IVL

report

IVL Swedish Environmental Research Institute

# Production optimisation in the petrochemical industry by hierarchical multivariate modelling

Magnus Andersson, Erik Furusjö, Åsa Jansson

B1586-B

June, 2004

IVL

**Organisation/Organization**

IVL Svenska Miljöinstitutet AB
IVL Swedish Environmental Research Institute Ltd.

**Adress/address**

Box 210 60
SE-100 31 Stockholm

**Telefonnr/Telephone**

+46 8 598 563 00

**RAPPORTSAMMANFATTNING**
**Report Summary**

**Projekttitel/Project title**

**Petrochemical process integration with hierarchical multivariate modelling**

**Anslagsgivare för projektet/**
**Project sponsor**

Energimyndigheten (Swedish National Energy Administration)
Nynäs Refining AB

**Rapportförfattare/author**

Magnus Andersson, Erik Furusjö, Åsa Jansson

**Rapportens titel och undertitel/Title and subtitle of the report**

Production optimisation in the petrochemical industry by hierarchical multivariate modelling

**Sammanfattning/Summary**

This project demonstrates the advantages of applying hierarchical multivariate modelling in the petrochemical industry in order to increase knowledge of the total process. The models indicate possible ways to optimise the process regarding the use of energy and raw material, which is directly linked to the environmental impact of the process. The refinery of Nynäs Refining AB (Gothenburg, Sweden) has acted as a demonstration site in this project. The models developed for the demonstration site resulted in:

- Detection of an unknown process disturbance and suggestions of possible causes.
- Indications on how to increase the yield in combination with energy savings.
- The possibility to predict product quality from on-line process measurements, making the results available at a higher frequency than customary laboratory analysis.
- Quantification of the gradually lowered efficiency of heat transfer in the furnace and increased fuel consumption as an effect of soot build-up on the furnace coils.
- Increased knowledge of the relation between production rate and the efficiency of the heat exchangers.

This report is one of two reports from the project. It contains a technical discussion of the result with some degree of detail. A shorter and more easily accessible report is also available, see IVL report B1586-A.

**Nyckelord samt ev. anknytning till geografiskt område eller näringsgren /Keywords**

Petrochemistry, petrochemical industry, crude oil distillation, process optimisation, energy conservation, hierarchical multivariate modelling, MVA, soft sensor

**Bibliografiska uppgifter/Bibliographic data**

IVL Rapport/report B1586-B

**Rapporten beställs via /The report can be ordered via**

Hemsida: www.ivl.se, e-mail: publicationservice@ivl.se, fax: 08-598 563 90 eller IVL, Box 210 60, 100 31 Stockholm.

# Summary

Thanks to methodological developments and enhanced computer capacity, there has been huge improvements in model based monitoring and optimisation technology. Increased computer capacity is also the reason why those techniques are now feasible tools for most industrial processes. The prospect is that these tools will give new optimisation opportunities even for such processes that are considered well optimised today.

This project demonstrates the advantages of applying hierarchical multivariate modelling in the petrochemical industry in order to increase knowledge of the total process. The models indicate possible ways to optimise the process regarding the use of energy and raw material, which is directly linked to the environmental impact of the process.

Three general objectives have been considered during the project.

- Reach a more effective production and thereby lower the energy demand and the environmental impact of the process.
- Obtain improved process economics through an increase in productivity and a decrease in energy and raw material consumption.
- Capture the present process knowledge of the individual operators in statistical and mathematical models, and thereby turning this knowledge into company knowledge.

The refinery of Nynäs Refining AB (Gothenburg, Sweden) has acted as a demonstration site in this project. The investigation was performed on data from the existing process, covering normal process variation. Since the basic distillation process is similar at most refineries, the general results of this project can easily be incorporated at other petrochemical sites.

Previous work in modelling of petrochemical processes has indicated that linear models are not adequate to describe the non-liner process behaviour. However, in this report it is shown that by using "smart" variable transformations, i.e. by including knowledge about process non-linearity in the data transformation strategy prior to the regression step, it is possible to use linear models to accurately describe the process and to predict the product quality.

During the course of the project, models and results have been presented to and discussed with the process operators. They have been able to verify that the models include known process variations and therefore have captured the personal knowledge of the operators. Previously unobserved variations were also discovered through interpretation of the process models.

The project has shown that:

- Steam flow to the AD and VD towers and the side strippers are not fully compensated for differences in production rate. This influences the product quality. It is suggested to add relative steam flow to operator process screens and to use this in process operation for enhanced control of the product quality.
- There are large oscillations present in the upper half of the AD tower. This leads to more inefficient process operation and higher consumption than necessary of resources and energy. The cause of the oscillations cannot be determined because of the high data compression used for some tags in the history database although three possible candidates are identified. It is recommended to increase the quality of data in the history database by decreasing compression for some process data.
- A hierarchical model of the entire process has shown large potential for increased yield of the most valuable product in combination with energy savings by optimisation of operating conditions.
- It is possible to predict product quality in the form of True Boiling Point (TBP) curves for at least four of the six distillation fractions. The models give prediction errors as low as 2.0 ºC for some of the products. For most of the products the accuracy of the model predictions is similar to the accuracy of the laboratory analysis method used today, which is run every 8 hours and gives results with approximately 4 hours delay. Since the predictions are made from the ordinary on-line measurements, they can be executed continuously and the models can act as soft sensors of the product quality. This leads to entirely new control possibilities. Key personnel at Nynäs estimates that the yield of the most valuable product could be increased by 0.5% absolute (approximately 5% relative) if the TBP soft sensors were implemented on-line, which agree well with the estimate from the hierarchical full process model that indicate 0.6% increase. The yield increase can be translated into energy savings by the same amount with respect to kg produced product. The economical benefits are also substantial; approximately 4 MSEK/year in increased income is a rough estimate by Nynäs.
- The models clearly show the gradually lowered efficiency of heat transfer to the crude oil and increased relative fuel consumption as an effect of soot build-up on the furnace coils. Chemical cleaning of the furnace has a large effect on the total relative fuel consumption, which is reduced from 13.5-14 kg per ton to approximately 11.5-12 kg per ton after cleaning. This can be used to determine when the cost of chemical cleaning can be motivated by a sufficient decrease in relative fuel consumption.
- The efficiency of the heat exchangers is significantly lower at higher production rate. Crude oil temperature differences at minimum and maximum feed rate ranges between 5 and 20°C, with larger differences at the end of the chain of heat

exchangers. This means that less temperature increase is required in the AD furnace at low production rate, which should be taken into account when optimising the process with respect to energy consumption.

In discussions with process operators and process engineers, possible benefits of putting some of the models developed in this project on-line was investigated. It was agreed that:

- Prediction of the TBP curves would give entirely new opportunities to run the process more efficiently and give on-line product quality control.
- PCA models can help monitor current process status and suggest how to steer the process into the most desirable state.

Nynäs current view is that it would be very valuable to put the TBP prediction models on-line and they see potential increases in yield that would correspond to energy savings and improved productivity. In a longer run it would also be interesting to use PCA models for process monitoring on-line but that is currently of lower priority.

The promising results obtained in this project show that it would be very interesting to make process models for other operating modes than the one studied in this project. There is at least two more production modes that are frequently used and where the effort would be worthwhile in our opinion.

## Preface

This project was accomplished within the Process Integration program of the Swedish National Energy Administration with additional sponsoring from Nynäs Refining AB. The tight collaboration with process operators at the demonstration site during the course of the project has been vital for the many interesting results achieved.

This report is one of two reports from the project. It contains a technical discussion of the result with some degree of detail. A shorter and more easily accessible report is also available, see IVL report B1586-A.

**Table of contents**

# 1 Introduction

The petrochemical industry is not commonly associated with terms like renewable energy and sustainability. Nevertheless it is fair to assume that the products of this industry will stay a commodity of our society for quite a long time. So even though the vision is that the use of non-renewable resources in time will be restricted, there is much reason to address issues like process optimisation, energy savings and reduced environmental impact of the petrochemical industry.

There is great potential for environmental improvement within the Swedish petrochemical industry. Along with the pulp, paper, metal, iron and steel industries the chemical industry is one of the largest industrial resource consumers of Sweden. 36 000 TJ of combustion energy per year is used by the Swedish petrochemical industry alone [1]. Although Swedish refineries today are well adapted regarding energy efficiency and emissions to the environment, they still have a considerable environmental impact, locally as well as globally.

Thanks to enhanced computer capacity, there has been huge improvements in model based monitoring and optimisation techniques. Increased computer capacity is also the reason why those techniques are now feasible tools for most industrial processes. The prospect is that these tools will give new optimisation opportunities even for such processes that are considered well optimised today.

Petrochemical sites are generally very large. Refining of crude oil comprises a number of process steps, e.g. fractionated distillation and mixing of fractions to obtain products with desirable qualities. This makes the overall process extremely complex, since all process steps affect each other through material and energy flows.

It is desirable to understand the effect that variations in crude oil quality, process disturbances and control parameters have on the final result, and to be able to optimise the process on account of material and energy consumption. This requires deep knowledge on how the specific process steps operate as well as their interactions with each other. Statistical modelling is an established method to give increased knowledge on processes. The fact that petrochemical industries generally are well documented by on-line instrumentation and highly automated makes it possible to extract a lot of data suitable for process modelling. Hierarchical multivariate modelling can thus be expected to give advantages for optimisation of refineries.

The refinery of Nynäs Refining AB (Gothenburg, Sweden) has acted as a demonstration site in this project. Since the basic process is similar at most refineries, the general results

of this project can easily be incorporated at other petrochemical sites. Naturally, more specific results such as the actual models will be site specific.

# 2 Objective

The intention with this project is to demonstrate the advantages of applying hierarchical multivariate modelling in order to increase knowledge of the total process in the demonstration site. The models should also indicate possible ways to optimise the process regarding the use of energy and raw material as well as the environmental impact of the process.

Three general objectives have been considered during the project.

- Reach a more effective production and thereby lower the energy demand and the environmental impact of the process.
- Obtain improved process economics through an increase in productivity and a decrease in energy and raw material consumption.
- Capture the present process knowledge of the individual operators in statistical and mathematical models, turning this knowledge into company knowledge.

It is also expected that unwanted variation in the product quality will be reduced as a result of the increased understanding of the process and the enhanced monitoring possibilities given by the models.

# 3 Scope and data

The scope of the study is to investigate data from the existing process, capturing current process variation in models and interpreting the effect it has on the product quality. Hence, retro fitting of the process is not considered.

The study is based on real process data from the demonstration site. Historical data from the on-line process documentation of one particular operational mode, from the period of June 2002 to June 2003, was used. Data close to a change in modes has been omitted from the study to make sure that transient effects of the change have disappeared. This corresponds to 110 days of data, in coherent periods of 2.5 days up to 11.5 days. The sampling frequency used here is 1/15 minutes and the sampled data corresponds to an average value over 15 minutes (from data sampled much more frequently, every 10 s for most variables). Part of the study was also conducted on data sampled with a higher

frequency to enable investigation of very short-term variation, see results under 6.2 below.

Data from laboratory analyses of the true boiling point (TBP) of four fractions from the distillation towers (see 5.6 below) have also been included in the study. GC data is available at a much lower frequency, approximately every 8 hours, so for models that include it the on-line process data has been selected in order to match the analysis data. Still, the process data value is an average over 15 minutes.

# 4 Methods

This section contains a description of the modelling methods used in this work. First, a brief description of standard multivariate statistical modelling methods is given, which is followed by a description of the multi-block versions of the algorithms. Finally, the problem with missing values in the data used for modelling is discussed and some published applications of multivariate statistical modelling to petrochemical distillation processes are presented.

## 4.1 Multivariate statistical methods for process modelling

Process modelling by multivariate statistical modelling methods, such as Principal Component Analysis (PCA) [2,3] and Partial Least Squares regression (PLS) [3, 4] and modifications thereof, are increasingly used and accepted in industry. This can be explained by their ability to handle the large amounts of process data generated in well-instrumented modern process industries and to extract relevant information from the data. There is a wide range of methods and applications of multivariate statistical modelling methods in process monitoring and optimisation, see [5,6,7] for an overview.

Typical requirements for models for process optimisation, process monitoring, fault detection and fault identification includes sensitivity to process state and deviations from normal operating conditions as well as easy model interpretation to detect the causes of deviations. PCA and PLS are powerful in both these respects compared to many other types of models. The ability to handle a large amount of collinear variables simultaneously increases sensitivity due to use of the covariance structure and noise reduction. The data dimensionality reduction accomplished by the use of so-called latent variables (principal components or PLS components) gives possibilities to model interpretation and facilitates graphical visualisation of the process state and the model.

No detailed account of the theory and algorithms are given here. Easily accessible information can be found in the classic book by Martens and Naes [3] and in several tutorials [2,4]. The following section contains some terminology and brief guidelines for

model interpretation for the reader who is not accustomed to multivariate latent variable modelling.

Scaling and centering of data prior to modelling is usually necessary to obtain satisfactory results. Unless otherwise stated, auto-scaling, i.e. mean centering and scaling to unit variance have been used in the models for different process sections. Scaling for multi-block models is discussed separately below.

Standard PCA and PLS analyses have been carried out using the SIMCA software (Umetrics, Umeå, Sweden). Multi-block modelling has been performed in Matlab using non-commercial code as described below.

### 4.1.1 Interpretation of PCA and PLS models

In the current context, PCA is a method primarily used for visualisation and interpretation of process data. No difference is made between different types of data i.e. process inputs or outputs.

PCA reduces the dimensionality of the data by finding latent (hidden) variables, called principal components (PC). The PCs are combinations of the original variables that are more efficient in describing the variation in the data than the original variables themselves. In fact, the PCs are found by searching for the directions of maximum variance in the data. Each PC is described by a set of *scores* and *loadings*.

The loadings describe the nature of the PCs, i.e. the relationships between the PCs and the original variables, which is strongly related to the covariance of the original data. Each PC has a set of loadings, one value for each original variable, that are often represented graphically.

- Variables with positive loading values have positive correlation with respect to the phenomenon modelled by the PC, i.e. high values of one of the variables with positive loadings is connected to high values of the other variables with positive loadings. The higher the value of the loading, the stronger the variation.

- Variables with negative loading have negative correlation with the variables with positive loadings with respect to the phenomenon modelled by the PC.

- Variables with loadings with small absolute values are not involved in the phenomenon modelled by the PC.

Loadings can be presented graphically for each component as column or line plots or for two PCs as a so-called loading scatter plot. In such a plot, each point represents a variable. Both versions are used in the interpretation of the models in this report.

Scores describe how the observations (samples or times) are located in relation to the new variables, i.e. principal components. For each PC there is one score value for each observation, just like there is one value of each original variable.

- Observations that have similar score values in all PCs are similar with respect to the original process variables.
- High score values for an observation means that it is strongly influenced by the phenomenon described by the corresponding loading, i.e. that the variables with high loadings are influenced upwards and the variables with negative loadings are influenced downwards by the phenomenon described by the PC.
- The opposite is true for large negative score values, i.e. the variables with high loadings are influenced downwards by the phenomenon described by the PC.

Scores can be represented graphically as line plots, showing the time trend of the phenomenon described by the PC, or for combinations of two PCs as scatter plots. The latter is often used for classification.

Loadings and scores can be presented in the same scatter plot, which is then usually called a bi-plot. Observations (visualised by the scores) that are close to a variable (visualised by loadings) in the plot have high values of that variable and low values of the variables on the opposite side of the graph.

PLS is used to find a quantitative relationship between two groups of variables: the explanatory variables, denoted X, and the responses, denoted Y. The relationship is found from training data but can then be applied to new data to predict values of the responses from the explanatory variables. Model interpretation can be used to learn about the nature of the relationship and which explanatory variables are important for the responses.

Scores and loadings can be interpreted in the same way as for PCA but will not be identical since the objective of PCA is to describe the directions of maximum variance in the data but the objective of PLS is also to find a relationship between X and Y. Thus, the loadings and scores describe phenomena that have large influence on the data and are important for the relationship between X and Y.

The predictive ability of the PLS models can be measured by different figures of merit. Frequently, two versions of explained variance and estimated prediction error are used based on calibration or validation data, respectively.

- $R^2$ is the fraction of explained variance in the Y training data. It can take values between 0 and 1 where 1 means perfect prediction and 0 no predictive ability at all.

- $Q^2$ is interpreted in the same way as $R^2$ but is calculated by a procedure where observations that are unknown to the model, i.e. not part of the training set, are predicted. This is accomplished by using separate validation data or by so-called cross validation. $Q^2$ is a better measure of predictive ability than $R^2$, which often over-estimates the accuracy of the model. A discussion on the cross validation method used in this study can be found in section 0 due to its tight relation to the results presented in that section.

- RMSEC (*root mean square error of calibration*) is a measure of the average prediction error (given in the same unit as the response variable) calculated from the training data.

- RMSEP/RMSECV (*prediction*, *cross validation*) are prediction error measures calculated using data unknown to the model (se $Q^2$ above) and are thus better measures of the true prediction error than RMSEC. In our opinion RMSEP or RMSECV from a proper cross validation scheme is the most informative way to measure model performance, since it gives the measure in the same unit as the property being predicted by the model and thus facilitates comparison with other methods and models.

## 4.2 Multi block models for multivariate process modelling

Standard multivariate statistical modelling methods, such as PCA and PLS, are efficient in handling large amounts of data. However, when applied to very large multi-step processes there is a risk that the increasing model size and complexity can decrease the usefulness of the model by hampering interpretation and making model maintenance difficult. If the data is organised in meaningful blocks, usually according to the sections of the process, so called multi block models [8] can be applied, which can increase the utility and interpretation abilities of the models significantly.

Multi block modelling methods uses a two level model structure: a sub-level that contains model structures for the individual blocks and a super-level that connects the blocks. A schematic of a process and a multi-block model structure is given in Figure 1. Common variation is modelled on the super level as components in the same way as in standard PCA and PLS. Super scores describe the variation modelled by each component and the contributions from each block to the variation are given by the so-called super weights. Block contributions can then be further interpreted by inspection of the block loadings and block scores. No details about the algorithms or properties of multi-block PCA and PLS models are given here. The interested reader is referred to several good descriptions in the literature [8,9,10,11,12].

The multi-level model structure increases interpretability. The common variations and interactions between process sections are modelled on the super level, while details about the contributions and effects in each section are obtained on block level. Multi-block PCA models can be used for process monitoring, fault detection, fault identification and process optimisation in a manner similar to ordinary PCA. The structure makes fault detection easier by providing easier interpretation of which process section is having problems on super level and details about the problems on block level. Multi-block PLS models can be used for predicting e.g. product properties from process data, similar to ordinary PLS, but with increased model interpretability.

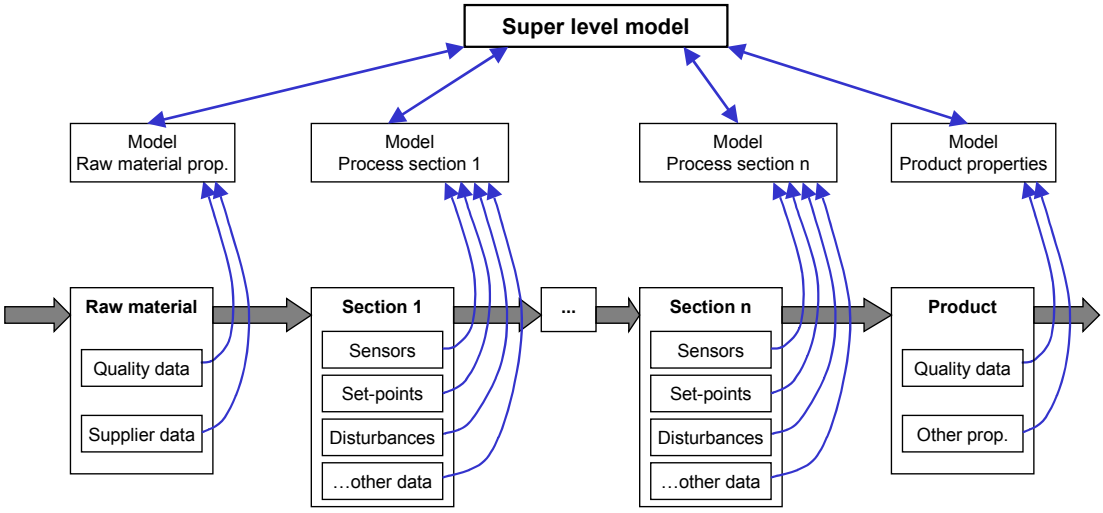



Figure 1. Schematic of a multi-block model applied to a multi-step process.

There are different versions of multi block models. In this report, the nomenclature of Westerhuis *et al* [8] is used. Briefly, the different model types relevant to the work described in this report are:

- Consensus PCA (CPCA) [13] finds common variations in all blocks, which are described by a super score. The super score is used for deflation of each block before calculation of the next component.

- Hierarchical PCA (HPCA) [11] is different from CPCA only in that super and block scores are normalised in the iterations instead of the loadings and weights. This leads to convergence problems and an unclear objective function [8].

- Hierarchical PLS (HPLS) [11] is an extension of HPCA where a PLS model cycle is performed between the augmented block scores and the response (Y) matrix in each iteration instead of the PCA model cycle performed on the augmented block scores in HPCA.

- Multi-block PLS (MBPLS) is different from HPLS, since a PLS iteration (not PCA) is used also in the modelling of the individual blocks. There are several different versions of MBPLS that have different properties due to the type of deflation of the X block used between the calculation of individual components. This is further discussed by Westerhuis and Smilde [10]. They conclude that the two methods most commonly used, deflation using block scores and deflation using super scores, both have drawbacks. The predictive ability of the model is not optimal in the first case, since information is lost in deflation that is not used to predict Y. In the second case, interpretability is hampered by the fact that information is transferred between blocks via the deflation. Westerhuis and Smilde suggest instead that only the Y block is deflated, which is claimed to avoid the drawbacks of the other methods. They show the theoretical advantages and demonstrate them by applying the method to data from a two-step tabletting process [10]. No other applications of this method have been published to our knowledge.

  The following nomenclature for MBPLS models is used in this report: MBPLS with block score deflation is denoted BPLS, MBPLS with super score deflation is denoted SPLS and MBPLS with only Y block deflation is denoted YPLS.

Multi-block modelling have been used successfully in a smaller application in a cracking unit by Wold *et al* [11]. Westerhuis and Coenegracht [14] describes an application from the pharmaceutical industry that demonstrates the advantages of multi-block modelling, although only two blocks are used. The advantages can be expected to be greater when studying a more complex process like in the project described in this report.

Due to the convergence problems and unclear objective function of HPCA and HPLS they are not used in the work described in this report. CPCA is used for multi-block PCA, while both BPLS and YPLS are used for multi-block PLS. SPLS is not used since the predictions for SPLS and YPLS have been shown to be identical [10] and interpretation ability is better in both BPLS and YPLS than in SPLS.

In addition to the normal variable scaling and centering that is usually applied in PCA or PLS of process data, block scaling can be used in multi-block modelling. If auto scaling of variables but no block scaling is used, the blocks are implicitly weighted according to the number of variables in each block. This is usually not desirable, since the importance of the block is usually not reflected in the number of variables. Commonly, all blocks are scaled to a common block variance, which means that all blocks have equal weight in the analysis. This is the approach used in the present work. It is also possible to use a priori information about the information content in different blocks and weight them accordingly but no such information was available in the present case.

All multi-block modelling have been carried out in the Matlab environment (Mathworks Inc., MA, USA). The implementations of CPCA, SPLS and BPLS from the Multiblock Toolbox (The Royal Veterinary and Agricultural University, Denmark, www.models.kvl.dk/source/) were used. Matlab code for YPLS [10] was kindly provided by Johan Westerhuis (Process Analysis & Chemometrics, University of Amsterdam, The Netherlands). It should be noted that the stand-alone multivariate modelling software SIMCA (Umetrics, Umeå, Sweden) has an implementation of hierarchical modelling that calculates the individual block models first and then the super level model based on the scores from the block models. This is not equivalent to calculating the full model simultaneously and was not used in the present work.

## 4.3 Missing data

There are implementations of PCA and PLS that can handle moderate amounts of missing data without serious effects on the estimated model. Values for the missing data are even estimated in the process [15,16]. The method is based on initially estimating the missing value by the mean of the variable. The estimate is then refined in an iterative process that can consume a significant amount of computer time for large data sets.

The software package SIMCA, which has been used for the standard (i.e. not multi-block) PCA and PLS calculations, uses a simplified missing data handling algorithm that require less computer time and has less favourable statistical properties. However, our experience is that, for the small amounts of missing data present in the current case, there is no significant difference between the results of the different algorithms.

For multi-block models, handling of missing data can be more difficult, since missing data is only estimated using the covariance structure in that particular block, not in the whole set of data. This means that accuracy is lost if there is correlation between blocks, which is the case for the data analysed in the present study. To overcome this problem and to save computer time, all missing values in the data was estimated from "un-blocked" data prior to multi-block analysis. The implementation found in The Missing Toolbox (The Royal Veterinary and Agricultural University, Denmark, www.models.kvl.dk/source/) was used. In the cases where data with a low time resolution have been used (e.g. in the multi-block PLS regressions with product property data), missing data were estimated from data with much higher time resolution which increases accuracy.

## 4.4 Multivariate modelling in the refinery industry

Multivariate statistical methods have been investigated and successfully applied for fluid catalytic cracking (FCC) units in a number of publications. Prantyasto and Qin used PCA

for sensor validation and process fault detection for a fluid catalytic cracking unit [17]. The authors show that PCA can be used to detect process faults at an early stage and reconstruct values from failing sensors. In addition, PCA has been used in combination with signed diagraphs to facilitate automatic fault identification [18].

Several authors have recognised the lack of good on-line sensors in distillation processes and the unsatisfactorily long response times for laboratory determinations of properties for products from the processes, which hampers quality control and efficient process control.

Chatterjee and Saraf [19] have studied software sensors for predicting product properties from a crude distillation unit. Their approach is based on a mixture of simplified first-principles equations and empirical relations. The method is dependent on a true boiling point (TBP) curve for the crude entering the column, which is not available on-line. Thus they devote considerable effort to estimating the current feed properties using laboratory test data for the product streams.

Shin, Lee and Park [20] have investigated different approaches for estimating product composition from distillation columns theoretically and using simulated data. The most interesting conclusions from their work is that linearisation of the problem by non-linear transformations of some variables can be advantageous. However, the transformations discussed require a significant amount of physical data about the system and cannot be used straightforwardly in a complex system like the one studied in this work. Further, the authors conclude that PLS is a good method to approach the problem and gives better performance than the other methods investigated.

# 5 Process description

The demonstration site in the present study is Nynäs Refining AB's refinery in Gothenburg, Sweden. At the refinery crude oil is fractioned into more desirable products through the process of distillation. The refinery has two distillation towers, one with atmospheric distillation (AD) and the other with vacuum distillation (VD), which is actually not performed at vacuum but at very low pressure. Bitumen, which is used in the making of asphalt, is the main product of this refinery but lighter products are also of importance, especially a product called D10 that is further processed in to special oils.

For modelling purposes, the process was divided into several logical blocks and measured parameters within each block were grouped together. These blocks are the basis of the following description of the process. They are listed below in the order of appearance in the process, but it should be noted that there is heat exchange between the warmer products and the cooler feed of crud oil, which of course has an effect on previous blocks.

A more detailed outline of the distillation process at the demonstration site is given in Appendix 1 – Process outline. The process variables used for modelling are listed according to block in Appendix 2 – Parameter lists.

## 5.1 Raw material

The raw material, i.e. the crude oil, is stored in a mountain chamber near the harbour and pumped through a pipeline into a cistern at the process site where it is preheated. The crude oil is fed into the process and additionally heated through a series of heat exchangers. In this way the excess heat energy in the products is recovered and the need for heating the crude oil in the furnace is reduced. The series of heat exchangers starts off with the coolest product from the AD tower and continues sequentially with warmer products. Finally there are several exchanges against the warmest of the products, bitumen.

Typical process parameters in this block are feed rate and temperatures. The intention was to also include properties of the crude oil from off-line analyses. However, they were never considered in the modelling on account of them being associated with too high inaccuracy and too low sampling frequency.

## 5.2 AD furnace

The furnace is divided in two parts, AD and VD furnace. In the AD furnace the crude oil is heated to the right temperature before it enters the AD tower. Each of the furnaces has two burners, one in the front and one in the back of the furnace. The fuel consumption in these burners is of course of high interest in the modelling work. It should be noted that the AD and the VD parts of the furnace are not completely isolated from each other. Heat transfers from one part to the other and they also have a mutual section in the top of the furnace where the furnace gas exits. The crude oil feed passes through this mutual section.

## 5.3 AD tower

Distillation, at atmospheric pressure, of the crude oil is performed in the AD tower. Steam is supplied at the base of the tower in order to force light molecules upwards. Three fractions are extracted in the AD tower: ADTOP, ADFR1 and ADFR2. The temperature in the tower is controlled by a return flow of ADTOP. Before the ADTOP fraction is sent to the storage tank, excess gas and water caused by the steam is removed. ADFR1 and ADFR2 passes through individual side strippers from which light molecules

are returned to the tower by addition of steam, before these fractions are sent to storage tanks.

Process parameters of particular interest in this block are the temperature profile through the tower, steam supply and fraction yields.

## 5.4 VD furnace

What is left of the crude oil after extraction of the three fractions in the AD tower needs further heating before it can enter the second distillation tower. This is done in the VD part of the furnace. As was mentioned in the description of the AD furnace, the VD furnace contains two burners for which the fuel consumption is of interest.

## 5.5 VD tower

Distillation, at very low pressure, of the residual feed from the AD tower takes place in the VD tower. As was the case for the AD tower, steam is supplied at the base of the VD tower. Another three fractions are extracted here, VDTOP, VDFR and bitumen. There are return flows to the VD tower of both VDTOP and VDFR. The VDTOP fraction is first separated from gas. Part of the VDFR fraction is refluxed for the purpose of level control in the tower, while the rest of it passes a side stripper on its way to the storage tank. In the side stripper steam is supplied in order to force lights molecules back to the tower. So there are two return flows of VDFR but with slightly different composition and they are returned to different levels in the VD tower. Bitumen is extracted at the bottom of the tower and before it reaches the storage tank it passes a series of heat exchangers, where the temperature of bitumen is lowered and the feed of crude oil is heated.

As for the AD tower, process parameters of particular interest in this block are the temperature profile through the tower, steam supply and fraction yields.

## 5.6 Product properties

The block for the product properties refers to the laboratory analyses of the four fractions ADFR1, ADFR2, VDTOP and VDFR. The curves of the true boiling point (TBP) for each of the fractions are analysed by gas chromatography (GC) three times a day. Figure 2 illustrates a typical TBP-curve, with boiling temperature plotted against weight percent of the fraction. These curves hold information on important qualities of the fractions, e.g. their viscosity, density and flash point. The TBP-curves also reveal the overlap between adjacent fractions. It is desirable to cut as clean fractions as possible from the crude oil. Parameters used in the modelling are the temperatures at 0.5, 1, 2, …, 99, 99.5 weight percent of each fraction. Consequently there are 101 variables per fraction.

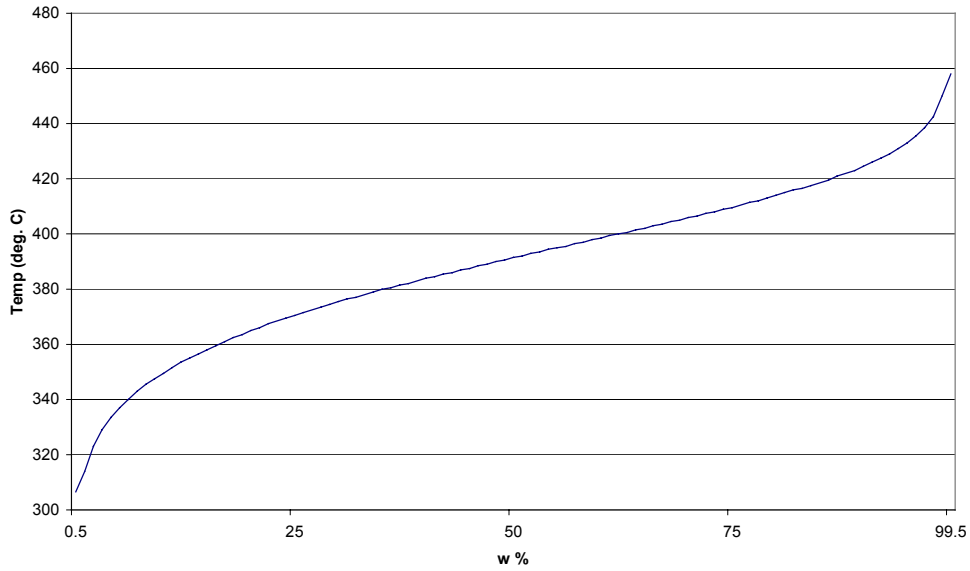



Figure 2 TBP-curve example where 50 weight percent of the molecules in the fraction have a boiling point above 390 °C.

# 6 Results and discussion

This section presents some results of the modelling efforts carried out in the project. It is not possible to show all possible interpretations. Instead the models and interpretations most relevant to improving process performance have been selected and are discussed below. The models discussed are of both PCA and PLS type as well as both single block and multi-block.

## 6.1 Effect of production rate

The second principal component in a PCA model based on data from the AD tower and GC TBP data from ADFR1 and ADFR2 clearly shows effects of production rate on the product properties. The scores for this component and the production rate are shown in Figure 3. It is clear that the component is heavily influenced by production rate but that there are other, more short-term, contributions also.

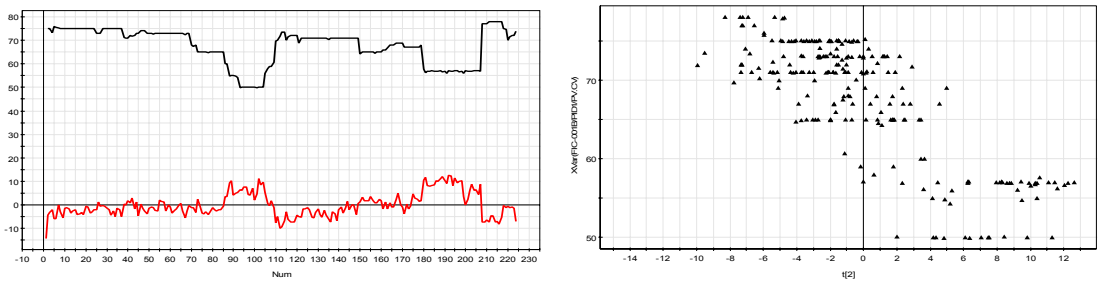



Figure 3 Left: Production rate as crude oil flow (black) and scores for the second principal component (red). Right: Production rate as crude oil flow as a function of scores for the second principal component.

The loadings show clearly that the component models the low-boiling ends of ADFR1 and ADFR2, i.e. the flash points of the two components, so that lower flash points are obtained when the production rate is high. The interpretation of the loadings of the process data is rather straightforward. It shows that lower flash points are obtained when higher production rates are not compensated with higher steam flows to C3 and C4. This also influences the temperatures of the ADFR1 and ADFR2 extraction points, so that these are higher when the production rate is higher. Note, however, that the steam flow to the AD bottom is compensated for production rate in the variation modelled by this component. The component does not contain any significant changes in yields of the fractions.

It should be noted, as is clear from Figure 3, that the production rate and side stripper steam flows are not the only factors contributing to varying flash points of ADFR1 and ADFR2. The effect of this component can be quantified as approximately ±3ºC for the 5% point of ADFR1 and approximately ±2ºC for the 5% point of ADFR2, which is a significant amount of their variation during production[1].

It is recommended that tools or routines for (semi) automatic compensation of steam flows to side strippers for production rate are introduced.

## 6.2 AD tower oscillations

The results discussed in this section are based on PCA of data from the AD tower only. The analysis showed that an oscillation with a period of approximately 85 minutes is present in a number of variables related to the top half of the AD tower. In a PCA model based on 15-minute process data from the AD tower, the oscillation is completely

[1] As noted in the discussion about prediction of TBP curves further down in this report, the standard deviations for the 5% point of ADFR1 and ADFR2 are 2.8ºC and 2.0ºC respectively.

dominating PC3, which explains about 8% of the variance in the data. This shows that the oscillation is a significant part of the total process variability.

No signs of a similar oscillation are found in the individual analysis of other process sections, which indicates that the phenomenon observed is local and does not influence other process sections significantly. This is confirmed by the multi-block analysis discussed elsewhere in this report[2].

Given that the period is only a few multiples of 15 minutes it is difficult to determine the frequency and possible phase shifts between variables accurately using this data. Thus, a more detailed analysis was undertaken based on data with higher time resolution: 30 seconds. For computer capacity reasons, the analysis of this data had to be restricted to a shorter time period. Results from one production period from 14 June to 17 June 2002 (12000 observations) are discussed here. Other periods have been investigated with very similar results.

Scores from a PCA model based on 30-second data from this period are shown in Figure 4. It is clear that in this case the oscillations are modelled by PC2, which explains 11% of the variance. Fourier Transformation (FT) of the score vector gives a power spectrum with a single peak at approximately 0.012 $min^{-1}$, which corresponds to a period of approximately 85 minutes. The loadings for this PC, shown in Figure 5, indicates that only a few variables are involved in the oscillation: Yield ADTOP, Flow ADTOP (volume and mass) to cistern, valve ADTOP to cistern, density ADTOP to cistern, flow gas C5-B, density ADFR1 to cistern, temperature ADFR1 extraction and temperature C3.

The first four of these variables are all related to ADTOP flow to cistern (the ADTOP yield is calculated from this flow and the crude flow). The density of ADTOP to cistern is also oscillating, which, since it cannot be attributed to temperature changes, indicates oscillating ADTOP product properties. The same is true for ADFR1 product properties, since ADFR1 density is oscillating but not the temperature of the ADFR1 flow to cistern.

Notably the temperature of the ADTOP reflux and the pressure in the top of the AD tower are not oscillating while the temperatures of the ADFR1 extraction and the corresponding side stripper C3 are. Also the flow of gas from the gas separator on the ADTOP stream, C5-B, is oscillating.

[2] When CPCA is applied to 15-minute process data from the full process, i.e. all process sections. The oscillations appear in the 7th principal component, which explains 7% of the variation in the data from AD tower block and less than 1% of the variation in the data from any other block. This, together with inspection of block scores, confirms the conclusion that the oscillation is not present in data from other process sections than the AD tower.

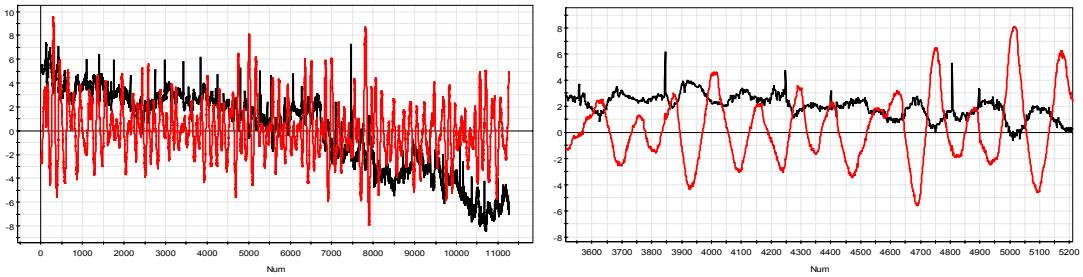



Figure 4. Scores from a PCA model based on 30-second data from the AD tower 14 June-17 June 2002: PC1 (black) and PC2 (red). Left: full period. Right: enlargement.

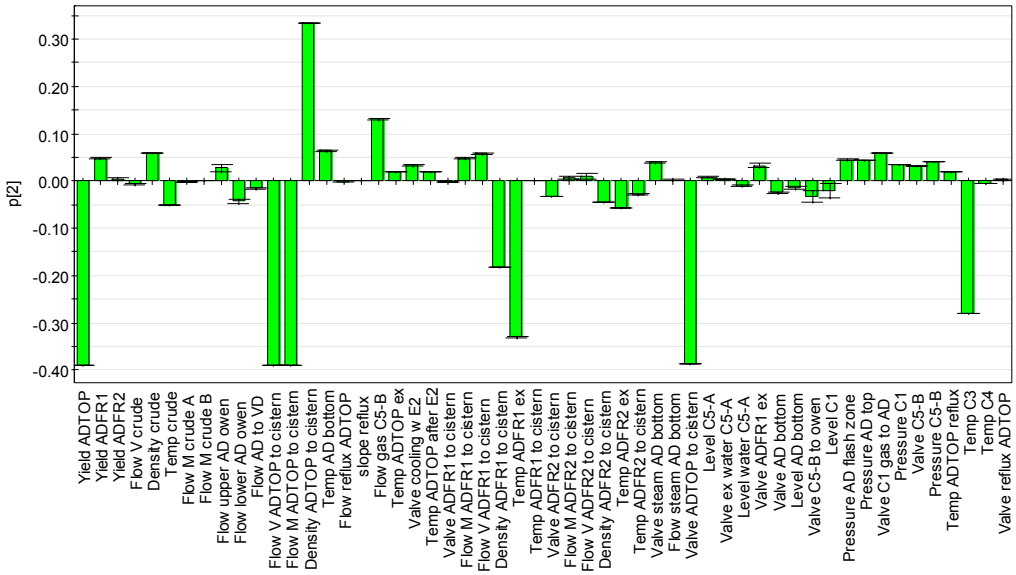



Figure 5. Loadings for PC2 from a PCA model based on 30-second data from the AD tower 14 June-17 June 2002.

In order to investigate more in detail the nature and the cause of the oscillations, dynamic PCA[3] [21] was applied to the 30-second data from the period in June 2002. The lag structure 1, 10, 20…110 observations, i.e. 0.5, 5, 10…55 minutes was used, since this covers more than half a period of the oscillation, which allows for identification of possible phase shifts between variables. In the dynamic model, the oscillations are modelled by PC4, which also contains some smaller contributions from other phenomena with a long time scale. The loadings of the lagged variables, however, clearly allows for identification of the variables having an oscillation with a period of approximately 80 minutes, as exemplified in Figure 6 for one variable, and for estimation of phase shifts by

[3] Dynamic PCA is standard PCA where the data is augmented with time lagged data, so that for each observation the value of a variable at that time and one or more earlier values of that variable are also included. This increases the number of variables but allows dynamics in the process to be identified.

comparing such plots for different variables. Note that only phase shifts of at least 5 minutes can be identified given the lag structure used.

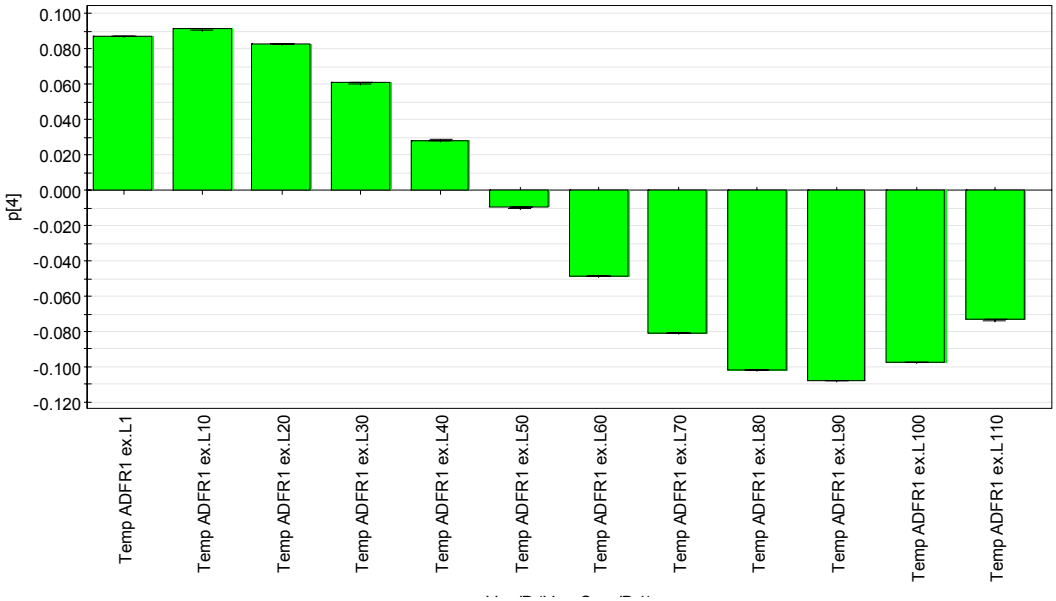



Figure 6. A part of the PC4 loading from dynamic PCA of the AD tower, only the lagged variables for ADFR1 extraction temperature are shown. L10, L20 etc. corresponds to lagging of 10 and 20 observations, i.e. 5 and 10 minutes, and so on.

The dynamic analysis confirms the results from the standard analysis and adds some extra information. First, there seems to be small phase shifts between the variables previously identified as oscillating. Secondly, more variables where the oscillation is only a minor contribution to the variance can be identified from the dynamic model. The findings are summarised in Table 1. It should be pointed out that interpretation related to the variables where the oscillation is only a minor contribution to the variation can be difficult due to the larger influence of other variations. In addition, the signals used for modelling are extracted from a process history database that stores the data in a compressed manner, which can hide smaller variations in the process[4] [22]. The uncertainty in the interpretation of the smaller effects in the lower half of the AD tower is further shown by the time lags estimated between the top and lower parts of the tower: +30 minutes to ADFR2 extraction and +50 minutes to the tower bottom. Both these lags are considered

[4] As an example the ADFR2 extraction temperature is only stored every 4 minutes or when the deviation is more than 1 degree from the previous value. Some example temperature data are shown for illustration in Figure 7 and Figure 8. It is clear that the data in the right half of both figures (ADFR2 extraction temperature, C4 temperature, ADTOP pressure) is influenced by compression. In the right half of Figure 7, oscillations are at least partly visible (and detected by dynamic PCA, cf. Table 1), while in the right half of Figure 8 no oscillations are visible which may indicate that they are not present or that they are hidden by the data compression. Thus, to a small extent the oscillations may influence more variables than discussed.

very long in relation to the residence time in the AD tower. Thus, no emphasis is put on interpretation of the oscillations in the lower part of the AD tower. All variables with major oscillations are found in the top half of the AD tower, *cf*. Table 1.

Table 1. Variables involved in the AD tower oscillations

| **Variable** | **Min loading [obs. lag]** [a] | **Mode**[b] | **Phase shift [min]**[c] |
|---|---|---|---|
| Flow gas C5-B | 70 | + | +10 |
| Yield ADTOP | 50 | +[b] | 0 |
| Flow ADTOP to cistern | 50 | + | 0 |
| Valve ADTOP to cistern | 50 | + | 0 |
| Density ADTOP to cistern | 50 | - | 0 |
| Temperature ADFR1 extraction | 50 | + | 0 |
| Temperature C3 side stripper | 40 | + | -5 |
| Density ADFR1 to cistern | 50 | + | 0 |
| Temp ADFR2 extraction[d] | 30 | + | -10 |
| | | -[e] | +30[e] |
| Density ADFR2 to cistern[d] | 30 | + | -10 |
| | | -[e] | +30[e] |
| Slope reflux[d] | 80 | - | +15 |
| | | +[e] | +55[e] |
| Temp. AD bottom[d] | 70 | - | +10 |
| | | +[e] | +50[e] |

[a] The lag giving the minimum absolute value of the loading for the variable, cf. L50 in Figure 6

[b] + denotes oscillation with Yield ADTOP and - a negative correlation, i.e. a phase shift close to 180°.

[c] Relative to Yield ADTOP and accounting for the mode, i.e. an alternative interpretation is to add/subtract 40 minutes to/from this value and change the mode.

[d] The oscillation are only a minor contribution to the variance in these variables.

[e] Alternative interpretation, see text.

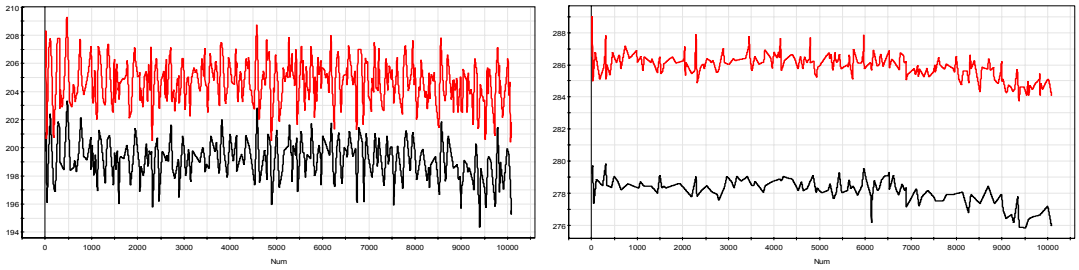


Figure 7. Data from the period 14-17 June 2002: temperatures in the extraction (red) and side stripper (black). Left: ADFR1/C3. Right: ADFR2/C4.

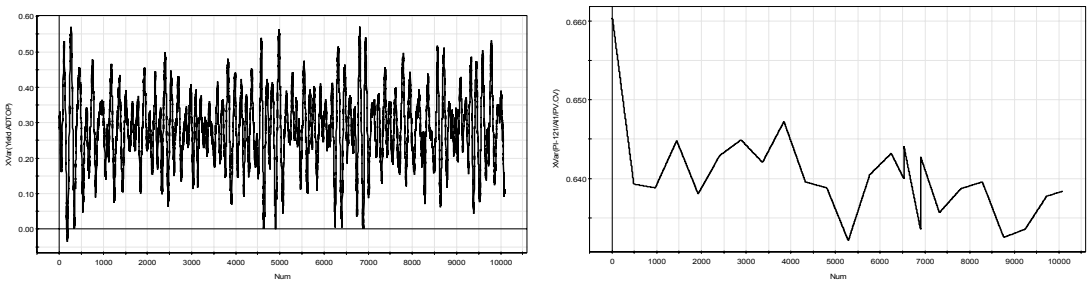


Figure 8. Data from the period 14-17 June 2002: the yield of ADTOP (left) and the ADTOP pressure (right).

Three possible explanations to the oscillations have been identified. The possible origins are the separator C5-A, the gas separator C1 or the side stripper C3.

In discussions with the process operators, it was put forward that the extraction of the ADTOP fraction from the water separator C5-A to tank (controlled by level control loop LIC-002) is irregular, which can give rise to the oscillations in ADTOP flow and yield. It has not been considered a problem since it was not known to influence the AD tower itself. The present analysis, however, strongly indicates such an influence. A possible mechanism may be that the irregular flow from C5-A causes pressure changes in C5-B and the ADTOP, which would then influence the rest of the tower. The fact that the ADTOP pressure does not show up as a part of the oscillations in PCA may be caused by data compression in the process history database. From the appearance of the ADTOP pressure signal (Figure 8 right) it can be suspected that some information in the signal is hidden by compression[5].

The pressure changes would then directly influence the equilibria governing the ADFR1 extraction temperature and with some delay to a small extent also the ADFR2 extraction

[5] It should be noted that the main contribution to the ADTOP pressure variation seen over a longer time span is the production rate in the AD tower, as discussed elsewhere in this report.

temperature. The changing extraction temperatures influence the composition of the products, which is reflected in their densities. The oscillating changes in product properties are not possible to detect from the GC TBP curves since sampling and analysis is done only once every 8 hours.

Another possible explanation to the oscillations is related to the gas flow from the gas separator C1 caused by poor pressure control in C1. If the pressure in C1 builds up during some time and is suddenly released by opening the valve for the gas flow to the AD tower, the amount of low-boiling components (including water) and possibly also the pressure in the AD tower can oscillate. Unfortunately the gas flow from C1 to the AD tower that would be the primary indicator is not logged. As discussed above, the pressure in the tower does not show any clear oscillatory behaviour but such behaviour can be hidden by data compression in the process history database. In an M.Sc. thesis about the control performance at the plant, it was pointed out that there is significant oscillatory behaviour in C1 but the problem discovered in that work were mainly related to level control and had significantly shorter time scale [23]. It should be noted, however, that the data studied in that work did not allow investigation of slow oscillations.

The third potential explanation to the oscillations is that the cause is related to the C3 side stripper. This is supported by the fact that the C3 temperature is "leading" the oscillations, about 5 minutes before the ADTOP flow. The oscillations in C3 would then influence the AD tower by either the flow from the tower to the side stripper or the gas flow in the opposite direction. It should be noted that the temperature fluctuations in C3 can be expected to influence the AD tower, which would show up in the ADFR1 extraction temperature and can be expected to influence the top of the AD tower rapidly through a gas flow.

Possible causes of the oscillations in C3 would be steam supply or the valve controlling the flow from the tower to C3. The steam flow to C3 is a strong candidate and data in the history database show indications of an oscillation but the signal is too heavily compressed to draw any firm conclusions. It can be noted that the C3 level is also oscillating, but with a period of approximately 2.5 minutes, i.e. substantially faster than the phenomenon investigated here[6].

[6] It can also be noted that the steam flow to the C4 side stripper and the level in C4 are oscillating more heavily than the corresponding variables in C3. The period for the steam flow is approximately 1 minute and the period for the level is approximately 6.5 minutes, so the phenomena are much faster than the one studied here.

### 6.2.1 Conclusions and recommendations

From the above results, it can be concluded that PCA and dynamic PCA are extremely powerful tools for identification of process disturbances such as oscillations. This is true for both variables where the oscillation is the dominating variation and variables where the oscillation is only a minor contribution or where the effect is partially hidden by database compression.

The oscillations have a marked effect on the AD tower and product streams. The ADTOP yield oscillates between 0 and 0.5% as shown in Figure 8. The density of this stream is oscillating between approximately 0.70 and 0.71 kg/dm$^3$. The temperature amplitude in the ADFR1 extraction, which is likely to influence product properties, is approximately 5°C, which is a major change. The standard deviation of the 50% point in the TBP curve for this fraction is 3°C. Due to the oscillating behaviour the effect is to some extent averaged out in the storage tank. However, the oscillations clearly contribute to a wider boiling point interval for ADFR1 with a lower flash point as a consequence. Further, process control is to a large extent based on the TBP curves determined by GC for samples taken every 8 hours. With the large oscillations present in the process, the precise timing of sampling has a large influence on the determined TBP curve and thus accurate process control is not possible. Thus, the oscillations can be expected to impair process performance with respect to energy consumption, process economy and product quality but it is very difficult to quantify the effect.

The PCA has pointed out some candidate causes to the oscillations but in order to determine the cause of the problem, it is necessary to have access to higher quality data for some of the variables measured in the process and stored in the history database. It is recommended to change the settings for data compression of the following variables: ADTOP pressure, ADTOP temperature, ADTOP reflux and steam flow to C3. Also, it would give valuable information if the flow from the AD tower to the C3 side stripper and the gas flow from C1 to the AD tower could be logged. When the necessary changes in the data history collection are done, it should be possible to identify the cause of the oscillation and to take the necessary measures to correct it, which would most likely mean to tune a control loop or repair a faulty valve.

## 6.3 Prediction of product quality

It is reasonable to believe that variation in product quality reflects variation in how the process is operated and changes in the raw material. It is also reasonable to believe that, under some circumstances, changes in raw material composition will be reflected in the process variables that are measured on-line. Therefore it can be possible to model product quality based upon the values of the process variables.

In order to investigate this further, PLS models (see 4.1 above) for TBP-curves of four fractions were calibrated. The advantage of these models over e.g. regular multiple regression models and artificial neural network models are that they can be used both for prediction and for interpretation of the underlying phenomena that causes the variation in the predicted parameters. This is valuable information for process optimisation purposes and, hence, equal emphasis is given to model interpretation, which is discussed in 6.4. A further advantage is the prediction diagnostics obtained by PLS that can be used to increase the reliability of the model by detecting an outdated model.

Previous work in modelling of petrochemical processes has indicated that linear models are not adequate to describe the non-liner process behaviour. However, in this report it is shown that by using "smart" variable transformations, i.e. by including knowledge about process non-linearity in the data transformation strategy prior to the regression step, it is possible to use linear models to accurately describe the process and to predict the product quality.

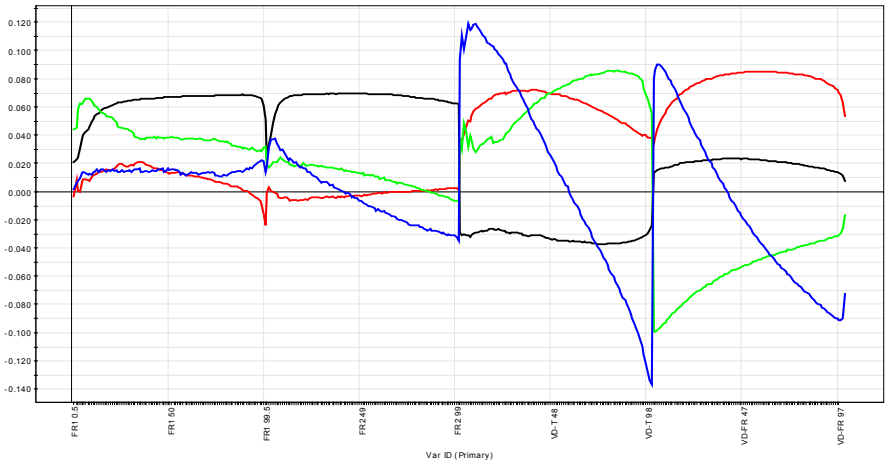



Figure 9 Loadings from a PCA model on TBP-curves of the fractions ADFR1, ADFR2, VDTOP and VDFR. The four components cover 90 % of the variation in the TBP-curves and were extracted in the following order: black, red, green and blue.

### 6.3.1 Sources of variation in TBP data

An initial PCA investigation of the TBP-curves of the fractions ADFR1, ADFR2, VDTOP and VDFR showed that, between the two towers, the quality of the products did not display any distinct co-variation. Variation in the TBP-curves of fractions from the AD and the VD tower were mainly described in separate components of the PCA model, see Figure 9. On the other hand, fractions from the same tower were highly correlated and only a few components were needed to explain a majority of the variance in data. Therefore the continued approach was to further investigate the fractions from each tower in separate PCA models, including the process variables related to the tower, and then

evaluate individual PLS models. One model is created for the fractions of the AD tower and one for the fractions of the VD tower.

### 6.3.2 Uncertainty of TBP reference data

When estimating the performance of a model prediction it is important to have an estimate of the error in the reference method for two reasons. 1) The prediction is compared to the reference value so the error estimate is actually influenced by the uncertainty in the reference value. 2) The purpose of modelling is usually to replace the reference method or to make the values from it available more frequently. It is then of great interest to know the relative performance of the methods.

Table 2. Estimates of errors in TBP reference data, based on Nynäs experience and the ASTM standard.

| Fraction | %[a] | Typical boiling point [ºC] | Error estimate [ºC][b] | Fraction | %[a] | Typical boiling point [ºC] | Error estimate [ºC] [b] |
|---|---|---|---|---|---|---|---|
| **ADFR1** | 5 | 165 | 2.0 | **VDTOP** | 5 | 270 | 2.8 |
| | 20 | 210 | 2.4 | | 20 | 305 | 3.1 |
| | 50 | 247 | 2.2 | | 50 | 336 | 2.2 |
| | 80 | 277 | 2.2 | | 80 | 363 | 2.2 |
| | 95 | 297 | 2.6 | | 95 | 388 | 2.6 |
| **ADFR2** | 5 | 260 | 2.8 | **VDFR** | 5 | 337 | 3.4 |
| | 20 | 297 | 3.0 | | 20 | 365 | 3.5 |
| | 50 | 322 | 2.2 | | 50 | 387 | 2.2 |
| | 80 | 349 | 2.2 | | 80 | 408 | 2.2 |
| | 95 | 370 | 2.6 | | 95 | 427 | 2.6 |
| **MIX D10** | 5 | 270 | 2.8 | | | | |
| | 20 | 298 | 3.0 | | | | |
| | 50 | 328 | 2.2 | | | | |
| | 80 | 355 | 2.2 | | | | |
| | 95 | 380 | 2.6 | | | | |

[a] The point on the TBP curve as percent mass evaporated. [b] Expressed as a standard deviation in order to facilitate comparison with model prediction error estimates.

The uncertainty of the reference TBP curves used for modelling, i.e. the TBP curves determined at the process laboratory by the ASTM D 2887-02 method, which is a GC method, was estimated based on the method specifications and discussions with the laboratory personnel. The standard gives estimates of reproducibility and repeatability. The actual error is believed to be between these two, since repeatability considers two

samples analysed in sequence on the same instrument, while reproducibility considers different laboratories. Hence, the two error estimates where pooled and the results shown in Table 2 were obtained. These are in accordance with the experience of the laboratory personnel but please note the discussion in connection to Table 3.

### 6.3.3 Estimation of prediction errors

The cross validation scheme used in the validation of the PLS-models is based on a *leave-one-production-period-out* approach, which means that the production periods (2.5-11.5 days, different number of samples) are used as cross validation segments. Thus, no data from a certain production period is present in the calibration data when the TBP curves for that production period is being predicted. This is believed to give better prediction error estimates than other cross validation schemes since there are major differences between operating conditions, and possibly raw material properties, between the periods but much smaller differences within periods. The larger differences are representative for future production and, hence, RMSECV based on leave-one-production-period-out should be realistic for future use of the model.

Other cross validation schemes with the same number of segments but where samples from all production periods were always present in the calibration set were also tried. They resulted in significantly lower RMSECVs closer to the RMSECs. The difference between RMSEC and RMSECV and the difference between the cross validation schemes show the importance of proper model validation to obtain realistic prediction error estimates.

### 6.3.4 AD tower

A six component PLS-model models 83% of the total variation in the calibration TBP-curves for ADFR2, with a $Q^2$ of 0.65. The predictive ability is worse for the ADFR1 TBP-curves, as discussed below. As illustrated in Table 3, the variation in the boiling points of the lighter parts of the fractions is smaller than the variation in the heavy parts, which makes it more difficult to model these with the same relative accuracy. Hence, the lighter part of these fractions has a weaker link to process data, at least for the samples analysed here or, alternatively, a higher uncertainty in the reference values[7]. It is possible that the oscillations in the AD tower, which were described earlier, has an effect on the lighter ends of the two fractions. Due to the relatively high frequency of these

[7] It is important to note that the prediction errors and the predictive performance are always measured against the reference values. This means that the measured prediction error is the sum of errors in the reference method and the model prediction, which means that a more uncertain reference method will give a higher apparent prediction error but not necessarily a higher true prediction error.

oscillations, it is impossible for a model based on the currently used data with samples only every eight hour to cover any phenomena linked to that disturbance. The sampling time of the GC samples are not known with enough accuracy to take such fast process changes into account.

The prediction error estimates are shown in Table 3 and Figure 10. The last column in the table shows the standard deviation of the data used for model calibration. By comparing this to the reference error, it is possible to estimate the part of the variance that is systematic (i.e. due to changes in the product properties and not measurement error). The first thing to note is that error estimates are higher than the standard deviation for the 5% and 20% points of ADFR2, which can only be explained by the fact that the error estimates are too high for these points. This confirms that the error estimates are rough but it is likely that they still show the approximate magnitude of the reference errors. For the other points of the ADFR2 fraction, the standard deviations of the data are much higher than the error.

Table 3. Prediction error estimates for PLS model for TBP curves of the fractions from the AD tower.

| Fraction | % | RMSEC | RMSECV[a] | TBP reference error | TBP reference std deviation |
|---|---|---|---|---|---|
| **ADFR1** | 5 | 2.5 | 2.8 | 2.0 | 2.8 |
| | 20 | 2.2 | 2.7 | 2.4 | 2.9 |
| | 50 | 1.9 | 2.7 | 2.2 | 3.0 |
| | 80 | 1.7 | 2.7 | 2.2 | 3.0 |
| | 95 | 1.6 | 2.8 | 2.6 | 3.1 |
| **ADFR2** | 5 | 1.4 | 2.1 | 2.8 | 2.0 |
| | 20 | 1.5 | 2.5 | 3.0 | 2.6 |
| | 50 | 1.7 | 2.5 | 2.2 | 3.3 |
| | 80 | 2.3 | 2.7 | 2.2 | 5.0 |
| | 95 | 3.1 | 3.2 | 2.6 | 7.5 |

[a]RMSECV is estimated by leave-one-production-period-out cross validation to ensure that realistic error estimates are obtained.

In general, the RMSECs are significantly lower than the RMSECV, although this is not true for the heavy end of ADFR2. For the ADFR1 TBP curve, the error estimates are of the same magnitude as the standard deviation in the data, which means that the present model is not useful for prediction of ADFR1 properties. It is likely that the difficulties, particularly in the light end, are due to the oscillating product properties identified elsewhere in this work (see 6.1). As noted in the discussion of the oscillations, the GC TBP data is sampled much too infrequently to be able to model this behaviour. The much lower

RMSECs of the heavy part of ADFR1 indicates that there is potential for model refinement, probably based on a more extensive data set covering several seasons.

For ADFR2, the results are much more encouraging with the exception of the light end that is difficult to predict. For the 50% point, the estimated prediction error is 2.5ºC, which is similar to the estimated reference method error (2.2ºC). If the result of the cross validation is corrected for the contribution from the reference method[8] error the prediction errors 1.2, 1.6 and 1.9ºC are obtained for the 50%, 80% and 95% point of ADFR2 respectively. This is lower than the errors of the reference method. Since the reference method error estimates are only approximate, these corrected prediction errors are also approximate but the calculation indicates that the accuracy of model prediction is in the same neighbourhood as the accuracy of the reference method [24] or even better. This can sound contradictory but is nothing strange if the methodology is considered as demonstrated empirically by Coates for NIR spectroscopy and PLS [25].

The good prediction of the heavy part of ADFR2 is confirmed by Figure 10. The predictions are not cross validation predictions but, as shown in Table 3, for the 95% point of ADFR2 visualised in the figure, RMSEC and RMSECV are approximately equal, which means that the cross validated predictions are about the same as those shown.

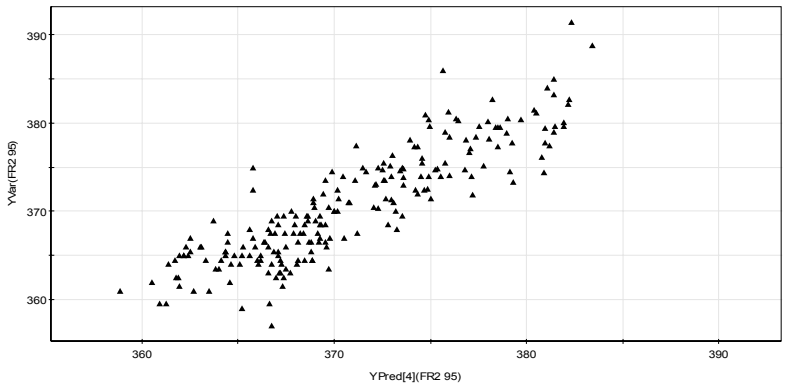



Figure 10. Predicted versus measured values for the 95% point of ADFR2.

The good predictive power indicates that the model describes realistic relationships between the process data and the product properties and, hence, that model interpretation can give valuable information about the process. The interpretation of the model is described in 6.4.2.

[8] As noted previously, the RMSECVs have contributions from both prediction errors and reference data errors. Since it is the prediction errors that are of interest, these can be corrected for the reference data errors if the magnitudes of these are known.

### 6.3.5 VD tower

Similarly to what is described for the AD tower above, models were developed to predict the properties of some products from the VD tower, i.e. the VDTOP and VDFR fractions. Models based on process variables from both the AD and the VD tower were found superior to models with only the VD tower variables as predictors. This is logical, since the products of the VD tower obviously depend on what is extracted from the crude oil in the AD tower. Models including all process blocks were also fitted. They did not improve the results, but merely made the model larger and interpretation more complex. The results presented here are associated with a YPLS-model with six components based on data from the blocks AD tower and VD tower and with the TBP curves of the products VDTOP and VDFR as response (Y) variables. The model's $R^2$ is 0.86 and the predictive power ($Q^2$) is 0.65. Calibration data comes from 2002 only, because prediction errors for observations in 2003 were very large. These large errors are most likely due to changes in process operation. During the winter stop 2002/2003 there were several changes in set points for variables in both towers. Since there are only two production periods in the data set from 2003, and the cross validation method used is leave-one-prediction-period-out (as discussed in 0), the large prediction errors of 2003 data are not surprising.

Prediction error estimates and the standard deviation of calibration data are shown in Figure 11. The result shows that the VDTOP fraction is easier to model than VDFR, which is displayed by the difference between the RMSEC and RMSECV curves. However, modelling of the lighter end of VDTOP is not useful since this part of the curve coincides completely with the standard deviation of the calibration data. Table 4 lists a selection of points from the curves along with the estimated reference error of the TBP analysis. Once again there are some reference errors that are higher than the standard deviation, which is an indication that the estimation of the reference errors is a bit rough, see discussion in 6.3.4 above. This concerns the 5% and 20% TBP's of VDTOP. Other than that, standard deviations are considerably higher than the reference error. It should also be noted that extreme data has been omitted from the model calibration, which will affect the standard deviation of the reference data.

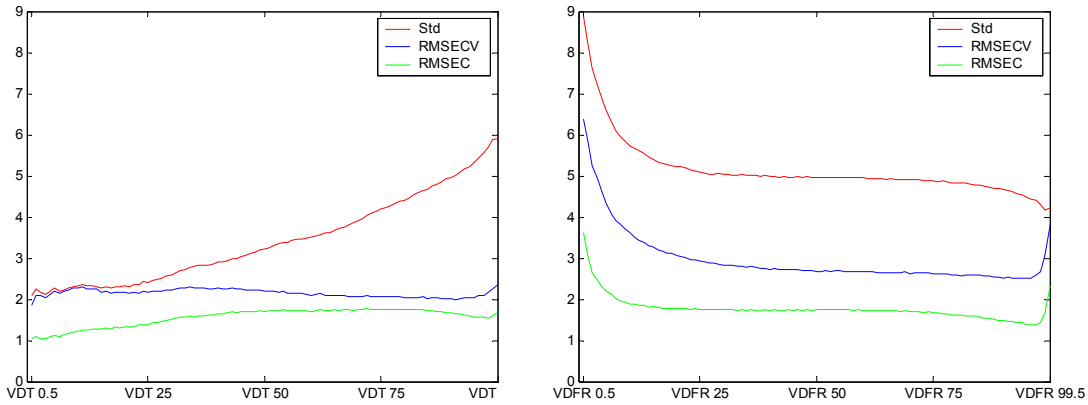



Figure 11 Standard deviation in calibration data, RMSECV and RMSEC for the PLS-model of the VD tower fractions VDTOP (left) and VDFR (right).

Table 4. Prediction error estimates for PLS model for TBP curves of the fractions from the VD tower.

| Fraction | % | RMSEC | RMSECV[a] | TBP reference error | TBP reference std deviation |
|---|---|---|---|---|---|
| **VDTOP** | 5 | 1.1 | 2.2 | 2.8 | 2.3 |
| | 20 | 1,4 | 2.2 | 3.1 | 2.4 |
| | 50 | 1.7 | 2.2 | 2.2 | 3.2 |
| | 80 | 1.8 | 2.1 | 2.2 | 4.2 |
| | 95 | 1.6 | 2.0 | 2.6 | 5.3 |
| **VDFR** | 5 | 2.2 | 4.3 | 3.4 | 6.7 |
| | 20 | 1.8 | 2.9 | 3.5 | 5.1 |
| | 50 | 1.7 | 2.7 | 2.2 | 5.0 |
| | 80 | 1.7 | 2.6 | 2.2 | 4.9 |
| | 95 | 1.4 | 2.5 | 2.6 | 4.5 |

[a]RMSECV is estimated by leave-one-production-period-out cross validation to ensure that realistic error estimates are obtained.

Apart from the lighter end of VDTOP and the final boiling points of VDFR in Figure 11, the prediction errors of both fractions are well below the standard deviation in the reference data. So the model can indeed be used for predictions of the majority of the two products, providing that the prediction errors are acceptable. For VDTOP, the prediction errors are in the order of 2°C across the entire curve. Hence, the model predictions of this fraction are as good as, or even better than, the reference method, Table 4. This is clearly acceptable! The model presents somewhat higher prediction errors for VDFR, ranging from 2.5°C to 6.4°C. These errors are generally higher than the reference errors, with the exception of VDFR 20% where the prediction error is slightly below the reference error. However, if a prediction error of about 2.5-3°C is satisfactory, a good model for the

greater part of VDFR is available, see Figure 11. The predictive ability of the model is further demonstrated in Figure 12, where the observations are plotted against predictions taken from the cross validation phase. The diagonal is drawn for easier evaluation of predictive ability.

Overall, the accuracy of the predictions of VDTOP is higher than for the fractions of the AD tower, while it is lower for VDFR. As VDTOP is one of the products in the mixed fraction D10, it is particularly interesting for prediction. VDFR, however, is usually the "slack" fraction for the production mode studied in this project, so good predictive ability of this fraction is not vital. The relationships between the process data and the product properties are interpreted in 6.4.3.

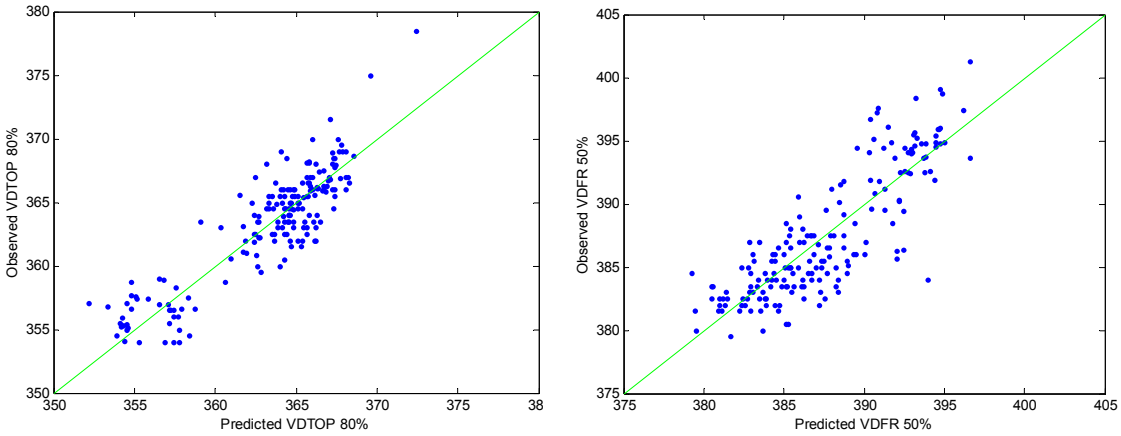



Figure 12 Observed vs. predicted during cross validation, VDTOP 80% (left) and VDFR 50% (right).

### 6.3.6 Discussion

The ability to calculate product quality from the on-line process parameters opens the door to a very interesting opportunity. Basically, it gives the process operators access to a soft sensor of the predicted parameters. Implementation of the models in the process monitoring system would result in new predictions at the same time as, and up to as frequent as, process data measured on-line is sampled. This gives the process operators valuable information on process performance much earlier compared to if they have to wait for the result of time-consuming analyses in the laboratory. In addition to predictive ability, the models can also be used for interpretation of the underlying phenomena that causes the variation in the predicted parameters. Thus they can be used to investigate how a change in process operation affects the product quality and which changes can be made without crossing the specified quality limits of the product.

The example from the demonstration site of this study, where the TBP-curves of the fractions were modelled, shows very good potential. The operators at the refinery have

stressed the importance of on-line sensors for viscosity, to be used on the final products. For them the possibility of getting the whole TBP-curves on-line means not only access to viscosity on-line, but also on-line estimates of all other product quality parameters that can be linked to the shape of the curves, e.g. density, flash point and contents of sulphur. Having this information on-line is a huge upgrade compared to getting a result every eighth hour with a five-hour lag from the sample time. This would definitely speed up the process of reacting to quality changes, since they would immediately find out if the product quality starts to deteriorate. In that way they can act in time and spare themselves of time-consuming and energy demanding re-distillation of product with quality outside of specification.

Key personnel at Nynäs estimates that the yield of the most important product, D10, could be increased by 0.5% absolute (approximately 5% relative) if the TBP soft sensors were implemented on-line. This can be translates into energy savings by the same amount with respect to kg produced D10 product. The economical benefits are also substantial; approximately 4 MSEK/year in increased income is a rough estimate by Nynäs.

## 6.4 Relations between process data and product quality

This section discusses models (PCA and PLS) based on both process data and product property data in the form of TBP curves for the product streams. Using both types of data in the same model gives possibilities to describe the relation between the two groups of variables. The PLS models have already been discussed in terms of predictive ability in 6.3 above.

For the AD tower, a PCA model of a combination of process data and GC TBP data is discussed first. After that, the PLS model that can predict the TBP curves of the ADFR1 and ADFR2 fractions based on 15 min averages of process data is presented and interpreted. For products from the VD tower the PLS model presented in 6.3.4 is interpreted.

### 6.4.1 PCA model of ADFR1, ADFR2 and process parameters

A PCA model of ADFR1, ADFR2 and the process variables measured in the AD tower showed that 6 components gave $R^2/Q^2$ values of 93/90%. Figure 13 shows loadings of the TBP curves for component 1-4, which cumulatively explain 88 % of the variance in the data.

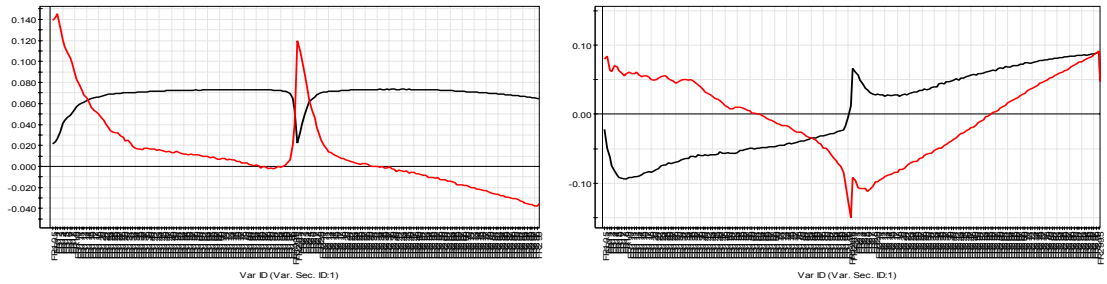



Figure 13 TBP curve loadings for ADFR1 and ADFR2. Left: component 1 (black) and component 2 (red). Right: component 3 (black) and component 4 (red).

The TBP part of the loadings (Figure 13) shows that the first four components describe the following variation in the data:

- Component 1: A general increase in boiling point for both ADFR1 and ADFR2 with less effect on the lighter ends of the TBP curves.
- Component 2: An increase of the boiling point of the lighter ends of both ADFR1 and ADFR2, i.e. flash point.
- Component 3: Temperature profile of both fractions apart from the lighter ends. Also the fractionation between ADFR1 and ADFR2. The way it is drawn here the component describes cleaner cut between the fractions.
- Component 4: A span change of the boiling point intervals. When the interval of ADFR1 is decreased the interval of ADFR2 increases and vice versa.

Since the PCA model also includes process variables it is possible to derive the relation between them and the four phenomena described above. This was investigated in detail but is not presented here due to the similarities between those results and the results of the PLS modelling presented below. Differences between the two models are pointed out when relevant.

### 6.4.2 PLS model of ADFR1 and ADFR2 from process parameters

The PLS model for prediction of product properties is interpreted in this section. Loadings of the response parameters for the four components of the PLS model, Figure 14, are somewhat different from those obtained for the TBP variables in the PCA model discussed above:

- Component 1 (black) describes a change in the whole TBP curve of both ADFR1 and ADFR2, except for the lighter ends that are not affected. The way it is drawn here, the component models a decrease in temperature of the TBP curves.

- Component 2 (red) describes a change in the whole TBP curve of both ADFR1 and ADFR2. The way it is drawn here, the component models an increase in temperature of the whole TBP curves.
- Component 3 (green) contains an increased boiling point for all of ADFR1 and the same for ADFR2 but with the greatest effect in the beginning of the curve (5-40 w%).
- Component 4 (blue) describes a shrinking of the boiling point interval of ADFR1 (with approximately unchanged average point), and decreased boiling point of the entire ADFR2 curve.

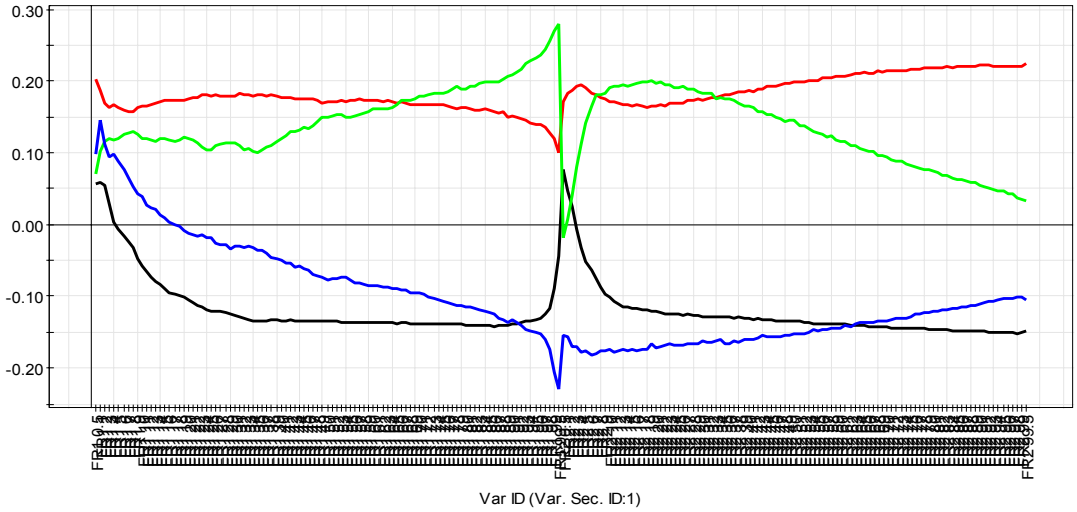



Figure 14 Response parameter loadings of a PLS model with four components (in the order black, red, green and blue). The first 101 variables represent ADFR1 (to the left) and the next 101 variables represent ADFR2 (to the right).

Comparing these results with those obtained with the PCA investigation above, it can be concluded that the second component in the PCA model covered only the lighter ends of both fractions. This is very important from a process optimisation view, since the flash point is an essential product quality factor. No single component in the PLS model describes the same variation, but by adding components 1 and 2 a new loading is obtained, which is only significant for changes of the lighter ends. This should be taken into consideration when interpreting the PLS model.

**The first principal component** explains 42% of X and 29% of Y. The component describes a decrease in the boiling points of both ADFR1 and ADFR2, with the exception of the lighter ends of the fractions, black curve in Figure 14. This component is significantly influenced by the feed flow rate, as can be seen by the similarity between the feed flow rate and the first score (scaled to enable comparison in the same chart) in Figure 15. It can be noted that the first component in a PCA model of only process parameters from the AD tower was also dominated by the feed flow rate (not shown here).

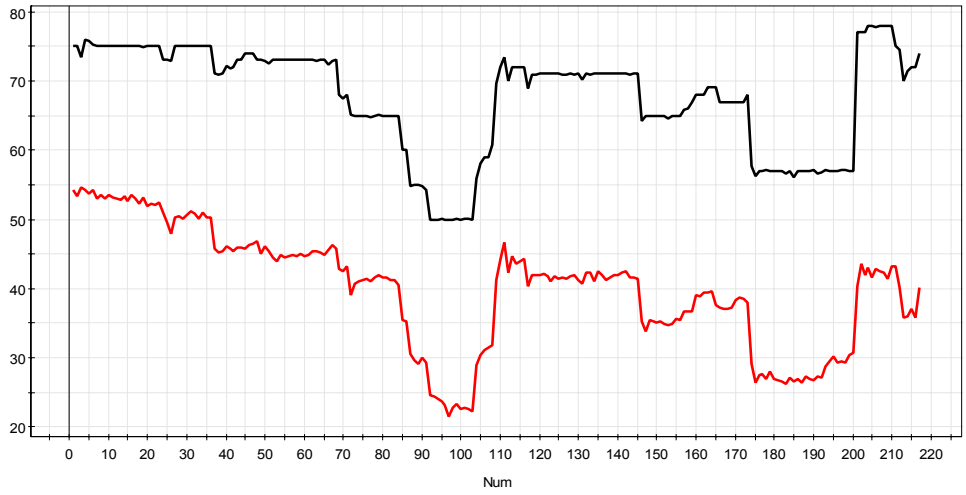



Figure 15 Feed flow rate (black) and a scaled version of the first score of the PLS model (red). Scaling: $-2*t_1 + 40$.

The following list indicates the process variables most important for the first component. Variables marked with (+) are associated with increased boiling points, i.e. lower score on the first component.

- High flow of crude oil in (+)
- Somewhat increased yields of ADFR1 and ADFR2 (but not ADTOP) (+)
- High temperature of ADFR2 to cistern (+)
- Pressure ADTOP (+)
- Steam flow to C3 and C4 (absolute, not relative product flows) (+)
- Low temperatures ADTOP extraction, ADTOP reflux and ADTOP after heat exchangers (-)
- Water level in C5-A

Some observations can be made:

- The feed flow rate affects the TBP of both ADFR1 and ADFR2 without any effect on the temperatures measured at the extraction point. This is difficult to explain.
- From other analyses of the data, the pressure at the top of the AD tower is known to closely follow the feed flow rate. Thus it is reasonable that this variable also is of importance for the component.
- The temperature in the top of and in the reflux to the top of the tower decreases without affecting the yield of ADTOP. This could be a result of more water in the reflux, on account of poor separation in C5-A, which also would increase the pressure.
- The reduction in temperature of ADFR2 flow to cistern could possibly be explained by the heat exchange involving the feed flow. The higher feed flow rate could result in more efficient cooling of the ADFR2 flow.

The interpretation of the first component is that the feed flow rate has an effect on the TBP of both ADFR1 and ADFR2 without causing major changes in the temperatures at the extraction points of these fractions. However, the temperature at the top and in the reflux to the top does decrease which could result from a poor separation of water in C5-A.

This has the following direct relations to economy and quality:

- The TBP curves of ADFR1 and ADFR2 are increased/decreased with the exception of the lighter ends.
- The addition of steam to the bottom of the AD tower relative the feed flow decreases slightly at higher feeds, i.e. the increased feed is not fully compensated for in the steam flows.

**The second principal component** explains 14% of X and 22% of Y and models an increase in TBP over the entire curves of both ADFR1 and ADFR2, red curve in Figure 14. Scores of this component display mostly long-term variation and also a large shift between the periods before and after the winter stop, see Figure 16.

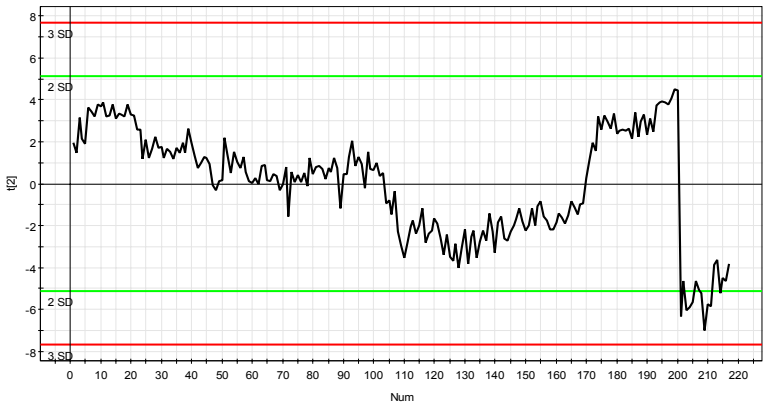



Figure 16 Scores of component 2 in the PLS model of the AD fractions.

Variables marked with (+) in the following list are associated with higher scores, i.e. higher values of the TBP curves, and vice versa.

- Increased yields of the fractions ADTOP, ADFR1 and ADFR2 (+) (i.e. decreased flow from the bottom of the AD tower)
- Increased temperatures at extraction of ADTOP, ADFR1, ADFR2 (+)
- Increased density of ADFR1 and ADFR2 (+)
- Decreased gas flow from C5-B to furnace (-)
- Decreased level of water in C5-A

It is logical that the fractions of the AD tower get higher boiling points when a larger part of the crude oil is extracted in this tower. The higher boiling points of ADFR1 and ADFR2 are naturally related to the increase in density for these fractions. Not

surprisingly, higher boiling point on ADFR1 and ADFR2 is associated with higher temperatures in the AD tower.

A possible interpretation of the second component is that this component models a manner of process operation with a higher yield from the AD tower than usual, which has the effect of increased boiling point of the AD fractions. This is done without any major changes in the steam supply to this tower. Only a small change in the steam supply to the bottom of the tower has been noted, where the steam flow relative the feed flow is somewhat decreased. The direct relations to economy and quality are:

- The entire TBP curves are increased, with the effect of e.g. increased flash point.
- No major changes in steam supply to the AD tower.

**The third component** (10% in X, 8% in Y) models an increase of the entire TBP curve of ADFR1 and an increase of ADFR2, mainly in the lighter end, *cf.* the green curve in Figure 14. The effect on ADFR2 is that the 50 w-% point is shifted towards a higher temperature. The scores in Figure 17 display long-term variation with only small effects from sample-to-sample variation.


Figure 17 Scores of component 3 in the PLS model of the AD fractions.

The following process parameters are of importance for the third component. As previously, the variables marked with + are associated with a high score, i.e. high values of these are associated with higher boiling points of ADFR1 and ADFR2.

- ADFR1 yield (+)
- Density of ADFR1 and ADFR2 (+)
- Steam flow to C3 relative ADFR1 flow (-)
- Temperature of ADFR1 extraction and in C3 (+)
- Temperature of ADFR2 extraction and in C4 (+)

The increased densities are consistent with the increase in boiling point of the fractions. There is no compensation of steam flow to C3 with respect to the increased yield of ADFR1[9].

The interpretation of the third component is that the component models a higher yield of ADFR1 without any increase in steam flow to C3. This increase in yield causes higher boiling points on the entire ADFR1 curve as well as on the lighter end of ADFR2, which is closest to ADFR1. Direct relations to economy and quality are:

- The boiling points of ADFR1 and ADFR2 are affected.
- Steam flow to C3 is not increased even though more ADFR1 is drawn from the tower.

**PLS component 4** (9% in X, 6% in Y) describes a reduction of the temperature span of the ADFR1 TBP curve and a decrease in boiling point of the entire ADFR2 curve, see the blue curve in Figure 14. The scores of this component contain a larger portion of short-term variation compared to previous components, particularly after the winter break, see Figure 18. It is possible that this is an effect of the oscillations in the AD tower discussed earlier.

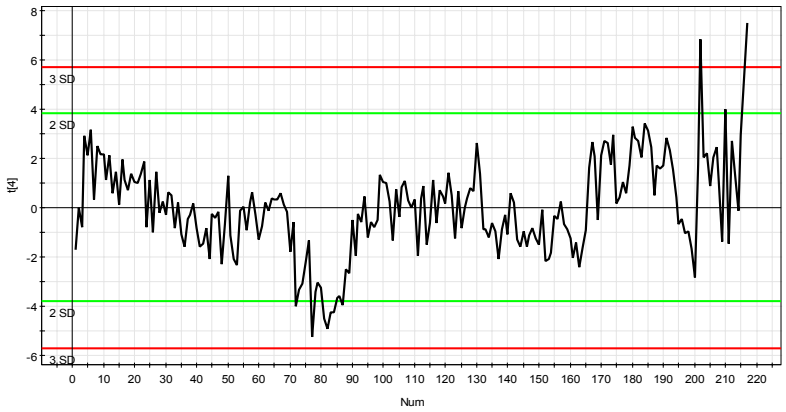



Figure 18 Scores of component 4 in the PLS model of the AD fractions.

The following process parameters have a large impact on the fourth component. As previously, the variables marked with + are associated with a high score, i.e. high values of these are associated with lower boiling points of ADFR2.

- Yield ADFR1 (-)

[9] The fact that the component is not influenced by other yield parameters than ADFR1 or the feed of crude oil would imply that this component is similar to the third component of the PCA model based only on process parameters (not shown here). This is not the case, however, which indicates either that part of the information in the third component of the process data model has already been extracted in earlier components of this model or that there are differences in variation depending on the different time scales. The later would mean that there exists short-term variation that is not modelled by the PLS model on account of the relatively low sample rate.

- Yield ADTOP (+)
- Temperature of ADTOP extraction and C4 (-)
- Water level in C5-a (+)

A possible interpretation of the fourth component is that the component models a shift between the yields of ADTOP and ADFR1 respectively. Lower yield of ADFR1 results in a shorter temperature span of the TBP curve of this fraction. This has an effect on ADFR2, which gets lower boiling points (also visible in the temperatures at the extraction point). There are no major changes in the magnitude of steam flows.

**Interpretation of regression coefficients**. Some observations can be made from the regression coefficients for different fractions of ADFR1 and ADFR2 (not shown)[10].

- The 5-95% points of both ADFR1 and ADFR2 have similar regression coefficients, indicating a great deal of co-variation among the responses. This was also evident in a PCA model of only the TBP curves (not shown here), where 84 % of the variation in ADFR1 and ADFR2 was captured by a single component.

- The feed flow of crude oil does not affect prediction of any of the responses.

- The ADFR1 yield is of great importance to prediction of all responses, including the boiling points of ADFR2.

- Densities of both ADFR1 and ADFR2 flow to cistern are important for the prediction of all responses.

- The relative steam flows to C3, C4 and the bottom of the AD tower have small coefficients for all responses. It would be fair to expect a causal link between steam flow to the side strippers and flash point of the fractions, but this is not reflected in the regression coefficients. This is possibly explained by the fact that enough steam to increase flash points to a desired level is usually used in the process during the period on which the model is based.

- The regression coefficient for the water level in C5-A is high for all response parameters. The only information contained in this variable is a change in set point

[10] Regression coefficients should be interpreted with caution, which has been demonstrated by Alison Burnham [26]. In general the regression coefficients do not display causal links because they are basically forced by the fact that they have to be orthogonal to (i.e. independent of) variation in process data not affecting the response parameters. Instead, the coefficients show the variables important for prediction of the response parameters, given the co-variance in the data. This is only the same as the true causal effect when total experimental design is applied to all X parameters. It should also be noted that the degree of explanation in the lighter ends (5 % points) of both fractions are quite bad, which is another reason to be careful in the interpretation of the regression coefficients for these points.

from 60 % to 70 % at 020820, see Figure 19. Can this really affect the process? It can at least be concluded that the reason for this variables large coefficient in the model is that there is a great shift in product characteristics of both ADFR1 and ADFR2 at about this time. Figure 19 also shows the scores of the first component of the PCA model on just TBP curves of ADFR1 and ADFR2 (84 % of explained variance).

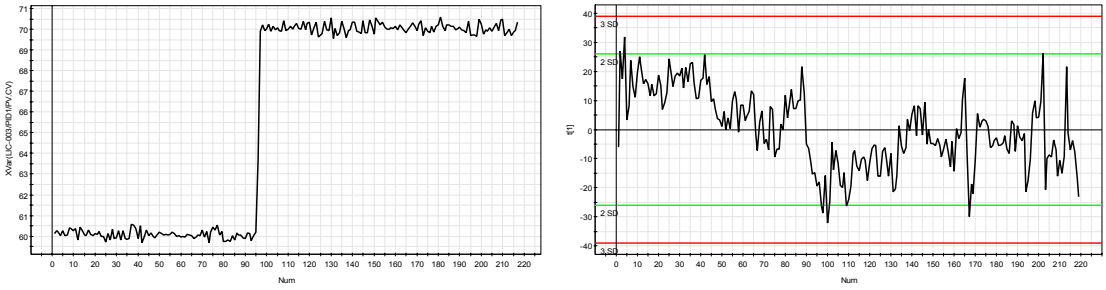



Figure 19 Variation in the water level in C5-A in the calibration data (left). The only significant change is a shift from 60 % to 70 % at sample 96 (020820). Right: the score for component 4 of the PCA model based on only TBP data from the AD tower with a large shift 020811-020819.

### 6.4.3 PLS model of VDTOP, VDFR from process parameters

The YPLS model for prediction of VD product properties is interpreted in this section. It was noted in 6.3.5 that predictive modelling of the fractions VDTOP and VDFR was improved when process parameters from both the AD and the VD tower was included. Figure 20 shows the relative importance of each process block for the six PC's of the model. As expected the operation of both the AD and VD towers have influence in the model, i.e. for the VD tower product properties, but the VD tower is more important.

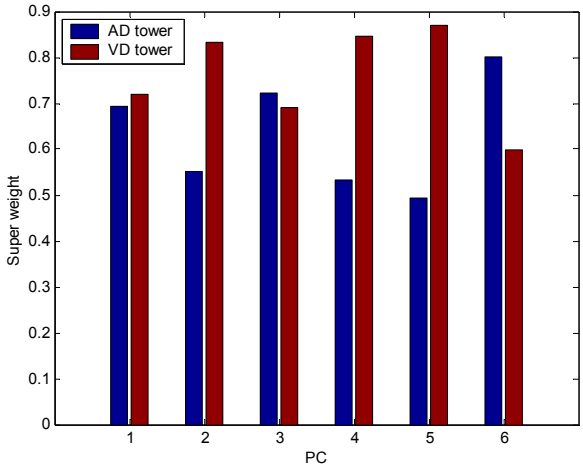



Figure 20 Super weights of the YPLS model for VDTOP and VDFR.

The following variations in the TBP curves are described separately by the model, see loadings of the six components in Figure 21. The percentages in parenthesis are the part of the TBP data explained by that component.

1. (30%) A general increase of boiling temperatures in VDFR and a small decrease in VDTOP, i.e. an effect on the cut between the fractions.

2. (30%) A general increase of the entire boiling point curves, both VDTOP and VDFR, with less effect in the lighter ends.

3. (11%) Increased boiling point intervals of both fractions in combination with less efficient fractionation between VDTOP and VDFR.

4. (5%) The major changes are in the lighter end and middle part of VDTOP. Only small changes to VDFR. Should be related to the amount of low-boiling compounds coming from AD tower.

5. (5%) Increased boiling points of VDTOP and VDFR, except in the heavy end of VDTOP.

6. (5%) The major change is in the heavy end of VDFR, i.e. the component models the cut between VDFR and bitumen.

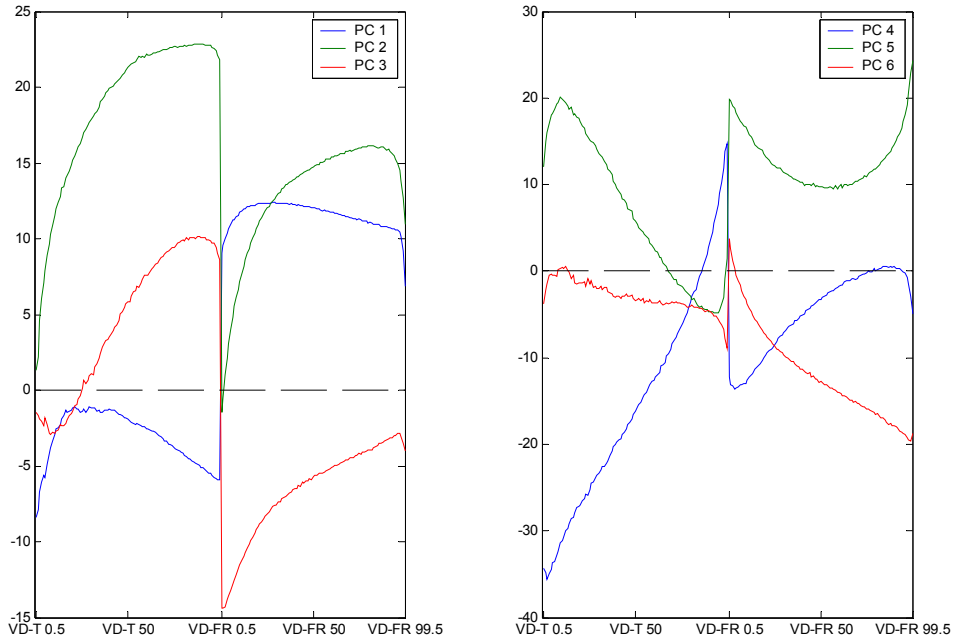



Figure 21 Loadings of the TBP curves, PC 1-3 (left) and PC 4-6 (right).

The relation between the variations in TBP curves and the process variables are interpreted component-wise below. Only the first three components are discussed here on account of them covering the three largest variations in data, cumulatively 71% of the total 86%. Due to its large influence on flash point, PC 4 would have been very interesting to interpret if VDTOP was one of the final products. For the production mode investigated here it is not, since VDTOP is blended with ADFR2 to form D10. Hence, changes in the lighter end of VDTOP are not the most important issue.

**The first principal component** is equally influenced by the process parameters in the AD and the VD tower, Figure 20. It is clear from the super score in Figure 22 that the component is strongly influenced by the crude oil feed rate. This is of course also confirmed by the process variable loadings (not shown). Note from the scores that besides the production rate effect there is also a long-term contribution to the variation modelled by this component.

From the loadings of the VD tower block (not shown), it can be concluded that steam supply to the bottom of the VD tower and to the C19 side stripper are not compensated for increased production rate. This is probably contributing to the effect seen in the VDTOP and VDFR properties. The high loading for VDFR extraction temperature is consistent with the increased boiling point of this fraction. It can also be noted that the reflux of VDFR does not seem to be compensated for increased production rate either, while VDTOP reflux seems to be compensated.

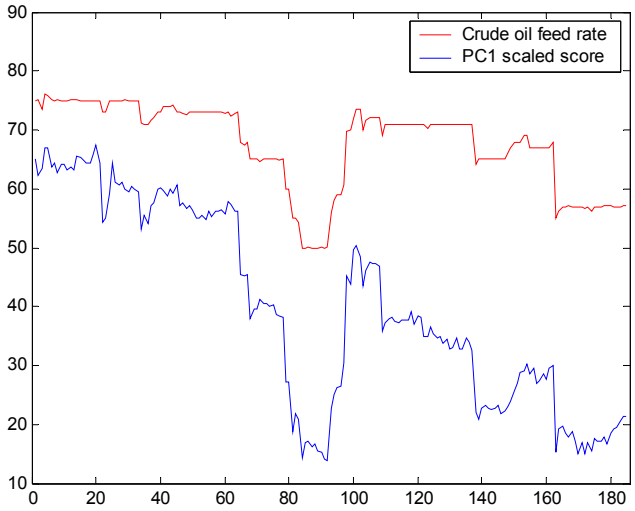



Figure 22 Crude oil feed rate (ton/h) and a scaled version of the super score of PC 1. Scaling: $T_{sup}$*300 +40.

**The second principal component** models mostly long-term variation as can be seen in Figure 23. It should be noted that great leaps in the super score occur not only between

production periods but also within single periods. The VD tower dominates the relative influence on this component, Figure 20, and is therefore the only block used for interpretation below.

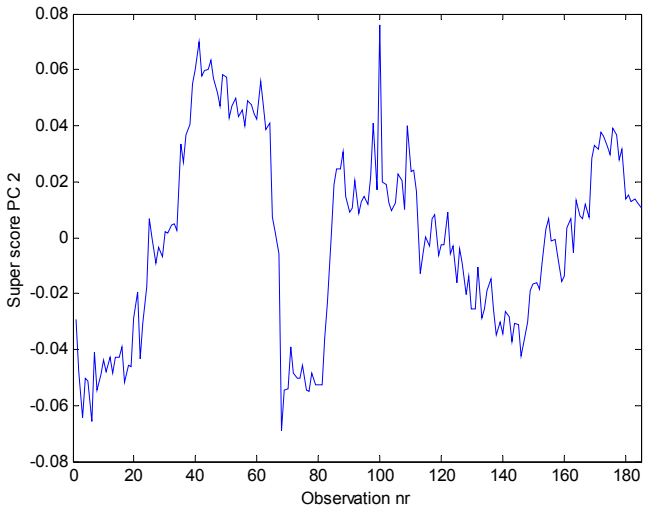



Figure 23 Super score of PC 2.

The following process parameters are of importance for the second component. The variables are marked with +/- to indicate if they are associated with a high/low super score, i.e. high values of the variables are associated with higher/lower values of the entire TBP curves for VDTOP and VDFR (although less effect on the light ends).

In the VD tower block:

- Yield VDTOP (+)
- Temp VD top and bottom (+)
- Temp VDTOP to cistern (+)
- Relative reflux VDTOP (-)

The loadings show that the increase in VDTOP yield is partly due to a decrease in yield of bitumen, which would explain the increase in VD tower temperatures and hence also the increase in TBP that is modelled by this component.

**The third principal component** also models mainly long-term variation, as did both PC 1 and PC 2. Once again great shifts in super score take place within single production periods. The process blocks are equally important for the predictions of this component, Figure 20.

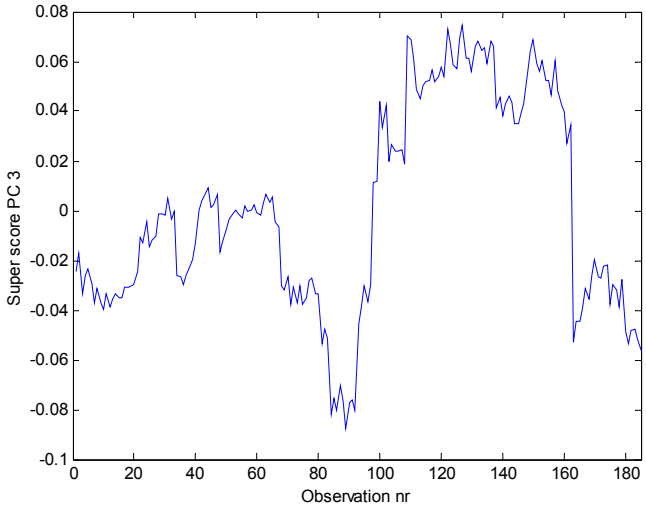



Figure 24 Super score of PC 3.

For the third component the following process parameters are most important. As previously the variables marked with +/- indicate if they are associated with a high/low super score. For PC 3 high super scores means increased boiling point intervals and less efficient fractionation between VDTOP and VDFR.

In the AD tower block:

- Yield ADFR2 (-)
- Crude oil feed rate and related (+)
- Temp ADFR1 extraction +cistern (+)
- Temp ADFR2 extraction + cistern (+)
- Temp side strippers C3 and C4 (+)
- Relative steam flow side strippers C3 and C4 (-)

In the VD tower block:

- Feed rate to VD tower (+)
- Temp VD top and bottom (+)
- Temp VDTOP all measurements except to cistern (+)
- Temp VDFR extraction (-)
- Temp VDTOP after E8 (+)
- Temp bitumen after E5 (+)
- Relative reflux VDFR (-)

The effects of production rate on the process AD tower is discussed in 6.1. It was concluded that the steam supply to the side strippers is not compensated for the increased production rate, which leads to lower flash points of ADFR1 and ADFR2. It can be concluded from the AD tower block loadings that the same phenomenon in the AD tower is modelled by this component. The variation in the VD tower is more difficult to interpret. It is clear that the reflux of VDFR is not compensated for production rate and hence extraction of VDFR, which changes the reflux ratio with the production rate. The

changes in extraction temperatures of VDTOP and VDFR are consistent with the changes in TBP curves for the fractions.

## 6.5 Interpretation of models of the full process

This section discusses some interpretations of a hierarchical model developed in this project. There are several other different interesting interpretations but they are not discussed here to limit the length of this report. All interpretations of the hierarchical model reflect the interactions of different process sections, since this is the main advantage of this type of model. The models are based on 9 blocks including both GC TBP data and process data. Thus they reflect both process state and its influence on product properties.

- TBP curve for ADFR1
- TBP curve for ADFR2
- TBP curve for VDTOP
- TBP curve for VDFR
- Incoming crude oil (heat exchangers)
- AD tower
- VD tower
- AD furnace
- VD furnace

### 6.5.1 Increases of yields and effects

Several components of the hierarchical model contain a large contribution from production rate, since this has a large influence on the process as has been shown in the individual models of process sections. When optimising a process on-line, the production rate is often not possible to adjust since it is given by the current market demand for the product or the supply of crude oil.

To facilitate optimisation (in particular with regard to D10 yield) at constant production rate, the first three components were rotated[11] to eliminate effects of changing production rate. The elimination reduces the degrees of freedom by one and the result is two new combinations of the components, denoted combination 1 and combination 2 below.

The interpretation of the first combination is (very briefly):

- Changes in yields of extracted fractions influence both the yield of D10 (+0.5%) and the product properties (see Figure 25 left).

- The changes also include a decrease of fuel consumption in the furnaces by a total of 2.2% (of the 950 kg/h normally used) according to some assumptions. One reason for

[11] Rotation in this context means that linear combinations of the components are formed.

this is probably that the temperature of the streams that are heat exchanged with the incoming crude oil changes.

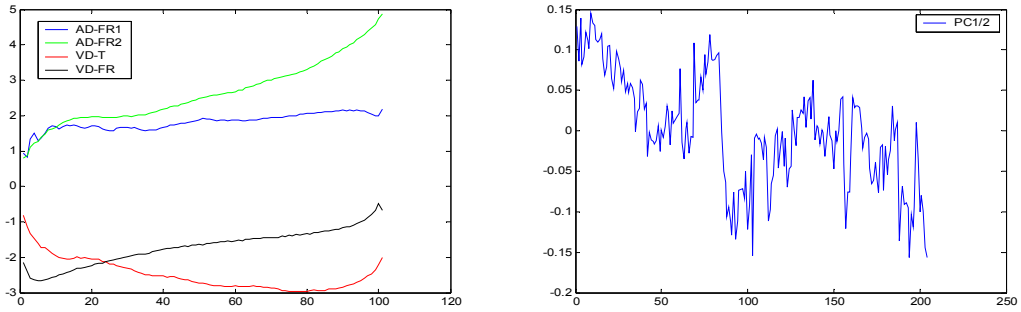



Figure 25. Left: the influence on TBP curves of the first combination scaled to reflect the magnitude as standard deviations (in degrees C). Right: the super scores for the first combination.

The interpretation of the second combination that eliminates production rate is (very briefly):

- Changes in yields of extracted fractions influence the yield of D10 (+0.5%) primarily by increasing the yield of VDTOP.
- The changes also influence the product properties (see Figure 26 left). When the D10 yield is increased, the boiling points of all product streams are increased (examples as one standard deviation: ADFR2 5% +1°C, ADFR2 50% +2°C, VDTOP 50% +0.5°c and VDTOP 95% +2°C).
- The fuel consumption in the furnace is increased by 1.5% with the increase in D10 along this direction.
- The primary effect on steam consumption is decrease of the flow to the AD tower (-85 kg/h). Other changes are much smaller.

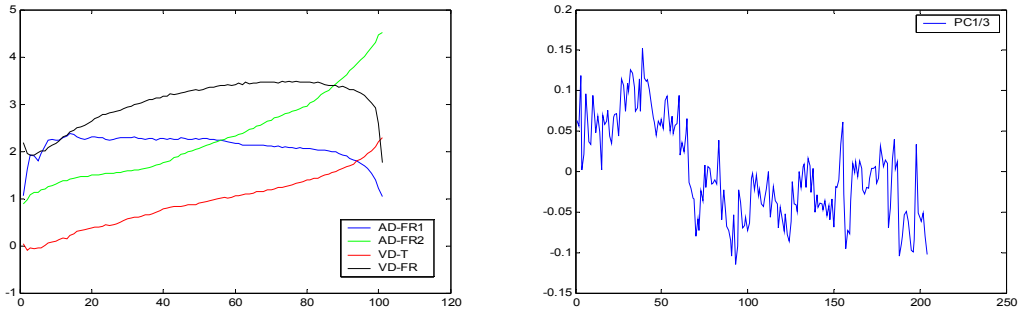



Figure 26. Left: the influence on TBP curves of the second combination scaled to reflect the magnitude as standard deviations (in degrees C). Right: the super scores for the second combination.

In order to investigate the possibilities to increase the yield of D10 with acceptable product properties or to save energy it is beneficial to investigate these components simultaneously in scatter plots. The model is visualised in Figure 27, Figure 28 and Figure 29 with different groups of variables in different plots not to make the figures to crowded. Note that (Figure 27):

- Movement along the top-left to bottom right diagonal influences D10 yield by the yields of VDTOP and ADFR1, which is the way yield optimisation is normally done by the process operators.

- Movement along the bottom-left to top-right diagonal does not influence the yield of D10 and changes in this direction can thus be used to optimise the process with respect to energy consumption and product quality at constant D10 yield.

The discussion below focuses on changes in these two directions and quantifies effects on yield, energy consumption and product properties. All quantified effects of changes in the discussion below are given as 2 standard deviations of the normal operation variation[12] (except for boiling point changes that are given as 1 standard deviation). Two standard deviations correspond to a change from the mean operating conditions to the extreme (95% interval end-point) in that direction alternatively a change from the extreme to the mean point. A movement from one extreme point to the other in that direction thus means that an effect twice as large as the one indicated below is encountered. It is important to note that no assumption is made that it is actually possible to move outside the domain that the process has been operated in during the year studied. In some cases this is probably not true but we choose to show a conservative estimate of the changes possible to make.

[12] This is accomplished by taking into account the contribution of that variable to the variation modelled by the component in question as well as the normal variability of the process in that direction. This means that the covariance is accounted for and that unrealistic combinations of process settings are excluded from the analysis.

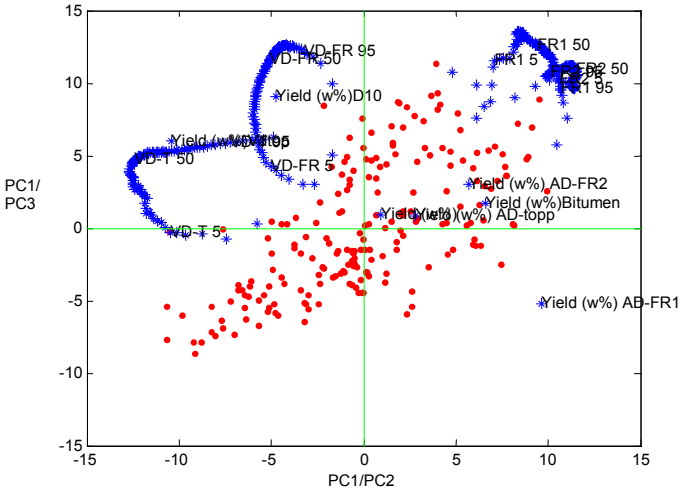



Figure 27 Loadings for the combinations where production rate has been eliminated. Only yields and product property data are shown in this figure. Red dots are scores showing the operation domain during June 2002-June 2003.

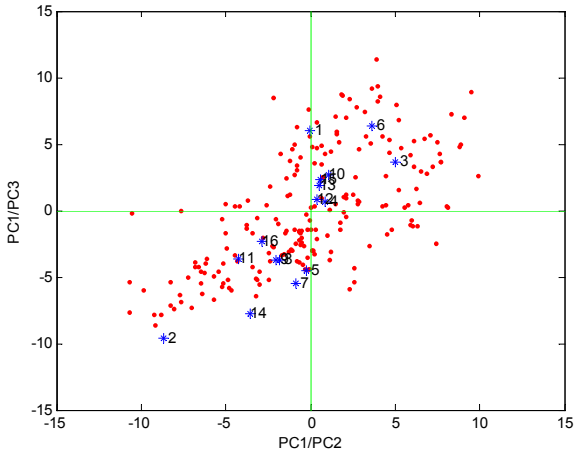



Figure 28. Loadings for the combinations where production rate have been eliminated. Only variables related to the furnaces are shown in this plot: 1 Temp crude oil before E1, 2 Temp crude oil after E9C, 3 Mass flow fuel AD-front, 4 Mass flow fuel AD back, 5 Steam flow from steam generator, 6 Temp AD furnace, 7 Temp feed flow AD tower, 8 air/fuel ratio AD front, 9 air/fuel ratio AD back, 10 Mass flow fuel VD front, 11 Mass flow fuel VD back, 12 Flow AD tower to VD furnace, 13 Temp VD furnace, 14 Temp after VD furnace, 15 air/fuel ratio VD front, 16 air/fuel ratio VD back.
Red dots are scores showing the operation domain during June 2002-June 2003.

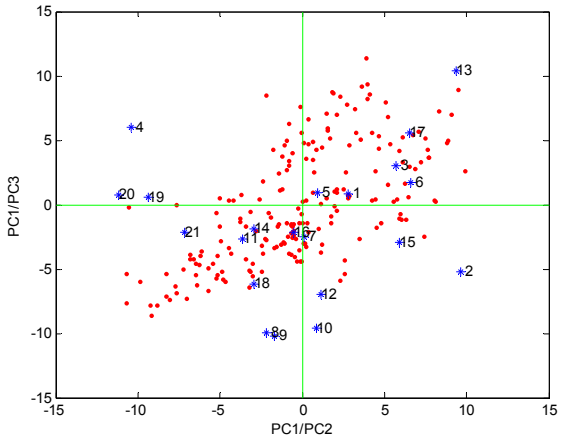



Figure 29. Loadings for the combinations where production rate has been eliminated: 1 Yield AD-TOP, 2 yield ADFR1, 3 yield ADFR2, 4 yield VDTOP, 5 yield VDFR, 6 yield Bitumen, 7 Temp AD bottom, 8 Temp AD top, 9 Temp ADFR1 extraction, 10 Temp ADFR2 extraction, 11 Relative reflux AD-TOP, 12 Steam flow AD bottom, 13 Steam flow C4, 14 Reflux VDFR, 15 Reflux VDTOP, 16 Steam flow VD bottom, 17 Steam flow C19, 18 Level VDFR extraction deck, 19 Temp VD bottom, 20 Temp VDTOP, 21 Temp VDFR extraction.
Red dots are scores showing the operation domain during June 2002-June 2003.

Movement along the top-left to low-right diagonal in Figure 27, Figure 28 and Figure 29 does influence the D10 yield as well as energy consumption according to the following.

- The D10 yield is increased by 0.6% mainly by increasing the yield of VDTOP at the expense of ADFR1. The change in yield along this direction of variation is clear from Figure 30.
- The changes on fuel consumption in the furnaces along this direction are non-significant.
- The main steam flow to be considered is the flow to the AD tower. It is decreased by approximately 10% when the D10 yield is increased. Other steam flows are not influenced significantly.
- The influence on product properties is shown in Figure 31. It can be noted that the AD tower fractions are not influenced significantly while the boiling point of the heavy end of VDTOP is increased by 4°C when the D10 yield is increased 0.6%. This is a natural consequence of the increased yield from the same raw material. The VDFR properties are also changed but this is less important from a process optimisation point-of-view.

It should be noted that the D10 yield is not increased more in this direction than in the combinations discussed above. This is a consequence of the fact that the process variation in this direction is weak as can be noted in e.g. Figure 27 and that the estimates are done based on the actual variation in the data. Thus, they are conservative in that it is assumed

that domains of operation not visited before are likely to be infeasible. If larger variations in this direction can be made without producing out of specification D10, the savings can be even larger than indicated.

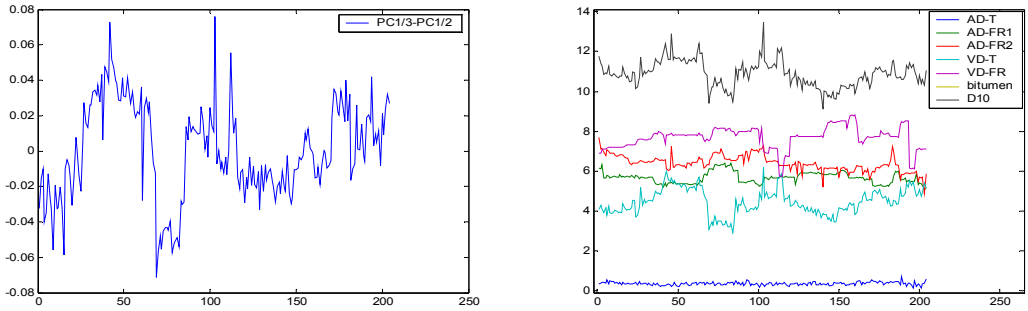



Figure 30. Left: "Score" for the top-left to low-right diagonal movement during June 2002-June 2003. Right: the yields of the different products during the same period. It is very clear that the component models D10 yield very well.

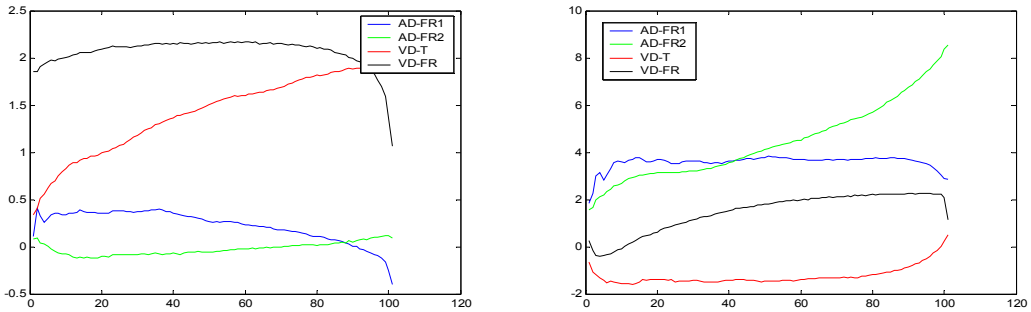



Figure 31. Quantified influence on TBP-curves of fractions from the AD and VD tower as one standard deviation of normal operating conditions (shown by red dots in Figure 27, Figure 28, Figure 29 above).
Left: movement along top-left to bottom-right diagonal in Figure 27, Figure 28, Figure 29
Right: movement along bottom-left to top-right diagonal in Figure 27, Figure 28, Figure 29

Movement along the bottom-left to top-right diagonal in Figure 27, Figure 28 and Figure 29 does not significantly influence the D10 yield but influences energy consumption and product properties according to:

- The total fuel consumption is decreased by approximately 4% when moving from the centre to the lower left part of Figure 27, Figure 28 and Figure 29.
- The steam flows are increased by approximately 5% when moving from the centre to the lower left part of Figure 27, Figure 28 and Figure 29.
- The changes in product properties are mainly for ADFR2 as visualised in the right part of Figure 31. Also VDTOP is influenced but in the other direction.

Thus, to summarise variation in this direction, it is possible to save fuel in the furnace (-4%) but it costs steam (+5%) and gives a lower boiling ADFR2 and a higher boiling VDTOP, which means that the boiling point range of D10 is increased.

The discussions in this section have shown large potential for increased yield of D10 and energy savings by optimisation of operating conditions. Due to fluctuations in the process, it is not likely that this change is possible to perform in reality to a full extent for long periods of time. However, the model clearly shows that there is a large fuel saving potential at constant or increased yield of D10 if the changes in product properties can be tolerated. Process maps, as the one shown in Figure 27 above, available on-line to the process operators would be a valuable tool to achieve the potential benefits.

### 6.5.2 Fuel consumption in furnaces

As a result of the hierarchical modelling, attention was drawn to fuel consumption in the furnace. The model reveals the relationship between fuel consumption in the AD and the VD furnace as well as their relation to other process parameters such as crude oil feed flow and the effect of the heat exchangers.

An interesting observation is that the relative fuel consumption in the AD furnace, *i.e.* kg fuel per ton crude oil, tends to increase with increased production rate whereas the opposite is true for the relative fuel consumption in the VD furnace, see Figure 32. The effect is about 1 kg per ton over the entire production rate span in both cases, not taking into account a period where the fuel consumption was extremely low in the AD furnace and extremely high in the VD furnace. However, looking at the total fuel consumption in the entire furnace, *i.e.* AD and VD together, it is clear that this has no obvious correlation to production rate, Figure 33 left. Apparently the feed rate induces a rearrangement of the energy input to the furnace while keeping it on basically the same level. This is examined in more detail later.

The effect of chemical cleaning of the furnace coils during the production stop in the winter 2002/2003 is clear, see Figure 32 and Figure 33. The total relative fuel consumption is significantly lower after furnace cleaning, ranging between 11.5 and 12 kg per ton compared to 12.5-14 kg per ton, Figure 33 left. It is apparent from Figure 32 that the cleaning mainly affects the AD part of the furnace. This is not strange since the two coils in the AD furnace passes through the furnace gas at the top of the furnace, where the build-up of soot on the coils are likely to be higher than in the rest of the furnace. The gradual build-up of soot in the furnace and the effect of chemical cleaning are easy to follow by looking at the total relative fuel consumption over time, Figure 33 right. Note that even though this time trend obviously displays the build-up of soot in the furnace over time, it is important to remember that this is not the only factor influencing fuel consumption in the furnace. Changes in fraction yields can affect the fuel

consumption, which was discussed in the previous section, Increases of yields and effects. There are also indications that the outside temperature affects the fuel consumption, especially at very low temperatures *i.e.* in wintertime, Figure 34.

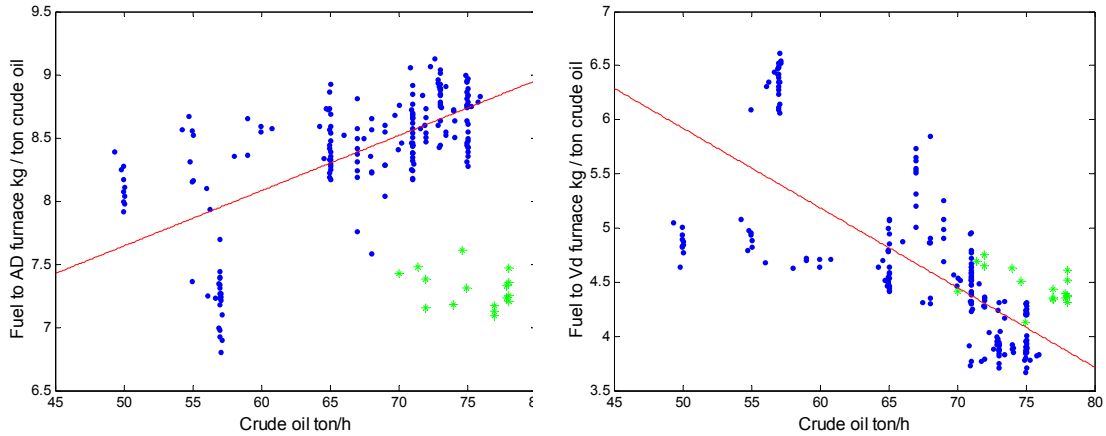



Figure 32 Relation between relative fuel consumption and crude oil feed rate, AD furnace (left) and VD furnace (right). Data before winter stop, blue dots and linear regression, data after winter stop, green stars.

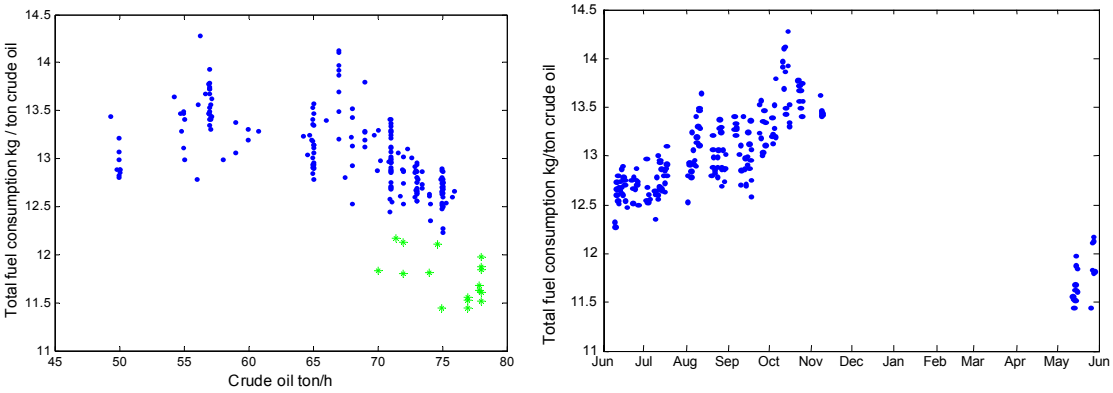



Figure 33 Relation between total relative fuel consumption and crude oil feed rate (left). Data before winter stop, blue dots, data after winter stop, green stars. Total relative fuel consumption, variation in time, (right).

The hierarchical model gives insight to why the fuel consumption in the AD furnace increases at high feed rates. The *Raw material* block shows that lower temperatures in the chain of heat exchangers are related to higher feed rate of the crude oil, see Figure 35. Hence, the crude oil entering the AD furnace is essentially colder at higher feed rates, which causes an increased consumption of fuel in order to reach the temperature set-point on the feed flow to the AD tower.

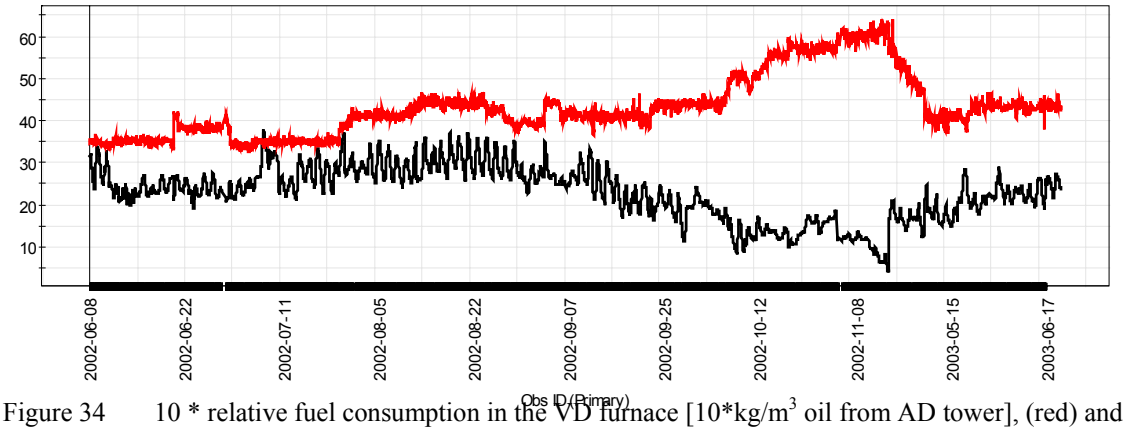



Figure 34 10 * relative fuel consumption in the VD furnace [10*kg/m$^3$ oil from AD tower], (red) and temperature on combustion air to furnaces [°C], (black). Note that the time line consists of several discrete periods that have been merged together. Data from 2003 starts at the noticeable shift in the black curve.

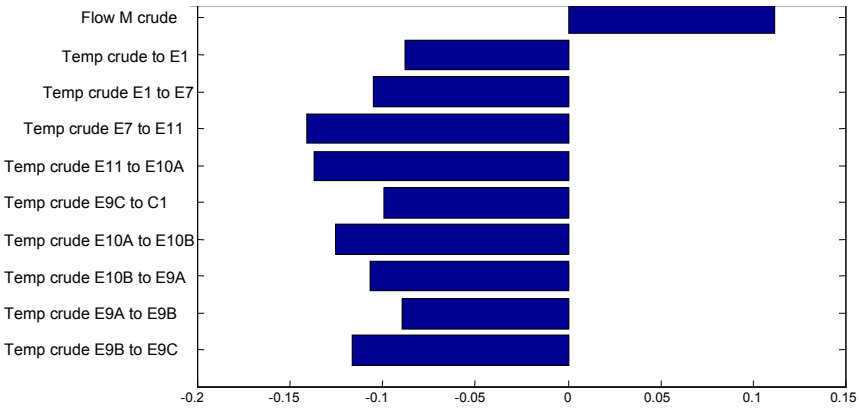



Figure 35 Loadings of the process block *Raw material* in the hierarchical model.

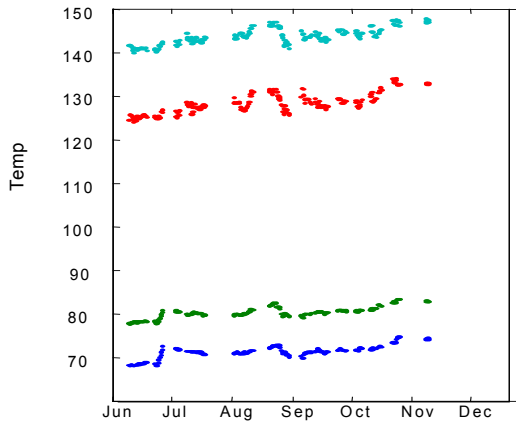



Figure 36 Temperature of crude oil before E1, E7, E11 and E10A.

Variation in crude oil temperature is enhanced in the chain of heat exchangers, Figure 36. Only the first four heat exchangers are shown here but the effect is apparent throughout the whole chain. There is approximately a difference of 5 °C in crude oil temperature after the first heat exchanger (E1) at minimum and maximum feed rate. By the time the crude oil passes the final heat exchanger the difference in temperature between minimum and maximum feed rate has increased to about 20 °C. The effect can be explained by less efficient heat exchange at higher feed rate, but also by the fact that the product streams, which the crude oil is heat exchanged against, are colder at higher production rates. This is true for both ADTOP (E1), VDTOP (E7) and bitumen (E10A).

The model also shows that the temperature in the bottom of the AD tower, from which the feed flow to the VD furnace comes, is higher at higher production rates. This can explain the reduced fuel consumption in the VD furnace.

Summarising the discussions above:

- Build-up of soot on the furnace coils, particularly in the AD furnace, gradually lowers the efficiency of heat transfer to the crude oil and increases the relative fuel consumption. From June to November 2002 there is a gradual increase of the total relative fuel consumption with approximately 1 kg per ton.

- Chemical cleaning of the furnace has a clearly visible effect on the fuel consumption. The total relative fuel consumption is reduced by at least 1.5 kg per ton after the chemical cleaning in the winter 2002-2003.

- At higher production rate the relative fuel consumption in the AD furnace increases, while the relative fuel consumption in the VD furnace decreases. The effect is about 1 kg fuel per ton crude oil over the entire production rate span.

- The effect of the heat exchangers is lower at higher production rate. Crude oil temperature differences at minimum and maximum feed rate ranges between 5 and 20 °C, with larger differences at the end of the chain of heat exchangers.

# 7 Conclusions and recommendations

The study has shown that steam flow to the AD and VD towers and the side strippers are not fully compensated for differences in production rate. This has been shown to influence product quality, e.g. flash points of ADFR1, ADFR2 and D10 products. It is suggested to add relative steam flows to operator process screens and to use this in process operation. Examples are kg steam to AD bottom per ton crude oil that enters the AD tower and kg steam to each side stripper per ton fraction extracted of the product

going to that side stripper. This simple measure would facilitate correct steam flows and thus, better control of the TBP curves of the products.

There are large oscillations present in the upper half of the AD tower. The decrease in performance is difficult to quantify, but the oscillations contribute to a larger boiling point interval for ADFR1 including a lower flash point and, perhaps more seriously, to make process control more difficult. This leads to more inefficient process operation and higher consumption than necessary of resources and energy. The reason is that, today, process control is mainly based on GC analysis of samples taken three times a day. If these samples are influenced by the oscillation, the action taken by operators will depend on were in the oscillatory cycle the process was when the sample was taken, which is clearly not optimal. The cause of the oscillations cannot be determined because of the high data compression used for some tags in the history database although three possible candidates are identified. It is recommended to increase the quality of data in the history database by decreasing compression for some process data (see 6.2.1).

It is possible to predict product quality in the form of TBP curves for ADFR1, ADFR2, VDTOP and VDFR from process data in real-time. The model accuracy is low for ADFR1 but higher for the more interesting fractions ADFR2 and VDTOP. The prediction errors (RMSECV) differ among the four products with highest accuracy for VDTOP, 2.0-2.2 °C, and lowest for the light end of VDFR, 4.3 °C at TBP 5%. For most of the products the accuracy of the model prediction are similar to or better than the accuracy of the GC method used today. In addition, TBP curves from GC analyses are obtained once every 8 hours with a delay of up to 5 hours from sampling, which makes process control using these curves difficult. Model predictions can be made available on-line in real-time.

On-line implementation of models for prediction of product quality would give entirely new possibilities for process control. This is discussed further in 6.3.6. The interpretation of the models gives some new insights but mainly confirms the knowledge kept by experienced process operators. This lends credibility to the models ability and stability and means that the goal of capturing operator knowledge in models and transfer it to company knowledge is reached.

The full process hierarchical model discussed in 6.5.1 has shown large potential for increased yield of D10 and energy savings by optimisation of operating conditions. Due to fluctuations in the process, it can be difficult to achieve the full theoretical benefit in reality for long periods of time. On-line prediction models for TBP curves of ADFR2 and VDTOP as well as process maps from PCA models would be a valuable tool to achieve the potential benefits. Key personnel at Nynäs estimates that the yield of the most important product, D10, could be increased by 0.5% absolute (approximately 5% relative) if the TBP soft sensors were implemented on-line. This agrees well with the estimate from the hierarchical full process model that indicates 0.6% yield increase. The

yield increase can be translated into energy savings by the same amount with respect to kg produced D10 product. The economical benefits are also substantial; approximately 4 MSEK/year in increased income is a rough estimate by Nynäs.

Build-up of soot on the furnace coils gradually lowers the efficiency of heat transfer to the crude oil and increases the relative fuel consumption. These effects can be clearly seen in the models developed. Chemical cleaning of the furnace has a large effect on the total relative fuel consumption, which is reduced from 13.5-14 kg per ton to approximately 11.5-12 kg per ton after cleaning. This can be used to determine when the cost of chemical cleaning can be motivated by a sufficient decrease in relative fuel consumption.

The effect of the heat exchangers is lower at higher production rate. Crude oil temperature differences at minimum and maximum feed rate ranges between 5 and 20°C, with larger differences at the end of the chain of heat exchangers. This means that less temperature increase is required in the AD furnace at low production rate, which should be taken into account when optimising the process with respect to energy consumption.

## 7.1 Future work

In discussions with process operators and process engineers, possible benefits of putting some of the models developed in this project on-line was investigated. It was agreed that:

- Prediction of the TBP curves would speed up the product quality control significantly.
- PCA models can help monitor current process status and suggest how to steer the process into the most desirable state.

Nynäs current view is that it would be very valuable to put the TBP prediction models on-line and they see potential increases in yield that would correspond to energy savings and improved productivity. In a longer run it would also be interesting to use PCA models for process monitoring on-line but that is currently of lower priority.

The promising results obtained in this project show that it would be very interesting to make process models for other operating modes than the one studied in this project. There is at least two more production modes that are frequently used and where the effort would be worthwhile in our opinion.

Another interesting future issue is process modelling based on data from designed process experiments, which would potentially give even more information than the models developed in this project based on normal operating data. An example of an effect that would be interesting to analyse is the effect of changing tower feed temperatures. The effects of changes in feed temperature cannot be analysed today, since the

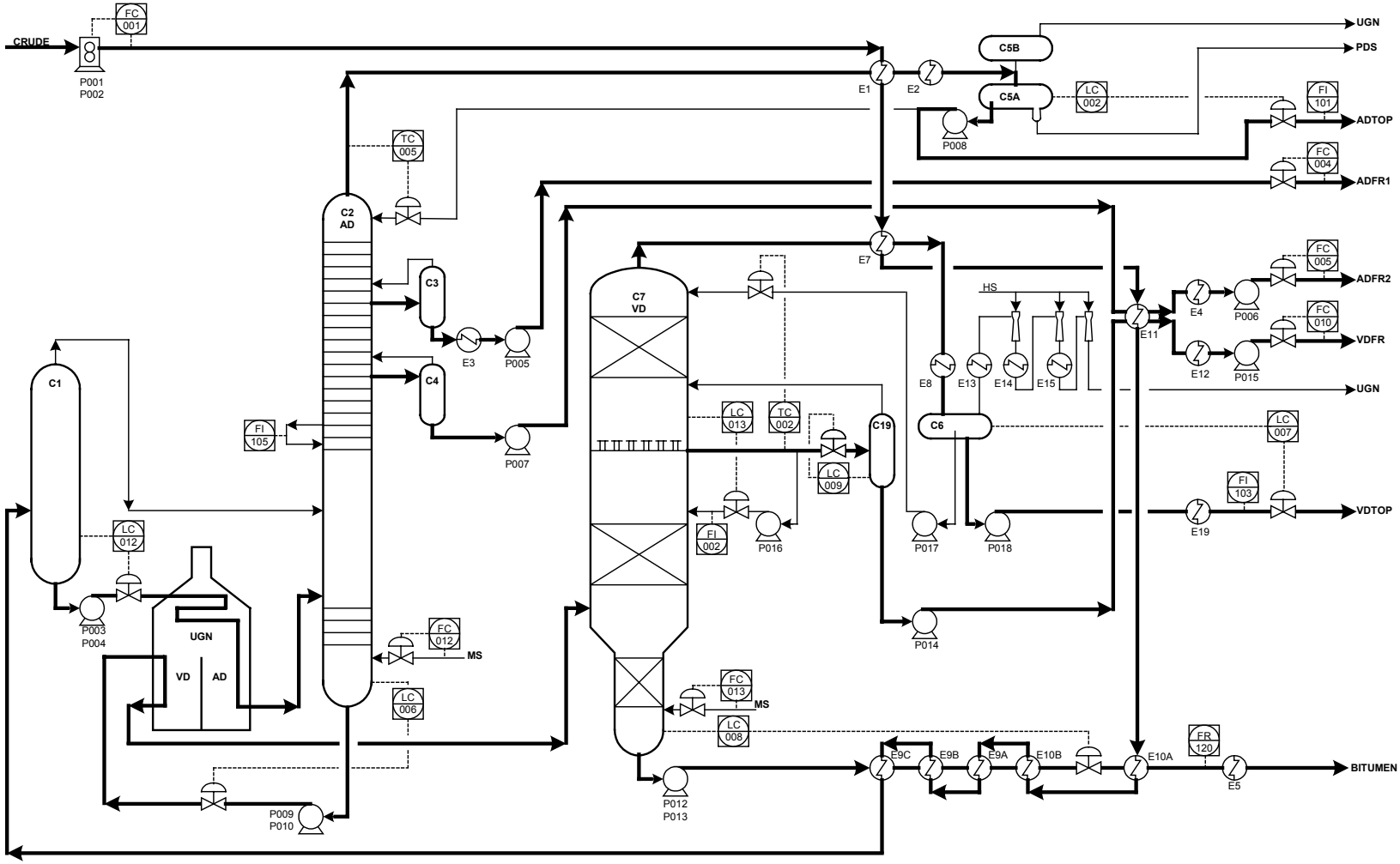
temperatures are controlled tightly, which leads to very low variability in the data. It is, however, not at all certain that the temperatures used are optimal from an energy consumption point-of-view, since it is reasonable to assume that energy in the furnace and the towers can replace each other at least to some extent.

# 8 References

1. Swedish National Energy Administration, Swedish Environmental Protection Agency, 2000. *EMIL2-Energianvändningen inom industrin*. (in swedish)
2. S. Wold, K. Esbensen and P. Geladi, *Principal Component Analysis*, Chemom. Intell. Lab. Syst. 2 (1987), 37-52
3. H. Martens and T. Naes, *Multivariate Calibration*, John Wiley & Sons, Chichester, 1989
4. P. Geladi and B R. Kowalski, *Partial least-squares regression: a tutorial* Anal. Chim. Acta 185(1986), 1-17
5. T. Kourti, *Process Analysis and Abnormal Situation Detection: from theory to practice*, IEEE Control Syst. Mag., October 2002, 10-25
6. B M. Wise and N. B. Gallagher, *The process chemometrics approach to process monitoring and fault detection*, J. Process Control 6(1996), 329-348
7. E L. Russel, L H. Chiang and R D. Braatz, *Data driven techniques for fault detection and diagnosis in chemical processes*, Springer-Verlag, London, 2000
8. J A. Westerhuis, T. Kourti and J F. MacGregor, *Analysis of multiblock and hierarchical PCA and PLS models*, J. Chemometrics 12 (1998), 301-321
9. S J. Qin, S. Valle and M J. Piovoso, *On unifying multiblock analysis with application to decentralized process monitoring*, J. Chemometrics 15 (2001), 715-742
10. J.A. Westerhuis and A. K. Smilde, *Deflation in multiblock PLS*, J. Chemometrics 15(2001), 485-493
11. S. Wold, N. Kettaneh and K. Tjessem, *Hierarchical multiblock PLS an PC models for easier model interpretation and as an alternative to variable selection,* J. Chemometrics, 10(1996), 463-482
12. J F. Mcgregor, C. Jaeckle, C. Kiparissides and M. Koutoudi, *Process monitoring and diagnosis by multiblock PLS methods*, AIChE J. 40 (1994),826-838
13. S. Wold, S. Hellberg, T. Lundstedt, M. Sjöström and H. Wold, *Proc. Symp. On PLS model building: theory and application*, Frankfurt am Main, 1987
14. J A. Westerhuis and P M J. Coenegracht, *Multivariate modelling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares*, J. Chemometrics 11 (1997), 379-392

15. B. Walczak. and D.L. Massart, *Dealing with missing data Part I.*, Chemom. Intell. Lab. Syst. 58 (2001), 15-27.
16. B. Walczak. and D.L. Massart, *Dealing with missing data Part II*, Chemom. Intell. Lab. Syst. 58 (2001), 29-42.
17. T. N. Pranatyasto and S J. Qin, *Sensor validation and process fault detection for FCC units under MPC feedback*, Control Eng. Practice 9 (2001), 877-888
18. H. Vedam and V. Venkatasubramanian, *PCA-SDG based process monitoring and fault diagnosis*, Control Eng. Practice 7(1999), 903-917
19. T. Chatterjee and D N. Saraf, *On-line estimation of product properties for crude distillation units*, J. Process Control 14 (2004), 61-77
20. J. Shin, M. Lee and S. Park, *Design of composition estimator for inferential control of distillation columns*, Chem. Eng. Comm. 178(2000), 221-248
21. Ku, W., R. H. Storer, and C. Georgakis, *Disturbance Detection and Isolation by Dynamic Principal Component Analysis*, Chemom. Intell. Lab. Syst. 30(1995), 179
22. N. F. Thornhill, M. A. A. S. Choudhuryb and S. L. Shahb, *The impact of compression on data-driven process analyses*, J. Process Control 14(2004), 389-398
23. J. Blom, *Utvärdering av reglerprestanda vid ett raffinaderi*, M.Sc. Thesis, Report LITH-ISY-EX-3096, Tekniska Högskolan i Linköping, Linköping, Sweden, 2001 (in swedish)
24. K. Faber and B. R. Kowalski, *Improved prediction error estimates for multivariate calibration by correcting for the measurement error in the reference values*, Appl. Spectrosc. 51(1997), 660-665
25. D. B. Coates, *Is near infrared spectroscopy only as good as the labotratory reference values? - An empirical approach*, Spectrosc. Europe 14(2002), 24-26
26. A J. Burnham, J F. MacGregor and R. Viveros, *Interpretation of regression coefficients under a latent variable regression model*, J. Chemometrics 15(2001), 265–284

# Appendix 1 – Process outline




Courtesy of Nynäs Refining AB.

# Appendix 2 – Parameter lists

This appendix contains the parameters used in the study. They are listed in Table A2-1 to Table A2- 5 according to process block.

Table A2-1. Parameters in the block *Raw material.*

| Variable name | Unit | Comments |
|---|---|---|
| Flow M crude | t/h | |
| Density crude | g/cm$^3$ | |
| Temp crude | ° C | |
| Flow V crude | m$^3$/h | |
| Flow M crude A | t/h | |
| Flow M crude B | t/h | |
| Pressure crude P2 | bar | |
| Pressure crude P1 | bar | |
| Temp crude to E1 | ° C | |
| Temp crude E1 to E7 | ° C | |
| Temp crude E7 to E11 | ° C | |
| Temp crude E11 to E10A | ° C | |
| Temp crude E9C to C1 | ° C | |
| Temp crude E10A to E10B | ° C | |
| Temp crude E10B to E9A | ° C | |
| Temp crude E9A to E9B | ° C | |
| Temp crude E9B to E9C | ° C | |

Table A2-2. Parameters in the block AD furnace.

| Varible name | Unit | Comments |
|---|---|---|
| Damper AD rear | % | |
| Damper combustion air AD furnace | % | |
| Damper in boiler | % | |
| Draught furnace | $mmH_2O$ | |
| Flow air AD front | $Nm^3/h$ | |
| Flow air AD front (rel) | $Nm^3/kg$ | |
| Flow air AD rear | $Nm^3/h$ | |
| Flow air AD rear (rel) | $Nm^3/kg$ | |
| Flow gas C5-B to furnace | kg/h | |
| Flow gas C5-B to furnace (rel) | kg/t | Relative *Flow M crude* |
| Flow lower AD furnace | $m^3/h$ | |
| Flow M crude | t/h | |
| Flow M crude to AD | $m^3/h$ | |
| Flow M fuel AD front | kg/h | |
| Flow M fuel AD front (rel) | kg/t | Relative *Flow M crude* |
| Flow M fuel AD rear | kg/h | |
| Flow M fuel AD rear (rel) | kg/t | Relative *Flow M crude* |
| Flow steam to all burners | kg/h | |
| Flow steam to all burners (rel) | kg/t | Relative *Flow M crude* |
| Flow upper AD furnace | m3/h | |
| Flow V fuel AD front | l/h | |
| Flow V fuel AD rear | l/h | |
| Flow w to boiler | t/h | |
| Gauge pressure combustion air AD furnace | mmH2O | |
| Level boiler | % | |
| Pressure combustion air AD furnace | mbar | |
| Pressure fuel to furnace | bar | |
| Pressure in boiler | % | |
| Pressure scr. gas to furnace | bar | |
| Set point valves AD upper and lower | % | |
| Smoke opacity | % | |
| Steam from boiler | t/h | |
| Steam from boiler (rel) | t/kg | Relative *Flow M fuel AD front + rear* |

| | | |
|---|---|---|
| Steam from boiler (rel c) | | Relative *Flow M crude* |
| Temp AD furnace | ° C | |
| Temp air to all burners | ° C | |
| Temp crude to AD tower | ° C | |
| Temp diff AD upper and lower | ° C | |
| Temp exhaust gas in furnace | ° C | |
| Temp fuel oil | ° C | |
| Temp furnace gas after boiler | ° C | |
| Temp furnace gas after economiser | ° C | |
| Temp furnace gas before boiler | ° C | |
| Temp furnace gas in chimney | ° C | |
| Temp lower AD furnace | ° C | |
| Temp scr gas to furnace | ° C | |
| Temp scr gas to furnace/ off gas | ° C | |
| Temp upper AD furnace | ° C | |
| Temp w after economiser | ° C | |
| Temp w before economiser | ° C | |
| Valve fuel return | % | |
| Valve fuel to AD burners | % | |
| Valve w to boiler | % | |

Table A2-3. Parameters in the block *AD tower*.

| Variable name | Unit | Comments |
|---|---|---|
| Density ADFR1 to cistern | $g/cm^3$ | |
| Density ADFR2 to cistern | $g/cm^3$ | |
| Density ADTOP to cistern | $g/cm^3$ | |
| Density crude | $g/cm^3$ | |
| Flow AD to VD | $m^3/h$ | |
| Flow gas C5-B | kg/h | |
| Flow gas C5-B (rel) | kg/t | Relative *Flow M crude B* |
| Flow lower AD furnace | $m^3/h$ | |
| Flow M ADFR1 to cistern | t/h | |
| Flow M ADFR1 to cistern set | t/h | |
| Flow M ADFR2 to cistern | t/h | |
| Flow M ADFR2 to cistern set | t/h | |
| Flow M ADTOP to cistern | t/h | |
| Flow M crude A | t/h | |
| Flow M crude B | t/h | |
| Flow reflux ADTOP | $m^3/h$ | |
| Flow reflux ADTOP (rel) | | Relative *Flow V ADTOP* |
| Flow reflux ADTOP (rel c) | | Relative *Flow V crude* |
| Flow steam AD bottom | t/h | |
| Flow steam AD bottom (rel) | | Relative *Flow M crude B* |
| Flow steam C3 | kg/h | |
| Flow steam C3 (rel) | kg/t | Relative *Flow M ADFR1 to cistern* |
| Flow steam C4 | kg/h | |
| Flow steam C4 (rel) | kg/t | Relative *Flow M ADF21 to cistern* |
| Flow upper AD furnace | $m^3/h$ | |
| Flow V ADFR1 to cistern | $m^3/h$ | |
| Flow V ADFR2 to cistern | $m^3/h$ | |
| Flow V ADTOP to cistern | $m^3/h$ | |
| Flow V crude | $m^3/h$ | |
| Level AD bottom | % | |
| Level C1 | % | |
| Level C3 | % | |
| Level C3 set | % | |
| Level C4 | % | |

| | | |
|---|---|---|
| Level C5-A | % | |
| Level water C5-A | % | |
| Pressure AD flash zone | bar | |
| Pressure AD top | bar | |
| Pressure C1 | bar | |
| Pressure C1 set | bar | |
| Pressure C5-B | bar | |
| Pressure C5-B set | bar | |
| Slope reflux | $m^3/h$ | |
| Slope reflux (rel) | | Relative *Flow V crude* |
| Temp AD bottom | ° C | |
| Temp ADFR1 ex | ° C | |
| Temp ADFR1 to cistern | ° C | |
| Temp ADFR2 ex | ° C | |
| Temp ADFR2 to cistern | ° C | |
| Temp ADTOP after E2 | ° C | |
| Temp ADTOP ex | ° C | |
| Temp ADTOP reflux | ° C | |
| Temp C3 | ° C | |
| Temp C4 | ° C | |
| Temp crude | ° C | |
| Valve AD bottom | % | |
| Valve ADFR1 ex | % | |
| Valve ADFR1 to cistern | % | |
| Valve ADFR2 to cistern | % | |
| Valve ADTOP to cistern | % | |
| Valve C1 gas to AD | % | |
| Valve C1 to furnace | % | |
| Valve C5-B | % | |
| Valve cooling w E2 | % | |
| Valve ex water C5-A | % | |
| Valve reflux ADTOP | % | |
| Valve steam AD bottom | % | |
| Yield ADFR1 | w% | |
| Yield ADFR2 | w% | |
| Yield ADTOP | w% | |

Table A2-4. Parameters in the block *VD furnace*.

| Variable name | Unit | Comments |
|---|---|---|
| Damper combustion air VD furnace | % | |
| Damper VD front | % | |
| Damper VD rear | % | |
| Flow AD to VD | $m^3/h$ | |
| Flow air VD front | $Nm^3/h$ | |
| Flow air VD front (rel) | $Nm^3/kg$ | Relative *Flow M fuel VD front* |
| Flow air VD rear | $Nm^3/h$ | |
| Flow air VD rear (rel) | $Nm^3/kg$ | Relative *Flow M fuel VD rear* |
| Flow M fuel VD front | kg/h | |
| Flow M fuel VD front (rel) | kg/t | Relative *Flow M crude* |
| Flow M fuel VD rear | kg/h | |
| Flow M fuel VD rear (rel) | kg/t | Relative *Flow M crude* |
| Flow steam to all burners | kg/h | |
| Flow steam to all burners (rel) | kg/t | Relative *Flow M crude* |
| Flow V fuel VD front | l/h | |
| Flow V fuel VD rear | l/h | |
| Gauge pressure combustion air VD furnace | $mmH_2O$ | |
| Pressure combustion air VD furnace | mbar | |
| Pressure fuel to furnace | bar | |
| Temp after VD furnace | ° C | |
| Temp air to all burners | ° C | |
| Temp fuel oil | ° C | |
| Temp VD furnace | ° C | |
| Valve fuel return | % | |
| Valve fuel to VD burners | % | |

Table A2-5. Parameters in the block *VD tower*.

| Variable name | Unit | Comments |
|---|---|---|
| Density VDTOP to cistern | g/cm$^3$ | |
| Flow AD to VD | m$^3$/h | |
| Flow M Bitumen to cistern | t/h | |
| Flow M crude B | t/h | |
| Flow M VDFR to cistern | t/h | |
| Flow M VDFR to cistern set | t/h | |
| Flow M VDTOP to cistern | t/h | |
| Flow reflux VDFR | m$^3$/h | |
| Flow reflux VDFR (rel c) | m$^3$/h | Relative *Flow M crude* |
| Flow reflux VDFR (rel) | | Relative *Flow V VDFR to cistern* |
| Flow reflux VDTOP | m$^3$/h | |
| Flow reflux VDTOP (rel c) | m$^3$/t | Relative *Flow M crude* |
| Flow reflux VDTOP (rel) | | Relative *Flow V VDTOP to cistern* |
| Flow steam C19 | kg/h | |
| Flow steam C19 (rel) | | Relative *Flow M VDFR to cistern* |
| Flow steam to ejectors | kg/h | |
| Flow steam VD bottom | kg/h | |
| Flow steam VD bottom (rel) | kg/t | Relative *Flow M crude* |
| Flow V Bitumen to cistern | m$^3$/h | |
| Flow V VDFR to cistern | m$^3$/h | |
| Flow V VDTOP to cistern | m$^3$/h | |
| Flow water to E5 | t/h | |
| Flow water to E5 (rel) | | Relative *Flow M crude* |
| Level C19 | % | |
| Level C6 | % | |
| Level VD bottom | % | |
| Level VDFR ex | % | |
| Pressure steam to ejectors | bar | |
| Pressure steam to ejectors set | bar | |
| Pressure VD tower | mmHg | |
| Pressure VDTOP after E7 | mmHg | |
| Temp Bitumen E10A to E5 | ° C | |
| Temp Bitumen E10B to E10A | ° C | |

| | |
|---|---|
| Temp Bitumen E9A to E10B | ° C |
| Temp Bitumen E9B to E9A | ° C |
| Temp Bitumen E9C to E9B | ° C |
| Temp Bitumen to cistern | ° C |
| Temp C19 bottom | ° C |
| Temp C6 | ° C |
| Temp steam to ejectors | ° C |
| Temp VD bottom | ° C |
| Temp VD top | ° C |
| Temp VDFR ex | ° C |
| Temp VDFR to cistern | ° C |
| Temp VDFR to cistern set | ° C |
| Temp VDTOP after E7 | ° C |
| Temp VDTOP after E8 | ° C |
| Temp VDTOP reflux | ° C |
| Temp VDTOP to cistern | ° C |
| Valve Bitumen to cistern | % |
| Valve cooling w E8 | % |
| Valve steam C19 | % |
| Valve steam to ejectors | % |
| Valve steam VD bottom | % |
| Valve VDFR reflux | % |
| Valve VDFR to C19 | % |
| Valve VDFR to cistern | % |
| Valve VDTOP reflux | % |
| Valve VDTOP to cistern | % |
| Yield Bitumen | w% |
| Yield VDFR | w% |
| Yield VDTOP | w% |

Table A2-6. Parameters in the block *Product properties*.

| Name | Unit | Comments |
|---|---|---|
| FR1 0,5 | w-% | ADFR1 consists of 101 parameters. |
| FR1 1,0 | w-% | |
| FR1 2,0 | w-% | |
| ⋮ | ⋮ | |
| FR1 99,0 | w-% | |
| FR1 99,5 | w-% | |
| FR2 0,5 | w-% | ADFR2 consists of 101 parameters. |
| FR2 1,0 | w-% | |
| FR2 2,0 | w-% | |
| ⋮ | ⋮ | |
| FR2 99,0 | w-% | |
| FR2 99,5 | w-% | |
| VD-T 0,5 | w-% | VDTOP consists of 101 parameters. |
| VD-T 1,0 | w-% | |
| VD-T 2,0 | w-% | |
| ⋮ | ⋮ | |
| VD-T 99,0 | w-% | |
| VD-T 99,5 | w-% | |
| VDFR 0,5 | w-% | VDFR consists of 101 parameters. |
| VDFR 1,0 | w-% | |
| VDFR 2,0 | w-% | |
| ⋮ | ⋮ | |
| VDFR 99,0 | w-% | |
| VDFR 99,5 | w-% | |