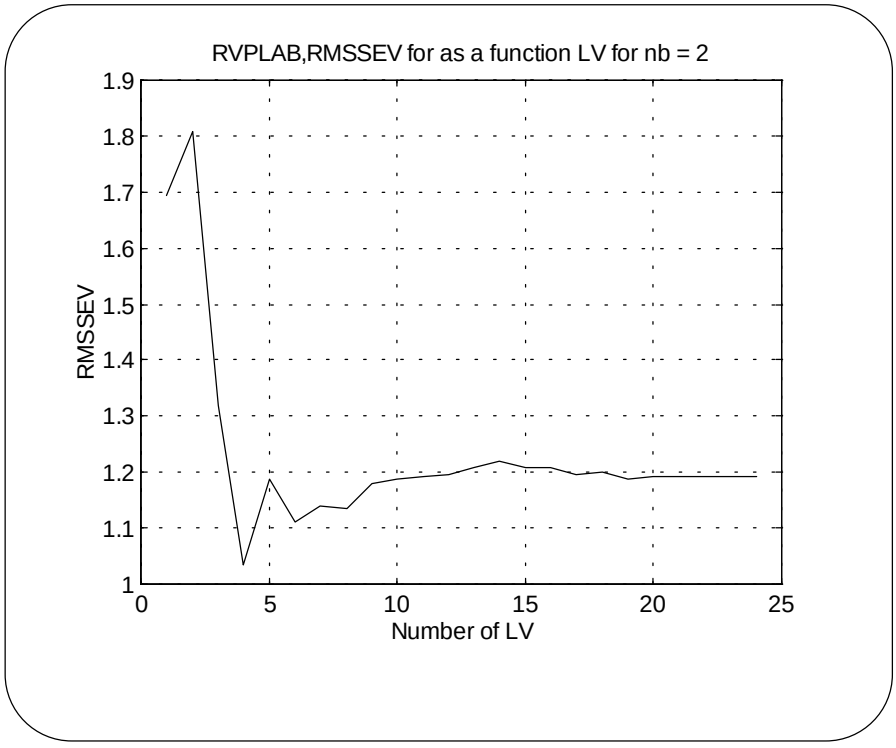

Figure B.2 : RVP Model, RMSSEV as a function of LV for nb =2.

The prediction error for this model is shown in figure B.3. Notice that the number of available RVP measurements are only 233. Figure B.4 shows a histogram plot of prediction error in calibration, which exhibit an approximate zero mean error.

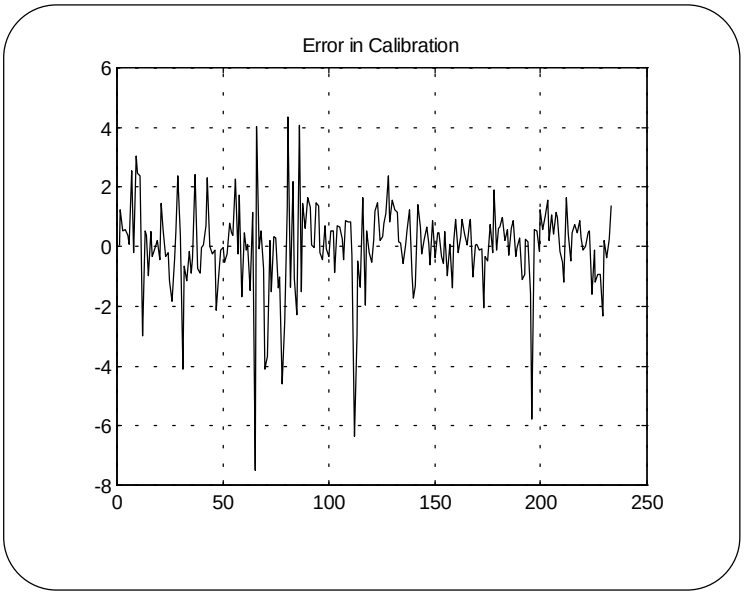

Figure B.3 : RVP Model, prediction error in calibration.

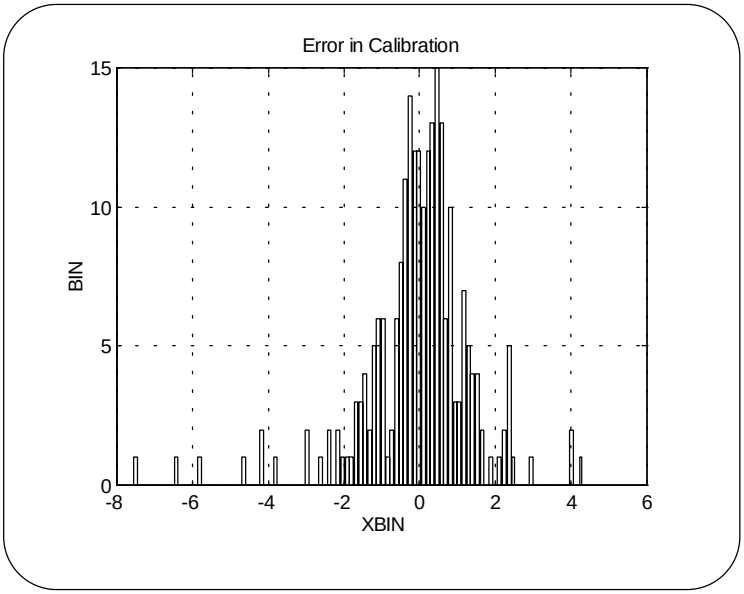

Figure B.4 : RVP Model, histogram plot of prediction error in calibration

Figure B.5 shows the open-loop simulation of the model by using calibration data set. Open-loop simulation is performed by letting the new predicted value of the output be used instead of measurement for prediction of the next output value. Figure B.6 shows the result for simulation of the model in which the actual output measurement is used for prediction of the next output. As it can be seen from figure B.5 and B.6, the model has captured the essential variation of RVP. Figure B.7 shows measured RVP at laboratory versus model predicted RVP in calibration.

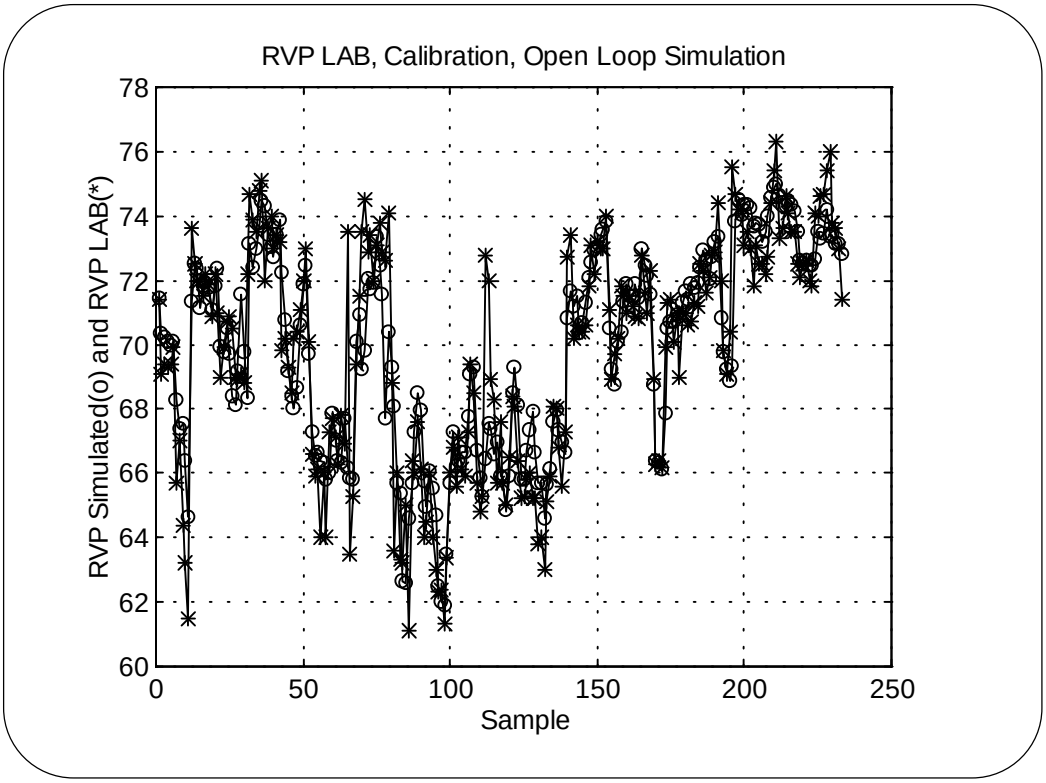

Figure B.5: RVP Model, open loop simulation in calibration

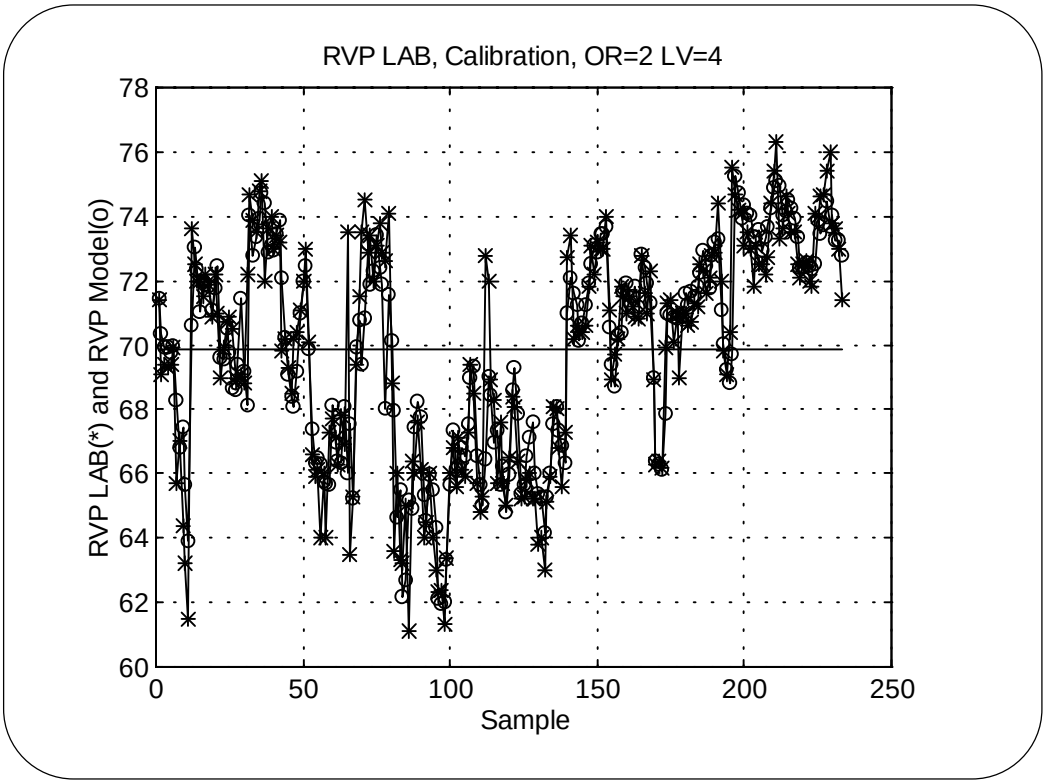

Figure B.6: RVP Model, prediction ability in calibration

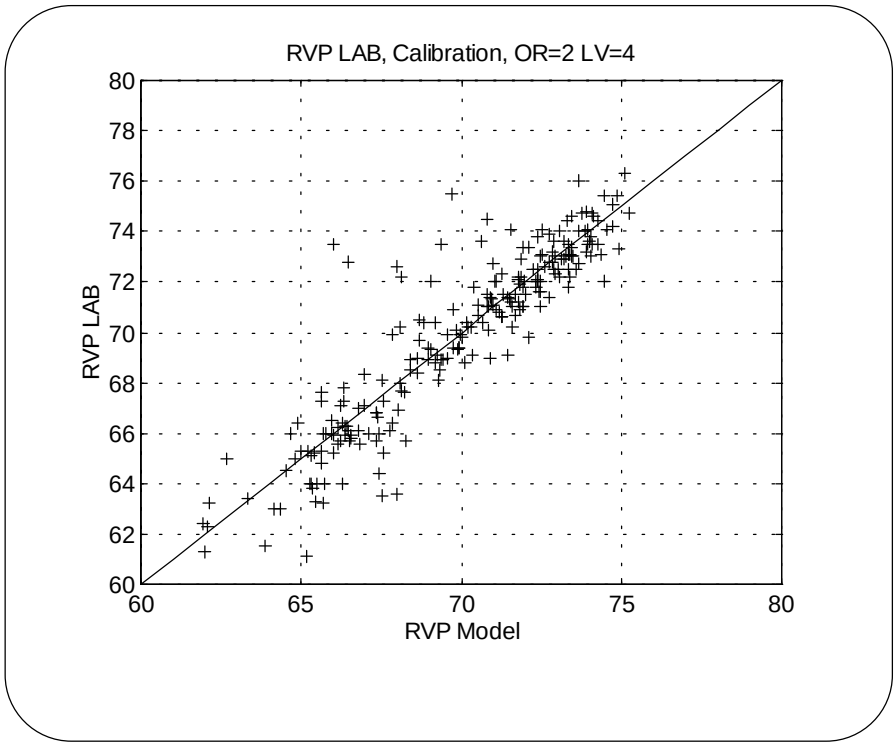

Figure B.7 : Measured RVP vs. model predicted RVP in calibration.

Selection of the best nb and LV is based on performance of the obtained model in validation. The important issue is to capture the maximum effect of input variables on prediction of output and obtain a model with minimum prediction error. The issues in validation of the selected model is discussed in the following subsection.

2.4 Validation

The validation is performed applying a completely distinct set of data. As mentioned before the input data in validation set consists of 3184 data set covering 6 months operation. In this period, after omitting the outliers, there are only 53 laboratory measurements of output RVP is remained.

Validation		Calibration	
RMSSEV	1.036	RMSSEC	1.452
RMSEAVGV	1.920	RMSEAVGC	3.510
RMSEZROV	1.708	RMSEZROC	2.025

Table B.2: RMSSE, average-model, and zero-model in validation and calibration.

The RMSSE in validation, the average-model RMSEAVGV and the zero-model RMSEZROV are shown in table B.2. It can be seen that the RMSSEV is less than average- and zero-model, indicating that the model has captured the essential variation both in input and output.

The prediction error in validation is shown in figure B.8. Figure B.9 shows a histogram plot of prediction error in validation.

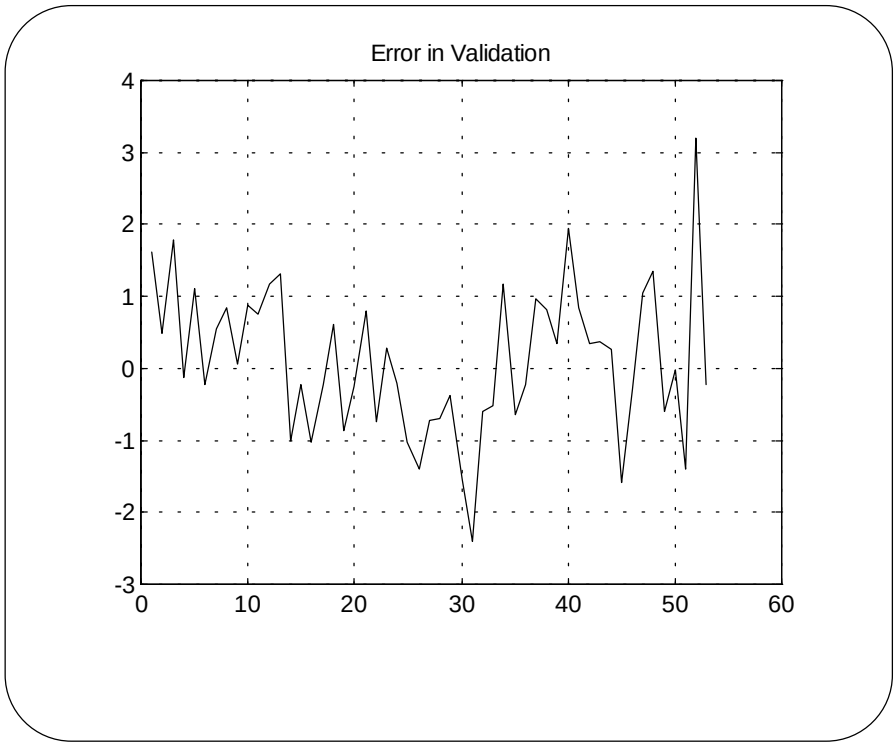

Figure B.8: RVP Model, prediction error in validation.

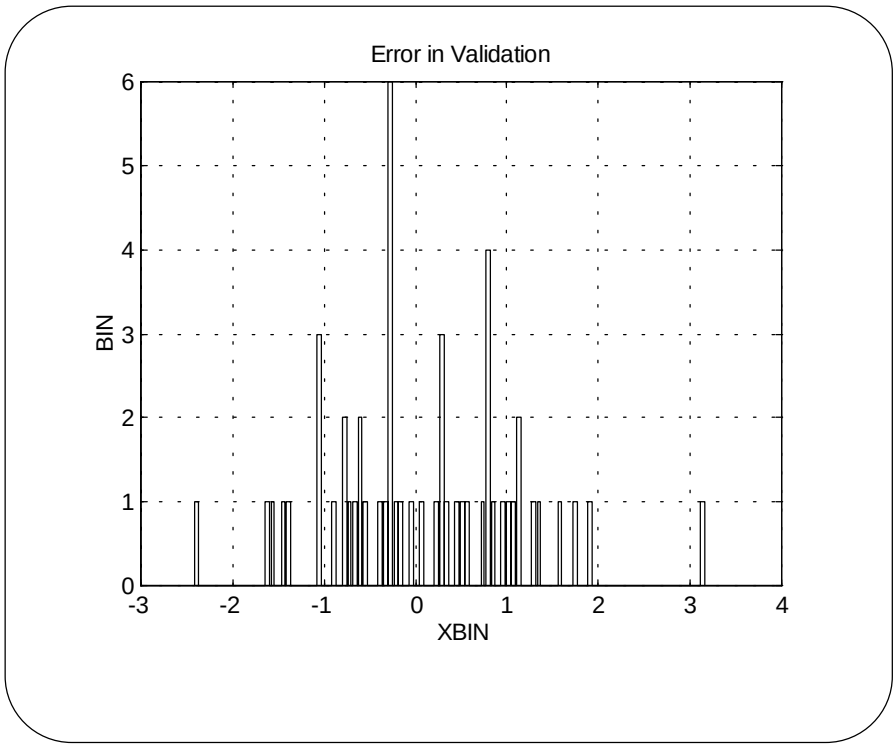

Figure B.9: RVP Model, histogram plot of prediction error in validation.

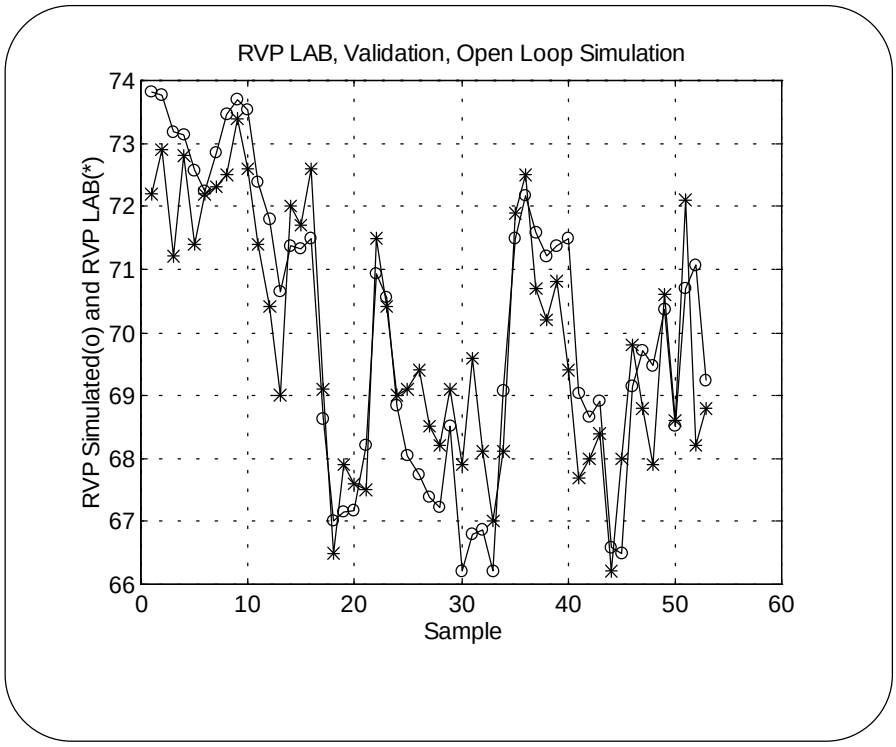

Figure B.10: RVP Model, Open Loop Simulation in validation

Figure B.10 shows the open-loop simulation in the validation, in which the predicted value of the output is used to predict the next output value. Open-loop simulation shows the predictability of the model during a period of operation without having the actual output measurement.
Figure B.11 shows the result of the simulation when the actual measurements are used. The result in figure B.11 can be better expressed in figure B.12, which shows measured versus model predicted RVP in validation.

As we can see the model has captured the essential variation in the data and the prediction ability is satisfactory.

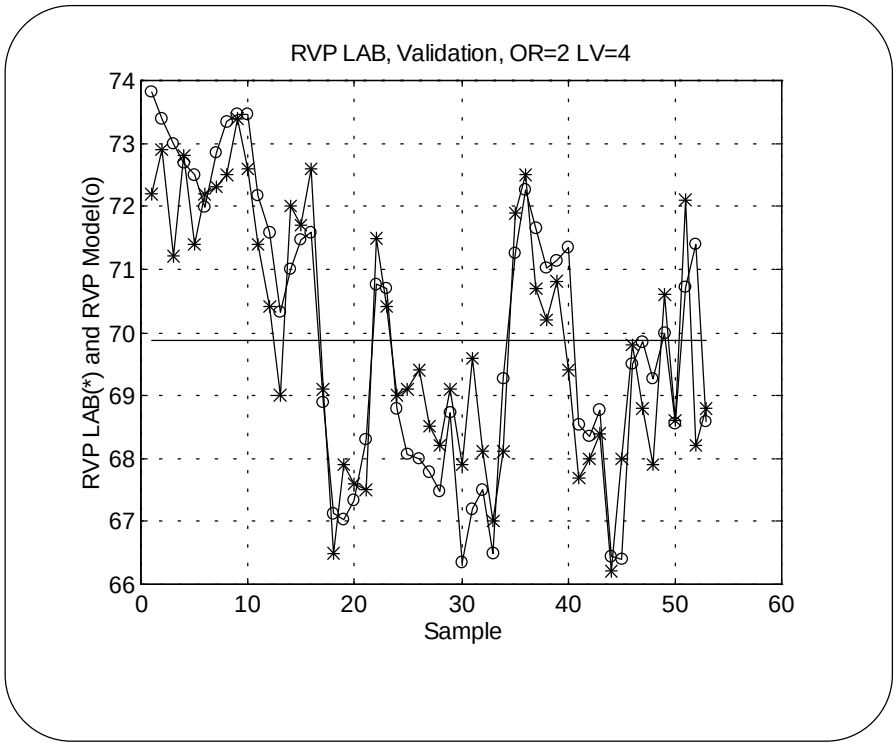

Figure B.11 : RVP Model, prediction ability in calibration

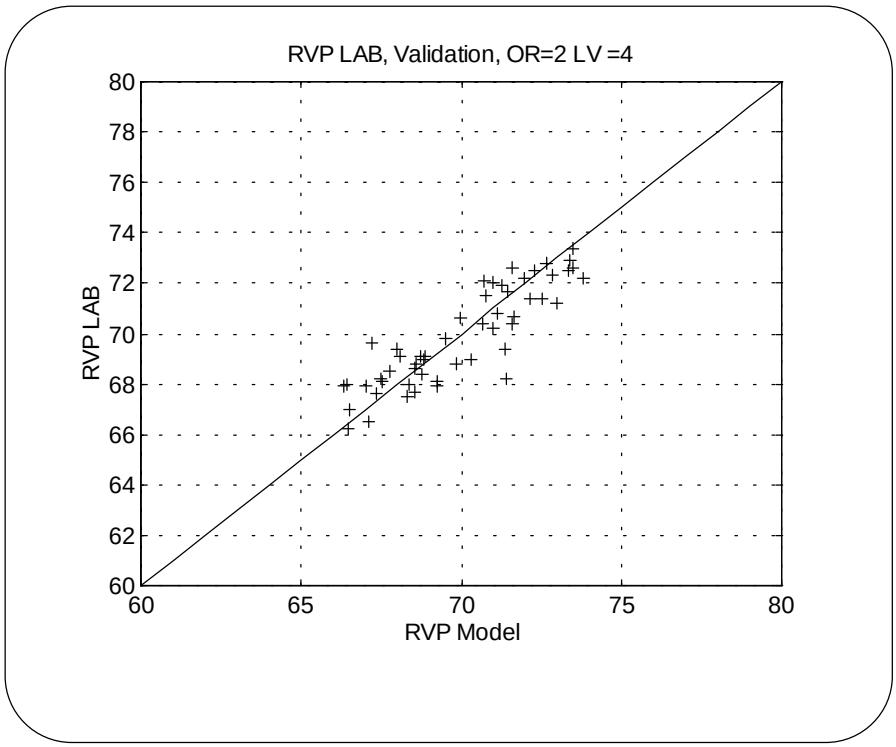

Figure B.12: RVP Model, prediction ability in calibration

3 RON Model

3.1 Introduction

In this section the model for prediction of Research Octane Number (RON) for isomerate product from isomerization unit will be presented.

3.2 Inputs an Output

The following input variables are used in the RON model.

1 Reactor Inlet temperature ^{0}C
2 Reactor A outlet temperature ^{0}C
3 Reactor B outlet temperature ^{0}C
4 Liquid Hourly Space Velocity (LHSV) 1/hr
5 H2 Consumption Sm^3/hr
6 Deisopentanizer (DIP) tray 8 temperature ^{0}C
7 DIP Bottom Flow Rate m^3/hr
8 DIP Feed Flow Rate
9 DIP Reflux Flow Rate
10 CP4703 top temperature (Pressure Corrected)
11 C-4703 Reflux Flow rate
12 C-4703 Feed Flow Rate

The total number of input data are 5699, and 4266 in calibration and validation data set respectively. These number of data set corresponds to 9 months operation data in calibration and 6 months operation data in validation. Notice that the data corresponding to the periods of process shutdown and outliers has been omitted. Regarding the output RON, there are only 238 and 100 laboratory measurements available in calibration and validation periods respectively.

3.3 Model Structure

Effective feedback control, based on manipulating the temperature of the reactors, has caused small variation in the RON quality, as it is shown in table B.3. Notice that this is the calibration data set that cover 9 months of operation.
It has been found that the type of model structure used in the case of RVP model is not suitable for prediction of RON. As it is shown in chapter 4, there is more variation in the RVP. This is particularly due to the control strategy in this unit, which is based on effective control of RON quality.

As it is discussed in chapter 5, linear interpolation is performed in order to estimate the missing RON values, which is consequently based on the assumption that the variation of RON from one day to another is small enough to permit a rough estimation of RON between two subsequent existing RON measurement.

Since we are applying interpolation in order to estimate the missing RON output, we will have an equal number of observations in both input and output data set. The number of A-parameters, and B-parameters are then determined by model order, and we will have the same number of na, and nb. The model structure is based on an ARX model, in which PLS is used for parameter estimation.

Calibration		Temperature		
	RON	Reactor Inlet	Reactor A Outlet	Reactor B Outlet
Average	87.29	142.96	189.10	159.80
Std. Deviation	0.47	2.21	2.28	2.52
Maximum	88.60	147.30	193.81	164.20
Minimum	85.70	100.27	132.52	112.18

Table B.3 : RON and reactor outlet temperatures in calibration data set.

It is important to emphasize that the interpolation is performed only in calibration data set. In the validation, we let the model apply its own predicted output, in order to predict the next output.

3.4 Calibration

Optimum number of model order nb, Latent Variable LV, and a set of suitable delay parameters k has been found by numerous recursive simulations.
The following delay parameters has been found for the input variables:

$$k = [\ 5 \quad 4 \quad 4 \quad 7 \quad 6 \quad 2 \quad 3 \quad 10 \quad 9 \quad 9 \quad 10 \quad 11\ ]$$

There is no delay for output RON.

Table B.4 shows the values of RMSSV obtained for the different nb and LV, in which the model order is changed from 1 to 25 and LV is changed from 1 to 25.

It can be seen that there is only one local minimum that appear already by a second order ARX model and LV=5. It can also be seen in figure B.13, which shows a plot of RMSSEV versus both LV and nb, and in figure B.14, which shows RMSSEV as a function of LV for nb=2.

It can be seen that the value of RMSSEV is less in the case of nb=1. However, the case with nb=2 is preferable since the captured variance in both inputs (X-block) and output (Y-block) are higher.

Hence, the model structure with nb=2, and LV=5 is chosen.

ARX Order	Min. RMSSEV	X-Block	Y-Block	LV
1	0.239	77.53	67.55	3
2	0.244	83.57	70.52	5
3	0.260	75.56	67.33	3
4	0.269	75.65	66.96	3
5	0.274	75.44	66.55	3
6	0.280	88.96	72.34	7
7	0.287	88.64	73.23	7
8	0.294	87.99	73.77	7
9	0.294	87.24	74.64	7
10	0.295	86.93	74.81	7
11	0.290	86.60	75.31	7
12	0.297	86.27	75.60	7
13	0.303	85.20	75.79	7
14	0.303	83.16	75.64	7
15	0.319	83.07	76.23	7
16	0.310	82.96	76.23	7
17	0.308	82.85	76.27	7
18	0.304	82.82	76.13	7
19	0.308	82.86	75.98	7
20	0.315	82.84	75.96	7

Table B.4 : RON model, min. RMSSEV for different value of ARX order and LV.

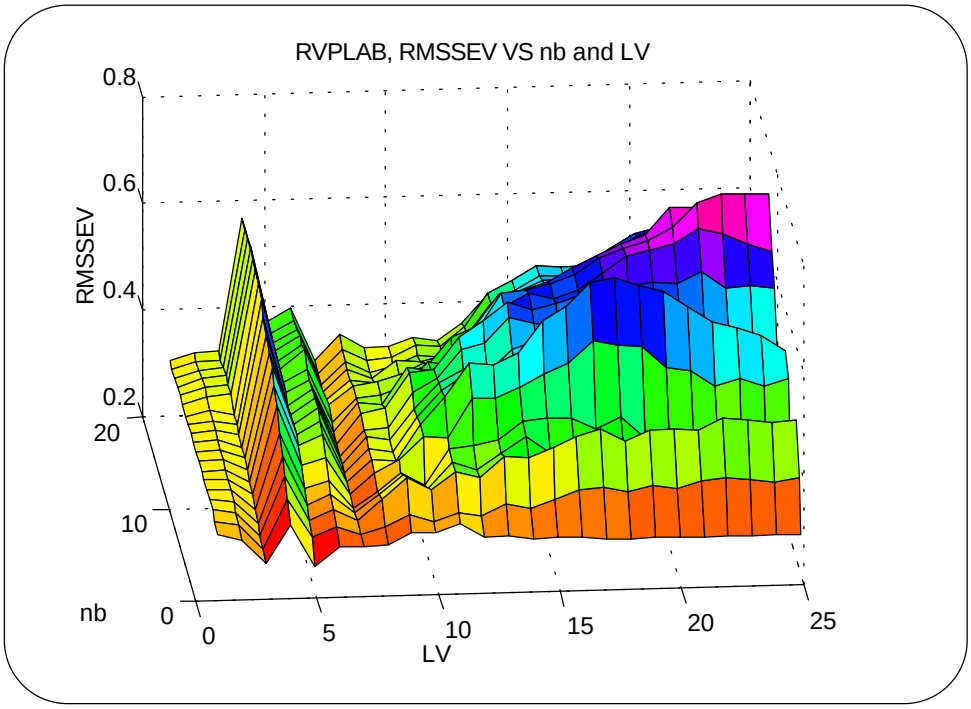

Figure B.13 : RMSSEV as a function of nb and LV.

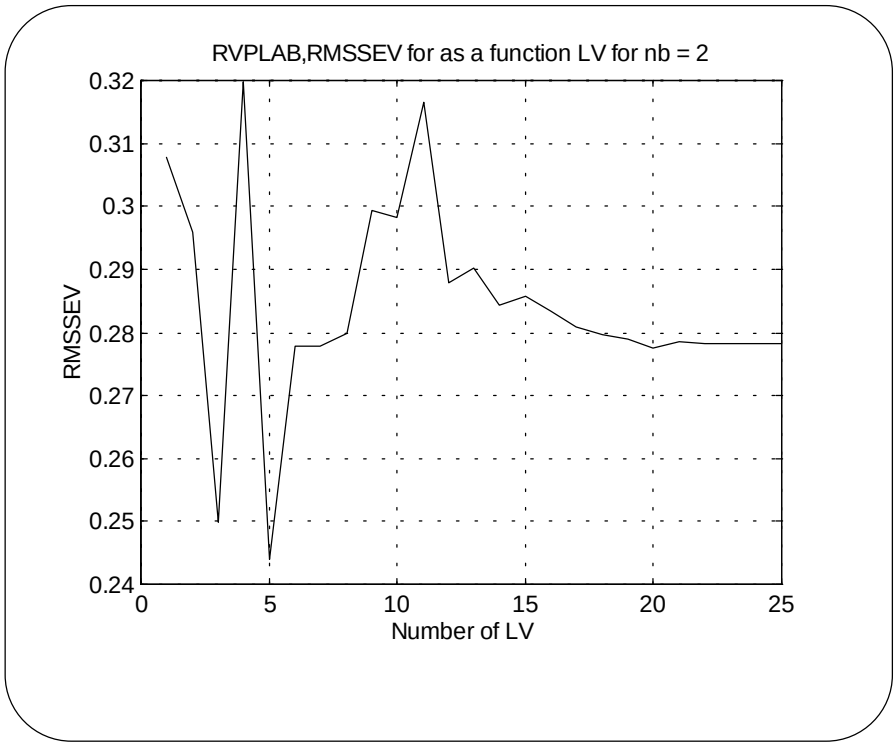

Figure B.14: RON model, RMSSEV as a function of LV for nb=2.

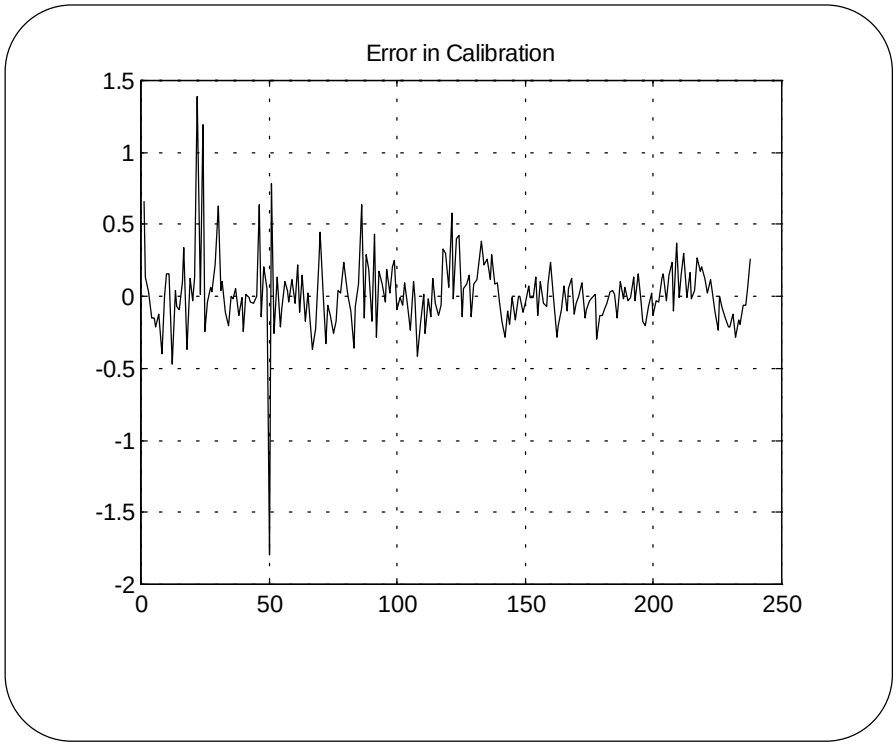

Figure B.15: RON model, prediction error in calibration.

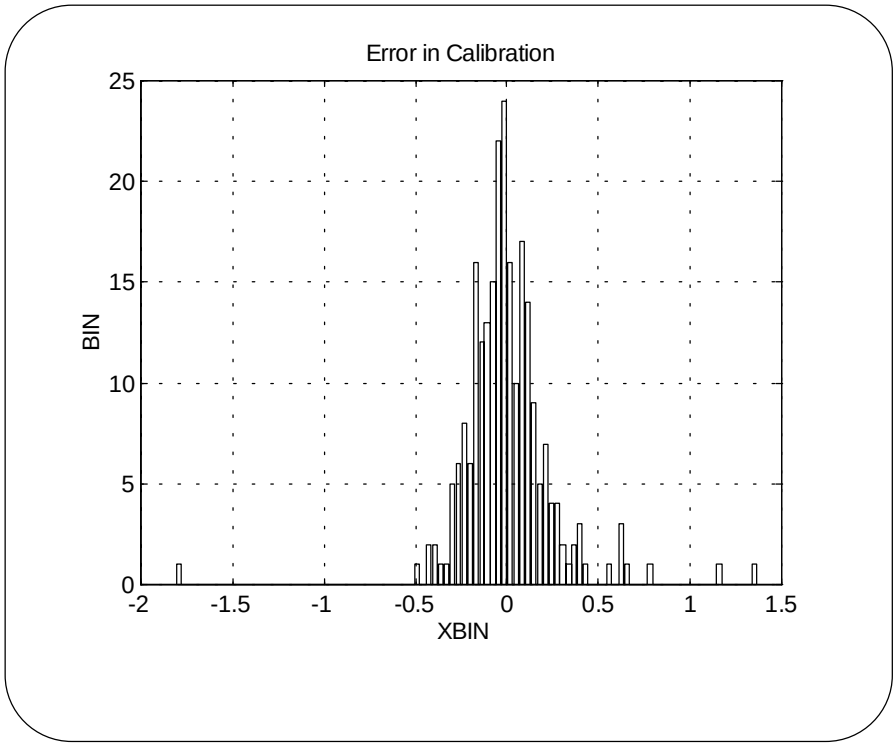

Figure B.16: RON model, histogram plot for prediction error in calibration.

The prediction error for this model is shown in figure B.15. The number of available RON measurements are only 238. Figure B.16 shows a histogram plot of prediction error in calibration, which exhibit an approximate zero mean error.

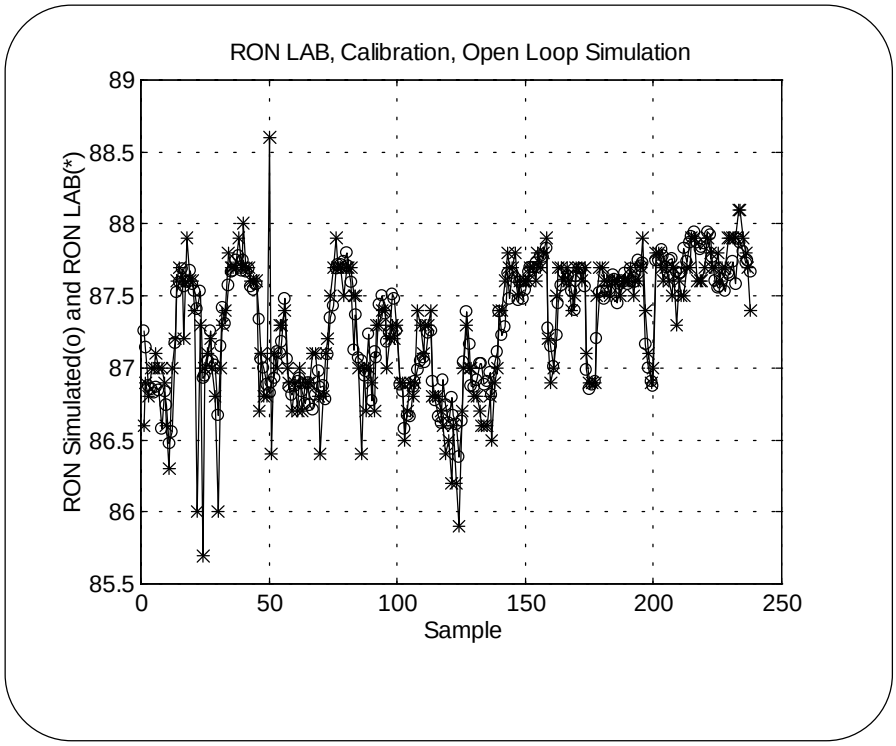

Figure B.17: RON model, open loop simulation in calibration.

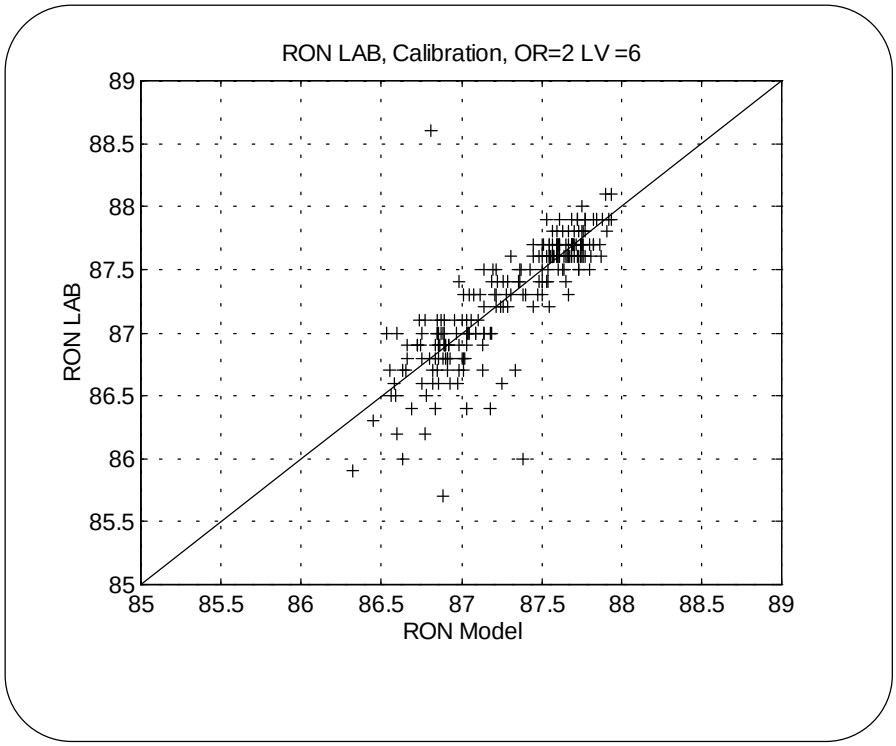

Figure B.18 Measured versus predicted RON in calibration.

Figure B.17 shows the open-loop simulation of the model by using calibration data set. Open-loop simulation is performed by letting the new predicted value of the output be used instead of measurement for prediction of the next output value. Figure B.18 shows measured versus model predicted RON in calibration.
As it can be seen from figure B.17 and B.18, the model has captured the essential variation in the data.

3.5 Validation

The validation is performed applying a completely distinct set of data. As mentioned before the input data in validation set consists of 4266 data set covering 6 months operation. In this period, after omitting the outliers, there are only 100 laboratory measurements of output RON is remained.

Validation		Calibration	
RMSSEV	0.244	RMSSEC	0.255
RMSEAVGV	0.324	RMSEAVGC	0.470
RMSEZROV	0.340	RMSEZROC	0.370

Table B.5 : RMSSE, average-model, and zero-model in validation and calibration.

The RMSSE in validation, the average-model RMSEAVGV and the zero-model RMSEZROV are shown in table B.5. It can be seen that the RMSSEV is less than average- and zero-model, indicating that the model has captured the essential variation both in input and output.

The prediction error in validation is shown in figure B.19. Figure B.20 shows a histogram plot of prediction error in validation.

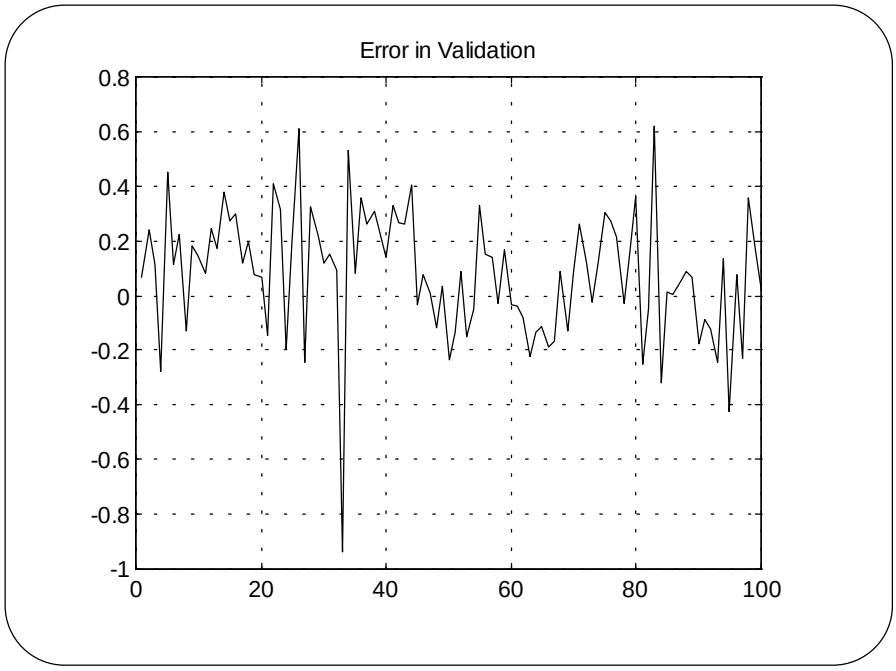

Figure B.19: RON model, prediction error in validation.

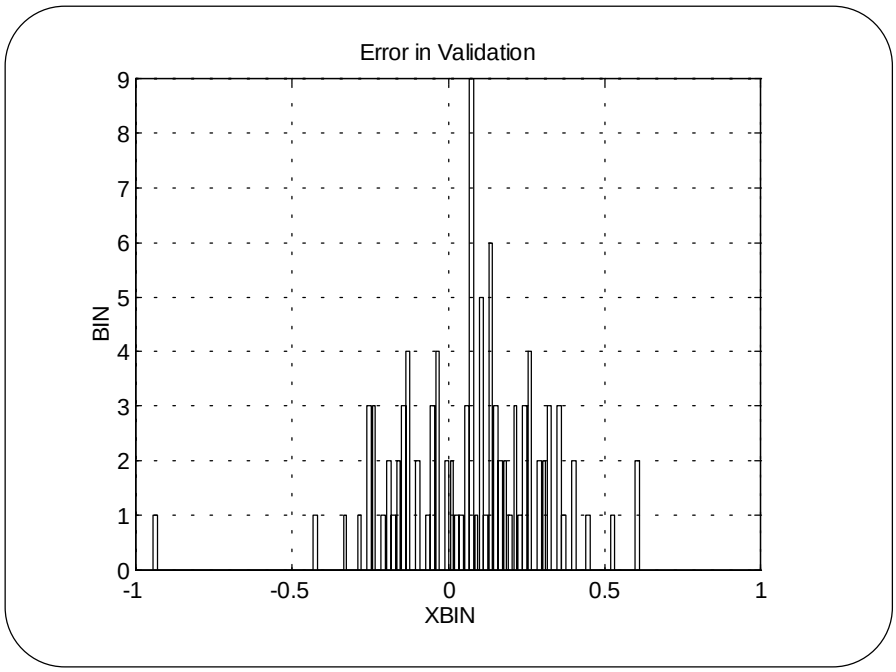

Figure B.20: RON model, histogram plot for prediction error in validation.

Figure B.21 shows the open-loop simulation in the validation, in which the predicted value of the output is used to predict the next output value. Open-loop simulation shows the predictability of the model during a period of operation without having the actual output measurement.

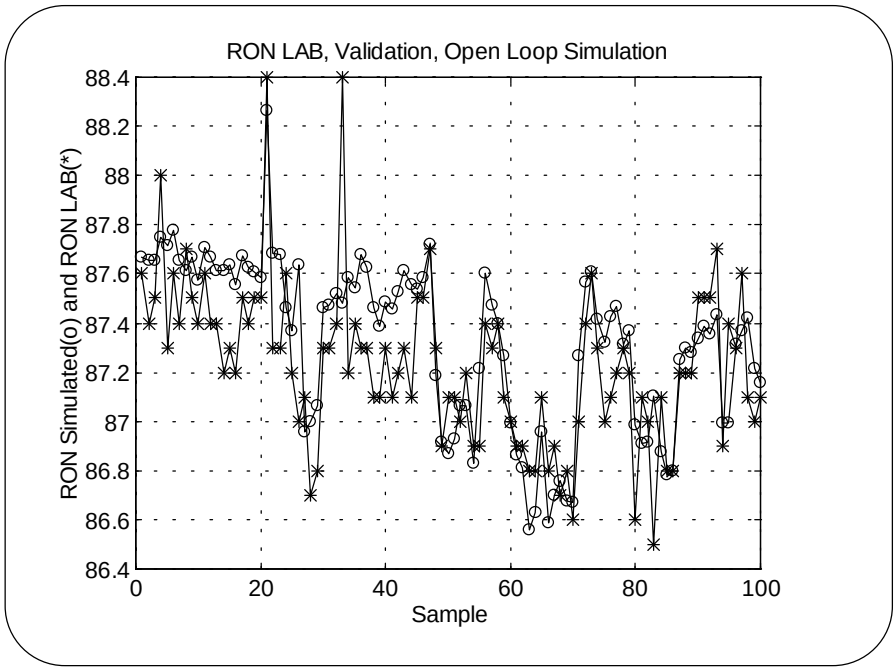

Figure B.21: RON model, open loop simulation in validation.

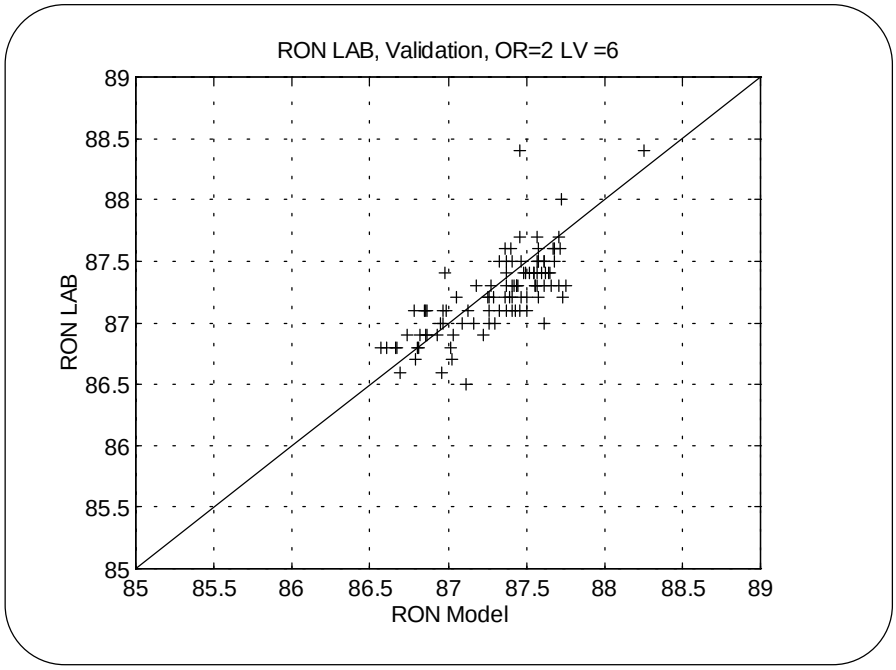

Figure B.22: Measured versus predicted RON in validation.

The result in figure B.21 can be better expressed in figure B.22, which shows measured versus model predicted RON in validation.
As we can see the model has captured the essential variation in the data and the prediction ability is satisfactory.

Appendix C

Multiple Period Blending Results

1 Product Order Number 1

Order Number 1: D92

Final Product		Volume	RON	Benzene	RVP
D92		2400	92	1.49	95
Components	Time Period	Volume m^3			
MTBE	0	0	0	0	0
Butane	0	237.93	93	0	460
Import	0	0	0	0	0
LVN	0	428.57	72.81	0	77
IC5	0	16	89	0	150
Isomerate (42)	0	280	89	1	70
Isomerate (23)	0	280	89	1	70
Reformate II	0	601.50	100.57	5	35
Reformate I+II	0	340	101	0	35
Reformate I	0	200	100	0	45
LVBN	0	16	86	0	125

Qualities and the volume of the final product and the blend components

2 Product Order Number 2

Order Number 2: D95

Final Product		Volume	RON	Benzene	RVP
D95		4800	95	0,61	95
Component	Time Period	Volumn m^3			
MTBE	2	0	0	0	0
MTBE	3	0	0	0	0
Butane	2	263,94	93	0	460
Butane	3	239,59	93	0	460
Import	2	0	0	0	0
Import	3	0	0	0	0
LVN	2	240	75	0	77
LVN	3	13,08	75	0	77
IC5	2	32	89	0	150
IC5	3	16	89	0	150
Isomerate (42)	2	453,88	89	1	70
Isomerate (42)	3	228,05	89	1	70
Isomerate (23)	2	0	0	0	0
Isomerate (23)	3	676,04	89	1	70
Reformate II	2	680	101	2.32	35
Reformate II	3	340	101	0	35
Reformate I+II	2	512,3	101	0	35
Reformate I+II	3	457,11	101	0	35
Reformate I	2	217,87	100	0	45
Reformate I	3	382,13	100	0	45
LVBN	2	0	0	0	0
LVBN	3	48	86	0	125

3 Product Order Number 3

Order Number 3: D98

Final Product		Volume	RON	Benzene	RVP
D98		2400	98	0,16	95
Component	Time Period	Volumn m^3			
MTBE	7	0	0	0	0
Butane	7	306,22	93	0	460
Import	7	0	0	0	0
LVN	7	0	0	0	0
IC5	7	0	0	0	0
Isomerate (42)	7	0	0	0	0
Isomerate (23)	7	395,85	89	1	70
Reformate II	7	0	0	0	0
Reformate I+II	7	1697,93	101	0	35
Reformate I	7	0	0	0	0
LVBN	7	0	0	0	0

4 Product Order Number 4

Order Number 4: S95

Final Product		Volume	RON	Benzene	RVP
S95		5400	95	0	95
Component	Time Period	Volume (m^3)			
MTBE	12	0	0	0	0
MTBE	13	0	0	0	0
MTBE	14	0	0	0	0
Butane	12	193,03	93	0	460
Butane	13	175,01	93	0	460
Butane	14	153,87	93	0	460
Import	12	0	0	0	0
Import	13	0	0	0	0
Import	14	0	0	0	0
LVN	12	301,82	75	0	77
LVN	13	11,99	75	0	77
LVN	14	239,48	75	0	77
IC5	12	0	0	0	0
IC5	13	0	0	0	0
IC5	14	176	89	0	150
Isomerate (42)	12	0	0	0	0
Isomerate (42)	13	0	0	0	0
Isomerate (42)	14	0	0	0	0
Isomerate (23)	12	9,38	89	0	70
Isomerate (23)	13	679,57	89	0	70
Isomerate (23)	14	0	0	0	0
Reformate II	12	0	0	0	0
Reformate II	13	0	0	0	0
Reformate II	14	0	0	0	0
Reformate I+II	12	0	0	0	0
Reformate I+II	13	0	0	0	0
Reformate I+II	14	0	0	0	0
Reformate I	12	1295,76	100	0	45
Reformate I	13	933,44	100	0	45
Reformate I	14	1230,65	100	0	45
LVBN	12	0	0	0	0
LVBN	13	0	0	0	0
LVBN	14	0	0	0	0

5 Product Order Number 5

Order Number 5: S98

Final Product		Volume	RON	Benzene	RVP
S98		3000	98	1,58	95
Component	Time Period	Volumn m^3			
MTBE	16	0	0	0	0
MTBE	17	0	0	0	0
Butane	16	191,39	93	0	460
Butane	17	191,39	93	0	460
Import	16	0	0	0	0
Import	17	0	0	0	0
LVN	16	0	0	0	0
LVN	17	0	0	0	0
IC5	16	0	0	0	0
IC5	17	0	0	0	0
Isomerate (42)	16	0	0	0	0
Isomerate (42)	17	0	0	0	0
Isomerate (23)	16	247,41	89	1	70
Isomerate (23)	17	247,41	89	1	70
Reformate II	16	0	0	0	0
Reformate II	17	0	0	0	0
Reformate I+II	16	1061,2	101	4.01	35
Reformate I+II	17	1061,2	101	0	35
Reformate I	16	0	0	0	0
Reformate I	17	0	0	0	0
LVBN	16	0	0	0	0
LVBN	17	0	0	0	0

6 Product Order Number 6

Order Number 6: G91

Final Product		Volume	RON	Benzene	RVP
G91		5400	91	0,25	90
Component	Time Period	Volume (m^3)			
MTBE	18	0	0	0	0
MTBE	19	0	0	0	0
MTBE	20	0	0	0	0
Butane	18	134,13	93	0	460
Butane	19	159,07	93	0	460
Butane	20	134,13	93	0	460
Import	18	0	0	0	0
Import	19	0	0	0	0
Import	20	0	0	0	0
LVN	18	376,7	75	0	77
LVN	19	514,13	75	0	77
LVN	20	376,7	75	0	77
IC5	18	0	0	0	0
IC5	19	0	0	0	0
IC5	20	0	0	0	0
Isomerate (42)	18	0	0	0	0
Isomerate (42)	19	280	89	1	70
Isomerate (42)	20	531,23	89	1	70
Isomerate (23)	18	531,23	89	1	70
Isomerate (23)	19	0	0	0	0
Isomerate (23)	20	0	0	0	0
Reformate II	18	0	0	0	0
Reformate II	19	0	0	0	0
Reformate II	20	0	0	0	0
Reformate I+II	18	0	0	0	0
Reformate I+II	19	846,8	101	0	35
Reformate I+II	20	0	0	0	0
Reformate I	18	757,93	100	0	45
Reformate I	19	0	0	0	0
Reformate I	20	757,93	100	0	45
LVBN	18	0	0	0	0
LVBN	19	0	0	0	0
LVBN	20	0	0	0	0

7 Product Order Number 7

Order Number 7: D92

Final Product		Volume	RON	Benzene	RVP
D92		2400	92	0,3	95
Component	Time Period	Volumn m^3			
MTBE	23	0	0	0	0
Butane	23	215,7	93	0	460
Import	23	0	0	0	0
LVN	23	389,59	75	0	77
IC5	23	0	0	0	0
Isomerate (42)	23	89,57	89	1	70
Isomerate (23)	23	633,82	89	1	70
Reformate II	23	0	0	0	0
Reformate I+II	23	6,79	101	0	35
Reformate I	23	1064,54	100	0	45
LVBN	23	0	0	0	0

8 Product Order Number 8

Order Number 8 : G95

Final Product		Volume	RON	Benzene	RVP
G95		3000	95	0,27	90
Component	Time Period	Volumn m^3			
MTBE	24	0	0	0	0
MTBE	25	0	0	0	0
Butane	24	142,28	93	0	460
Butane	25	131,72	93	0	460
Import	24	0	0	0	0
Import	25	0	0	0	0
LVN	24	26,56	75	0	77
LVN	25	23,33	75	0	77
IC5	24	0	0	0	0
IC5	25	51,63	89	0	150
Isomerate (42)	24	348,97	89	0	70
Isomerate (42)	25	560	89	1	70
Isomerate (23)	24	248,63	89	1	70
Isomerate (23)	25	0	0	0	0
Reformate II	24	0	0	0	0
Reformate II	25	0	0	0	0
Reformate I+II	24	733,56	101	0	35
Reformate I+II	25	733,32	101	0	35
Reformate I	24	0	0	0	0
Reformate I	25	0	0	0	0
LVBN	24	0	0	0	0
LVBN	25	0	0	0	0

9 Product Order Number 9

Order Number 9 : D98

Final Product		Volume	RON	Benzene	RVP
D98		2400	98	2	95
Component	Time Period	Volumn m^3			
MTBE	28	0	0	0	0
Butane	28	290,65	93	0	460
Import	28	0	0	0	0
LVN	28	0	0	0	0
IC5	28	0	0	0	0
Isomerate (42)	28	332,63	89	1	70
Isomerate (23)	28	0	0	0	0
Reformate II	28	893,47	101	5	35
Reformate I+II	28	0	0	0	0
Reformate I	28	883,25	100	0	45
LVBN	28	0	0	0	0

10 Product Order Number 10

Order Number 10 : D95

Final product		Volume	RON	Benzene	RVP
D95		4800	95	1,18	95
Component	Time Period	Volume m^3			
MTBE	32	0	0	0	0
MTBE	33	0	0	0	0
Butane	32	258,18	93	0	460
Butane	33	233,13	93	0	460
Import	32	0	0	0	0
Import	33	0	0	0	0
LVN	32	50,34	75	0	77
LVN	33	26,91	75	0	77
IC5	32	0	0	0	0
IC5	33	115,87	89	0	150
Isomerate (42)	32	156,4	89	1	70
Isomerate (42)	33	870,4	89	1	70
Isomerate (23)	32	762,41	89	0.37	70
Isomerate (23)	33	0	0	0	0
Reformate II	32	0	0	0	0
Reformate II	33	1076,71	101	0	35
Reformate I+II	32	1172,67	101	3.72	35
Reformate I+II	33	76,98	101	0	35
Reformate I	32	0	0	0	0
Reformate I	33	0	0	0	0
LVBN	32	0	0	0	0
LVBN	33	0	0	0	0

11 Product Order Number 11

Order Number 11 : G91

Final Product		Volume	RON	Benzene	RVP
G91		5400	91	0,21	90
Component	Time Period	Volume (m^3)			
MTBE	36	0	0	0	0
MTBE	37	0	0	0	0
MTBE	38	0	0	0	0
Butane	36	134,13	93	0	460
Butane	37	147,47	93	0	460
Butane	38	147,47	93	0	460
Import	36	0	0	0	0
Import	37	0	0	0	0
Import	38	0	0	0	0
LVN	36	376,7	75	0	77
LVN	37	376,39	75	0	77
LVN	38	376,39	75	0	77
IC5	36	0	0	0	0
IC5	37	0	0	0	0
IC5	38	0	0	0	0
Isomerate (42)	36	0	0	0	0
Isomerate (42)	37	586,17	89	1	70
Isomerate (42)	38	0	0	0	0
Isomerate (23)	36	531,23	89	1	70
Isomerate (23)	37	0	0	0	0
Isomerate (23)	38	586,17	89	0	70
Reformate II	36	0	0	0	0
Reformate II	37	689,97	101	0	35
Reformate II	38	689,97	101	0	35
Reformate I+II	36	0	0	0	0
Reformate I+II	37	0	0	0	0
Reformate I+II	38	0	0	0	0
Reformate I	36	757,93	100	0	45
Reformate I	37	0	0	0	0
Reformate I	38	0	0	0	0
LVBN	36	0	0	0	0
LVBN	37	0	0	0	0
LVBN	38	0	0	0	0

12 Product Order Number 12

Order Number 12 : S98

Final Product		Volume	RON	Benzene	RVP
S98		3000	98	0,44	95
Component	Time Period	Volume m^3			
MTBE	40	0	0	0	0
MTBE	41	0	0	0	0
Butane	40	191,39	93	0	460
Butane	41	191,39	93	0	460
Import	40	0	0	0	0
Import	41	0	0	0	0
LVN	40	0	0	0	0
LVN	41	0	0	0	0
IC5	40	0	0	0	0
IC5	41	0	0	0	0
Isomerate (42)	40	39,45	89	0	70
Isomerate (42)	41	156,4	89	1	70
Isomerate (23)	40	207,96	89	1	70
Isomerate (23)	41	91	89	1	70
Reformate II	40	172,57	101	5	35
Reformate II	41	0	0	0	0
Reformate I+II	40	888,63	101	0	35
Reformate I+II	41	1061,2	101	0	35
Reformate I	40	0	0	0	0
Reformate I	41	0	0	0	0
LVBN	40	0	0	0	0
LVBN	41	0	0	0	0

13 Product Order Number 13

Order Number 13 : S95

Final Product		Volume	RON	Benzene	RVP
S95		5400	95	1,65	95
Component	Time Period	Volume (m^3)			
MTBE	42	0	0	0	0
MTBE	43	0	0	0	0
MTBE	44	0	0	0	0
Butane	42	174,05	93	0	460
Butane	43	191,43	93	0	460
Butane	44	191,43	93	0	460
Import	42	0	0	0	0
Import	43	0	0	0	0
Import	44	0	0	0	0
LVN	42	28,47	75	0	77
LVN	43	11,61	75	0	77
LVN	44	11,61	75	0	77
IC5	42	0	0	0	0
IC5	43	0	0	0	0
IC5	44	0	0	0	0
Isomerate (42)	42	280	89	1	70
Isomerate (42)	43	747,22	89	1	70
Isomerate (42)	44	747,22	89	1	70
Isomerate (23)	42	341,88	89	1	70
Isomerate (23)	43	0	0	0	0
Isomerate (23)	44	0	0	0	0
Reformate II	42	0	0	0	0
Reformate II	43	849,74	101	5	35
Reformate II	44	513,52	101	5	35
Reformate I+II	42	0	0	0	0
Reformate I+II	43	0	0	0	0
Reformate I+II	44	336,21	101	0	35
Reformate I	42	959,23	100	0	45
Reformate I	43	0	0	0	0
Reformate I	44	0	0	0	0
LVBN	42	16,37	86	0	125
LVBN	43	0	0	0	0
LVBN	44	0	0	0	0

14 Product Order Number 14

Order Number 14 : D98

Final Product		Volume	RON	Benzene	RVP
D98		2400	98	2	95
Component	Time Period	Volumn m^3			
MTBE	47	0	0	0	0
Butane	47	306,22	93	0	460
Import	47	0	0	0	0
LVN	47	0	0	0	0
IC5	47	0	0	0	0
Isomerate (42)	47	0	0	0	0
Isomerate (23)	47	395,85	89	0	70
Reformate II	47	1697,93	101	2.83	35
Reformate I+II	47	0	0	0	0
Reformate I	47	0	0	0	0
LVBN	47	0	0	0	0

15 Product Order Number 15

Order Number 15 : D95

Final Product		Volume	RON	Benzene	RVP
D95		5400	95	2	95
Component	Time Period	Volume (m^3)			
MTBE	48	0	0	0	0
MTBE	49	0	0	0	0
MTBE	50	0	0	0	0
Butane	48	185,2	93	0	460
Butane	49	191,43	93	0	460
Butane	50	182,12	93	0	460
Import	48	0	0	0	0
Import	49	0	0	0	0
Import	50	0	0	0	0
LVN	48	0	0	0	0
LVN	49	11,61	75	0	77
LVN	50	110,77	75	0	77
IC5	48	0	0	0	0
IC5	49	0	0	0	0
IC5	50	32	89	0	150
Isomerate (42)	48	719,49	89	0.57	70
Isomerate (42)	49	747,22	89	0	70
Isomerate (42)	50	453,2	89	0	70
Isomerate (23)	48	0	0	0	0
Isomerate (23)	49	0	0	0	0
Isomerate (23)	50	0	0	0	0
Reformate II	48	0	0	0	0
Reformate II	49	849,74	101	4.24	35
Reformate II	50	0	0	0	0
Reformate I+II	48	849,67	101	3.76	35
Reformate I+II	49	0	0	0	0
Reformate I+II	50	381,35	101	5	35
Reformate I	48	0	0	0	0
Reformate I	49	0	0	0	0
Reformate I	50	640,56	100	2.64	45
LVBN	48	45,63	86	0	125
LVBN	49	0	0	0	0
LVBN	50	0	0	0	0

16 Product Order Number 16

Order Number 16 : D98

Final Product		Volume	RON	Benzene	RVP
D98		2400	98	2	95
Component	Time Period	Volumn m^3			
MTBE	52	0	0	0	0
Butane	52	246,6	93	0	460
Import	52	0	0	0	0
LVN	52	0	0	0	0
IC5	52	80	89	0	150
Isomerate (42)	52	58,8	89	1	70
Isomerate (23)	52	0	0	0	0
Reformate II	52	0	0	0	0
Reformate I+II	52	245,01	101	0	35
Reformate I	52	1641,59	100	2.89	45
LVBN	52	128	86	0	125

17 Product Order Number 17

Order Number 17 : G95

Final Product		Volume	RON	Benzene	RVP
G95		4800	95	1,14	90
Component	Time Period	Volume (m^3)			
MTBE	57	0	0	0	0
MTBE	58	0	0	0	0
MTBE	59	0	0	0	0
Butane	57	133,2	93	0	460
Butane	58	138,04	93	0	460
Butane	59	139,11	93	0	460
Import	57	0	0	0	0
Import	58	0	0	0	0
Import	59	0	0	0	0
LVN	57	76,61	75	0	77
LVN	58	51,32	75	0	77
LVN	59	120	75	0	77
IC5	57	32	89	0	150
IC5	58	0	0	0	0
IC5	59	16	89	0	150
Isomerate (42)	57	280	89	1	70
Isomerate (42)	58	366,38	89	1	70
Isomerate (42)	59	193,62	89	1	70
Isomerate (23)	57	156,4	89	0	70
Isomerate (23)	58	156,4	89	0	70
Isomerate (23)	59	156,4	89	0	70
Reformate II	57	0	0	0	0
Reformate II	58	0	0	0	0
Reformate II	59	0	0	0	0
Reformate I+II	57	0	0	0	0
Reformate I+II	58	0	0	0	0
Reformate I+II	59	0	0	0	0
Reformate I	57	921,79	100	5	45
Reformate I	58	887,85	100	0	45
Reformate I	59	974,87	100	0	45
LVBN	57	0	0	0	0
LVBN	58	0	0	0	0
LVBN	59	0	0	0	0