CHALMERS

Process Stream Data Analysis: Data Reconciliation and Gross Error Detection for Process Integration Studies

Master's Thesis

CRISTINA MURCIA MAYO

Department of Energy and Environment
Division of Industrial Energy Systems and Technologies
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2015

MASTER’S THESIS

Process Stream Data Analysis: Data Reconciliation and Gross Error Detection for Process Integration Studies

Master’s Thesis

CRISTINA MURCIA MAYO

SUPERVISOR(S):

Anders Åsblad and Eva Andersson

Chalmers Industriteknik – Industriell Energi

EXAMINER

Simon Harvey

Department of Energy and Environment
Division of Industrial Energy Systems and Technologies
CHALMERS UNIVERSITY OF TECHNOLOGY

Göteborg, Sweden 2015

Process Stream Data Analysis: Data Reconciliation and Gross Error Detection for Process Integration Studies.
Master's Thesis
CRISTINA MURCIA MAYO

Department of Energy and Environment
Division of Industrial Energy Systems and Technologies
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone: + 46 (0)31-772 1000

Cover:
Image of the Oil Refinery from Preem in Lysekil. (©Cision)

Chalmers Reproservice
Göteborg, Sweden 2015

Process Stream Data Analysis: Data Reconciliation and Gross Error Detection for Process Integration Studies.

Master's Thesis
CRISTINA MURCIA MAYO
Department of Energy and Environment
Division of Industrial Energy Systems and Technologies
Chalmers University of Technology

ABSTRACT

One of the major challenges for energy companies is to adapt their process plants to the continuous improvements of available technologies, so as to make their old plants as competitive and cost-efficient as the new ones. Along these lines, process stream data was recently collected for analysing opportunities for improved process integration of the Hydrocracker Unit of a major oil refinery located in Lysekil on the West Coast of Sweden. However, inconsistencies in the process data measurements, e.g. energy balances that do not add up, made the study cumbersome. For analysing heat exchanger networks it is essential to establish sets of process data with balanced heat balances for the existing heat exchangers. The aim of this thesis project was to develop a computer-based solution for systematic analysis, identification and correction of the "raw" data obtained from process data measurements in order to acquire such a consistent set of data.

With this purpose, a tool for Data Reconciliation and Gross Error Detection for process stream data was developed using Visual Basic in Microsoft Excel. The tool is based on the Modified Iterative Measurement Test. A second tool, which is easier for handle large data sets and especially designed for networks with non-linear constraints was also developed. This second tool is only able to solve Data Reconciliation problems, so it is targeted for sets of data where there are exclusively random errors.

Both developed tools were used to analyse the data set collected from the refinery's Hydrocracker Unit with the purpose of generating a consistent set of data with balanced heat exchangers. The solution proposed is an energy balanced network, where from the 32 temperature measurements, all the reconciled values, except two, are within the specified bounds indicated. The two reconciled temperatures outside the bounds are the ones in which the presence of a gross error has been confirmed. Since this is a preliminary study, the solution proposed must be taken as a recommendation.

Key words: data reconciliation, gross error detection, Modified Iterative Measurement Test, hydrocracker unit.

Contents

Preface

This thesis work was carried out with Eva Andersson and Anders Åsblad at Chalmers Industriteknik – Industriell Energi CIT-IE) as supervisors. Without their support it would have been much more difficult and complicated to achieve the objectives of the project. I am especially grateful for their active involvement in the project day after day, for patiently listening and answering my many questions and for encouraging and motivating me in the hardest moments. I also want to express my gratitude to Stefan Heyne for his valuable input, particularly in the turning points, making more clear in which direction I should keep going. Also, I would like to thank my examiner Professor Simon Harvey for the useful contribution regards the literature review and the thesis structure. I am especially grateful for his dedication and time invested in the follow-up meetings, which were really helpful for the improvement of this thesis work.

Last but not least, thanks to all the members of the IEST division at Chalmers. You made my working hours pass faster, by keeping me awake with the coffee at nine, by adding sugar and cream in every Friday "fika", and by making my weekends in the office not so lonely.

Göteborg June 2015

Cristina Murcia Mayo

Notations

Abbreviations:

bpd	Barrels per day (158.987 liters per day)
BGLR	Bounded Generalized Likelihood Ratio
CCR	Catalytic Reforming Platformer
CDU	Crude Distillation column
CLPS	Cold Low Pressure Separator
DR	Data Reconciliation
FCC	Fluid Catalytic Cracker
GE	Gross Error
GED	Gross Error Detection
GLR	Generalized Likelihood Ratio Test
HCU	Hydrocracker Unit
HX	Heat Exchanger
ICR	Hydrocracker unit
ISO	Isomerisation
IMT	Iterative Measured Test
LCT	Linear Combination Technique
MEROX	Mercaptan Oxidation Unit
MHC	Mild Hydrocracker
MIMT	Modified Iterative Measurement Test
MMP	Modified Method of Pseudonodes
MP	Method of Pseudonodes
MSCS	Modified Serial Compensation Strategy
MSEK	Million Swedish kronor
MT	Measurement Test
NHTU	Naphtha Hydrotreating Unit
PHA	Polycyclic Aromatic Hydrocarbons
POLY	Polymerization
SC	Screened Combinational
SRU	Sulphur Recovery Unit
SSCS	Simple Serial Compensation Strategy
SSU	Synsat Unit
TGTU	Tail Gas Treatment Unit
VBU	Visbreaker Unit
VDU	Vacuum Distillation Unit
VGO	Vacuum Oil Gas

Symbols:

a	Measurement adjustment vector
b	Vector of the independent terms of the linear constraints
d	Modified adjustment vector
r	Vector of balance constraint
x	Vector of the reconciled values
y	Vector of the measured values
A	Incidence matrix for the linear constraints
ΔQ	Relative variation of the heat load
$Q_{meas.}$	Heat load calculated using the measured values
Q_{rec}	Heat load calculated using the reconciled values
V	Covariance matrix of r
W	Covariance of the measurement adjustment vector

α	Level of significance
ε	Random error
δ	Systematic or gross error
σ	Standard deviation
β	Modified level of significance
Σ	Covariance Matrix of Measurements

Subscripts:

lb_i	Lower bound of the measurement i
n_i	Reference number for the measurement i
ub_i	Upper bound of the measurement i
x_i	Reconciled value for the measurement i
y_i	Measurement i
FCp	Heat flow capacity [MJ/s ˚C]
H_0	Null hypothesis
H_1	Alternative hypothesis
Q_i	Heat load for the utility i
W_{jj}	Diagonal value from the $\overline{W}$ for each measurement
Z_i	Measured test statistics for the measurement i
$Z_{1-\alpha/2}$	Threshold criterion for a level of significance equal to α
$Z_{1-\beta/2}$	Modified threshold criterion for a level of significance equal to β
Z_{max}	Maximum measurement test statistic

1 Introduction

Todays the world is moved by petroleum. Crude oil is one of the most important fossil fuel in modern society, and all of us make use of it daily in any of its myriad forms.

Even if you ride a bike instead of using the car, and you try to avoid taking the plane or any transport run by oil, you are taking benefit from the petroleum or any of its by-products. Roads, lubricants, oils, plastics, gas for the barbecues, house heating and even cosmetic products are made with it. According to the International Agency (2015), the world oil demand for the first trimester of 2015 was around 93.45 million of barrels per day, that is, around 14857 million of litres per day (1bbl≈$0.159m^3$).

When crude oil is extracted from a well, it is a useless component, and it is necessary to process it in order to separate it into different fractions. The set of processes needed are carried out at refineries.

Since oil is a natural and non–renewable resource, whose reserves according to BP (British Petroleum) has been estimated to run out in about 53 years, it is essential to design the most efficient refining process so that the largest amount of products with the desired quality can be obtained. In addition to this, the scenario is continuously changing, legislations are becoming stricter and the market is looking for new environmentally friendly technologies. For that reason, it is important to collect data not only for monitoring the process but for studying the improvements' opportunities and adapting the plants to this new reality.

With the last purpose, process stream data was collected in the Preem refinery in Lysekil few years ago. This data set is the one used as a test bench for the study.

1.1 Background

Industrial chemical processes are usually complex, involving different units such as reactors, heat exchangers, splitters, etc., and connected by mass and energy flows. With the aim stated in the introduction, large amounts of process variables (flowrates, temperatures, pressures, fluid levels, compositions, etc.) are continuously measured and recorded. However, measurements are never 100% precise, resulting in incoherencies that complicate and, in some cases, corrupt or even invalidate the results from the studies done based on the collected data.

This thesis focuses on analysing the stream data collected for pinch analysis studies and making the adjustments so that the energy balances are satisfied and no inconsistencies are present in the data.

One element used in pinch analysis, which aims at identifying the potential for energy savings by minimising the thermal energy consumption and maximising the amount of heat recovered, is the grid diagram. It is a useful and powerful representation of the heat exchanger network. Figure 1.1 is an example of heat exchanger network, where the values from the temperatures are readings from the measurements. In this case, the example consists only of three streams, one hot stream and two cold streams. Hot streams need to be cooled from the start temperature to the target temperature, in this case T16 and T14, respectively. Conversely, cold streams need to be heated from the start temperature to the target temperature. These changes can be achieved by using heaters or coolers. But since the purpose is to minimize the primary energy consumption; what can be done instead is to use heat exchangers to recover heat from the hot stream to the cold stream, as in Figure 1.1.

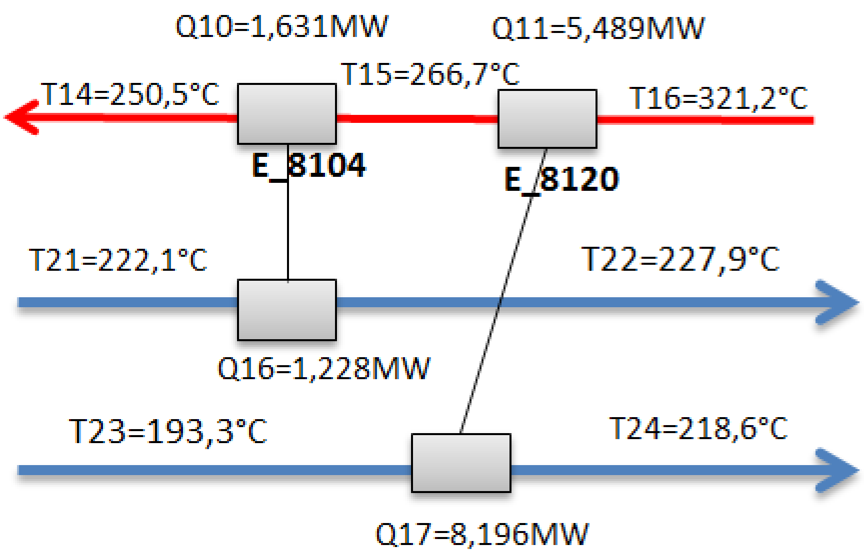

Figure 1.1 Simplified example of a Heat Exchanger Network. E_8104 and E_8120 are representations of heat exchangers. The number of digits of the temperatures and heat loads does not reflect precision.

If the losses to the surrounding are neglected, the amount of heat transferred from the hot stream must be equal to the amount of heat absorbed by the cold stream. For the case of the Figure 1.1, in the heat exchanger E_8104;

$$Q_{10} = Q_{16}, \tag{1.1}$$

and following the same reasoning in E_8120;

$$Q_{11} = Q_{17} \tag{1.2}$$

As can be seen, none of the energy balances, equations (1.1) and (1.2), are satisfied. For the HX E_8104 the absolute difference is small but comparing in relative terms, the heat transferred from the hot stream is 32,8% time larger than the heat absorbed by the cold stream. This variation is even more significant in the HX E_8120 the difference between both heat loads is 2.7 MW, and the relative variation reaches 49,3%, which is significant. The same principle can be applied for mass flows, e.g. assuming that there are no leaks in the system, the flowrate going into a heat exchanger must be the same as the flowrate going out.

There are many causes that can lead to errors in the measurements, for instance, the fact that processes in industry are not steady-state but vary in time or time lags when changes in the process are performed. Both may contribute to increase the difference between the true value and the measurement. However, the main reason why this happens is because during the measuring, processing and transmission of the measured signal, process measurements are usually corrupted by errors.

Industrial plant systems are complex, and a heat exchanger is just one of the multiple units involved in a process. Thus, identifying the errors in the measured data and mitigate their effects must be a priority.

There are various methods to achieve this objective, from applying “engineering judgment” to balance the network manually, to building a computer model and using the values from the model instead of the measurements. The problem of the former is that it is difficult to do but is even more complicated to judge how reliable the solution is. And the problem of the second one is that it is expensive and requires many man-hours. The solution proposed in this thesis is to apply statistical methods for detecting the presence of gross errors and reconcile the process variables using the resulting data set, which is free of gross errors.

1.2 Purpose and objective

The purpose of this thesis is to develop a Microsoft Excel tool for systematic analysis, identification and correction of the “raw” data obtained from process data measurements in order to acquire a consistent set of data. The aim is not to develop a tool for controlling the process but implement a systematic procedure to reconcile data that can be used for process integration studies. This involves processing data so that energy balances are satisfied for all relevant parts of the process system to be analysed. The analysis should also aim at identifying and eliminating “outliers”, i.e. data points that are most likely due to measurements errors.

With this purpose a study about the different methodologies that rely on the test statistics for detecting the gross error has been conducted. It is also within the scope of the thesis, to achieve a general knowledge about the oil refining process and the processes performed in every stage, due to the data used as a test bench for the study is from the Preem refinery.

1.3 Scope and limitations

The set of process data that was used was collected from the Preem refinery in Lysekil five years ago, in 2010. The project only includes the analysis and error detection of the data and does not include collection of the data nor the final use of the reconciled data for process integration study purposes. The system studied was assumed to be at steady-state, which means that all the system variables are constant, because there is a flow through the system. The task of this thesis was to implement a computer-based solution for data reconciliation and gross error detection using one of the existing methodologies and test it later for the data set collected from the refinery in Lysekil. This process includes specifying the inputs needed for the reconciliation problem, in other words, decide the measurements and define the constraints that characterize the process.

Due to limited time resources not all the process units of the Preem Refinery was analysed. The study was restricted to Data Reconciliation and Gross Error Detection for a simplified sub-network of the Hydrocracker Unit (HCU).

The commercial spread-sheet software Microsoft Excel was used for calculation purposes.

2 Preem Refinery

Data Reconciliation and Gross Error Detection computer based tools were tested with data collected from the hydrocracker unit (HCU) of Preem refinery in Lysekil. For that reason, this chapter aims to present Preem as a company, provide a general overview of the refining processes, and finally focus on the studied unit.

2.1 Introduction to Preem

At the moment there are five refineries running in Sweden. Two of them owned by Nynäs, situated in Gothenburg and in Nynäshamn, another one in Goteborg owned by St1 and the other two owned by Preem.

Preem is the largest fuel company in Sweden, supplying more than half of Sweden's industrial companies and one third of the small companies with heating and energy products. With capacity of more than 18 million m^3 of crude oil every year provided by two refineries; one in Gothenburg (125,000 bpd) and the other in Lysekil (220,000bpd), they account for 80 percent of the Swedish refinery capacity and 30 percent of the Nordic refinery capacity. Their market is basically Europe but Preem also exports to Africa and North America. They supply companies and customers with gasoline, diesel, lubricants, heating oil and renewable fuels. With a service network close to 600 fuelling stations they provide more than half of the fuel oil and diesel that is consumed in Sweden, and almost 40% of the gasoline, Preem (2015).

Figure 2.1 The location of the two Preem refineries are indicated on the map (in blue)

The raw data used for this thesis was collected in 2010 from the Preem refinery in Lysekil, which is the Scandinavia's largest facility with capacity to refine 11,4 million tons of crude oil per year.

2.2 Oil refining process

Oil refining is the process whereby the extracted crude oil from the well, which at that moment is essentially unusable, is transformed into useful products for different uses.

The development of the current oil industry dates from the middle of the 1800s, where an increasing need of producing large quantities of kerosene for lighting, which was cheaper and better than whale oil, stimulated the industry to adapt to the new demand. However, its importance increased with the invention of the combustion engine, and the urgency of new fuels for cars and planes.

Nowadays, the production of transportation fuels is by far the most important fraction but other non-fuel products, such as sulphur, propane, propylene and some others feedstocks for the chemical and petrochemical industries are also produced. Figure 2.2 shows the wide variety of products that are obtained from the oil refining. From the light gases within one to four hydrocarbons in length, to heavier fractions of large hydrocarbons chains, with large boiling points, is exploited.

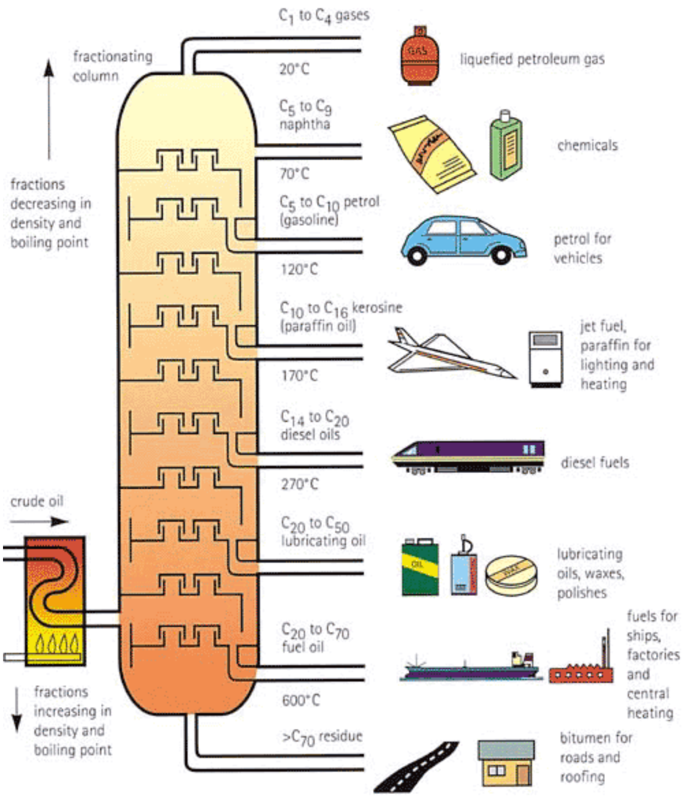

Figure 2.2 Range of products obtained from the oil refining. (World Fuel Limited)

Products from the top of the fraction column are the more volatile ones, with boiling temperatures around 20°C at atmospheric pressure, whereas the ones from the bottom are the heaviest, with boiling points up to 600°C.

The liquefied petroleum gas, or refining gas, is the one with smallest molecules and with such a low boiling point that is still a gas when it is removed from the fractionator. Propane and butane are some examples, and they are used for running barbecues or heaters, as they can be stored in bottles under high pressure. Moving down from the top, the next product is naphtha. It is not so useful as fuel as it is for the petrochemical industry, representing around 70% of the feedstock to that industry sector. Naphtha can be cracked to form alkynes that can be turn into polymers for making plastics. The middle lengths products are fuels; petrol, kerosene and diesel oil are obtained from the fractionator column in this order, and they are by large margin the ones with higher value in the market. Finally, the less volatile products are the lubricants oils and waxes, the fuel oil used for ships and factories as well as for central heating, and some solid residues as bitumen used for paving the roads and to cover the roofs.

However, nowadays, one of the biggest challenges of the energy companies is how to reduce the environmental impact and meet environmental legislations, which are becoming more and more strict. It has become a priority need to develop and redesign the processes so that less pollutants are emitted. In the oil industry, the focus is on reducing the emissions of hydrocarbons, carbon monoxide, sulphur and nitrogen oxides, among others contaminants.

Due to constant changes in the external conditions, every refinery is different regarding the market conditions, national legislation, feedstock, required products…etc. Figure 2.3 is a scheme of the major units operations and activities carried out in Preem refinery.

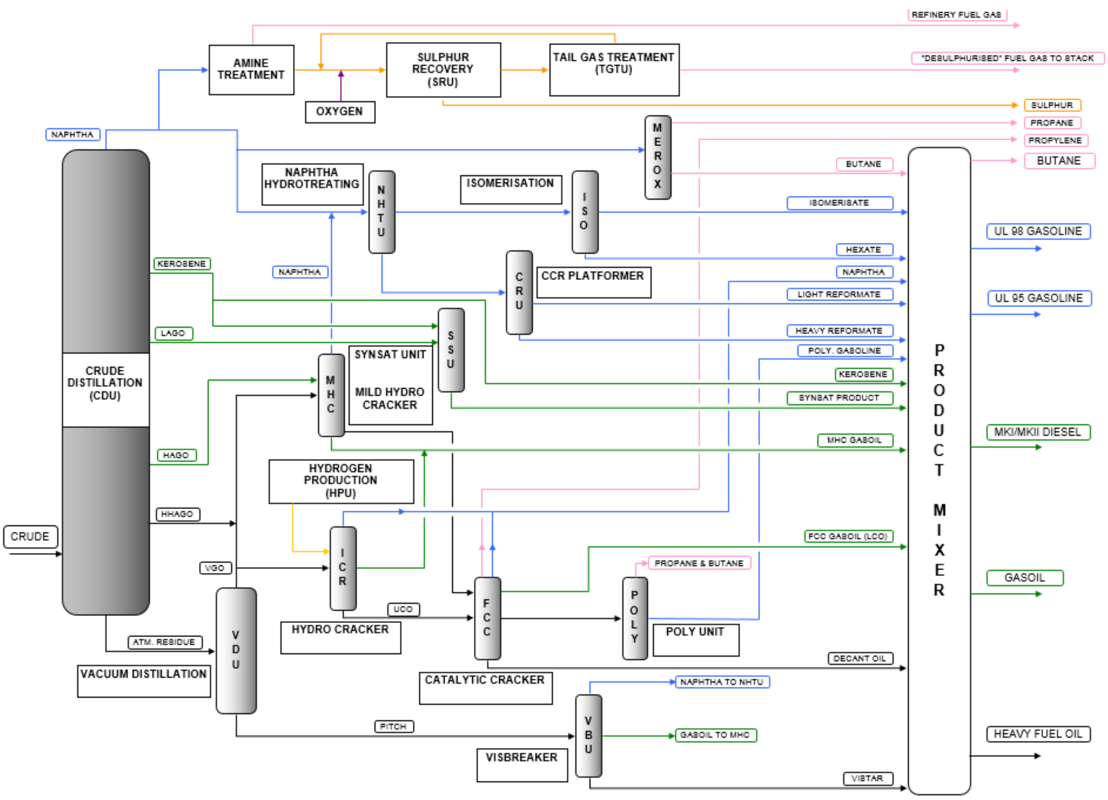

Figure 2.3 Overview flow sheet of Preem's Refinery in Lysekil.

The process of refining crude oil into usable petroleum products can be explained by classifying the different process into two phases, Joint Research Centre (2013). The first phase consists on the separation of the different components from the crude oil using the difference in the volatility. The second phase, is where the molecular structure of the hydrocarbon is changed by breaking, combining and reshaping the molecules.

The first phase is carried out in the Atmospheric Crude Distillation column (CDU) and in the Vacuum distillation (VDU). In the former unit, the crude oil is heated to elevated temperatures and exposed under atmospheric pressures to separate the various fractions according to the boiling range. From the overhead of the column is obtained the lighter and non-condensable refinery fuel gas which also contains hydrogen sulphide and ammonia gases. From the bottom of the column, the heavier non-vaporized components are sent to the vacuum distillation (second unit) for further separation. In this case, the feed is subjected to very low pressure in order to increase the volatilization and separation whilst avoiding thermal cracking.

As stated previously, the most volatile product from the CDU contains sulphurs. For that reason this stream is sent to the Amine treatment unit where the hydrogen sulphide (H_2S) and the carbon dioxide (CO_2) are removed by the use of aqueous solutions of various alkyl-amines. Thereafter, the hydrogen sulphide is introduced in the Sulphur Recovery Unit (SRU) where the sulphur is recovered using the Claus process. The tail gases from the SRU that still contain hydrogen sulphur are introduced in the Tail Gas Treatment Unit (TGTU) that reduces the sulphur vapour and sulphur dioxide into hydrogen sulphur. This last one is returned to the SRU for further sulphur recovery. There is also some naphtha stream that is send directly to the Mercaptan Oxidation Unit (MEROX), where it is washed in a concentrated alkaline extraction column under elevated pressure in order to remove the mercaptans, and decrease the content of odorous and corrosive components.

Intermediate products of the distillation column are processed in different units, depending on the impurities' content. The most common ones are sulphur, nitrogen, oxygen, halides and metals. In order to remove them, the different products have to be exposed under a huge amount of hydrogen at high pressure and temperature and in presence of a catalyst, which depend on the feed composition. Those units can be identified in the Figure 2.3 with the following names: Naphtha Hydrotreating (NHTU), Synsat Unit (SSU), Mild Hydrocracker (MHC) and Hydrocracker (ICR or HCU). Moreover, almost all the units/processes described before also reduce the molecular weight by cracking and separating the light fractions from the heaviest ones.

Thereafter, the most volatile fractions are sent to the Isomerisation (ISO) or to the Catalytic Reforming Platformer (CCR). The aim of these two processes is to increase the octane index. Whereas, the less volatile are send to the Fluid Catalytic Cracker (FCC) where the hydrocarbons molecules are cracked using zeolites catalysts. The mild-weight products from the last process are later sent to Polymerization Unit (POLY), where the propene and butene is converted to a high octane gasoline blending components. Finally, the heaviest fractions resulting from the VDU are sent to the Visbreaker Unit (VBU) where its viscosity is reduced by breaking the large hydrocarbons molecules. At the end of all the processes the streams are mixed in the Product Mixer in order to obtain the final products which the compositions required in the market.

2.3 Hydrocracker unit (ICR/HCU)

The hydrocracker unit, as indicated by its name, is a conversion process that combines the catalytic cracker and the hydrogenation. The former process uses a zeolitic catalyst, such as nickel, palladium or platinum, which tends to break alkanes chains into two or more alkane and alkene products, generally making molecules between five and ten carbons in chain length. The second one is used for reducing or to saturate organic compounds with hydrogen. Thanks to the association of both processes, the heavy factions of distillate are decomposed through the breaking of carbon-carbon bounds and the subsequent or simultaneous hydrogenation, resulting in lighter products with higher commercial value. For instance, the dodecane can result in octane plus butene, decane and ethene, or butene, pentane and propane, see Figure 2.4.

$$C_{12}H_{26} \leftrightarrow C_8H_{18} + C_4H_8$$

$$C_{12}H_{26} \leftrightarrow C_{10}H_{22} + C_2H_4$$

$$C_{12}H_{26} \leftrightarrow C_4H_8 + C_5H_{12} + C_3H_6$$

Figure 2.4 Possible resulting components from the braking of a dodecane molecule.

In general terms, the hydrocracking is conducted under substantial pressure, between 6 and 12 MPa, and between 300 and 450°C, Joint Research Centre (2013). Depending on the size of the unit, the feedstock and the desired products, three type of units can be defined; single-stage once-through, single-stage recycle and two-stage recycle. The first mentioned only uses fresh feed and is able to achieve rates of conversions between 80-90%, whereas the second one, that differs from the former due to it recycles the unconverted oil, achieve bigger rates around 97-98%. The two-stage recycle hydrocracker has same conversion rate as the previous, but it is especially indicated for feedstocks with a very high refractory such a deasphalted oil, since it uses two reactors in series. In the present case, Preem is using a single-stage hydrocracker unit, as the one in the Figure 2.5.

The main feeds of process are Vacuum Oil Gas (VGO) from the heavy products of the crude distillation column, referred as fresh feed in the Figure 2.5, and substantial quantities of hydrogen. As stated previously, in order to break and split the large carbon chains, the VGO must be heated. Thereby it is passed through various heat exchangers in line in order to preheat the feed before it is introduced into the Reactor Heater, where the temperature is increased markedly. Afterwards, it is sent to the first reactor, where the impurities are removed. Sulphur and the nitrogen reacts with hydrogen resulting in sulphuric acid (H_2S) and ammonia. In order to accelerate the reaction, metallic catalysts containing for example platinum and nickel are used. In the first beds of the reactor also takes places the saturation of olefins and partial saturation of Polycyclic Aromatic Hydrocarbons (PHA). The second reactor is where the cracking takes place. Large hydrocarbons molecules break resulting in smaller and lighter compounds. Thanks to the introduction of the hydrogen the free bonds are filled by hydrogen molecules. In this case, the cracking process depends on an acid catalyst. The most commonly used are the zeolites.

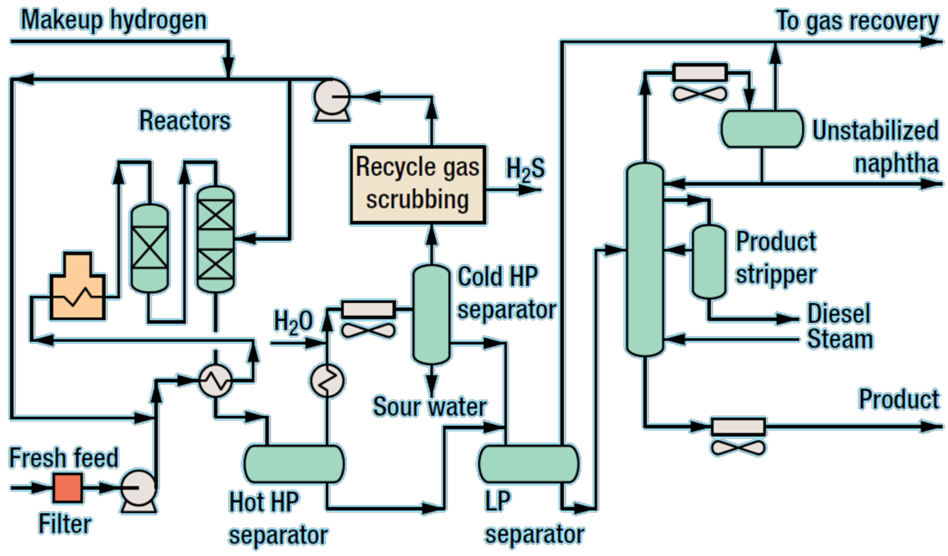

Figure 2.5 Overview flowsheet of an hydrocracker unit.

After that, the product is cooled down and sent to a high pressure separator where the lighters gases are extracted. The heavier fractions are directly sent to the low pressure separator but the lighter ones previously passed through a cold high pressure separator. Since hydrogen is fed to the reactors in excess with the purpose of purifying the hydrogen rich gas from impurities, there is much unconverted hydrogen in the outflows from the reactors. The purpose of the cold high pressure separator is to extract the pure hydrogen for recycling it. The resulting stream from the low pressure separator is sent to a propane/butane stripper for further separation. From this column, as an overhead product, light naphtha is obtained. As a side-steams, kerosene and diesel are obtained and from the bottom, and unconverted oil as a residual.

In order to break and crack the molecules, the feedstock needs to be heated until high temperatures and cooled down again before supplying the next unit or sending to store some of the products. This double variation in the temperatures makes it a suitable process for performing process integration. With that purpose, some of the process variables were measured. The reader is referred to the Appendix 1 for consulting the data collected.

3 Data Reconciliation and Gross Errors Detection.

As stated in the introduction, data collected is often corrupted by errors produced during the measuring, resulting in a difference between the measured value and the true value. This difference can be expressed as the sum of random errors and systematic errors,

$$y = x + \varepsilon + \delta \tag{3.1}$$

where x is the true value of the measured variable, ε is the unknown random error and δ corresponds to the systematic error.

The information explained in this chapter is mostly based on the following sources: Carl Knopf F (2012) and Natasimhan S.,Jordache C. (1999).

3.1 Random Errors (ε)

Random errors are small in magnitude, and neither the value nor the sign can be predicted with certainty. As their name indicates, they are unpredictable and unavoidable and basically depend on the method and the instrument used for measuring. Examples of causes of random errors are, for instance, the intrinsic accuracy of the measurement tool, the electric noise in the circuit of an electrical instrument or the environment conditions. Random errors are fluctuations scattered around the true value, as the sample is affected by almost the same amount of negatives and positives errors. Therefore, random errors add variability to the data but since the arithmetic mean *(E)* tend to approach zero, they don't rebound on the average performance.

It can be assumed that random errors follow a Gaussian Normal Distribution with average 0.

$$E(\varepsilon) = 0$$

$$var(\varepsilon) = E(\varepsilon^2) = \sigma^2 \tag{3.2}$$

In the equation (3.2), σ is the standard deviation of the measurement, which quantifies the amount of variation and dispersion of the data set. If σ approaches 0, data points are really close to the mean, whereas, if the value of σ is large, data points are spread out over a wide range and away from the true value.

Unfortunately, the standard deviation is unknown so, in most of the cases, the estimated standard deviation from a sample is used instead. The sample standard deviation is calculated according the equation (3.3), where N is the number of observations, y_i is the value of the ith observation and $\bar{y}$ is the arithmetic average of the N observations.

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{3.3}$$

$$cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i, \varepsilon_j) = 0 \tag{3.4}$$

It is also assumed that random errors present in the measurements of two different process variables, for instance i and j, are statistically independent, so the correlation between both variables equals to 0, equation (3.4).

Although random errors do not represent a big problem, there are some techniques for improving the accuracy of measurements and reduce the effect of them when the measured data is used in further studies. Data Reconciliation (DR) is one example. This method makes use of process model constraints and obtains the estimates of process variables by adjusting the process measurements so that estimates satisfy the constraints. In other words, it is a constrained optimization problem where the objective function is the equation (3.5) and the constraints are the energy and mass balances equations, see equation (3.6). For the specific case of this thesis work, as one will see later, the reconciled values are only constrained by energy balances.

$$Minimize \sum_{\substack{i, measured \\ variable}} \left(w_i(y_i - x_i)\right)^2 \tag{3.5}$$

$$Subject\ to\ h_k\left(x_i, u_j\right) = b\ k = 1, \ldots., K \tag{3.6}$$

For the above equations, y_i refers to the measured variable, x_i is the reconcile value, w_i is a weight that reflects the measurement's accuracy, $h_k\left(x_i, u_j\right)$ are the constraints, generally mass and energy balances, and u_j is the estimated value of non-measured variable, which will be discussed later on. Finally *b* refers to the independent terms of the constraints, generally it is equal to zero.

It must be highlighted that data reconciliation assumes that no gross errors are present in the measured data. Otherwise, the reconciled values can be very inaccurate and even unfeasible.

3.2 Gross Errors (δ)

Gross errors, also called systematic errors, are significant and systematic deviations from the true value caused by non-random events. Unlikely random errors, their sign must be considered, which means that if the measurement is repeated under identical conditions, the value obtained will be the same and so the contribution of the gross error too. Gross Errors are caused by instrument malfunctioning as bias or drift, miscalibration, corrosion of the sensors, or even process leaks. The most common gross errors are represented in the Figure 3.1.

It is important to detect and remove them from the data, otherwise, the reconciled values will be corrupted and therefore, the balanced network invalidated. For that reason, a tool for analysing, detecting, and removing the errors must be used. Thereby, the adjustments are smaller and the quality of the resulting reconciled values is higher. With this purpose, a wide range of algorithms that include gross error detection as a preceding step to data reconciliation have been developed. Some of them are presented on the Section 3.4

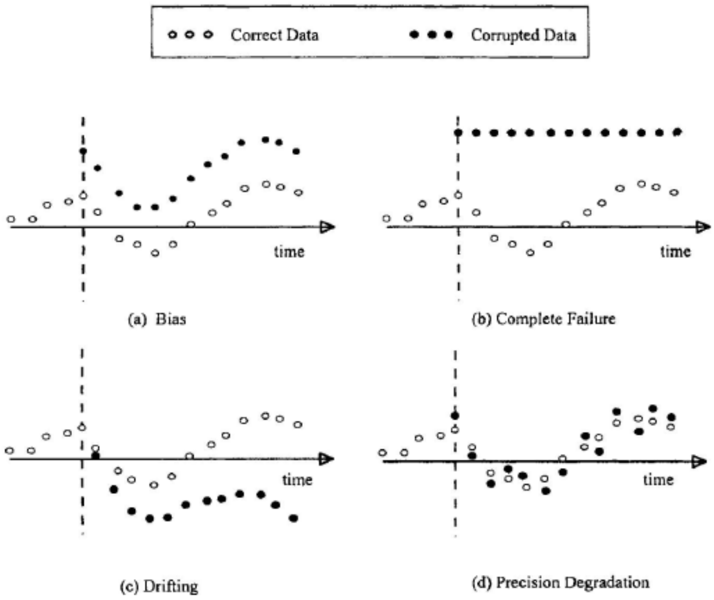

Figure 3.1 Most common Gross Errors caused by instrument type's faults. (Natasimhan S.,Jordache C,1999)

3.3 Process Variable types.

Adding complexity to the problem, not always all the variables from a data set are measured and even when they are, sometimes the reconciled value cannot be computed. Data collection is a complex procedure that is considered out of the scope on the present thesis work because of limited time. However, understanding the different types of variables depending on the data available is essential for applying data reconciliation and gross error detection.

In processing plants, hundreds of process data such as flow rates, temperatures, pressures, stream compositions… are present, and depending on the point of interest of the study, one data or another is measured. However, not all the data is accessible, in most cases there are some process variables whose measurement could not be performed and the values are unknown. The usefulness of data reconciliation depends on the proportion between the measured variables and the unmeasured, and where these last are located in the grid diagram.

When computing the reconciled network for a data set where unmeasured variables are present, it is necessary to divide the calculation process in two steps. Firstly, it is performed a reduced reconciliation problem where only the measured variables are involved, removing the unmeasured variables from the objective function and the constraints. The second step consists of estimating the values for the unmeasured variables using the original set of constraints and the reconciled values obtained in the first step. In this case, when all the unmeasured variables can be estimated, the unmeasured variables are called observable. However, there is the possibility that infinite values can satisfy the constraints making impossible to calculate the estimates. In this case, the unmeasured variables are called unobservable.

In fact, the reconciled values can only be successfully computed for the process variables for which measurement are known and, at the same time, are constrained by usable material or energy balance. These process variables are called redundant measured variables. Conversely, the reconciled values for non-redundant variables, which are the ones whose measured value is known but they don't appear in a usable process constraint, cannot be calculated. Since they are not constrained the reconciled value will be always equal to the measured value.

To sum up, a variable is observable when it can be estimated by using the measurements and the process constraints. Thus, measured variables are always observables. Derived from this concept, a measured variable is redundant when it is observable even when its measurement is removed.

3.4 Multiple Data Reconciliation and Gross Error Detection strategies

The problem of data reconciliation was first introduced in 1961 by Kuehn and Davidson. Since then, multiple researchers have published their studies among which Mah, R.S.H.(1976), and Narasimhan S.(1999) stands out. This section aims to present some strategies for data reconciliation and gross error detection which apply to steady-state systems and linear constraints. It is also included within this section to give a general outlook of the main differences and note the similarities and differences in their performances.

In order to compare the system on the same basis, it has been assumed that only gross errors caused by biases are present in the data set. Moreover, since the aim of this thesis work is to develop a generic methodology that can be apply in any data set collected is, the methodologies that are going to be studied are designed for Multiple Gross Error Detection.

The vast majority of strategies for identifying and locating gross errors made use of test statistics, and are based on the principle that gross errors on measured data result in the violation of the model constraints. However, as stated before, the measurement contains also some random errors so it cannot be expected that the constraints are strictly satisfied. In order to detect just the outliers resulting from the gross errors, since random errors are small in value and do not represent a problem for data reconciliation mechanisms, a concept called normalized error has been defined. The normalized error is the difference between the measured value and the expected mean divided by its standard deviation. In order to identify gross errors, a confidence interval of the normalized errors can be defined. So that the values that fall outside $(1\text{-}\alpha)$ the confidence interval, where α is the level of significance, are identified as outliers and thus, likely to contain a gross error.

Hypothesis testing is the chosen methodology in all cases to perform the analysis. Two hypotheses are defined; the null hypothesis, Ho, is that no errors are present in the data, whereas the alternative hypothesis, H_1, is that one or more gross errors are present. However, depending on the strategy, the calculation of the test statistic and the threshold criterion value varies.

Basically, multiple gross error detection strategies can be classified within the following categories: simultaneous, serial or combinational. All this methods are further explained in Chapter 7, Natasimhan S.,Jordache C. (1999).

Simultaneous error detection strategies

In this case, the algorithm attempts to identify all the gross errors present in the data simultaneously or in a single iteration. For these algorithms the corresponding test statistics for each variable are calculated simultaneously and tested against the test criterion. Those measurements whose test statistics exceed the test criterion are likely to contain errors. The Measured Test (MT) applied for all the measurements and the Generalized Likelihood Ratio Test (GLR), are the main ones within this category,

Serial error detection strategies

In contrast to the simultaneous strategy, serial techniques identify the errors one by one by doing multiple iterations of the same procedure. Within this group, the different techniques can be sub-classified depending on whether the principle is elimination or compensation. In the former, every time that a gross error is detected, the corresponding measurement is removed before calculating again the new estimates and new stats. The Iterative Measured Test (IMT), or its powerful version Modified Iterative Measurement Test (MIMT), which allows also setting bounds to the variables, are an example. In the second one, as its name indicates, instead of eliminating the wrong measurement from the data set, it uses a compensation system based on the identified type error, location and the estimated magnitude. Simple Serial Compensation Strategy (SSCS), Modified Serial Compensation Strategy (MSCS), and Bounded Generalized Likelihood Ratio (BGLR) are included within this group. In the last one, the bounds of the variables can be included as inequality constraints, turning it into a complex quadratic programming optimization method.

Combinational error detection strategies

Last but not least, there is the combination strategy, which is based on nodal imbalanced tests. The method consists of reducing the system into small subsystems by aggregating two or more nodes (in a process graph, the nodes are generally units, tanks and junctions in the process flow sheet). In that manner, the process variables that are in between the merged nodes are deleted. In that way, less variables are involved in each test. This procedure is done many times until all the process variables have been tested. If the nodal test is rejected, it means one or more of the incident streams, i.e. the ones going in or out from the node, have a gross error, whereas if the test is not rejected it can be considered that all incident streams are free of errors. Unlike previous strategies, the method identifies the measurements that according the criterion are correct. The remaining ones, those that cannot be confirmed, are the ones suspected of having a gross error. In this case, no judgment can be made for the process variables in the streams that are interconnecting the aggregated nodes, and so when the hypothesis is rejected, no information is obtained regarding which process variables contain the error. For that reason, multiple nodal combinations must be done. The Linear Combination Technique (LCT), Method of Pseudonodes (MP), and Modified Method of Pseudonodes (MMP) are examples of combinational methods.

Not all the mentioned techniques have the same efficiency or accuracy, and most of the time, their performance also depends on the characteristics of data set. Although several articles and studies have been conducted comparing their performances, unfortunately; none of them provides a clear answer. But they do give some advices and conclusions that are presented in the following. However, before that, it must be pointed out that not only not detecting a gross error in a measured data when it is present is a problem (*Type II error*), but also the opposite, falsely detecting a gross error when there actually none

(*Type I error*), is a sign of bad performance In fact, a method that detects all the gross errors present in a data set but that also end up in some false alarms is not useful. Therefore, when designing a methodology, the power of the test, which is the probability of correct detection, must be balanced against the probability of false detection.

According to Natasimhan S.,Jordache C. (1999) simultaneous techniques seem to be the less accurate, resulting in too many mispredictions. That happens because the variables are all related through the constraints and the test statistics make use of the constraints residuals. In this way, a gross error can be spread among other variables and may cause the test statistic of a good measurement to exceed the test criterion. Or what is even worse, test statistics of good measurements exceed the test criterion but the one containing the gross error does not. This is known as smearing effect.

Serial and Combinational algorithms perform better because the test statistics are calculated again after each error detection. Nevertheless, combinational strategies results in less mispredictions than serial ones, but in return, *Type II errors* are more frequent due to partial or complete cancellation. Moreover, since combinatorial methodologies involve multiple nodal combinations, a methodology reducing the number of combinations is required. On top of that, computational time is longer compared to the serial strategies and combinational algorithms are more difficult to apply to non-linear systems.

Due to these arguments, it was decided not to implement a combinational strategy within the present work. Therefore, the discussion is between the different serial strategies. According to Natasimhan S.,Jordache C. (1999), identical results are obtained from using the serial elimination strategy in conjunction with the measured test or in conjunction with GLR test, both use the same principle. Thus, the modified serial compensation performs equal to serial elimination, and both will lead to the same results. For the purpose of the present work, the MIMT appears to be the best option due to its simpler implementation.

The final discussion is between the MIMT and the modified serial compensation (MSCS). The only difference between both is that MSCS also is capable of detecting leaks, i.e. the measurement differs from the true value not because the process variable has been wrong measured but because there is a leak in the process. However, this benefit does not make a difference in the present study since it has been assumed that there are no leaks and that the flow is always constant. In accordance, both algorithms can be considered as good choices. This conclusion is also ratified by Shert R.W., Heenan A. W, (1986) which conclude that the MIMT and the SC algorithms are the more accurate ones, detecting around 80% of the gross errors and achieving a reduction of total errors around 60%.

There is one advantage of the modified iterative measurement test that has not been taken into consideration until now, but that has made the MIMT the algorithm of choice in the present work; the method allows also defining bounds for the measurements. In that manner, the reconciled values will be inside the range defined by the bounds, avoiding the optimal but infeasible solutions.

3.4.1 Modified Iterative Measurement Test (MIMT)

The Modified Iterative Measurement Test (MIMT) is a serial elimination algorithm for Data Reconciliation and multiple Gross Error Detection. As its name indicates it is based on the iterative use of the Measurement Test (MT) outlined by Mah and Tamhane (1982).

The whole procedure can be divided into two parts. The first part consists of performing error detection and deleting the wrong measurements from the collected data set. The second part is to perform data reconciliation using this new data set.

Therefore, the data set used for reconciling the network has only random error, and so, equation (3.5) and equation (3.6) defined in the Section 3.1 can be used. The difference is that in this case, both equations are used in its matrix form.

$$Min\ (y-x)^T W (y-x) \tag{3.7}$$

$$Subject\ to \quad Ax - b = 0, \tag{3.8}$$

where y is the vector of the measured values, x is the vector of the reconciled values and W is a diagonal weighted matrix. In equation (3.8), A is the incidence matrix for the linear constraints and b is the vector of the independent terms of the linear constraints. The incident matrix A shows the relationship between the process variables (each column), and the energy or mass balances that constraint the system (each row). If the entry is equal to zero, that means that the process variable is not affected by that constraint. In this case, A is a full rank matrix since all the constraints are linearly independent.

As explained before, random errors follow a normal distribution with mean zero and known variance σ_{ii}^2. If it is assumed that the measurements of the process variables are independent, it is logical to say that their random errors are independent too. In accordance, the Covariance Matrix of Measurements Σ, equation (3.9) can be defined.

$$\mathrm{Cov(y)} = \Sigma = \begin{bmatrix} \sigma_{11}^2 & 0 & \dots & 0 \\ 0 & \sigma_{22}^2 & \dots & 0 \\ \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & \sigma_{ii}^2 \end{bmatrix} \tag{3.9}$$

The standard deviation σ_i is a measure that quantifies the degree of dispersion of a data set. If the standard deviation is close to zero, it means that the data points tend to be together, in other words, the precision is higher. In contrast, a higher value of standard deviation implies larger variation on the measurements, so less precision. The weighted matrix W defined on the equation (3.7) attempt to give higher weight to the accurate measurements and lower to the inaccurate ones. Therefore, the weighted matrix is inversely proportional to covariance matrix, and Data Reconciliation problem can be expressed from a statistical point of view as:

$$Min\ (y-x)^T \Sigma^{-1} (y-x) \tag{3.10}$$

$$Subject\ to \quad Ay - b = 0 \tag{3.11}$$

By using Lagrange multipliers methodology, which is not going to be explained here, but is briefly summarized in Appendix 2, the analytical solution to the data reconciliation problem can be calculated using the equation (3.12).

$$x = y - \Sigma A^T (A \Sigma A^T)^{-1} (Ay - b) \tag{3.12}$$

The reconciled values obtained from the equation (3.12) are only valid if the collected data set is free of gross error. For that reason, a preceding step detecting and deleting the measures suspected to contain gross errors is necessary. And it is exactly what the Modified Iterative Measurement Test algorithm attempts to do by performing multiple Measurement Tests as a preceding step to data reconciliation.

The Measurement Test is a technique for detecting systematic errors derived from the basic principle of normalized error explained in Section 3.4.1. In this case, instead of using the normalized errors, it uses the vector of measurement adjustments, defined as the difference between the measured value and the reconciled value. But the core principle is exactly the same.

Assuming that no systematic errors are present in the data, i.e. the H_0 is true and the adjustment vector follows a Normal distribution with mean zero and covariance $\overline{W}$. Equation (3.13) and equation (3.14) show two different forms for computing the vector of measurements adjustments, and also for computing its covariance matrix $\overline{W}$.

$$a = y - x \tag{3.13}$$

$$a = \Sigma A^T V^{-1} r \tag{3.14}$$

$$\overline{W} = \mathrm{cov(a)} = \Sigma A^T V^{-1} \Sigma \tag{3.15}$$

$$r = Ay - b \tag{3.16}$$

$$V = cov(r) = A \Sigma A^T \tag{3.17}$$

$$Z_{a,i} = \frac{|ai|}{\sqrt{\overline{W}_{ii}}} \quad \text{j=1,2,..n} \tag{3.18}$$

In the stated equations, r is the vector of balance constraint and V is the covariance matrix of r, and $\overline{W}_{ii}$ is the diagonal value from $\overline{W}$ for each measurement.

Applying equations (3.12)-(3.17), which are using the defined adjustment vector and its covariance, a test statistic with normal distribution $N(0,1)$, under the assumption of $H_{0,}$ is drawn using the equation (3.18) for each measurement i. Nevertheless, with the purpose of maximizing the power of detecting a single gross error a modification of the equation. (3.18) is done. The modification consists of pre-multiplying the vector of adjustments by the invers of the covariance matrix of measurements Σ, resulting in the following equations:

$$Z_{a,i} = \frac{|di|}{\sqrt{W_{ii}}} \quad \text{i=1,2,..n} \tag{3.19}$$

$$d_j = \Sigma^{-1} a \tag{3.20}$$

$$W = cov(d) = A^T V^{-1} A^T \tag{3.21}$$

To perform the Measurement Test, is also necessary to calculate the threshold criterion $Z_{1-\alpha/2}$, defined as the critical value of the standard normal distribution for a given level of confidence α (generally $\alpha = 5\%$). This implies that it is acceptable to have a 5% probability of incorrectly rejecting the null hypothesis. However, according to Natasimhan S.,Jordache C. (1999), as multiple tests are performed using the same critical value, the probability of rejecting the test even though there are no gross errors present in the sample (*Type I Error*) is higher than the one defined by α.

In order to keep the probability lower or equal to α, a modified level of significance β is used ($Z_{1-\beta/2}$). This parameter is calculated using equation (3.22), where m is the number of simultaneous multiple tests computed in each iteration. Note that the value of m will decrease every time a variable is removed from the data set.

$$\beta = 1 - (1-\alpha)^{\frac{1}{m}} \tag{3.22}$$

With the m test statistics and the modified threshold criterion computed, the measurement test can be performed. All the test statistics, one for each process variable that has been measured, are tested against the modified threshold criterion. The ones that exceed the later are those with a high probability of containing a gross error. The higher value of the measurement statistic test, the greater the probability of having a gross error.

In other words, if one or more tests statistics are above the threshold the collected data has at least one gross error and it is located in any of the measurements whose test statistic exceeds the threshold. But since a measurement with gross error sometimes can affect the nearby measurements, causing their test statistics to exceed the criterion, it would be a big mistake to say that all the measurement whose test statistic is higher than the threshold are wrong. For that reason, if there are more than one test statistics that exceeds the criterion, the test statistics can be used as indicators, but never to confirm the location of a gross error, without doing another test.

In order to not commit such an error, only the measurement with the highest test statistic that is suspected to contain gross error is removed from the data set each iteration. Then, using the reduced data set, the reconciled values, the measurements test statistics and the threshold criterion are calculated again, and a new measurement test is conducted. The algorithm terminates when the maximum measurement test statistic does not exceed the test criterion, which means that, the resulting network is free of gross errors. Since only one measurement is eliminated at each stage the possibility of make a *Type I error* decreases.

The Modified Iterative Measurement Test includes also another improvement for avoiding infeasible reconciled values. Lower and Upper bounds can be included in the algorithm. In that way, when a measured value with gross error is removed from the data set, the resulting estimated values must be inside the bounds. Otherwise, the measurement will be re-incorporated into the data set and the next one with the highest test statistics will be removed. The whole procedure is schematised in the Figure 3.2.

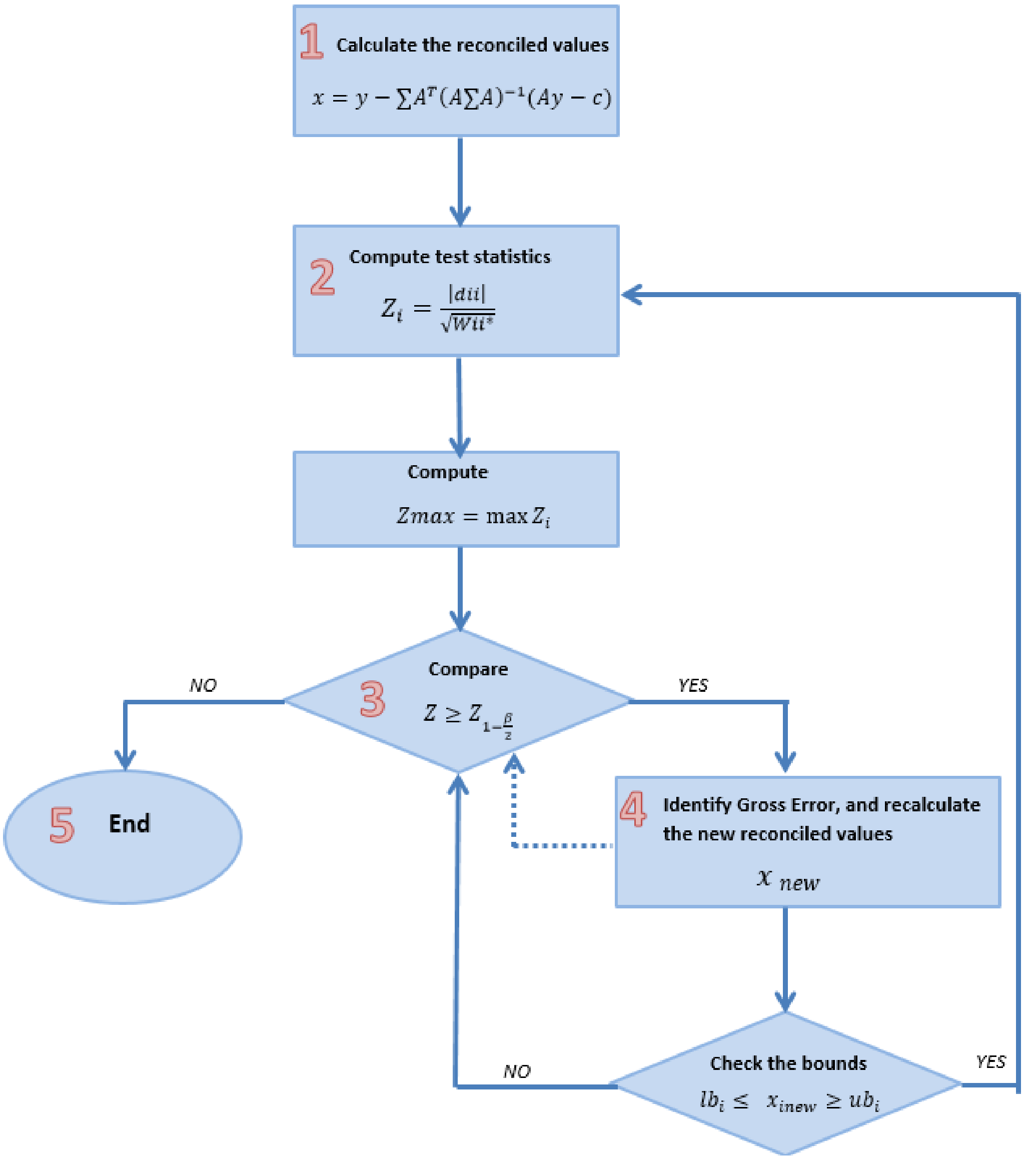

Figure 3.2 MIMT scheme, adapted from Kim, I., Kang, M. S., Park, S., Edgar, T. F. (1996)

For easy understanding, the procedure explained is described step by step below.

Step 1: Solve the initial reconciliation problem. Compute the reconciled values x, the vector of adjustments a, and the modified vector d.

Step 2: Compute the measurement test statistics Z_i for all the measurements in the data set

Step 3: Calculate the threshold criterion $Z_{1-\beta/2}$ using the 2-tailed t test and compare it with the maximum measurement test statistics ($Z_{\max}$) from the step above.

- ➔ If $Z_{\max} \leq Z_{1-\beta/2}$ no gross errors are detected. Go to *Step 5*.
- ➔ If $Z_{\max} > Z_{1-\beta/2}$ the measurement corresponding to $Z_{\max}$ is likely to contain gross error. Take it out from the data set by using nodal aggregation (treated as

unmeasured variable). If two or more $Z_{\max}$ are equal select the one with lower i index.

Step 4: Compute again the vectors x, a, and d.

- ➔ If $lb_i \leq x_i \leq ub_i$ the measurement is confirmed to have a gross error. Return to Step 2 for the reduced set of measurements
- ➔ If $lb_i \geq x_i$ or $x_i \geq ub_i$ so the reconciled values are unrealistic, put back the deleted measurement and return to *Step 3*. Identify next $Z_{\max}$ among the remaining measurements

Step 5: The presence of a gross error is confirmed for those measurements that have been permanently removed from the data set. In this case, the final reconciled values are those obtained from the *Step 4*. In case there is no gross error or was not able to verify its presence, the reconciled values are the ones obtained in *Step 1*.

If the test statistics of the last iteration are all below the threshold criterion, the last reconciled values, which are the solution to data reconciliation, are correct and can be used in further studies. In contrast, if any of the test statics exceeds the threshold criterion, the algorithm was not able to remove all the measurements with gross errors, so the reconciled values are corrupted. In this case, in order to detect the remaining gross errors in the collected, further investigation around the measurements with test statistics above the criterion needs to be done.

3.4.2 Limitations of the method

The Modified Iterative Measurement Test unfortunately has some limitations.

Limitation 1 comes from the fact that the method uses nodal aggregation for deleting those measurements that are likely to contain gross errors or for working with unmeasured variables. The procedure consists of eliminating the process variable from the objective function and also from the constraints, resulting in a reduced reconciliation problem where it is no present. This is done by aggregating the nearest neighbour nodes of the process variable.

For instance, if the measurement T15 (Figure 3.3 on the left), which is equal to 266.7ºC is suspected to have a gross error, a reduced network as the one represented in the image on the right will be obtained by eliminating the flow which contains the suspected process variable and merging both heat exchangers. For the original network the energy constraints are Q_{10}=Q_{16} and Q_{11} =Q_{17}. By aggregating both heat exchangers, the constrained equations are reduced to Q'_{10}=Q_{16} +Q_{17}. The reconciled values will be those resulting from solving the reduced network, in which the wrong temperature does not appear in the objective function nor in the constraints. Thanks to nodal aggregation, the reconciled values can be calculated anyway, without being corrupted by the wrong measurements.

However, if a gross error is detected in the measurement T23, as it is the start temperature, and so it is located in one of the sides of the network, it is not possible to do nodal aggregation since the flow where the process variable is cannot be eliminated, as it is not within two units or nodes.

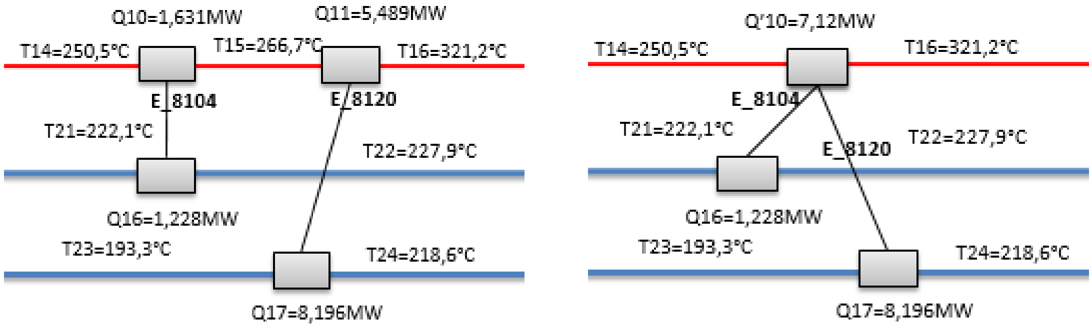

Figure 3.3 Detail of the network from the HCU before (left) applying nodal aggregation and the resulting network afterwards (right).

Limitation 2 is not specific to this methodology, but it should be also taken into consideration and reminded in this point. Neither reconciled values nor gross error can be computed and detected in non-redundant variables. As mentioned before, since they do not appear on the constraints, the reconciled value is always equal to the measured.

4 Methodology

The aim of this thesis work is to implement an application in Excel to perform data reconciliation and gross error detection and test it with the data collected from the hydrocracker unit of Preem refinery in Lysekil. This chapter starts with presentation of the study case, then it is explained why and how the tools have been developed, and finally, using two simply test cases, the tool's performances have been compared. It is also included within this chapter an explanation of the assumptions made for adapting the data collected to the requirements of the tools.

4.1 Case study: Heat exchanger network from HCU

Stream data was collected from the Hydrocracker unit at Preem refinery in Lysekil with the purpose of studying process integration possibilities. However, the presence of many inconsistencies made it difficult to use "engineering judgement" for balancing the heat exchanger network manually.

The starting point of this thesis is the imbalanced heat exchanger network from the Hydrocracker Unit. But, in order to simplify the study, only half of the streams involved have been taken into account. For the complete data set the reader is referred to the Appendix 1.

The simplified network consists of a total of 10 streams, 5 hot stream and 5 cold streams, 9 heat exchangers, 3 coolers and 1 heater. Based on the data provided (measurements of the temperatures, total heat loads of the streams and heat flow capacities) the simplified network has been represented, see Figure 4.1. To facilitate the identification, a reference number from top to bottom and left to right has been assigned for the temperatures and heat loads of the heat exchangers and utilities.

As the reader can see from the data attached in the Appendix 1, temperatures T5, T7, T14 and T26 have been measured more than one time. For these cases, the value used is the average of the measurements available. Moreover, measurements of temperatures T2, T28, T29 and T32 were not performed and the values used are instead estimations. For this reason, a second study has been conducted. The aim of this second analysis is to minimize the weight of the adjustment, which is the difference between the estimate and the reconciled value, for the estimates in the objective function.

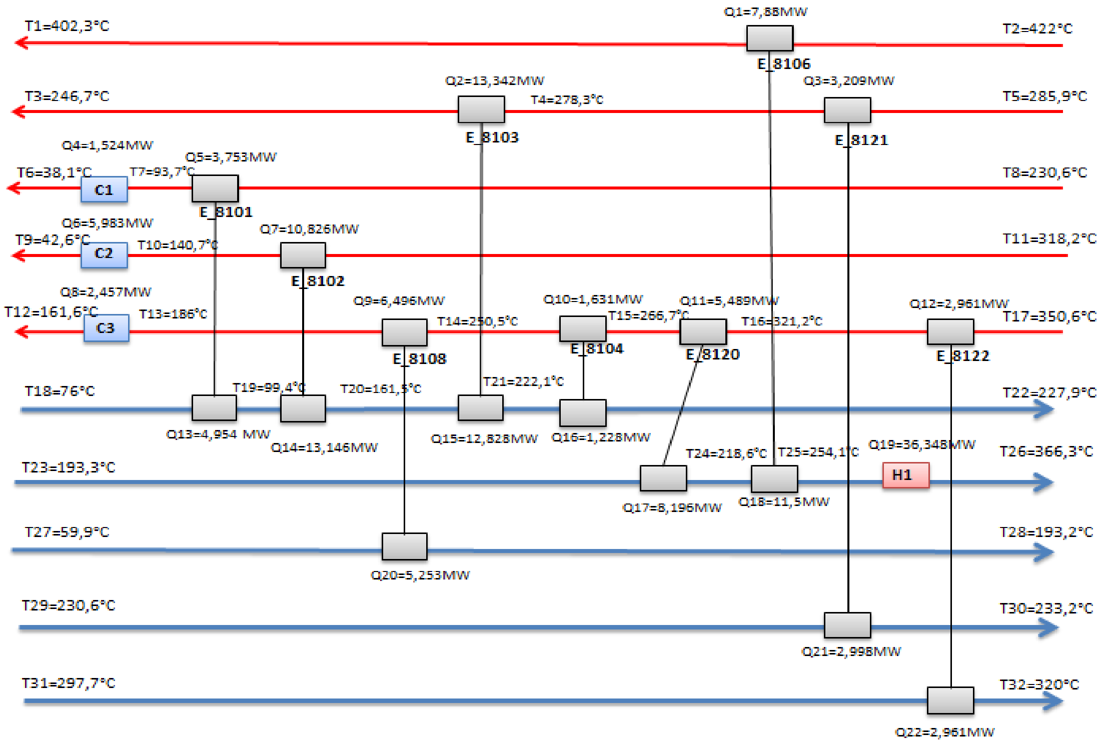

Figure 4.1 Scheme of the imbalanced heat exchanger network from the data set provided. All exchangers and utilities are referenced with a letter and a number. Letter C is used for the coolers, E for the heat exchangers and H for the heaters.

To simply the study only temperatures are assumed to be adjustable variables. Mass flows and the corresponding heat capacity flow rates, are assumed to be fixed. Therefore, the network will only be balanced from an energetic point of view, which means that the only constraints of the system will be the energy balances in the 9 heat exchangers. With that purpose, the heat losses have been neglected, and so it has been considered that heat transferred from the hot stream must be equal to heat absorbed by the cold stream, in the nine heat exchangers. The heat transferred can be calculated using the equation. (4.1) between both sides of the heat exchanger:

$$Q = \Delta H = FCp\Delta T, \tag{4.1}$$

where ΔH is the enthalpy change *(MJ/s)*, F is the flow rate *(Kg/s)* , Cp is the heat capacity ($MJ/\ Kg\,^{\circ}C$) and ΔT is the difference between target temperature and start temperature *(˚C)* .

In this case it has been assumed that there are no leaks in the network or changes in the stream composition. Thus, the value of the heat flow capacity *(FCp)* is constant and the same for the overall flow. This assumption might lead to an over or under detection of the gross errors but had to be made in order to be able to implement the Modified Iterative Measurement Test.

Given all the above considerations, the constraints for the new network are specified below, equations (4.2)- (4.10):

HX E_8101: $0.027415*(T_8 - T_7) + 0.21169*(T_{18} - T_{19}) = 0$ (4.2)

HX E_8102: $0.06099*(T_{11} - T_{10}) + 0.21169*(T_{19} - T_{20}) = 0$ (4.3)

HX E_8103: $0.422219*(T_4 - T_3) + 0.21169*(T_{20} - T_{21}) = 0$ (4.4)

HX E_8104: $0.100707*(T_{15} - T_{14}) + 0.21169*(T_{21} - T_{22}) = 0$ (4.5)

HX E_8106: $0.400011*(T_2 - T_1) + 0.323954*(T_{24} - T_{25}) = 0$ (4.6)

HX E_8108: $0.100707*(T_{14} - T_{13}) + 0.039409*(T_{27} - T_{28}) = 0$ (4.7)

HX E_8120: $0.100707*(T_{16} - T_{15}) + 0.323954*(T_{23} - T_{24}) = 0$ (4.8)

HX E_8121: $0.422219*(T_5 - T_4) + 1.152982*(T_{29} - T_{30}) = 0$ (4.9)

HX E_8122: $0.100707*(T_{17} - T_{16}) + 0.132771*(T_{31} - T_{32}) = 0$ (4.10)

As one can see, temperatures T6, T9, T12 and T26 do not appear in the constraints. Therefore, in accordance to what is explained in the Section 3.4.2 its reconciled value will be always equal to the measured value *(limitation 2)*, there is no redundancy.

Unfortunately, the studied network is also affected by *limitation 1*, more than half of the measured process variables are start or target temperatures, and so nodal aggregation cannot be performed. Thus, only intermediate measurements T4, T14, T15, T16, T19, T20, T21 and T24 can be removed from the data set, which means that, the MIMT can only confirm the presence of a gross error in these 8 measurements.

With the aim of not leaving the rest measurements away from the analysis, the test statistic has been used as indicator of the probability of containing a gross error. The probability is higher the further above the corresponding test statistic is from the threshold criterion. However, as mentioned before, since all the process variables are inter-connected by the constraints a gross error in one measurement can affect the test statistic of a good measurement, so by doing this, one assume the risk of ending up in wrong conclusion. Having the test statistic above the threshold criterion is a necessary criterion but not sufficient for confirming the presence of a gross error.

4.2 Implementation of the tools with Excel

Microsoft EXCEL is a spreadsheet application used mostly for storing datasheets, do systematic calculations and basic graphic representations. However, EXCEL includes some other powerful features, converting it in a cheaper, competitive and more user friendlier alternative to the high-level technical computing language of Matlab, especially for unfamiliar users with the latter.

When the project started, the main purpose was to implement the Modified Iterative Measured Test using VBA in order to get a systematic and fast solution for the data reconciliation and gross error detection problem. However, the VBA, which stands for Visual Basic for Applications, is an event-driven programming language that enables to automate task by building user-defined functions, creating macro-driven applications or developing customs add-ins, but which requires some time to get familiar with. In contrast, Excel has an add-in tool called Solver, which is especially designed to determine the optimal solutions that satisfy a set of constraints. This feature is already implemented and thus, much easier to use for solving data reconciliation. The only drawback is that within this tool error detection is not included.

For the reasons stated before, it was decided to develop both computer based tools; a program for reconcile the raw data set using Solver, and the program wanted from the beginning, based on the MIMT methodology and which is able also to detect the gross errors present in the data set.

The aim of this chapter is to present and explain the two computer-aided solutions implemented with EXCEL. Explain how they work, which are the inputs and outputs for each, and the modifications made in order to adapt the tools for meeting the requirements of the case study.

4.2.1 Tool 1 for Data Reconciliation

As mentioned above Solver is a simply and helpful tool for the optimization of engineering procedures where the user just needs to specify the objective function and define the constraints, whether if these are linear or not. This simply-to-use technique became even more powerful in conjunction with VBA. Thereby, the whole procedure can be automated.

The computer-based tool has been designed with the aim of providing with a reconciled network based on the assumption that gross errors are not present on the measurements. The objective function is defined by default and cannot be modified, since it is always minimizing the weighted least square difference between the measured value and the reconcile value for all the measured variables, equation (3.10).

For using the tool, the input needed is: the measurements for all the process variables, the weight of the measurement, and the constraints. The weight works as a criterion of importance, a higher weight must be defined for the measurements that are more trustable and a lower weight for the ones less reliable. In this manner, the adjustment of the process variables with higher weight will be smaller and so the reconciled value closer to the measurement.

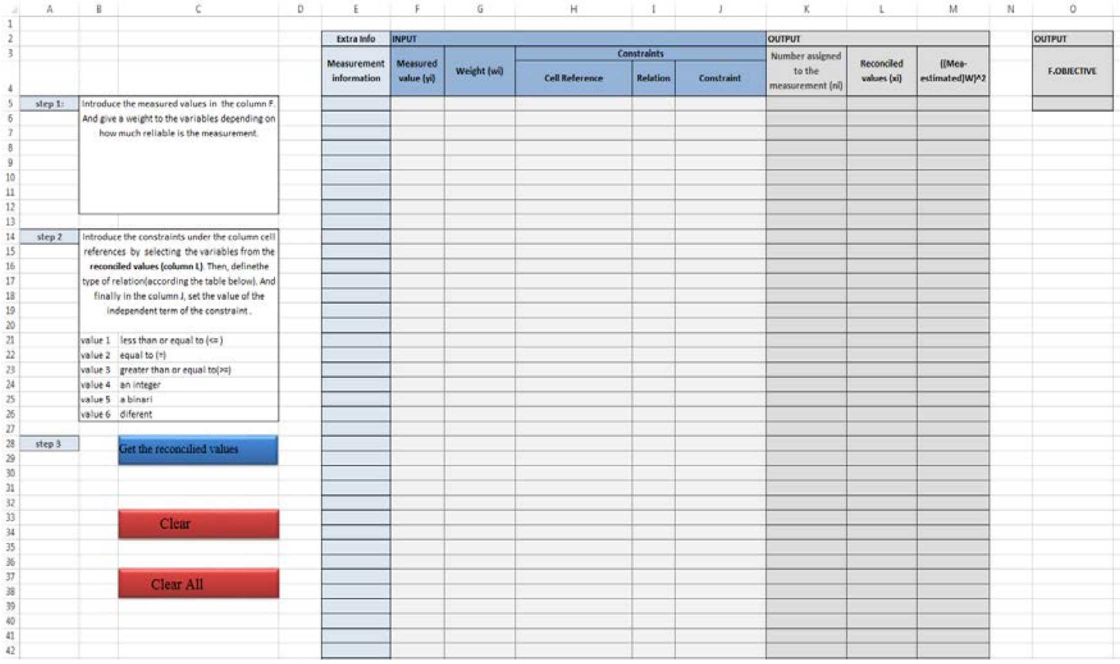

Figure 4.2 Main screen of the computer-based solution for Data Reconciliation (Tool1).

As stated before, one of the main advantages of this tool is that almost any kind of constraints can be defined. For that reason, it is important to read carefully the instructions on the left of the worksheet before defining the constraint. It is also crucial to write the constraints referred to the reconciled values and not to the measured, since the first are the variables of the problem. For instance, for introducing the constraint of the HX E_8101, which is equal to 0,027415*$(T_8 - T_7)$+0,21169* $(T_{18} - T_{19})$ =0, the left side of the equation must be written in the column called cell reference on the left, and the right side in the column called constraint in the right. The column in between is for defining the relation type. In this specific case the relation type is an equality and according to the instructions, the number 2 must be written. See Figure 4.3

INPUT				
Measured value (yi)	Weight (wi)	Constraints		
		Cell Reference	Relation	Constraint
402,3	2	0,027415*(T8-T7)+0,21169*(T18-T19)	2	0

Figure 4.3 Detail of the input data columns of the Tool 1. As an example, the constraint for the heat exchanger E_8101, equation (4.2), has been defined.

Once the input data is all introduced, it is enough to press the blue button called "Get the reconcile values" on the left side of the screen. Thanks to a simply Visual Basic code, the input data is transferred to Solver, where the reconciled values are calculated and transferred again to the main screen of the tool, where they are listed in the corresponding output column. The method chosen in Solver is the GRG (Generalized reduced Gradient), which is suitable for linear and nonlinear constraints.

4.2.2 Tool 2 for Data Reconciliation and Gross Error Detection

As stated on Section 3.4.1, the Modified Iterative Measurement Test algorithm is a reliable and effective method for data reconciliation and gross error detection.

Using the VBA programming language, the algorithm shown in the Figure 3.2 has been implemented. As a first step, the user only has to enter in the program the measured values and its bounds. And then, define the constraint matrix. The aim is to obtain the reconciled values for the balanced network, using a data set where the process measurements with gross errors have been deleted.

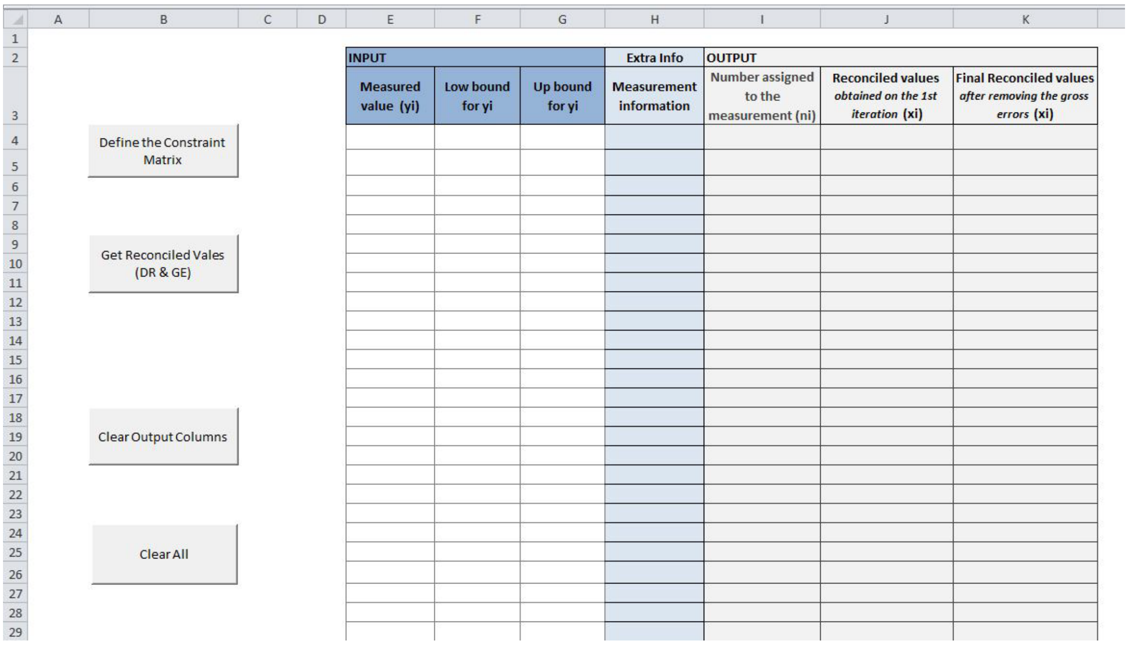

Figure 4.4 Main screen of the computer-added solution for Data Reconciliation and Gross Error Detection with MIMT (Tool2).

The Figure 4.4 shows the main screen of the Tool 2. There are in total seven columns. The three of the left are the inputs, and the four of the right the outputs, separated in the middle by an extra column, where extra information about the corresponding measurement, e.g. indicate if the measurement corresponds to a temperature, flow, composition etc… can be added.

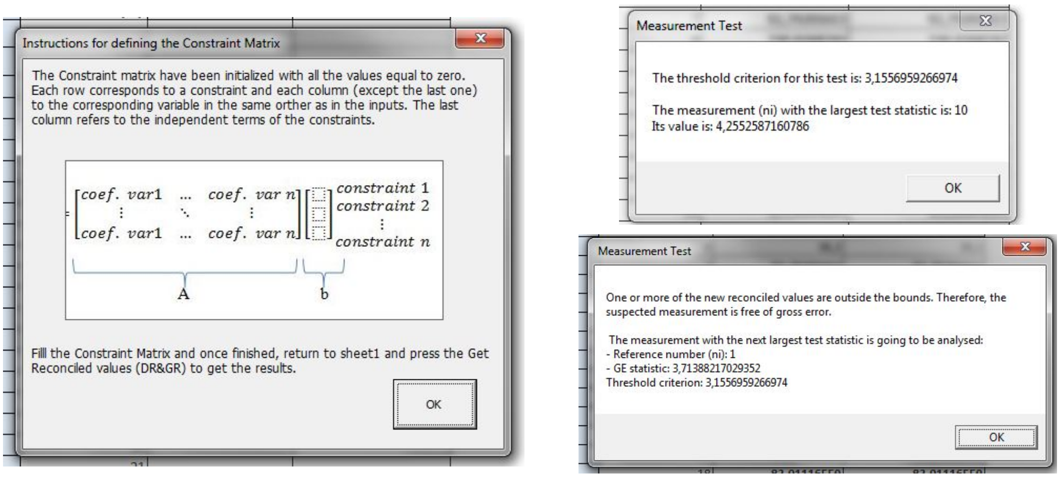

Figure 4.5 Examples of some message boxes that are shown during the simulation of the Tool 2.

Since MIMT is a complicated methodology the tool is also more difficult to use than the first one. For that reason, detailed step-by-step instructions, are provided below.

Step1- Fill the input columns with the measured values and their lower and upper bound.

Step2- Press the button called "*Define the Constraint Matrix*" on the left side. Enter the number of constraints in the Input box that is shown and press OK.

Step3- Write the constraints following the instructions that appear in the message box. See Figure 4.5 in the left.

Step4- Return to the main screen sheet and press "*Get the reconciled values (DR&GE)*".

The reconciled values computed in the step 1 using the original data set, and the reconciled values computed in the last iteration using a reduced data set, which are the final solution, will appear on the corresponding output columns. As stated before, the solution proposed by this methodology must be taken as a recommendation and must be always interpreted before its use. Because depending on the data set (input), three different situations can occur:

If the data set is free of gross error, the reconciled values calculated on the first step are already the solution to the data reconciliation problem, and the same values are listed in the solution column ("*Final Reconciled Values*"*).* For this case, all the test statistics will be below the threshold criterion.

If the data set has some gross errors, two different solutions can be obtained:

- The first, is when the tool is able to detect and eliminate all gross errors from the data set. In this case, the statistics of the last iterations are all below the threshold, and the tool only provides the reconciled values for the correct measurements.

- The second one, which is the worst case, happens when the MIMT is not able to detect the exact location of the gross errors. Therefore, the reconciled values will be calculated using a data set with gross error and so the values provided are corrupted. In this case, one or more test statistics of the last iteration exceed the threshold.

As one can see, it is highly important to follow the performance of the method and especially to know the values of the test statistics. For that reason, an information box appears every time a new statistic is compared with the threshold criterion (Figure 3.2 step 3) and also every time that a measurement is removed from the data set and the new reconciled values are calculated. See the examples on the right in Figure 4.5. With the same purpose, there is an extra sheet, called Data, when the adjustments, its covariance, the test statistics from the first iteration and the test statistics from the last iteration can be consulted.

Assumptions for this specific thesis work

In contrast with the Tool 1, since this method also performs gross error detection based on statistical test, it is necessary to specify the values of the standard deviation for all the measurements. However, as the reader can see from the data attached in the Appendix 1, the standard deviation of the measurements is not available, and cannot be calculated because in almost of the cases, there is only one measurement for each process variable. In order to solve the problem and achieve the initial goal, a method for estimating the standard deviation has been developed based on the definition of standard deviation.

The standard deviation is an estimation of the dispersion of the measurements, in other words, an indicator of how wide is the range of values that the measurement can take. Based on the theory, the measurements of a process variable follow a Normal Distribution *N* (μ, σ) with mean μ and standard deviation σ. According to this, and as it is shown on the Figure 4.6, the 68% of the measurements will fall within the range defined by plus-minus one standard deviation, and the 95% within the range defined by plus-minus two times the standard deviation.

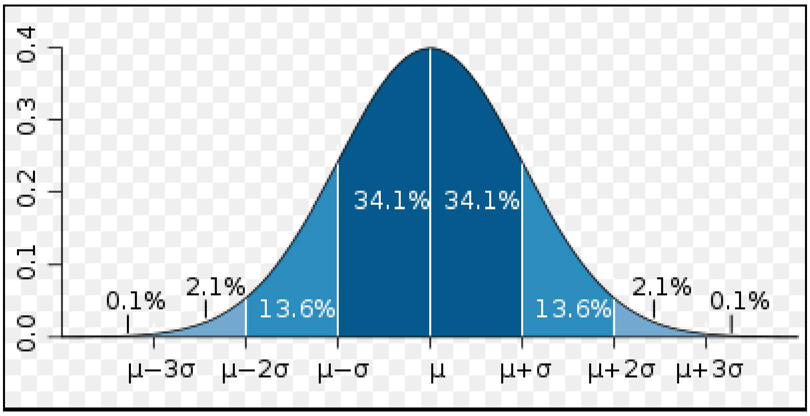

Figure 4.6 Plot of a normal distribution, where each band has a width of 1 standard deviation (Wikipedia)

Based on the operating conditions of the different process units and the accuracy of the measuring instruments, a lower and upper bound for each measurement has been estimated. Outside this range, assuming a confidence level of 95%, the measurement is considered incorrect. It has been considered as true, that this range is the same as the one defined by plus minus two times the standard deviation, equation 4.11. Accordingly, standard deviation can be calculated as the upper bound minus the lower bound divided by four.

$$[\mu - 2\sigma, \mu + 2\sigma] = [lower\ bound, upper\ bound] \quad (4.11)$$

This rough estimation tends to relax the system and under detect the gross errors (*Type II Error*), since the ranges defined are wider and so the values for the standard deviation are higher. However, considering that same procedure is used for all the measurements, it will not lead to incorrect detection of the error location.

This developed procedure might be very useful also in further studies because most of the times, as in this case, the values of the standard deviation for the measurements are unknown. Nevertheless, process variables are always constrained by the requirements of the process, thereby, is easier to have an estimation of the bounds than of the standard deviations.

4.3 Comparison of the two tools.

After the implementation and before their use with the real data set from Preem Refinery, both tools have been tested. With that purpose, two test cases based on an example from teaching material from the National University of Singapore were performed. The wordings and the solutions are attached in the Appendix 3. The stream network in both cases is the same, the only difference is that in the first case the data set is free of gross errors, whereas in the second one, a gross error has been introduced in stream 2. The process includes five units and ten streams, and the aim is to reconcile the stream values so the mass balances are satisfied. The results of both performances are shown in the Table 4.1 and Table 4.2.

Table 4.1 *The table show the results obtained from the two computed based tools where the data set is free of gross errors (test case 1).*

INPUT: case 1		**OUTPUT**		
Measured value (yi)	**Standard deviation (σi)**	**Reconciled values Tool 1 (xi)** DR with Solver	**Reconciled values Tool 2 (xi)** MIMT	
			1st Iteration (step 1)	**Last Iteration** (final solution)
100	5	92.38546574	92.38546575	92.38546575
90	2	92.38546574	92.38546575	92.38546575
45	2	43.83285975	43.83285973	43.83285973
50	2	48.55260599	48.55260601	48.55260601
120	10	127.0063428	127.006343	127.006343
40	5	39.67553776	39.6755378	39.6755378
38	5	38.77819907	38.77819914	38.77819914
10	5	11.42712354	11.42712354	11.42712354
50	5	51.1026613	51.10266134	51.10266134
100	10	89.88086037	89.88086048	89.88086048

In the first test case (Table 4.1), reconciled values obtained with both tools are exactly the same. Since no errors are present, it is only a data reconciliation problem, and although the methodology is different, the principle behind both tools is the same, so the results must be also the same.

In the second case (Table 4.2), Tool 1 calculates the reconciled values using all the data, consequently the gross error in the stream 2 is spread out, affecting all the reconciled values. However, in Tool 2, based on the reconciled values of the first step, the test statistics are calculated and Measured Tests are performed. The gross error introduced in the stream 2 is successfully detected and the corresponding measurement, in this case the flow rate, is removed from the data set. Using this new data set and the new reconciled values, the test statistics are calculated again (2nd iteration), since all of them are below the threshold criterion the simulation ends. For this simple example, the performance of Tool 2 is excellent. The final solution is more adjusted, the values of the reconciled flow rates are closer to the measurements.

Table 4.2 *The table show the results obtained from the two computed based tools where an error in the flowrate of the stream 2 has been introduced (test case 2).*

INPUT: case 2		**OUTPUT**		
Measured value (yi)	**Standard deviation (σi)**	**Reconciled values Tool 1 (xi)** DR with Solver	**Reconciled values Tool 2 (xi)** MIMT	
			1st Iteration (step 1)	**Last Iteration** (final solution)
100	5	104.3804579	104.3804579	95.95993355
110	2	104.3804579	104.3804579	removed
45	2	49.91866893	49.91866886	45.64641063
50	2	54.46178897	54.46178901	50.31352291
120	10	131.4219961	131.4219962	128.3221929
40	5	39.02617705	39.02617703	39.48203045
38	5	37.93403009	37.93403014	38.52663958
10	5	11.88167602	11.88167608	11.56257869
50	5	50.90785307	50.90785311	51.04460913
100	10	88.84188316	88.84188325	89.57124872

Both tools performed as expected, Tool 1 solves DR based on the data set provided, whereas Tool 2 detects gross errors and uses the resulting data set for solving DR. It must be highlighted that the detection of errors is doubly beneficiary because, on one side, incorrect measurements are detected, but on the other side, the quality of the reconciled values obtained is higher due to uncorrected values having been removed from the data set, so the reconciled values tend to be closer to the measured values.

However, one of the biggest challenges is about how to deal with large sets of data and complicated networks, where many units and process are involved, and how to do it faster. This is where Tool 1 gains advantage. Solver add-in was exactly designed for solving an optimization function subjected to some constraints, which is exactly the core of data reconciliation. Although it does not guarantee an optimal solution, if the user has a basic knowledge about the process that has been measured, is less probable to commit a mistake. Moreover, it is effortless to use, more powerful, and less time consuming. Additionally, Solver accepts any kind of constraints; linear, non-linear and even inequalities, avoiding having to use complicated linearized methods as a previous step to data reconciliation. Finally, another advantage is that the standard deviation of the measurements neither the bounds are needed, it is enough with setting a weight for the measurements depending on its accuracy.

In conclusion, any of the techniques met, in one way or another, the initial goal: develop a tool for systematically reconciliation of measured process data for its use in future process integration studies. The first is specially indicated for high quality and large data sets free of gross error, in which not all the constraints are linear. In contrast, the second one is more suitable for linear systems and smaller, bad or unknown origin of the data set.

5 Results and discussion

In this chapter the results of the data reconciliation problem obtained by the two implemented tools are presented, analyzed individually and compared.

In order to make the discussion easy and more understandable, this chapter also includes a summary of previous explanation about the simulations made and how the discussion of the results has been conducted.

5.1 Illustration of use of the two tools based on the case study

The need to create a systematic tool for analysing stream data arose from the purpose of studying opportunities for process integration in the hydrocracker unit at Preem's refinery in Lysekil.

Since two different tools are used, two sets of reconciled values are obtained for a same set of data. The results obtained using Tool 1 are only the solution to the data reconciliation assuming that no gross errors are present on the data set, and may differ from the ones obtained with Tool 2 if any measurement with gross error is present. It is interesting to compare those differences and examine if same conclusions can be derived from the results obtained by the two tools, or in contrast, the deductions differ.

In order to compare the solutions on the same basis, DR & GED, after obtaining the reconciled values from the Tool 1, a gross error detection study applying engineering judgment was performed. In order to do so, the attention was mainly focused in two parameters. First, it was considered that those temperatures whose reconciled value is outside of the range defined by the bounds could be suspected of containing an error. By definition, see Section 4.2.2, Figure 4.6, if the reconciled value is outside the range the measurement is incorrect. In addition, the resulting balanced network is infeasible. And secondly, special attention was paid to the variables whose reconciled value differs from the initial temperature (the measurement) by more than 5ºC. According to theory (see Sections 3.1 and 3.2), random errors are small in magnitude, whereby adjustments to data assumed to contain random errors only should also be small in magnitude. The choice of 5ºC as maximum allowable adjustment was judged to be a reasonable choice comparing it with the bounds defined.

Based on these two standards some hypotheses were formulated. These hypotheses are compared with the results obtained using Tool 2. Hypotheses are verified or rejected based on the values of the test statistics.

For setting the hypotheses, the Smearing Effect was taken into account. Process variables are all inter-related by the constraints, so an error in one of them might lead the test statistic of a good measurement to exceed also the test criterion (*Type I Error*). Thus, if the test statistics of two process variable that are connected by a heat exchanger are above the criterion, there is a high probability that one of them has a gross error, but not both. For this reason, measurements suspected that are directly connected by an energy balance were grouped together in the same hypothesis.

Moreover, as it has been mentioned in Section 4.1, measurements of the temperatures T2, T28, T29 and T32 are estimates, meaning that the initial values used for these process variables are less accurate than the other ones. In order to provide with a balanced network as closely as possible to the original one, a second analysis where a smaller weight in the objective function is given to the four estimated measurements was undertaken. In this manner, the difference between the initial value and the reconciled value has less repercussion than the same difference in the temperatures in which the initial temperature is measured.

Table 5.1 *Input data used for both analysis. The temperatures highlighted are the ones which bounds differ in both analysis (estimates measurements).*

Reference (ni)	**INPUT:**	**Analysis 1**		**Analysis 2**	
	Measured value (yi) [˚C]	**Lower bound for yi [˚C]**	**Upper bound for yi [˚C]**	**Low bound for yi [˚C]**	**Up bound for yi [˚C]**
T1	402.3	401.3	403.3	401.3	403.3
T2	422	418	426	412	432
T3	246.7	241.7	251.7	241.7	251.7
T4	278.3	273.3	283.3	273.3	283.3
T5	285.9	282.8	288.8	282.8	288.8
T6	38.1	33.1	43.1	33.1	43.1
T7	93.7	88.7	98.7	88.7	98.7
T8	230.6	227.6	233.6	227.6	233.6
T9	42.6	37.6	47.6	37.6	47.6
T10	140.7	135.7	145.7	135.7	145.7
T11	318.2	313.2	323.2	313.2	323.2
T12	161.6	156.6	166.6	156.6	166.6
T13	186	181	191	181	191
T14	250.5	245.6	255.6	245.6	255.6
T15	266.7	261.7	271.7	261.7	271.7
T16	321.2	316.2	326.2	316.2	326.2
T17	350.6	345.6	355.6	345.6	355.6

T18	76	71	81	71	81
T19	99.4	94.4	104.4	94.4	104.4
T20	161.5	156.5	166.5	156.5	166.5
T21	222.1	217.1	227.1	217.1	227.1
T22	227.9	224.9	230.9	224.9	230.9
T23	193.3	188.3	198.3	188.3	198.3
T24	218.6	213.6	223.6	213.6	223.6
T25	254.1	249.1	259.1	249.1	259.1
T26	366.3	363.3	369.3	363.3	369.3
T27	59.9	54.9	64.9	54.9	64.9
T28	193.2	183.2	203.2	183.2	203.2
T29	230.6	225.6	235.6	220.6	240.6
T30	233.2	228.2	238.2	228.2	238.2
T31	297.7	292.7	302.7	292.7	302.7
T32	320	315	325	310	330

The final goal is to provide a list of the reconciled values and include a recommendation of which measurements should be repeated or removed because a gross error has been detected. The reader is encouraged to follow the discussion of the results and to understand the outputs of both tools, as the computer aided solutions implemented are intended to be used in further studies.

5.2 Analysis 1: Original data set

Table 5.2 lists the results obtained from both tools and their variation in temperature from the measured value. As can be seen, the reconciled process variables are practically the same in both cases. However, this can occur because no gross errors are present on the data set, or because the MIMT algorithm was not able to detect the exact location of the gross error, which is the current case. None of the suspected measurements are removed from the data set, therefore, the reconciled values of Tool 2 are also calculated using the initial data set.

Table 5.2 *Results of the first analysis obtained for each tool. The values of the last column (ΔT), are the absolute variation between the measured values and the reconciled values from the Tool 1. Temperatures highlighted are the non-redundant ones.*

Rerence (ni)	**Measured value (yi) [˚C]**	**Reconciled values (xi) [˚C] Tool 1**	**Reconciled values (xi) [˚C] Tool 2**	**ΔT [˚C] ($\lvert T_{meas}-T_{tool1} \rvert$)**
T1	402.3	402.019	402.014	0.3
T2	422	426.519	426.573	4.5
T3	246.7	245.776	245.774	0.9
T4	278.3	279.199	279.200	0.9
T5	285.9	285.917	285.909	0.0
T6	38.1	38.109	38.100	0.0
T7	93.7	92.824	92.792	0.9
T8	230.6	230.941	230.927	0.3
T9	42.6	42.612	42.600	0.0
T10	140.7	138.287	138.248	2.4
T11	318.2	320.711	320.652	2.5
T12	161.6	161.637	161.600	0.0
T13	186	190.705	190.721	4.7
T14	250.5	246.491	246.499	4.0
T15	266.7	263.928	263.991	2.8
T16	321.2	322.934	322.826	1.7
T17	350.6	350.968	350.963	0.4
T18	76	83.044	83.011	7.0
T19	99.4	100.931	100.900	1.5
T20	161.5	153.489	153.453	8.0
T21	222.1	220.152	220.123	1.9

T22	227.9	228.446	228.445	0.5
T23	193.3	199.657	199.698	6.4
T24	218.6	218.000	217.988	0.6
T25	254.1	248.252	248.313	5.8
T26	366.3	366.273	366.300	0.0
T27	59.9	58.018	58.053	1.9
T28	193.2	200.577	200.589	7.4
T29	230.6	230.647	230.672	0.0
T30	233.2	233.107	233.128	0.1
T31	297.7	298.205	298.178	0.5
T32	320	319.468	319.522	0.5

Note that as foretold before, the non-redundant temperatures; T6, T9, T12 and T26 (highlighted in Table 5.2) have the same reconciled value as the measurement. These temperatures cannot be reconciled nor checked for an error, thus, they are left out of the study.

5.2.1 Discussion of the results obtained from Tool 1

Reconciled temperatures obtained using Tool 1 are listed in the third column of Table 5.2.

The heat loads for the balanced network were calculated, the values are listed together with the ones from the imbalanced network in Table 5.3. The highest relative variation is found in the cold stream of heat exchanger E-8104 with a relative variation of 43% from the imbalanced network. However, even though the percentage is high, the consequence for the network is small. The heat absorbed by the cold stream rises from 1.227 MW to 1.756MW. The gap between the temperatures of the heat exchangers is small, therefore any change in the temperatures, even if it is small, will affect in greater proportion. In fact, the reconciled start temperature (T21) differs by less than 2 ºC from the measured one, and the difference in the target temperature (T22) is even less, it is only 0.5 ºC.

Table 5.3 *The table shows the heat duties from the original network and the heat duties computed with the reconciled process variables for each side of the heat exchangers, and for the coolers and heaters. It is also calculated the relative variation between both values. Higher relative differences are highlighted.*

Utility	**Hot stream**			**Cold stream**		
	Qmeas. [MW]	**Qrec. [MW]**	**AQ [%]**	**Qmeas. [MW]**	**Qrec. [MW]**	**AQ [%]**
HX_8101	3.753	3.786	1%	4.954	3.786	24%
HX_8102	10.826	11.126	3%	13.146	11.126	15%
HX_8103	13.342	14.112	6%	12.828	14.112	10%
HX_8104	1.631	1.756	8%	1.228	1.756	43%
HX_8106	7.880	9.800	24%	11.500	9.800	15%
HX_8108	6.496	5.618	14%	5.253	5.618	7%
HX_8120	5.489	5.942	8%	8.196	5.942	27%
HX_8121	3.209	2.836	12%	2.998	2.836	5%
HX_8122	2.961	2.823	5%	2.961	2.823	5%
C1	1.524	1.500	2%			
C2	5.983	5.835	2%			
C3	2.457	2.927	19%			
H1				36.348	38.233	5%

The next significant relative change is equal to 28% and occurs in the cold side of heat exchanger E-8120. For this case the heat load varies from 8.196 MW to 5.942 MW, which means a decrease of more than 2.2 MW, which is the largest absolute change. In contrast with the previous case, the reconciled start temperature of the HX, T23, is 6 ºC higher than the measured value. Since the reconciled value for T23 is outside the bounds, according to the criterion defined in the section 5.1, this measurement is suspected of having a gross error.

The hot stream of the heat exchanger E-8106 has also a relevant variation, around 25%. The heat recovered has increased by almost 2 MW. However, the start and target temperatures for this HX are T2 and T1, respectively, T2 being one of the estimated measurements. Therefore, if the adjustment is significant, it may be considered the possibility that the estimation made is incorrect. According to Tool 1, the correct values is 426.573 ˚C, 4.5 ˚C higher than the estimate.

The same relative variation is reached for the heat recovered by the cold stream in the heat exchanger E-8101, which decrease from 4.954 MW to 3.787 MW.

The substantial differences found in the heat loads, suggest that there are one or more gross errors in the initial measurements. Five reconciled temperatures are outside the range defined by the bounds, which according to the standards defined, suggest the presence of gross errors. T18 and T20 are the ones farthest from the limits, thus more likely to be wrong. However, as can be seen from the Figure 5.1, both temperatures are in the same stream and they are inter-connected by the energy balances of HX E-8101 and HX E-8102 through the temperature T19. According to the theory about smearing effects, the presence of a gross error in the measurement T18 can affect the test statistics of the process variables near by the latest, and vice versa. Thereby, there is a high probability that one of the temperatures was wrongly measured, but not both. Following this reasoning hypotheses have been set out.

The first hypothesis $H_{1,A}$ made is that the measurement of T18 or T20 has a gross error. The same theory can be applied for the reconciled temperatures T2, T23 and T25, which are also outside the range. All three temperatures are inter-connected by the HX E-8106 and for the temperature T24, thus an error in one of them may be spread out to the rest. The second hypothesis $H_{1,B}$ is that any of the initials values for temperatures T2, T23 and T25 contain a gross error.

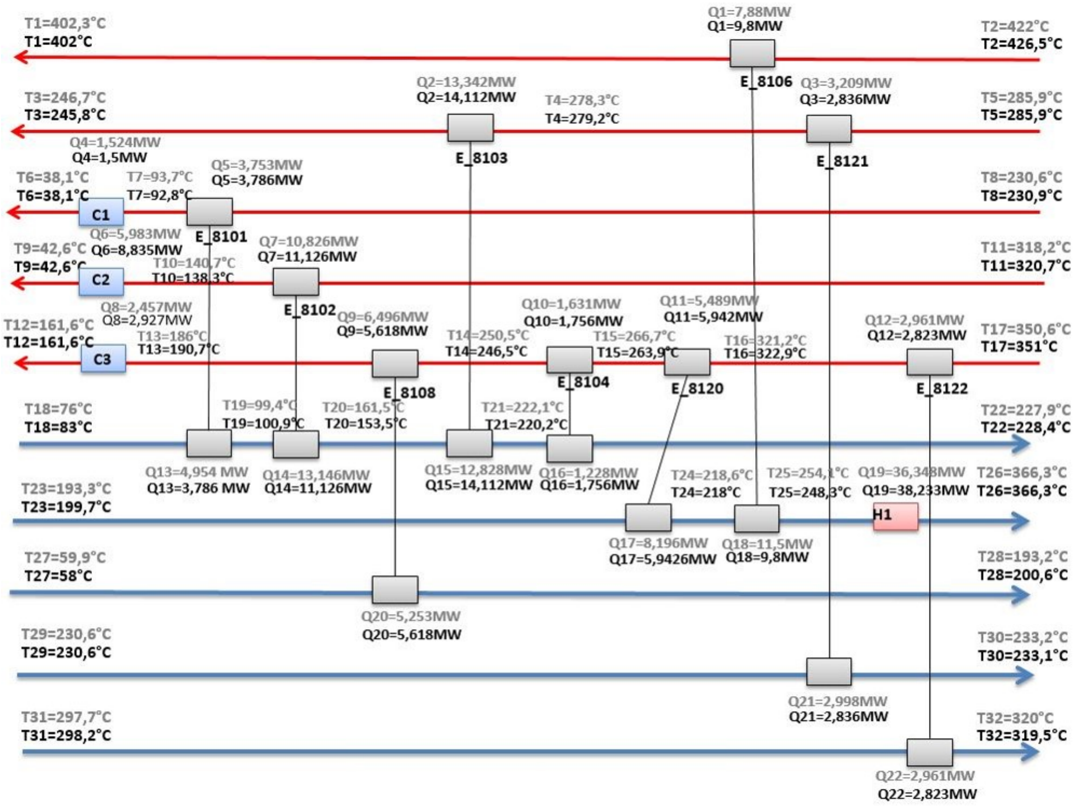

Figure 5.1 Balanced HX network proposed by the Tool 1. Values in grey correspond to the initial imbalanced HX network (starting point), and values in bold are the reconciled ones.

Focusing on the differences found between the measured values and the reconciled values (second standard), similar hypotheses can be drawn. The largest variation is 8 ºC, and it is reached for the temperature T20, which is already included in the hypothesis $H_{1,A}$. Measurements T18 and T28 have adjustments of around 7 ºC. The first is also included in $H_{1,A}$, whereas, for the second one a new hypothesis should be added. Thus, hypothesis $H_{1,C}$ is set out since the initial estimate value for T28 is wrong.

The conclusions extracted from the first simulation are:

- Hypothesis $H_{1,A}$: T18 or T20 has a gross error
- Hypothesis $H_{1,B}$: T2, or T23 or T25 has a gross error.
- Hypothesis $H_{1,C}$: the estimated value for T28 is not correct.

It is interesting to notice that, besides T28, the remaining suspected measurements and heat exchangers with higher relative variation in the heat loads are all concentrated in the first hot stream and in the first and second cold streams. This could mean that the probability of a gross error in any of the flow measurements is quite high.

5.2.2 Discussion of the results obtained from Tool 2

Reconciled values obtained using Tool 2 are exactly the same as the ones obtained using Tool 1. However, some of the final test statistics exceed the threshold criterion, thus it can be confirmed that the solution provided is affected by at least one gross error which means that the reconciled values are not valid.

When the calculations are performed with Tool 2, a set of 10 measurement are suspected to contain errors. However, the presence of a gross error cannot be confirmed in any of them. The reason is because nodal aggregation can only be applied to T20, all the other measurements are affected by *limitation 1*, and the resulting reconciled values obtained for this case are outside the bounds, and so, the solution is unfeasible. Due to these reasons, the exact location of the gross error is unknown and none of the measurements suspected to contain errors are confirmed and eliminated from the data set. After 10 iterations, the final reconciled values are the same as the ones computed in the first step of the first iteration, and so, they are also the same as the ones provided by the Tool 1. In consequence, the analysis of the results is conducted using the values of the test statistics, Table 5.4.

Table 5.4 *The table shows the values of the Measured Test for all the process variable. The test statistics from the first iteration are listed in the third and seventh column. The final test statistics, after all the iterations, are in the columns forth and eight. The word "check" means that the corresponding process variable was attempted to remove from the sample, and the number indicates the order.*

Reference (ni)	**Measured value (yi) [˚C]**	**Zi (1st iteration)**	**Zi (last iteration)**	**Reference (ni)**	**Measured value (yi) [˚C]**	**Zi (1st iteration)**	**Zi (last iteration)**
T1	402.3	3.714	check_4	T17	350.6	0.338	0.338
T2	422	3.714	check_5	T18	76	3.476	check_9
T3	246.7	0.543	0.543	T19	99.4	0.740	0.740
T4	278.3	0.524	0.524	T20	161.5	4.097	check_3
T5	285.9	0.042	0.042	T21	222.1	1.019	1.019
T6	38.1	0.000	0.000	T22	227.9	0.768	0.768
T7	93.7	3.476	check_7	T23	193.3	3.465	check_10
T8	230.6	3.476	check_8	T24	218.6	0.323	0.323
T9	42.6	0.000	0.000	T25	254.1	3.714	check_6
T10	140.7	4.255	check_1	T26	366.3	0.000	0.000
T11	318.2	4.255	check_2	T27	59.9	3.063	3.063
T12	161.6	0.000	0.000	T28	193.2	3.063	3.063
T13	186	3.063	3.063	T29	230.6	0.042	0.042
T14	250.5	2.473	2.473	T30	233.2	0.042	0.042
T15	266.7	2.563	2.563	T31	297.7	0.338	0.338
T16	321.2	1.394	1.394	T32	320	0.338	0.338

The measured test is performed for the modified level of significance β, using equation 3.22 and setting $\alpha = 5\%$ and $m = 32$. Using these values, the threshold criterion $Z_{1\text{-}\beta/2}$ is equal to 3.155. Therefore, any of the measurements with a test statistic that exceeds $Z_{1\text{-}\beta/2}$ is suspected to contain a gross error. The larger the measurement test statistic is, the higher the probability there is a gross error. Non-constrained measurements (T6, T9, T12 and T26) have a test statistic equal to zero since the reconciled value is the same as the measured, leading to an adjustment equal to zero.

Therefore, for the conducted study measurements T10, T11, T20, T1, T25, T2, T8, T18, T17 and T23 are likely to contain a gross error. The stated order goes from larger to smaller measured test, thus, the first ones have a higher probability of containing a gross error than the last ones. These results confirm hypothesis $H_{1,A}$. Test statistics for the process variables T18 and T20 exceed the threshold, so they are likely to contain a gross error. In the same line of reasoning as before, measurements T7 and T8 must also be included in this hypothesis since they are related to T18 and T20 by the heat exchanger E-8101. Hypothesis $H_{1,B}$ is also confirmed, T2, T23 and T25 have measured test above the criterion. In this case T1 must be added to $H_{1,B}$, since it is directly connected with T2 by the HX E_8106. Hypothesis $H_{1,C}$ is rejected, because the test statistic for T28 is below the criterion, and new hypothesis must be set out for the measurements T10 and T11. Hypothesis $H_{1,D}$: T10 or T11 has a gross error.

5.2.3 Comparison and conclusions of the first analysis

Three hypotheses are set out in the section 5.2.1. Based on the test statistics from Tool 2, two hypothesis are confirmed and one is rejected, $H_{1,C}$. Moreover, three measurements have been included to the hypotheses already drawn, and a new hypothesis for T10 and T11 has been set out.

- Hypothesis $H_{1,A}$: T18, or T20, or T8 or T7 has a gross error
- Hypothesis $H_{1,B}$: T1, or T2, or T23 or T25 has a gross error.
- Hypothesis $H_{1,D}$: T10 or T11 has a gross error.

Results showed that the data collected from the hydrocracker unit has at least one gross error located in any of the 10 measurements listed in the hypotheses. However, since the exact location is unknown it is not possible to formulate data for a balanced heat exchanger network. The reconciled values calculated with both tools are corrupted by the presence of gross errors and their use in process integration studies can lead to erroneous conclusions.

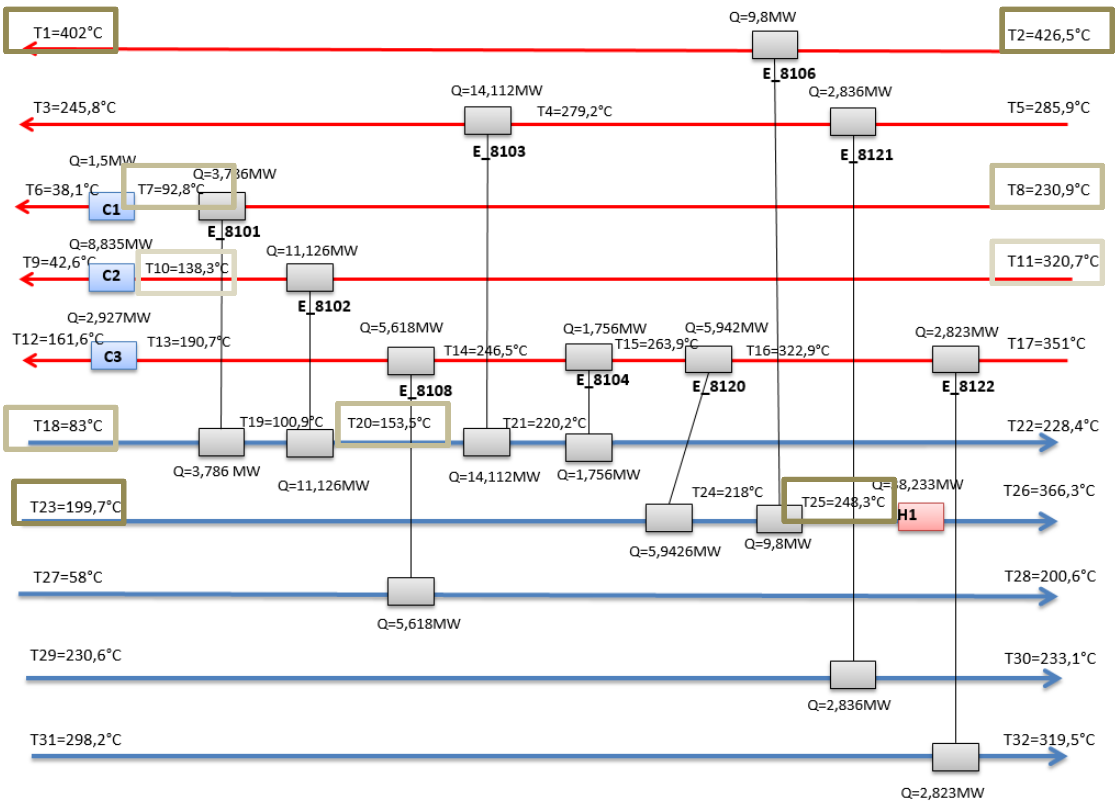

Figure 5.3 Temperatures suspected of having a gross error are squared. Different color has been used for each hypothesis.

5.3 Analysis 2: larger bound values for the estimate measurements

The aim of this second analysis is to find the exact location of the gross error and to formulate data for a valid balanced network for process integration studies of the HCU. This is not a new study but a continuation of the former, in which a distinction between the temperatures whose initial value is a measurement and those four temperatures (T2, T28, T29 and T32) whose initial value is estimated was made. This second analysis is done by establishing a criterion of importance where the adjustment for the estimates have less weight in the objective function, equation (3.7), than the adjustment of the measurements.

According to the theory, the weight can also be understand as the inverse of the standard deviation. Thereby, setting a smaller weight is equal to defining a higher standard deviation. And as it is shown in Figure 5.4, a higher standard deviation means also a wider range. With this purpose, using equation (4.11), a variation of 10 K around the measured value was defined for the four variables whose measurements were estimated, while the original values were retained for the other variables. Temperature T28 is not affected since the width range by default was already equal to 20 K.

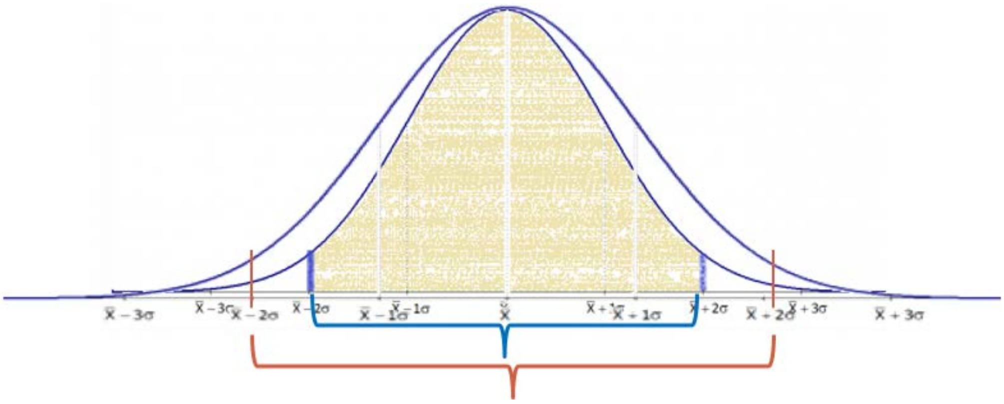

Figure 5.4 Consequences of increasing the standard deviation for the estimates. The original range for the estimates is represented in blue and the new one in red.

The range of values for which the reconciled value of a process variable is considered correct is now wider. Therefore, the reconciled value is less restricted and it can adopt a value further away from the estimates. If the adjustment, which is the difference between the estimate and the reconciled value, increases significantly, it is because the reconciled value tends to a value which is away from the estimate. In this case, it is highly probable that the estimation made was wrong. This is a way to validate the assumptions made for the estimates.

The results of the simulations for both tools are detailed in Table 5.5. In this case, the reconciled values obtained in both simulations are not the same.

Table 5.5 *The table shows the reconciled values from the analysis 2 obtained from both tools. The rows highlighted correspond to the estimates temperatures.*

Rerence (ni)	**Measured value (yi) [˚C]**	**Reconciled values (xi) [˚C] Tool 1**	**Reconciled values (xi) [˚C] Tool 2**	**ΔT [˚C] ($\|T_{meas}-T_{tool1}\|$)**	**ΔT [˚C] ($\|T_{meas}-T_{tool2}\|$)**
T1	402,3	402,204	402,204	0,096	0,096
T2	422	431,548	431,562	9,548	9,562
T3	246,7	245,815	244,609	0,885	2,091
T4	278,3	279,174	280,350	0,874	2,050
T5	285,9	285,904	285,915	0,004	0,015
T6	38,1	38,100	38,100	0,000	0,000
T7	93,7	92,792	93,270	0,908	0,430
T8	230,6	230,927	230,755	0,327	0,155

T9	42,6	42,598	42,600	0,002	0,000
T10	140,7	138,251	140,398	2,449	0,302
T11	318,2	320,648	318,502	2,448	0,302
T12	161,6	161,598	161,600	0,002	0,000
T13	186	190,750	190,689	4,750	4,689
T14	250,5	246,550	246,442	3,950	4,058
T15	266,7	264,472	264,632	2,228	2,068
T16	321,2	322,493	322,503	1,293	1,303
T17	350,6	350,731	350,734	0,131	0,134
T18	76	83,005	79,322	7,005	3,322
T19	99,4	100,894	97,127	1,494	2,273
T20	161,5	153,445	removed	8,055	-
T21	222,1	219,978	219,725	2,122	2,375
T22	227,9	228,504	228,378	0,604	0,478
T23	193,3	197,899	197,923	4,599	4,623
T24	218,6	215,936	215,913	2,664	2,687
T25	254,1	252,169	252,164	1,931	1,936
T26	366,3	366,300	366,300	0,000	0,000
T27	59,9	58,039	58,065	1,861	1,835
T28	193,2	200,634	200,539	7,434	7,339
T29	230,6	230,714	231,050	0,114	0,450
T30	233,2	233,178	233,088	0,022	0,112
T31	297,7	297,877	297,877	0,177	0,177
T32	320	319,296	319,293	0,704	0,707

5.3.1 Discussion of the results obtained from Tool 1

Reconciled values obtained using Tool 1 are listed in Table 5.5. The heat loads have been also calculated and listed in Table 5.6.

The heat exchangers with largest relative differences in the heat loads coincide with the ones highlighted in the first analysis. However, in this case, the largest variation, which is almost 50%, is found in the heat transferred by the hot stream in HX E-8106. More precisely, it is allocated between the temperatures T1 and T2, T2 being one of the estimated temperatures.

Table 5.6 *The table shows the heat duties from the original network and the heat duties computed with the reconciled process variables for each side of the heat exchangers, and for the coolers and heaters. It is also calculated the relative variation between both values. The largest relative variation have been highlighted.*

Utility	**Hot stream**			**Cold stream**		
	Qmeas. [MW]	**Qrec. [MW]**	**AQ [%]**	**Qmeas. [MW]**	**Qrec. [MW]**	**AQ [%]**
HX_8101	3,753	3,787	1%	4,954	3,787	24%
HX_8102	10,826	11,124	3%	13,146	11,124	15%
HX_8103	13,342	14,084	6%	12,828	14,084	10%
HX_8104	1,631	1,805	11%	1,228	1,805	47%
HX_8106	7,880	11,738	49%	11,500	11,738	2%
HX_8108	6,496	5,620	13%	5,253	5,620	7%
HX_8120	5,489	5,843	6%	8,196	5,843	29%
HX_8121	3,209	2,842	11%	2,998	2,842	5%
HX_8122	2,961	2,844	4%	2,961	2,844	4%
C	1,524	1,499	2%			
C	5,983	5,834	2%			
C	2,457	2,936	19%			
H				36,348	36,973	2%

Following the standards defined, four measurements are now suspected of having a gross error. The number of reconciled temperatures that fall outside the bounds have been reduced from five to two. Only process variables T18 and T20 remain outside the bounds, which means that hypothesis $H_{1,A}$ is verified again.

Measurements with largest adjustments (T2, T18, T20 and T28) are the same than in the previous analysis and even the reconciled values coincide, besides for T2. In this case, the reconciled temperature increases from 426.52 ˚C to 431.55 ˚C, reaching an adjustment of 9.5˚C, only 0.5 ˚C below the new upper bound. Therefore, an increase in the acceptable temperature range for T2 has led to a disproportionate increase in the reconciled temperature, which otherwise has helped to stabilize the network. Such behaviour confirms the hypothesis $H_{1,B}$, the estimation of the measured value of T2 is not correct. In this case anything can be stated from hypothesis $H_{1,D}$.

In general, the reconciled values calculated by Tool 1 are closer to the measurements, but for the ones that varies the differences are larger.

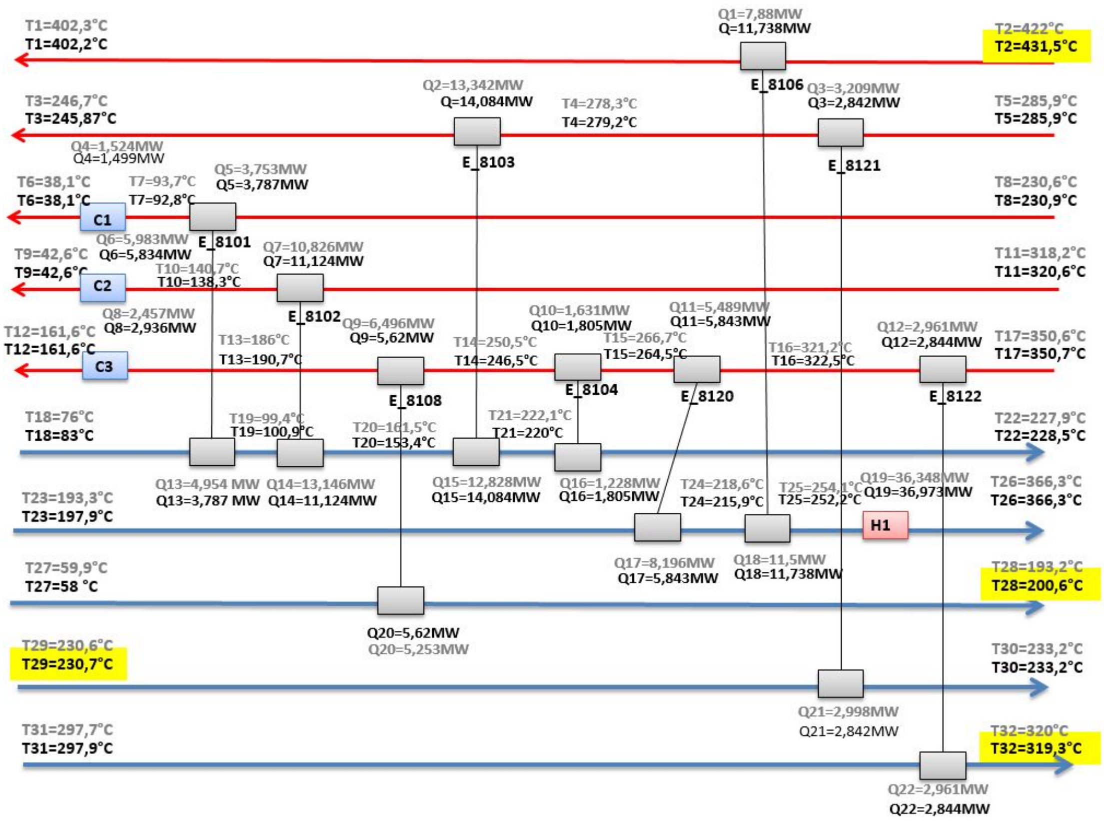

Figure 5.5 Scheme of the reconciled HX network proposed by the Tool 1. Values in grey correspond to the initial imbalanced HX network, and values in bold are the reconciled ones. The four estimated temperatures are the ones highlighted in yellow.

5.3.2 Discussion of the results obtained from Tool 2

When setting wider ranges for the estimates, Tool 2 confirms the location of a gross error in the measurement T20. The solution provided is the balanced network represented in the Figure 5.6, for which the reconciled temperatures are suitable for further analysis of the HCU.

As one can see from Table 5.7, when the first iteration of the simulation is performed, six measurements, T10, T11, T20, T7, T8 and T18, are suspected to contain a gross error. The test statistics of all of them, in the mentioned order, exceed the threshold criterion ($Z_{1-\beta/2}$ = 3.155). Accordingly, the measurements T10 and T11 are the ones with a higher probability of being wrong. However, as before, both measurements cannot be removed from the data set due to *limitation 1*. For this reason, the following measurement in the list T20, which is the first suspected intermediate process variable, is removed. In contrast to what happens in the analysis 1, the reconciled values calculated using the reduced data set are within the new bounds. Therefore, the solution is feasible, and consequently, the measurement is permanently removed from the data set. The test statistics calculated for the next iteration using the new reduced data set, forth and eight columns of Table 5.7, are all below the new test criterion ($Z_{1-\beta/2}$ = 3.043). Thereby, no more measurements are suspect of having a gross error, which confirms that the reduced data set is free of gross errors. At this point, the simulation ends.

Table 5.7 The table shows the values of the Measured Test for all the process variables. Test statistics from the first iteration are listed in the third and seventh column. Final test statistics are listed in the columns forth and eight. The word "removed" means that the corresponding process variable has been removed from the sample.

Reference (ni)	**Measured value (yi) [˚C]**	**Zi (1st iteration)**	**Zi (last iteration)**	**Reference (ni)**	**Measured value (yi) [˚C]**	**Zi (1st iteration)**	**Zi (last iteration)**
T1	402,3	2,145	2,149	T17	350,6	0,174	0,175
T2	422	2,145	2,149	T18	76	3,474	1,840
T3	246,7	0,522	1,254	T19	99,4	0,737	1,257
T4	278,3	0,515	1,225	T20	161,5	4,103	removed
T5	285,9	0,025	0,102	T21	222,1	1,094	1,226
T6	38,1	0,000	0,000	T22	227,9	0,853	0,675
T7	93,7	3,474	1,840	T23	193,3	2,632	2,646
T8	230,6	3,474	1,840	T24	218,6	1,510	1,522

T9	42,6	0,000	0,000	T25	254,1	2,145	2,149
T10	140,7	4,251	1,254	T26	366,3	0,000	0,000
T11	318,2	4,251	1,254	T27	59,9	3,082	3,043
T12	161,6	0,000	0,000	T28	193,2	3,082	3,043
T13	186	3,082	3,043	T29	230,6	0,025	0,102
T14	250,5	2,442	2,508	T30	233,2	0,025	0,102
T15	266,7	2,134	1,983	T31	297,7	0,174	0,175
T16	321,2	1,425	1,433	T32	320	0,174	0,175

The reconciled value for T20 has been calculated using the constraint derived from the energy balance in the HX- E8102, see equation (4.3).

$$10.8625 = 0.21169 * (T_{20} - 97.1266)$$

$$T_{20} = 148.44°C$$

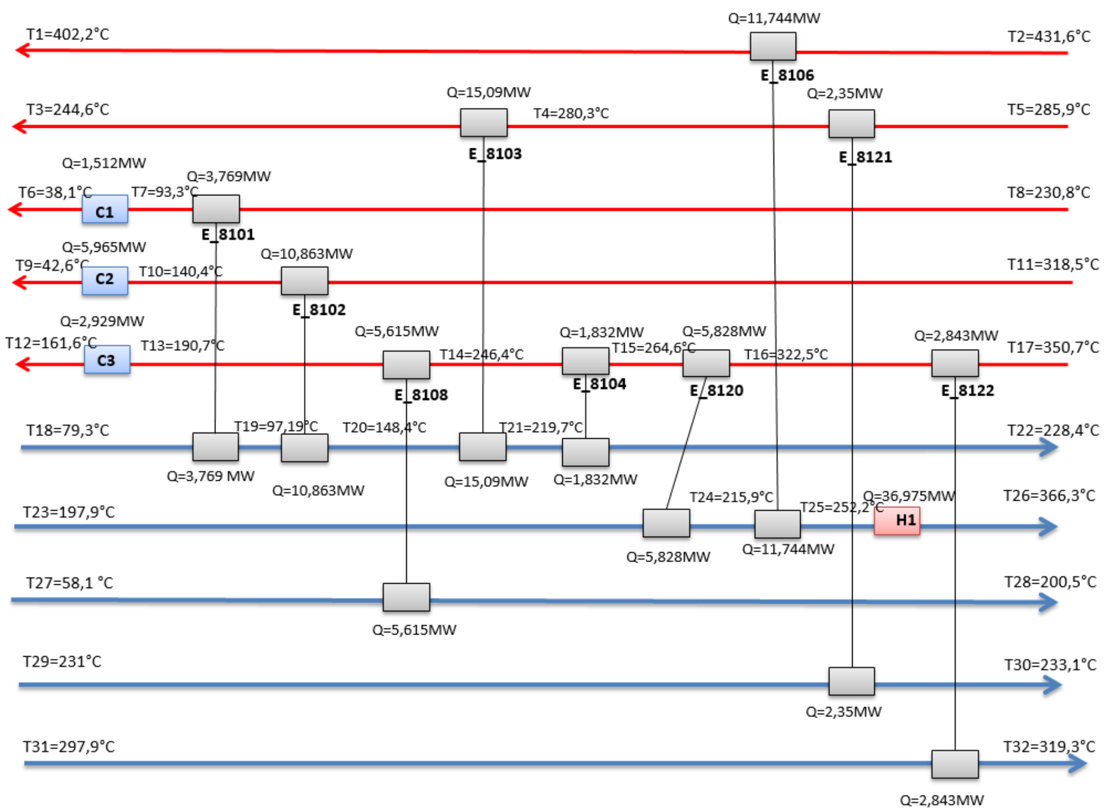

Figure 5.6 Scheme of the reconciled HX network proposed by the Tool 2.

The same procedure can also be done for HX_E8103, using equation (4.4), since T20 is the target temperature for the HX_E8102 but also the start temperature of HX_E8103. The reconciled temperature obtained is obviously the same.

$$15.0902 = 0.21169 * (71.2845 - T_{20})$$

$$T_{20} = 148.44℃$$

The heat loads computed with the new reconciled values are quite similar to the ones discussed in the previous section. The biggest variations are found in the heat exchangers around the measurements T2 and T20. The heat transferred in the heat exchanger E-8106 by the cold stream is according to the reconciled temperatures 11.096MW, 3.215MW more than the transferred in the original network. And the heat absorbed in the HX_E-8102 changes from 12.828 MW to 15.09MW.

Besides this, the resulting balanced heat exchanger network, shown in the Figure 5.6, is perfect for our initial purpose, a process integration study of the hydrocracker. It is not corrupted with gross errors and besides it is feasible, since the reconciled temperatures are within the bounds.

5.3.3 Comparison and conclusions of the second analysis

The reconciled values obtained in both tools are quite similar, with the exception of T20, which has been computed as explained in Section 5.3.2. Only four temperatures have a reconciled value that differs significantly from those obtained using Tool 1. The biggest variation is equal to 5℃ and is obviously reached for the temperature T20. The rest are around 3.7℃ for measurements T19 and T18 and 2.1℃ for T10 and T11. In consequence, the heat loads are practically the same as the ones computed using Tool 1 in the previous section. The largest relative variations are in the heat exchangers HX-E8102 and HX- E8103 next to the measurement T20. For the first, the heat absorbed has varied from 14.0844 MW to 15.0903 MW, and for the second, from 1.805MW to 1.832 MW.

The solution to data reconciliation and gross error detection confirms the hypothesis $H_{1,A}$, and rejects the hypothesis $H_{1,D}$, since the measured test statistics for T10 and T11 are below the threshold once the measurement T20 is removed. This example illustrates perfectly the smearing effect and how an error in one measurement is spread out over the others, altering their statistic tests. In this case, the three temperatures are directly related by the heat exchanger E_8102. According to the values of the test statistics, the Hypothesis $H_{1,B}$ has also been rejected however, it is not ruled out the possibility that the estimation of T2 is incorrect. Within the four estimates; T2, T28, T29 and T32 only the temperature T2 presents a notable increase, almost reaching the extreme value of the range. Whereas for the rest, the reconciled values are practically the same as the ones obtained for the first analysis, even closer in the case of the temperature T29. Nevertheless, the reconciled temperature for T28 has also a significant adjustment, around 6 ℃. The reason why this process variable retains the same value is because the initial bounds where already defined as plus and minus 10 ℃ around the estimate. Therefore, the changes made for the second analysis has no effect in T28. But, in contrast with T2 the reconciled temperature for T28 is accepted since it is within the bounds provided. Thereby, from the four estimations, only the one for the temperature T2 is incorrect.

5.4 Final results

The results of this thesis indicate that the data set from the hydrocracker unit from Preem refinery in Lysekil contains a gross error in the target temperature of the heat exchanger E_8102, T20. Furthermore, the estimation made in the data collection about the value of the measurement T2 is incorrect.

For this resulting network, the largest adjustments are in the intermediate process variables. Start and target temperatures have reconciled values quite close to the measurements, which means that the heat loads for the overall streams are more or less preserved. As one can see from the Table 5.7, besides the first hot stream (st_1) and the second-to-last cold stream, the rest relative variation on the heat loads are equal or below 5%. For both mentioned streams, the high rate can be easily explained. In the former the start temperature is T2, and as has been demonstrated the initial estimate is not correct, neither is the initial heat load. In the latest, the start temperature T29 is also an estimate and since the size of the heat exchanger is quite small, any variation will lead to a high relative difference. Moreover, both streams have only one heat exchanger which means that the variation cannot be compensated.

Table 5.7 The table shows the heat loads from the overall flows. The streams have been numbered from up to down starting in the number 1.

Stream	**T start [˚C]**	**Ttarget [˚C]**	**Qmeas. [MW]**	**Qrec. [MW]**	**AQ [%]**
st_1	431.6	402.2	7.88	11.744	49%
st_2	285.9	244.6	16.551	17.44	5%
st_3	230.8	38.1	5.277	5.282	0%
st_4	318.5	42.6	16.809	16.827	0%
st_5	350.7	161.6	19.034	19.047	0%
st_6	79.3	228.4	32.156	31.554	2%
st_7	197.9	366.3	56.044	54.546	3%
st_8	58.1	200.5	5.253	5.615	7%
st_9	231	233.1	2.998	2.35	22%
st_10	297.9	319.3	2.961	2.843	4%

Therefore, this is the balanced heat exchanger network proposed for studying the process integration opportunities of the hydrocracker hydrogen unit. The values of all the reconciled temperatures and heat loads are shown the Figure 5.7.

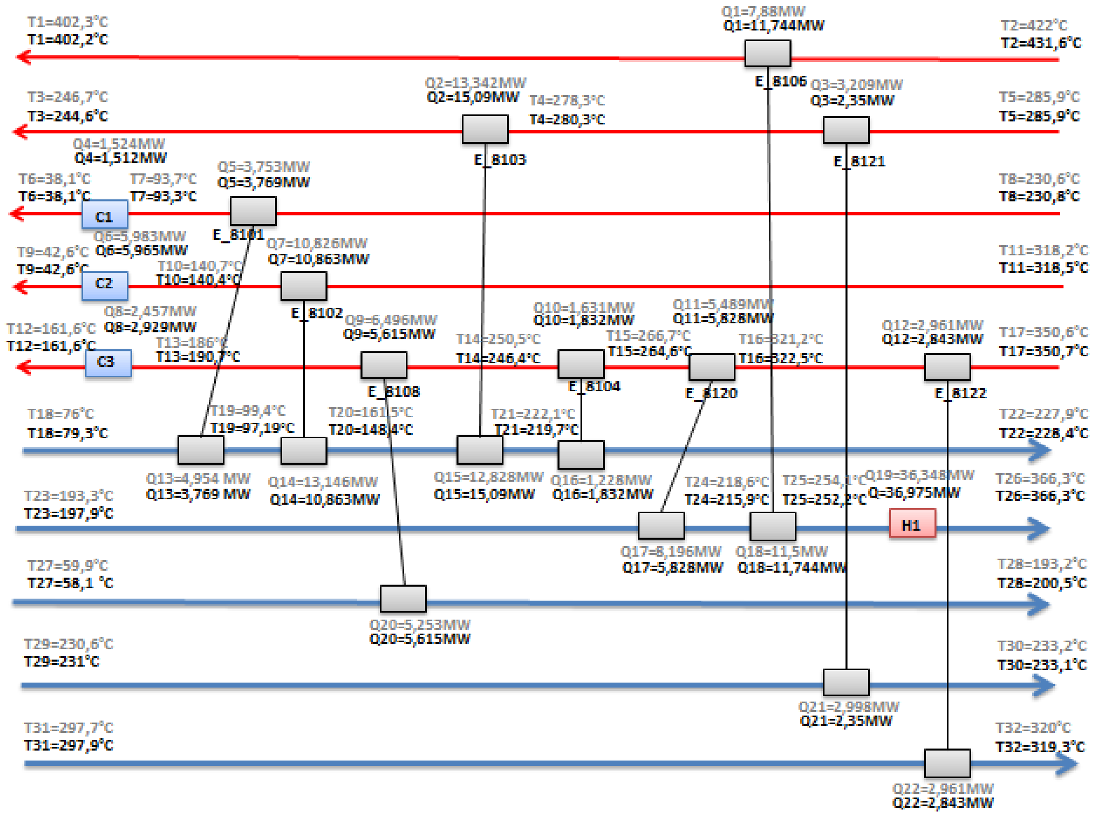

Figure 5.7 Scheme of the final reconciled network. The values in bold refers to the new reconciled values and the values in grey are the ones from the original data set (measurements).

6 Conclusions

6.1 Tool 1

The aim of Tool 1 is to provide a solution to the problem of data reconciliation, but taking into account the constraints and the boundaries of the system. If the initial condition is met, that is, none of the initial measurements contain a gross error, the reconciled values computed are correct and can be used in further studies without need of any other verification. The developed computer-based solution even allows the use of non-linear constraints, or inequalities. However, sometimes, when the quality of data from measurements is low, as is the case of this thesis work, this assumption could not be made. For this case, the reconciled values must be taken only as a guideline.

In this study, based on the trend of the results and the significant differences between the original network and the reconciled one, three hypotheses have been drawn. The aim was to find out if the reconciled values obtained for the cases where there is an initial error, can serve as indicators of the origin of the error. For this particular case, the conclusions drawn are similar to the final conclusions, two of three hypothesis are afterwards confirmed. However, without the help of the Tool 2, it would have been impossible to reject or confirm any hypothesis. Furthermore, the hypotheses are worded in a very general manner, with a total of six measurements under suspicion, which hampers from taking any decision and reaching a final conclusion.

6.2 Tool 2

The purpose of Tool 2 is to find a solution to the problem of data reconciliation, but unlike the previous one, in this case, the tool must be able to provide with a valid solution even if the data collected contains gross errors.

Before its use with the real data collected from the Preem refinery, the tool was tested using a simply mass flow network. The results obtained for this test case are excellent. The error introduced is detected and the reconciled flowrates are calculated with a reduced data set where the former has been removed. The simulation is fast and the final results are accurate, easy to interpret and can be directly used for process integration studies. But the same does not happen for the studied network. In the study case, error detection becomes more cumbersome. The final conclusion, is the resulting judgement of an iterative analysis, which includes two analyses and multiple intermediate hypotheses before arriving to the final result. The fact that the quality of the data collected is quite low, i.e. 4 of the measurements are estimates since the real value was not available, have further complicated the analysis.

The performance of the MIMT for the study case is extremely limited. From the 32 measured process variables, the existence of a gross error can merely be successfully confirmed in 8 intermediate measurements, which is only the 25% of the cases. The usefulness of the Tool is improved when including the values of the statistics test in the analysis, and when more than one simulation is performed. For that reason, it is crucial that the user has a good understanding of the tool and also of the network that is going to be studied in order to get a trustworthy outcome.

The conclusions set out from the resulting network of HCU would probably had been a bit different if another error detection methodology had been used. However, some of the final conclusions seems to be supported. It has been confirmed that there is a high probability that the estimate value for T2, which corresponds to the outflow temperature from the first reactor, is not correct.

In conclusion, if the data set is free of gross errors, the solution is accurate. Otherwise, the results must to be analysed in line with the test statistics of the last iteration. If all of them are below the threshold, the resulting reconciled network is free of gross errors and the measurements with gross errors are the ones deleted from the data set. However, if one or more test statistics from the last iteration exceed the threshold, this means that the reconciled network has been calculated using a data set with gross errors, thus, the reconciled network is invalid. In this case, test statistics can be used to find possible measurements with gross error, but they never can be used to confirm the exact location of the gross error.

In contrast with the Tool 1, the potential advantage is that the user is aware of the existence or not of gross errors in the collected data. And therefore, whether or not the reconciled process variables are valid or corrupted because they have been computed using a data set with gross errors.

7 Future studies

To fully understand the potential of both computer-based solutions developed in this thesis it is suggested to make further performance tests.

In this case it has been demonstrated that Tool 1 performs excellently for reconciling networks where no gross errors are present and where all the constraints are linear. But, the main advantage of working with Excel Solver is that non-linear constraints, inequalities and even equations that only involve integers can be defined as constraints. It is highly recommended to further work into this area especially for reconciling networks where the composition of the streams varies and so the heat transfer coefficient is not constant. It could also be interesting to restrict the degrees of freedom by including the operating conditions as constraints of the reconciled temperatures. So if for instance, the feed temperature for the reactor has to be higher than 350 ºC, the corresponding reconciled value should also be higher than 350 ºC.

This thesis work has shown that there is a relation between the likelihood that a variable contains a gross error and the fact that the corresponding reconciled value falls outside the range defined by the bounds. The same reasoning can be applied for the process variables with large adjustments. It is suggested further investigation along this lines, in order to known how much related this ideas are. It could be helpful to define standards or a guideline for identifying measurements suspected of having a gross error by comparing them with their corresponding reconciled values.

Tool 2 has even more potential of improvement. It has been shown that nodal aggregation is difficult to perform especially where there are so many streams and just a few constraints. In order to solve this problem, the analyses conducted was performed using the test statistics and engineering judgement. In order to reduce resorting to engineering judgement as much as possible and drawing erroneous conclusions it is essential to have a deeper understanding of how the tool performs. Thus, it is recommended to test the tool with multiple data sets for which the user knows *a priori* where the gross errors are located. In that manner, it is easier to get familiar with the tool and gain a better understanding of how it works before using it with the collected data. Improving the actual tool, the efforts could also focus on developing a linearization methodology so that the tool can also be used for systems with non-linear constraints.

For further investigations it is suggested to try to implement other algorithms such as the Modified Serial Compensation (MSC) or the Generalized Likelihood Ratio (GLR) as according to the literature review they are also good options. Both methodologies use a compensation strategy to neutralize the effect of the gross error instead of eliminating the corresponding measurement. Therefore, since almost all the limitations found were because nodal aggregation could not be performed, they are a potential alternative.

Another research could go through integrating both programs, so data reconciliation is solved using the Solver functionality whereas for the detection of the errors it is used the Modified Iterative Measurement Test. Indeed, this could be the easier manner for developing a data reconciliation and gross error detection tool for networks with non-linear constraints.

8 References

British petroleum (BP). Statistical Review of World Energy. Available at: http://www.bp.com/en/global/corporate/about-bp/energy-economics/statistical-review-of-world-energy/review-by-energy-type/oil/oil-reserves.html

[Accessed June 7, 2015]

Carl Knopf F. (2012): *Modeling, Analysis and Optimization of Process and Energy Systems.* John Wiley & Sons, 2012, Inc.428pp

Svensson E. (2014): *Course compendium for Industrial Energy Systems*. Department of Heat and Power Technology , Chalmers University of Technology, Göteborg, Sweden, 2014, 152 pp.

International Energy Agency (IEA), 2015. Oil Market Report. Available at: https://www.iea.org/oilmarketreport/omrpublic/

[Accessed May 25, 2015]

Joint Research Centre. (2013): *BAT Reference Document for the Refining of Mineral oil and Gas.* Institute for Prospective Technological Studies (IPTS), pp 35-113.

Kemp I. (2006): *Pinch Analysis and Process Integration.* Butterworth-Heinemann, Great Britain, 2006, 390pp.

Kim, I., Kang, M. S., Park, S., Edgar, T. F. (1996): *Robust Data Reconciliation and Gros Error Detection: The Modified MIMT using NLP.* Pergamom,1997.

Mah R., Stanley G., Downing D.: *Reconciliation and Rectification of Process Flow and Inventory Data.* Department of Chemical Industrial Engineering, Northwestern University, Illinois. Vol. 15, No. 1, May 1976

Natasimhan S.,Jordache C. (1999): *Data Reconciliation and Gross Error Detection: An Intelligent use of Process Data.* Gush Publishing,1999, 372 pp.

Preem website, 2015. Available at: https://www.preem.se/en/in-english/

[Accessed February 27, 2015]

Shert R.W., Heenan A. W.: *Gross error Detection and Data Reconciliation in Steam-Metering Systems.* Vol. 32, No. 5, May 1986

Walkenbach J.(2010): *Excel VBA Programming for DUMMIES.* Wiley Publishing, 2010, 371 pp.

World Fuel Limited, Available at: http://world-fuel.me/HTML/button_09_selected_html/button_09_selected.htm

[Accessed March 10, 2015]

9 Appendix

Appendix 1: Raw data set from the HCU provided by Preem

Appendix 2: Lagrange Multipliers, Carl Knopf F (2012)

Appendix 3: Test cases

9.1 Raw data set from the HCU provided by Preem

This appendix contains the data set from the Hydrocracker Unit provided by Lysekil Preem´s Refinery. Since there are a lot of streams and units involved on the system, for the purpose of this thesis not all of them are included in the analysis. A reduced network has been defined instead.

Table 9.1 *The table shows the heat duties for each stream involved in the HCU process. The streams highlighted are the ones included within the thesis work.*

Number of stream	**Name of stream**	**Tstart** [˚C]	**Ttarget** [˚C]	**Q** [MW]	**ΔT** [˚C]	**FCP** [MJ/Kg]
810-1	Feed to V-8101	76	227.9	32.15567	15	0.21169
810-2	R-8101 Feed	226.5	388.3	64.57364	10	0.399095
810-3	R-8101 to R-8102	422	402.3	7.880214	15	0.400011
810-4	R-1802 to V-8102	422.7	239.9	92.38248	10	0.505375
810-5A	V-8102 OH	224.4	163.9	14.93252	10	0.246819
810-5B	V-8102 OH. air he	130	56.8	23.36435	10	0.319185
810-6	V-8103 OH	233.7	74.6	2.064449	10	0.012976
810-7	T-8120 OH	82.3	38.2	11.18599	10	0.253651
810-8	T-8120 to T-8121	193.3	367.8	56.5299	10	0.323954
810-9	T-8121 OH	130.5	80.1	27.76428	10	0.550879
810-10	Gasoil pump around	285.4	246.7	16.33986	15	0.422219
810-11	Kerosene to tank	230.6	38.1	5.277317	15	0.027415
810-12	Diesel to tank	318.2	42.6	16.80884	15	0.06099
810-13	UCO to tank and to FCC	350.6	161.6	19.03361	15	0.100707
810-14	UCO to tank	161.6	83.4	2.50653	15	0.032053
810-15	V-8105 to T-8120	59.9	193.2	5.253278	15	0.039409

810-16	Hydrogen from compressors to R8101	92.5	201.3	13.98158	15	0.128507
810-17	T-8122 heater	230.6	233.2	2.997753	15	1.152982
810-18	T-8123 heater	297.7	320	2.960784	15	0.132771

Table 9.2 *The table shows all the measured process variables involved in the study case. There stream highlighted are those whose measurements were unavailable and so the value used is an estimate.*

Name of the flow	**Name of the variable**	**Measured temperatures from Preem (yi)**	**Lower bound for yi** [˚C]	**Upper bound for yi** [˚C]	**FCP** [MJ/Kg]
810-3	T1	402.3	401.3	403.3	0.400011
810-3	T2	422	418	426	0.400011
810-10	T3	246.7	241.7	251.7	0.422219
810-10	T4	278.3	273.3	283.3	0.422219
810-10	T5	286.3 /285.4	282.8	288.8	0.422219
810-11	T6	38.1	33.1	43.1	0.027415
810-11	T7	93.8 /93.6	88.7	98.7	0.027415
810-11	T8	230.6	227.6	233.6	0.027415
810-12	T9	42.6	37.6	47.6	0.06099
810-12	T10	140.7	135.7	145.7	0.06099
810-12	T11	318.2	313.2	323.2	0.06099
810-13	T12	161.6	156.6	166.6	0.100707
810-13	T13	186	181	191	0.100707
810-13	T14	248.9 /252.4	245.6	255.6	0.100707
810-13	T15	266.7	261.7	271.7	0.100707
810-13	T16	321.2	316.2	326.2	0.100707

810-13	T17	350.6	345.6	355.6	0.100707
810-1	T18	76	71	81	0.21169
810-1	T19	99.4	94.4	104.4	0.21169
810-1	T20	161.5	156.5	166.5	0.21169
810-1	T21	222.1	217.1	227.1	0.21169
810-1	T22	227.9	224.9	230.9	0.21169
810-8	T23	193.3	188.3	198.3	0.323954
810-8	T24	218.6	213.6	223.6	0.323954
810-8	T25	254.1	249.1	259.1	0.323954
810-8	T26	364.5 /364.1 /368.9 /367.8	363.3	369.3	0.323954
810-15	T27	59.9	54.9	64.9	0.039409
810-15	T28	193.2	183.2	203.2	0.039409
810-17	T29	230.6	225.6	235.6	1.152982
810-17	T30	233.2	228.2	238.2	1.152982
810-18	T31	297.7	292.7	302.7	0.132771
810-18	T32	320	315	325	0.132771

9.2 Lagrange Multipliers.

Lagrange multiplier is a strategy for converting equality constrained optimization problem into an equivalent unconstrained problem. For the purpose of this thesis, the Lagrangian function can be expressed as:

$$Min\ L(x,\lambda) = f(x_i) - \sum_k \lambda_k h_k(x_i), \quad (9.1)$$

where $f(x_i)$ is the objective function, λ_k are the Lagrange multipliers, one for each equality constraint $h_k(x_i)$.

The aim of this appendix is to briefly explain how are computed the solution to the data reconciliation problem, i.e. the reconciled values. Therefore the starting point is the unconstrained data reconciliation equations presented in Section 3.5, and written down:

$$Min\ (y-x)^T \Sigma^{-1} (y-x)$$

$$Subject\ to \quad Ay - b = 0$$

Where y refers to the measured values, x to the reconciled values, Σ to the variance-covariance matrix, A is the incidence matrix for the linear constraints and b is the independent terms of the constraints. For the explanation, it is going to be assume that the vector of independent terms is zero (b=0), which is the general case. It is also convenient to define an adjustment vector a, which is the difference between the reconciled and measured value (a = y-x). Accordingly the data reconciliation problem can be written as:

$$Min\ (a)^T \Sigma^{-1} (a) \quad (9.2)$$

$$Subject\ to \quad Ay + Aa = 0 \quad (9.3)$$

Based on the last, and using the general Lagrange formulation equation (9.1), the data reconciliation problem can be expressed as:

$$Min_{a,\lambda}\ L = (a)^T \Sigma^{-1} (a) - 2\lambda^T \ (Ay + Aa) \quad (8.4)$$

Or:

$$Min_{a,\lambda}\ L = (a)^T \Sigma^{-1} (a) - 2\lambda^T Ay \ - 2\lambda^T Aa$$

Using the transpose property $\lambda^T Aa = (Aa)^T \lambda = A^T a^T \lambda$, the formula before can be written as:

$$Min_{a,\lambda}\ L = (a)^T \Sigma^{-1} (a) - 2\lambda^T Ay \ - 2A^T a^T \lambda \quad (9.5)$$

In order to solve the equation (9.5) the conditions are:

$$\frac{\partial L}{\partial \lambda} = 0\ and\ \frac{\partial L}{\partial a} = 0 \quad (9.6)$$

So for the first,

$$\frac{\partial L}{\partial \lambda} = Ay + Aa = 0 \quad (9.7)$$

And for the second one,

$$\frac{\partial L}{\partial a} = 2\Sigma^{-1} a - 2A^T \lambda = 0 \quad (9.8)$$

Which results in,

$$a = \Sigma A^T \lambda \tag{9.9}$$

Combining both equations, and substituting equation (9.9) into equation (9.7),

$$Ay = -A\Sigma A^T \lambda \tag{9.10}$$

Solving equation (9.10) for λ,

$$\lambda = -(A\Sigma A^T \lambda)^{-1} Ay \tag{9.11}$$

And finally by substituting equation (9.11) into equation (9.9),

$$a = -\Sigma A^T (A\Sigma A^T)^{-1} Ay$$

And using the definition of the adjustment vector a,

$$x = y - \Sigma A^T (A\Sigma A^T)^{-1} Ay \tag{9.12}$$

If the assumption made at the beginning is not true, and the vector of the independent terms of the constraint matrix is different to zero, the equation (9.12) can be expressed as:

$$x = y - \Sigma A^T (A\Sigma A^T)^{-1} (Ay\text{-b}) \tag{3.12}$$

9.3 Test cases

With the purpose of testing the performance of the two developed computed-based tools, two test cases have been conducted. For the first, the measurements of the process variables are free of gross errors, whereas from the second one, a gross error on the process stream 2 has been introduced.

These test cases were based on set homework problems for the Autumn 2012 offering of the elective course CN4205R "Process Systems Engineering" for 4th year Chemical Engineering Students at the National University of Singapore (NUS).

The process network, which is the same for both test cases, includes 5 units and 10 process variables. Since all the measurements are flowrates, data reconciliation is only constrained by mass balances. The scheme of the network is represented in Figure 9.1.

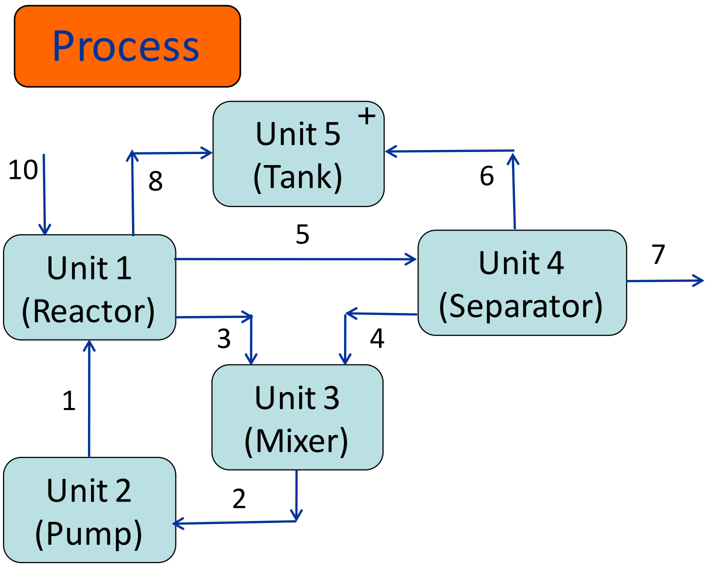

Figure 9.1 *Scheme of the process network studied.*

Accordingly, the mass balances constraints are:

Unit 1: $+x_1 - x_3 - x_5 - x_8 + x_{10} = 0$ (9.13)

Unit 2: $-x_1 + x_2 = 0$ (9.14)

Unit 3: $-x_2 + x_3 + x_4 = 0$ (9.15)

Unit 4: $-x_4 + x_5 - x_6 - x_7 = 0$ (9.16)

Unit 5: $+x_6 + x_8 - x_9 = 0$ (9.17)

Using equations (9.13)-(9.17) and following the same procedure that is defined in the Section 4.2.2 the incidence matrix *A* is:

1	0	-1	0	-1	0	0	-1	0	1
-1	1	0	0	0	0	0	0	0	0
0	-1	1	1	0	0	0	0	0	0
0	0	0	-1	1	-1	-1	0	0	0
0	0	0	0	0	1	0	1	-1	0

And the vector of independent terms is identically zero (b=0).

9.3.1 TEST CASE 1: Data Reconciliation

Table 9.3 *Measurements of the process variables*

Stream	yi	σii
1	100	5
2	90	2
3	45	2
4	50	2
5	120	10
6	40	5
7	38	5
8	10	5
9	50	5
10	100	10

From the data above, matrixes y and Σ and can be defined:

$$y = \begin{bmatrix} 100 \\ 90 \\ 45 \\ 50 \\ 120 \\ 40 \\ 38 \\ 10 \\ 50 \\ 100 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 25 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 100 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 25 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 25 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 25 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 100 \end{bmatrix}$$

And using equation (3.12), the reconciled values are calculated:

$$x = \begin{bmatrix} 92{,}38547 \\ 92{,}38547 \\ 43{,}83286 \\ 48{,}55261 \\ 127{,}0063 \\ 39{,}67554 \\ 38{,}7782 \\ 11{,}42712 \\ 51{,}10266 \\ 89{,}88086 \end{bmatrix}$$

9.3.2 TEST CASE 2: Data Reconciliation and Gross Error Detection

Same data set is used for the test case 2 with the only difference that a gross error has been introduced in the stream number 2.

Table 8.4 *Measurements of the process variables*

Stream	yi	σii
1	100	5
2	110	2
3	45	2
4	50	2
5	120	10
6	40	5
7	38	5
8	10	5
9	50	5
10	100	10

The constraint matrix A and the covariance matrix Σ are the same as before. The new measured vector y is:

y =

100
110
45
50
120
40
38
10
50
100

The first step is to solve the reconciled values with the new data set and compute the adjustment vector a, and the modified vector d

a =

-4,38046
5,619542
-4,91867
-4,46179
-11,422
0,973823
0,06597
-1,88168
-0,90785
11,15812

d=

1,226517
-1,27348
0,046966
-0,07045
0,117414
0,953366
-0,15244
-1,14083
1,105805
0,035025

and then the covariance wii

0,036162	-0,02399	-0,01217	-0,01182	-0,00035	0,000208	0,00027	-0,00015	6,23E-05	8,31E-05
-0,02399	0,100063	-0,07607	-0,07386	-0,00221	0,001299	0,001688	-0,00091	0,00039	0,000519
-0,01217	-0,07607	0,088244	0,085683	0,002561	-0,00151	-0,00196	0,001055	-0,00045	-0,0006
-0,01182	-0,07386	0,085683	0,090452	-0,00477	0,002805	0,003647	-0,00196	0,000842	0,001122
-0,00035	-0,00221	0,002561	-0,00477	0,00733	-0,00431	-0,00561	0,003018	-0,00129	-0,00172
0,000208	0,001299	-0,00151	0,002805	-0,00431	0,016654	0,00565	0,012342	-0,011	-0,00134
0,00027	0,001688	-0,00196	0,003647	-0,00561	0,00565	0,011345	4,49E-05	0,005695	-0,00574
-0,00015	-0,00091	0,001055	-0,00196	0,003018	0,012342	4,49E-05	0,01536	-0,0123	-0,00306
6,23E-05	0,00039	-0,00045	0,000842	-0,00129	-0,011	0,005695	-0,0123	0,016699	-0,0044
8,31E-05	0,000519	-0,0006	0,001122	-0,00172	-0,00134	-0,00574	-0,00306	-0,0044	0,007465

Then, the next step is to calculate the test statistics:

Zstats =

0,921416
4,441248
4,139467
3,708855
1,334106
0,301843
0,024774
0,6073
0,281017
1,291477

And the threshold value $Z_{1-\beta/2} = 2.8$ for $\alpha = 0.05$. $Z_{a,2}$ is the largest test statistic that exceeds the threshold value.

So the measurement 2 is temporarily removed from the data set, by aggregating node 2 and node 3. The constraint matrix for the reduced system is:

A =

1	-1	0	-1	0	0	-1	0	1
-1	1	1	0	0	0	0	0	0
0	0	-1	1	-1	-1	0	0	0
0	0	0	0	1	0	1	-1	0

Using the same procedure before, the reconciled values are calculated for the reduced reconciliation problem, leading to:

a =

4,040066
-0,64641
-0,31352
-8,32219
0,51797
-0,52664
-1,56258
-1,04461
10,42875

x =

95,95993
45,64641
50,31352
128,3222
39,48203
38,52664
11,56258
51,04461
89,57125

Since no information about bounds is provided, it is assumed that these values are realistic (there are no negative flows which is the only possible indication of an unrealistic reconciled value). In order to be sure that no more errors are present, the new test statistics are computed. It is also computed again the threshold value (see below) since the number of measurements in the data set is now equal to 9. For $\alpha = 0.05$, $\beta = 0.005683$ and so $Z_{1-\beta/2} = 2.76$. None of the test statistics exceeds the value so, the procedure is terminated.

$z_{a,j} =$

0,926702
0,926702
0,413527
0,975291
0,160629
0,198022
0,504448
0,323363
1,207276

In conclusion, serial elimination strategy detect correctly the error location introduced in the process stream number 2.