KNOWLEDGE EXTRACTION FROM PUBLISHED PAPERS IN LITERATURE FOR THE CATALYTIC METHANOL PRODUCTION FROM SYNTHESIS GAS USING DATA MINING TOOLS

by

Dennis Moskov

B.S., Process Engineering, Hamburg University of Applied Sciences, 2014

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Chemical Engineering
Boğaziçi University

2016

ACKNOWLEDGEMENTS

First of all I want to express my gratitude towards everyone who made my graduation possible, professors who gave lectures, that I took, personnel who helped me with technical issues, friends and people I met, which brought some color into the daily academic life. I could not possibly thank everybody individually, but want to highlight some persons who deserves special gratitude.

My thesis supervisor, Prof. Ramazan Yıldırım deserves high appreciation for his guidance, perpetual encouragement and support. He made me see achievements in points I considered unimportant and contributed highly to my understanding of catalyst related matters.

I want to thank my thesis co-supervisor, Assist. Prof. Mustafa Gökçe Baydoğan for his technical support, idea proposals and knowhow contributions. He offered a splendid range of possible solutions to several problems and was always eager to discuss those.

I also want to thank Çağla Odabaşı, Elif Can and Dilara Saadetnejad for sharing their experience, helping with organizational matters and lightening up the days at the lab.

A special thanks goes to my wife Gözde Kavak Moskov for her endless support and encouragement to work steadily on this thesis. She stood by my side during all phases of the work and was of great help mentally and intellectually.

Last but not least, I want to thank my parents Lidia Lams-Moskov and Karl Hermann Lams for their understanding and never ending support during all my time in Turkey.

ABSTRACT

KNOWLEDGE EXTRACTION FROM PUBLISHED PAPERS IN LITERATURE FOR THE CATALYTIC METHANOL PRODUCTION FROM SYNTHESIS GAS USING DATA MINING TOOLS

In this work, a database for methanol synthesis was constructed from published literature to extract knowledge and to build models to help the future studies. The database was built with 357 data points showing the effects of 28 input variables such as catalyst preparation and reaction conditions on CO_X conversion, methanol selectivity and methanol yield as response variables. Multiple linear regression (MLR), decision trees and random forest (RF) were independently applied to model CO_X conversion, methanol selectivity and methanol yield, in the R 3.2.3 environment. MLR and regression trees were not successful in interpretable results. Classification trees were applied using discretized response variables; the misclassification errors for training responses were 22 %, 32 % and 25 % for CO_X conversion, methanol selectivity and methanol yield, respectively. Considering that those ratios were much higher for testing (73 %, 83 % and 77 % respectively), it was decided that this method may be used to deduce empirical observations but it is not capable to predict unseen experiments. Random forest yielded the best results in terms of goodness of fit with the $R^2_{adj.}$ as high as 0.95; hence, it was used to extract information for variable importances. Reduction time and catalyst preparation method were found to be most important variables for CO_X conversion while reaction temperature and CO/CO_2 ratio were found to be key variables for methanol selectivity; reduction temperature and CO amount in the feed composition were found to be most significant variables for the methanol yield. However, despite the low prediction RMSE values between 0.12 and 0.34, RF was also not able to predict unseen experiments successfully and generated nearly random results. As a result, it was concluded that the data mining tools have been successful for descriptive tasks like significance analysis and rule deduction, but failed in predicting the conversion, selectivity and yield.

ÖZET

VERİ MADENCİLİĞİ ARAÇLARI KULLANILARAK LİTERATÜRDE YAYINLANAN MAKALELERDEN SENTEZ GAZINDAN KATALİTİK METANOL ÜRETİMİ KONUSUNDA BİLGİ ÇIKARIMI

Bu çalışmada, literatürde yayınlanmış makalelerden metanol sentezi konusunda bilgi çıkarımı yapmak ve gelecek çalışmalarda kullanılabilecek modeller geliştirmek için bir veri tabanı oluşturulmuştur. Bu veri tabanı, katalizör bileşimi, katalizör hazırlanması ve reaksiyon koşulları gibi 28 girdi değişkeninin, çıktı değişkenleri olarak CO_X dönüşümü, metanol seçimliliği ve metanol verimine etkisini gösteren 337 veri noktası içermektedir. CO_X dönüşümü, metanol seçimliliği ve metanol verimi için birbirinden bağımsız olarak R 3.2.3 ortamında çoklu doğrusal regresyon (ÇDR), karar ağacı ve rastgele orman (RO) yöntemleri uygulanmıştır. ÇDR ve regresyon ağaçları yorumlanabilir sonuçlar üretmede başarısız olmuştur. Sınıflandırma ağaçları ise çıktı değişkenleri kategorik hale getirilerek uygulanmış, öğrenmede yanlış sınıflandırma hataları dönüşüm, seçimlilik ve verim için sırasıyla %22, %32 ve %25 olmuştur. Test aşamasında bu oranlar çok yüksek olduğundan (sırasıyla %73, %83 ve %77) sınıflandırma ağaçlarının bir takım ampirik gözlemler elde etmek için kullanılabileceğine, ancak bunların daha önce görülmemiş deneyleri tahmin etme yeteneğine sahip olmadığına karar verilmiştir. RO tekniği ile üretilen model ise 0,95 gibi yüksek $R^2_{adj.}$ değerleri ile istatistiksel açıdan başarılı bulunmuş ve değişkenlerin göreceli önemleri ile ilgili bilgiler elde etmek üzere kullanılmıştır. CO_X dönüşümü için indirgeme süresi ve katalizör hazırlama yönteminin en önemli değişkenler olduğu görülürken, metanol seçimliliği için temel değişkenlerin reaksiyon sıcaklığı ve CO/CO_2 oranı olduğu gözlenmiş, verim için ise, indirgeme sıcaklığı ve besleme bileşimindeki CO miktarının etkili oldukları saptanmıştır. Öte yandan, tahmin için ortalama karekök hata değerleri 0,12 ve 0,34 gibi düşük bir aralıkta olmasına rağmen; RO, daha önce görülmemiş deneyleri başarılı bir şekilde tahmin edememekte ve neredeyse rastgele sonuçlar vermektedir. Sonuç olarak veri madenciliği araçlarının, analiz ve kural çıkarımı gibi açıklayıcı görevlerde başarıyla kullanılmasına rağmen; dönüşümü, seçimliliği ve veriminin tahmininde başarısız olduğuna karar verilmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

B	Number of random variables
c_i	Class i
C	Number of predefined classes
$d_2(x_i, x_j)$	Dissimilarity between data point i and j
d_{ij}	Gower distance between data point i and j
$d_{ij}^{(k)}$	Distance between data point i and j in variable k
f_i	Relative frequency of class i
$F(x)$	Average sum of dissimilarities
$Gini(S)$	Gini index of data set S
$Gini(S_i)$	Gini index of subset S_i
$Gini_{split}$	Quality of a split into k subsets S_i
k	Variable k
k	Number of subsets S_i
m	Number of random variables for random forest
n	Total number of data points (size of database)
n	Number of unique values for a variable
n_i	Size of subset S_i
p	Total number of variables (dimensionality of database)
P	Pressure
R^2	Coefficient of determination
$R^2_{adj.}$	Adjusted coefficient of determination
R^2_{CV}	Cross validation coefficient of determination
s_i	Number of data points belonging to class i
S	Selectivity
S	Dataset
S_i	Subset i
x	Molar fraction
x	Value of a variable
x_i, x_j	Input variables i and j in multiple linear regression

x_{ik}, x_{jk}	Values of data points i and j for variable k
X	Matrix of inputs
X	Conversion
X^t	Transpose of matrix of inputs
y_i	Observed response
$\bar{y}$	Average of observed responses
$\hat{y}_i$	Estimated (fitted) response
$\hat{y}_{CV,i}$	Response predicted by cross validation
Y	Vector of outputs
Y	Yield
$\hat{Y}$	Vector of estimated (fitted) responses
z	Standardized value of a variable x, (z-score)
β	Vector of partial regression coefficients
β_j	Partial regression coefficient j
$\hat{\beta}$	Estimated (fitted) regression coefficient vector
$\delta_{ij}^{(k)}$	Comparability of data points i and j in variable k
$\Delta H_{r,298}$	Standard enthalpy change of reaction at 298 K
ε	Error term in multiple linear regression
ε_i	Residual error for data point i
$\bar{\varepsilon}$	Mean of residual errors
ϵ	Vector of residual errors
μ	Mean value of variable x
σ	Standard deviation
σ^2	Variance
ω_k	Weight of variable k

LIST OF ACRONYMS / ABBREVIATIONS

ADP	Ammonia-driving deposition-precipitation method
ANN	Artificial neural network
Approx.	Approximately
BASF	A chemical company
C5.0	A decision tree algorithm based on the early ID.3 algorithm
Cal.	Calcination
CART	Classification and regression trees
Comp.	Composition
CT	Classification tree
DME	Dimethyl ether
Etc.	Et cetera
GA	Genetic algorithm
GHSV	Gas hourly space velocity
HAl	Hierarchical meso–macroporous alumina
HM	Honeycomb monolith
MeOH	Methanol
MLR	Multiple linear regression
MSS	Model sum of squares
OF	Open-cell foam
PAM	Partitioning around medoids
PB	Packed bed
Prep.meth.	Preparation method
PRMSE	Prediction root mean squared error
PSS	Prediction sum of squares
RBFN	Radial basis function network
Red.	Reduction
RF	Random forest
RMSE	Root mean squared error
RSS	Residual sum of squares
RT	Regression tree

Rxn.	Reaction
SBA-15	A mesoporous silica
SD	Standard deviation
SPM	Split plate micro mixer
STY	Space time yield
SVM	Support vector machines
TSS	Total sum of squares
UAl	Unimodal mesoporous alumina
VAM	Valve assisted micro mixer
WGS	Water gas shift reaction

1. INTRODUCTION

Methanol is a light, colorless, flammable liquid at room temperature and is the base material for a lot of goods and materials we deal with on an everyday basis. It is a key material in the production of chemicals like formaldehyde, acetic acid and olefins, which in turn are used for the production of plastics, synthetic fibers, paints, resins, solvents, polyester and a various other products. Beside the use as a base chemical, pure or blended methanol is also used as transportation fuel due to a number of different reasons like availability, superior vehicle performance compared with gasoline, less toxic emissions and a lower price than gasoline or ethanol. This is in fact the fastest growing segment of the methanol marked. Another use of methanol is in the transesterification of triglyceride to biodiesel. Due to environmental regulation and a rising awareness of the need to counter pollution, the biodiesel production is growing steadily, so a more environmentally friendly fuel can be used in diesel engines. Furthermore methanol is a very good hydrogen carrier with more hydrogen atoms than any other stable liquid at normal conditions, which makes it an ideal candidate for the use in fuel cells, which are used to power vehicles, to provide backup power or as portable energy source for personal use. Taking into account of all these possible uses of methanol and several others, it is an easy forecast that methanol will stay to be a major chemical in everyday life and will play an even larger role in the energy sector in the future (Methanol Institute, 2011).

Industrial methanol is mainly produced from synthesis gas containing CO, CO_2 and H_2 which can be produced from various feedstocks like natural gas, biomass, coal and other fossil fuels. The first industrial plant for the catalytic methanol synthesis from synthesis gas was built in the early 1930's by BASF and operated at 300°C and 200 bar, using a zinc/chromium-oxide catalyst. Since then the process has undergone several changes and modifications regarding the technologies, operation conditions and catalysts. Today methanol is produced mainly by a heterogeneous catalytic process, operating at 250 – 300°C and 50 to 100 bar, using $Cu/ZnO/Al_2O_3$ catalyst. Besides the wide implementation of the existing process, significant amount of research regarding catalyst optimization by using different structured support or addition of promoter metals, catalyst preparation, ideal process condition, feed

composition and alternative methods including different kinds of reactors or process structures to achieve higher CO_X conversion, selectivity towards methanol and longevity of the catalyst is still going on. (Lee, 1990).

Academic and industrial research in methanol production from synthesis gas is started before the first industrial plant was build and is still continuing. Therefore a vast amount of research results have been accumulated in literature over the years. Using this accumulation in literature, information based models can be constructed to analyze past studies and determine the relations and significances of chemical and physical attributes of the reaction. Considering the huge amount of data that can be derived from the literature, an artificial agent is needed to process it thoroughly. Data mining tools and methods can be used to analyze the data, construct a model and extract previously unknown information and relationships. In addition, predictions of outcomes for future experiments can be done. On the other hand, although the data mining can be a mighty tool, the human interaction is always inevitable.

In this study a data base for catalytic methanol synthesis from synthesis gas was constructed using published literature and models were build using various data mining tools to extract useful information of the reaction and predict CO_X conversion, methanol selectivity and methanol yield. In Chapter 2, a literature review of methanol synthesis form synthesis gas of the last 10 years, the theory of data mining techniques used for the modeling, the validation of the models and a review of data mining research in catalysis are presented. Chapter 3 describes the construction of the database in a detailed manner and explains the implementation of the data mining algorithms. The results of each model and their discussion are situated in Chapter 4. Finally Chapter 1 contains the conclusion of this study and recommendation for future research on this topic.

2. THESIS BACKGROUND

This chapter covers the theoretical background of methanol synthesis from syngas, various researches of the last decade and their findings of how physiochemical and catalytic properties influence the reaction results, the theory of data mining, details of methods used in this work and an overview of data mining research in the field of catalysis.

2.1. Methanol Synthesis from Synthesis Gas

As already mentioned in Chapter 1, methanol is mainly produced from synthesis gas which traditionally consisted of CO and H_2 but over the years addition of CO_2 to the feed up to total exclusion of CO was also studied. A catalyst is used to enhance the reaction conditions and make the process feasible. During the synthesis several reactions take place and produce desired and undesired products; the equations 2.1, 2.2 and 2.3 are the main reactions in the process.

CO hydrogenation:

$$CO + 2H_2 \leftrightarrow CH_3OH \qquad \Delta H_{r,298} = -90.7\ kJ/mol \tag{2.1}$$

CO_2 hydrogenation:

$$CO_2 + 3H_2 \leftrightarrow CH_3OH + H_2O \qquad \Delta H_{r,298} = -40.9\ kJ/mol \tag{2.2}$$

Water gas shift reaction (WGS):

$$CO + H_2O \leftrightarrow CO_2 + H_2 \qquad \Delta H_{r,298} = -42\ kJ/mol \tag{2.3}$$

All the reactions above are exothermic and reactions 2.1 and 2.2 are accompanied by a decrease in volume. Therefore in principle the methanol formation is favored by high pressures and low temperatures (Bertau *et al.*, 2014).

2.2. Literature Survey of Methanol Synthesis from Synthesis Gas

The researches reported in the literature were done on various aspects of the methanol synthesis; to give a more detailed understanding of these researches, a review of the last 10 years is sorted by the variables that were changed and is presented next. Since there is a significant amount of published papers, instead of presenting every one of them an exemplary selection is presented.

2.2.1. Feed Composition

Most of the research takes place at a fixed feed composition (with CO, CO_2 or mixed feeds); only few of researches examined the influence of varying CO_X/H_2 and CO/CO_2 ratios.

One team investigated the CO_2 free case and variated the CO/H_2 ratios in the feed using Cu/ZnO catalysts for the hydrogenation. CO amount in the range of 0 – 100 % was tried; the highest methanol signal was observed at the ratio of 2, which corresponds to a CO content of 67 % (Vesborg *et al.*, 2009).

A kinetic study on $Cu/ZnO/AL_2O_3/ZrO_2$ catalysts showed that the addition of CO_2 to the feed enhances the methanol production as long as the CO_2 fraction does not exceed 0.25 % (Lim *et al.*, 2009). The authors attributed that effect to a decrease of dimethyl ether (DME) production with increasing CO_2 because of a decreased WGS reaction and therefore less water production, which is needed to produce DME. Since the CO hydrogenation decreases faster than the WGS reaction, the effect is reversed after a threshold value is reached. A thermodynamic modeling study of isothermal and adiabatic cases, conducted by other researchers, also leads to very similar results (Iyer *et al.*, 2015).

Another team of researchers compared the effectiveness of binary and ternary catalysts witch the components Cu, ZnO and Al_2O_3 in different arrangements in the absence and presence of 0.61 vol. % CO_2 in the synthesis gas. The results showed that if CO_2 is excluded, a high space time yield (STY) is achieved with a Cu/Al_2O_3 catalyst and a moderate STY with a $Cu/ZnO/Al_2O_3$; all other constellation have insignificant yields. If CO_2 is added the STY

of the $Cu/ZnO/Al_2O_3$ catalyst raises by approximately 100 % while the STY of the Cu/Al_2O_3 catalyst decreases to an insignificant number. Hence it can be deducted that the catalyst has to be chosen according to the feed composition or vice versa (Santiago *et al.*, 2012).

2.2.2. Base Metal and Support

In the catalytic methanol synthesis, the most commonly used catalyst is the $Cu/ZnO/Al_2O_3$ with a Cu content of ~60 %, ZnO of ~30 % and Al_2O_3 of ~10 %; the form of each species can vary from pure metal to metal oxides, bimetallic compounds and other forms. Experiments involving the addition of other species, substitution of present species and construction of novel structured catalysts have been also done to increase the conversion, methanol selectivity and catalyst stability. Taking into account of all these experiments, it is hard to distinguish the base metals from the support material since it is considered differently in different researches. In general, it is established that the catalyst activity is proportional to the Cu metal area and is enhanced by the interaction with the ZnO part of the catalyst. The Al_2O_3 is contributing to the catalysts stability (Waugh, 2012).

Using different preparation methods and calcination conditions, Mierczynski *et al.* (2011) constructed various Cu/support catalyst with the support being ether monoxides or binary Zn and Al oxides with different proportions. The highest activity was observed on a $ZnAl_2O_4$ binary oxide with the ratios of Zn/Al = 0.5 and Cu/ZnO = 0.72. This was explained by the fact that this binary oxide was the one with the highest surface area among the binary oxides and therefore contributed to the high activity. Even though the Al_2O_3 mono oxide has higher surface area, it lacks the synergetic effect with the ZnO and therefore yields a lower activity.

One study gradually substituted the Al component with Zr and tested the different Cu/Zn/Al/Zr catalysts. It was found that the exposed Cu surface area and dispersion have a peak at the Zr/(Al+Zr) atomic ratio of 0.3, which was also the catalyst with the best performance in conversion and methanol selectivity (Gao *et al.*, 2013). A different study confirmed the enhancing Zr effect by testing $Cu/ZnAl_2O_4$ catalyst with and without added ZrO_2. The results demonstrated an improvement in CO conversion from 10.9 to 14 % and in methanol selectivity from 56 to 90 % if 5 wt. % ZrO_2 was added to the catalyst (Mierczynski *et al.*,

2014). A different team investigated Cu/Zn/MO_x catalysts with M being Al, Zr, Ce or CeZr. They found the CuO–ZnO–ZrO_2 catalyst to yield best results in a pure CO_2 hydrogenation; a CO_2 conversion rate of 23 % and a methanol selectivity of 33 % were obtained (Angelo *et al.,* 2015).

A study addressing the structure of Cu/ZnO/Al_2O_3 catalyst, prepared in form of mesoporous aerogels and xerogels, was conducted by Guo *et al.* (2006). Although the aerogels showed better activity than the xerogels, the best aerogel achieved a catalytic activity of 22 % only, which is low compared to industrial standard catalyst. Another research compared commercial Cu/ZnO/Al_2O_3 catalyst with Cu/ZnO/SBA-15 catalyst with different Cu/Zn ratios (SBA-15 being a mesoporous silica). The best SBA-15 supported catalyst had a Cu/Zn ratio of 0.5 and yielded a CO conversion of 15.6 % and a methanol selectivity of 95.6 %. Most of the other SBA-15 catalyst showed conversion values less than 10 % and a selectivity of less than 90 %. The commercial reference catalyst chowed a slightly better result with a conversion of 17.5 % and a selectivity of 96 % (García-Trenco and Martínez, 2013). The research team of Witoon *et al.* (2015) examined copper-loaded hierarchical meso–macroporous alumina (Cu/HAl) catalyst and compared it with copper-loaded unimodal mesoporous alumina (Cu/UAl) catalyst. While no significant difference was observed in CO_2 conversion, the methanol selectivity of the HAl catalyst was 5 % higher than that of the UAl catalyst. This was attributed to the presence of macropores, which diminished the occurrence of side reactions by shortening the mesopores diffusion path length. The Cu/HAl catalyst also exhibited higher stability than the Cu/UAl catalyst due to the fast diffusion of water out from the catalyst pellets, enhancing the oxidation of metallic copper to CuO.

A Pd/CeO_2 catalyst was used in sulfur containing syngas to investigate catalyst poisoning and hence stability of the catalyst. It was found that CeO_2 containing catalysts lead to a stable reaction activity, whereas the Al_2O_3 counterparts lead to a decline in reaction activity. The effect was contributed to the ability of CeO_2 to convert H_2S to SO_X and therefore to prevent the active metal from being poisoned by sulfur. Consequently an active metal/CeO_2 catalyst was proposed for sulfur containing synthesis gas (Ma *et al.,* 2009). Further, Yoo *et al.* (2013) investigated Cu/$Ce_{1-x}Zr_xO_2$ catalysts with x having the range between zero and one. The team found that the methanol yield depended on the exposed Cu surface

area and the amount of oxygen vacancies in the support. The amount of oxygen vacancies was highest at x = 3; this was also the value that yielded the best activity.

In other studies, experiments with other base metals and supports were conducted but focused on other variables like promoter addition or reaction conditions and not on the comparison of the metals/supports among each other.

2.2.3. Promoter

Promoters are elements or compounds of small quantity added to the catalyst to increase the catalytic activity, selectivity or stability. Promoters can be of different kind i.e. nobble metals, transition metals, non-metals and other compounds like oxides. Most research focused on adding promoters to the commercial catalyst $Cu/ZnO/Al_2O_3$ or an alteration of it. Few researchers conducted studies on promoting catalyst different than the commercial one.

One team of researchers conducted experiments on ZnO by adding 1, 2 and 3 wt. % of gold. While the selectivity to methanol was always 100 %, the activity increased with the increasing amount of gold; the addition of 1 % increased the conversion by 55 %. Since the increase in conversion was minimal when 2 % Au were added and increased by 18 % when 3 % Au were added, the team concluded that the increase in conversion is not a linear function of the amount of added Au. The increase in conversion was attributed to the increase of oxygen vacancies at the Au/ZnO interface (Strunk *et al.*, 2009). A similar study conducted by Pasupulety *et al.* (2015) dealt with promoting $Cu/ZnO/Al_2O_3$ catalysts with 0.5 – 3 wt. % of gold and led to slightly different results. It was found that decreasing order of CO_2 conversion was obtained with the order of Au %: 1 wt. % > 0.5 wt. % > 0 wt. % > 3 wt. %. According to the researchers, one of the reasons for this irregularity is the Cu/Au ratio, which is optimum at a value of approx. five.

A study comparing the Au and Pd promoted and unpromoted $Cu/ZnO/Al_2O_3$ catalyst in CO hydrogenation showed that best results in conversion (15 %) and methanol selectivity (69 %) were obtained if 5 wt.% Pd was added to the catalyst. Au promotion increased the selectivity compared to unpromoted catalyst but decreased the conversion. The increase in

selectivity using Au promoted catalyst was explained with the creation of a gold – copper interface which had a synergetic effect, and the increase in activity and selectivity with Pd promoted catalyst was explained by the spillover effect of H_2 between Pd and CuO (Mierczynski *et al.,* 2014). Pd promotion was also studied in combination with Cu promotion on SiO_2 supported catalysts; the Pd/(Pd+Cu) ratio was varied and the activity and selectivity towards methanol was measured in a CO_2 hydrogenation. It was deducted that, due to a synergetic effect of the two promoters, each ratio of a bimetallic promoter yielded higher result than monometallic promotion. The best results were achieved with the Pd/(Pd+Cu) atomic ratios between 0.25 – 0.34. That yielded a conversion of 6.7 % and a selectivity of 34 %. The team figured out that the alloy formation between Pd and Cu plays a crucial role in the bimetallic promotion (Jiang *et al.,* 2015).

Several studies investigated the effects of various kinds of promoters rather than of their amount. One of these studies conducted a factorial design study by adding metal oxides in various combinations to a commercial $Cu/ZnO/Al_2O_3$ catalyst. The metal oxides were Mn, Mg, Zr, Cr, Ba, W and Ce oxides. The oxide additives were found to influence the catalytic activity, Cu dispersion, Cu crystallite size, surface composition of the catalyst and its stability. Mn and Zr promoted catalysts yielded high performance for methanol synthesis from syngas. Whereas the Mn doping contributed to a higher activity, the Zr doping yielded a better stability of the catalyst. Therefore a bimetallic promotion was suggested (Meshkini *et al*., 2010). Similarly, another team of researchers testes several metal oxides and colloidal silica as promoters on a commercial $Cu/ZnO/Al_2O_3$ catalyst. They tried Mn, Ga, Zr and Si oxides and compared their activity and stability among each other. Catalyst promoted only by SiO_2 exhibited the best stability by inhibiting sintering of the catalyst whereas Mn doped catalyst showed no significant increase in stability, which is in accordance with the previously mentioned study (Samei *et al.,* 2012). A third study with other promoters was conducted by (Gao *et al.,* 2013). They used Mn, La, Ce, Zr and Y as promoters for a commercial $Cu/ZnO/Al_2O_3$ catalyst. It was found that the CO_2 conversion depends on the exposed Cu surface area and the methanol selectivity increases linearly with the increase of the proportion of strongly basic sites to the total basic sites. It was observed that Y modified catalysts had the highest total number of basic sites and Zr modified catalysts had the highest proportion of strongly basic sites. Y modification yielded the highest CO_2 conversion while Zr

modification gave the highest selectivity. The highest methanol yield was achieved with the Y promoted catalyst.

One research team studied the effect of mono oxide and binary oxide Ti and Zr promotion on a CuO/ZnO catalyst. They conducted that while all additions favored the CO_2 conversion and methanol selectivity, the mixed oxide yielded the best results. The mixed oxide promoted catalyst increased the methanol yield by 30 % compared with the not promoted catalyst. Furthermore a non-linear increase in catalyst activity was observed with the increase of Cu surface area (Xiao *et al.,* 2015). In another research, Cu/Zn/Al/Zr catalysts were modified with fluorine. With the introduction of fluorine, a strong decrease in Cu surface area and a significant increase in the proportion of strongly basic sites was observed. That led to a slight decrease in CO_2 conversion and a remarkable increase in selectivity towards methanol, which finally contributed to a higher methanol yield than that of not promoted catalyst (Gao *et al.,* 2014).

Experiments with Au and Ag promoted $Cu/CrAl_3O_6$ catalyst were conducted, and an improvement in activity was observed with silver promoted catalysts, which was assigned to the silver chromate formation. No gold chromate formation was confirmed using the gold promoted catalyst and hence no improvement in activity (Maniecki *et al.,* 2009). A similar but more comprehensive study was conducted comprising Cu, Au and Ru promoted spinel type support ($FeAlO_3$, $ZnAl_2O_4$ and $CrAl_3O_6$). The unpromoted supports showed no activity in methanol formation. An addition of 5 wt.% Cu increased the activity drastically. The authors anticipate that catalyst activity increases with the donor-acceptor ability for oxygen atoms, which increases with the introduced Cu amount. The activity of the Cu doped catalyst was found as follows: $CrAl_3O_6 > FeAlO_3 > ZnAl_2O_4$. The addition of a second promoter (Ru or Au) to the Cu promoted catalysts showed lower activity in Ru promoted catalysts than in Au promoted catalysts. This was explained by the significantly increased amount of adsorbed carbon oxide on the catalyst surface of the gold promoted catalyst, which led to an increase in catalyst activity. The final order of methanol formation rate was found to be as shown in equation 2.4 (Maniecki *et al.,* 2009).

$$\begin{aligned} &5\%\ Cu/CrAl_3O_6 > 1\%\ Au–5\%\ Cu/FeAlO_3 > 1\%\ Ru–5\%\ Cu/FeAlO_3 > \\ &1\%\ Ru–5\%\ Cu/ZnAl_2O_4 > 5\%\ Cu/FeAlO_3 \approx 5\%\ Cu/ZnAl_2O_4 \end{aligned} \quad (2.4)$$

2.2.4. Catalyst Preparation

A vast amount of research is made on the catalyst preparation. The catalyst preparation basically consists of three important steps, which are preparation of the base material, calcination and reduction. During all of this steps the process can be manipulated to yield different catalysts. Most of the methanol synthesis catalysts are made by a coprecipitation method i.e. all ingredients of the catalyst are processed simultaneously. One part of the researches deals with different preparation methods like incipient wetness impregnation, sol-gel method and others and compares the resulting catalyst and its performance. Another part of the studies investigates different parameters during the preparation process like precipitation time, temperature and pH value. Some experiments regard different source materials for the catalyst ingredients. After the catalyst is prepared, in most cases, it undergoes a heat treatment called calcination. This treatment can be manipulated in terms of temperature and time, which is also a field of research. Finally an activation of the precursor to give the active catalyst is done. This is called reduction. Studies are conducted to learn the influence of reduction time, temperature and medium.

In one study, filament like ZnO and rod like ZnO were prepared by a hydrothermal process via ammonia-evaporation-induced impregnation of ZnO and used as a component in CuO/ZnO catalysts. These were compared to a CuO/ZnO catalyst prepared by conventional coprecipitation method. Characterizations showed that CuO/ZnO catalyst with filament-like ZnO exhibited stronger interaction between ZnO and Cu and had more oxygen vacancies and hence yielded the best results with a 16.5 % conversion of CO_2 and 78.2 % selectivity towards methanol. CuO/ZnO with rod like ZnO showed the weakest conversion (8.0 %) and selectivity (61.8 %) (Lei *et al.*, 2015). Another team of researchers prepared Cu/ZnO/SBA-15 catalysts by an ammonia-driving deposition-precipitation method (ADP) to confine the Cu/ZnO active parts inside the ordered mesoporous SBA-15 silica. Additionally a sample was prepared by impregnation and a $Cu/ZnO/Al_2O_3$ catalyst was prepared by coprecipitation for comparison. Catalyst prepared by the ADP approach showed 14 times higher activity than the one prepared by impregnation and almost as high as the reference coprecipitated catalyst (García-Trenco and Martínez, 2013). Another team prepared Cu/ZnO catalysts by citrate decomposition and coprecipitation and studied the effect of the different preparation methods. In that study the catalyst prepared by citrate decomposition achieved

higher activity than the one prepared by coprecipitation. The team contributed that effect to the higher interdispersion of Cu and ZnO particles (Karelovic *et al.,* 2012).

One research group prepared a Cu-Zn precipitate and added alumina with different morphologies using the reverse precipitation method to study the impact of the morphologies on the final catalyst. Two alumina morphologies were prepared by mixing aluminum nitrate with ammonia and sodium hydroxide respectively. One morphology was prepared as pseudo-boehmite and one as γ-Al_2O_3. A commercial Al_2O_3 alumina was used as reference. The catalyst prepared by precipitating aluminum nitrate with ammonia showed the highest Cu surface area and dispersion and therefore exhibited the highest conversion. The researchers explain it with the smaller CuO crystallites and because of that a higher Cu/Zn interaction, which plays a key role in methanol synthesis. The selectivity was almost 100 % with all tested catalysts (Wang *et al.,* 2010). Another team focused on a similar research; preparing $Cu/ZnO/Al_2O_3$ catalysts by coprecipitation using different precipitating agents, namely sodium carbonate, ammonium carbonate, potassium carbonate and sodium hydroxide. Also the influence of using novel continuous split plate (SPM) and valve assisted (VAM) micro mixers for the catalyst preparation were compared with the conventional batch method. It was deducted that Na_2CO_3 and $(NH_4)_2CO_3$ increased the Cu surface approximately 3-fold independently of the preparation method compared, whereas using different preparation methods had different effects on each precursor. Coinciding with these observation the methanol production rate was higher for Na_2CO_3 and $(NH_4)_2CO_3$ using any method and for K_2CO_3 using the micro mixers. Furthermore the decrease of Cu surface area before and after reaction using $(NH_4)_2CO_3$ was approximately 7 %, whereas conventional Na_2CO_3 showed only a slightly better decrease of 4 four %. The conclusions were that in the conventional batch process $(NH_4)_2CO_3$ can be used as a sodium free source and achieve similar results in activity and stability like the catalyst from the Na_2CO_3 precursor. Furthermore a change to continuous micro mixers would improve the methanol production rate significantly regardless of the precursors (Simson *et al.,* 2013).

The team of Frei *et al.* (2014) investigated the influence of precipitation and ageing temperatures on a $Cu/ZnO/ZrO_2$ catalyst prepared by coprecipitation and its activity. They found that precipitation temperature and ageing temperature influenced the pre-catalysts morphology. Higher temperature led to a higher crystallinity and lower temperature to a

more amorphous precursor phase. The amorphous catalysts showed a higher Cu surface area. Surprisingly the increase in catalyst activity of the more amorphous catalysts was only in the range of 10 – 25 %. They concluded that the morphology transforms to a more uniform phase under contact with the feed gas, under reaction conditions and then develops its active state. One study on Cu/ZnO catalysts prepared by coprecipitation experimented with different initial solution concentrations, stirring rates, ageing time and ageing temperature. It was shown that all these variables effect the catalysts morphology. In general nonlinear dependencies were found. Best catalytic activity was achieved with a 1M initial solution, a stirring rate of 500 rpm, an ageing time between 0.5 and 0.75 h and an ageing temperature between 40 and 60°C (Farahani *et al.*, 2014). A similar study on Cu/ZnO/Al_2O_3 catalysts prepared by coprecipitation under different conditions was conducted and included the parameters like precipitation pH, precipitation temperature, ageing time and calcination temperature. It was confirmed that all these parameters had an effect on the catalyst structure. Some general rules were deducted like a minimum ageing time of 20 min so the Cu/Zn particles can form bimetallic structures which are important for the methanol synthesis. It was also deducted that calcination temperature higher than 300 °C is contra productive in terms of activity. The best catalyst had a catalytic activity of 125 % compared with a commercial catalyst. It was concluded that following values are best for catalyst preparation: pH 6 – 8, precipitation temperature of 70 °C, ageing time of 20 – 60 min and calcination temperature of 200 – 300 °C (Baltes *et al.*, 2008).

One study was conducted on CuO-ZnO-Al_2O_3 catalyst using sol-gel and coprecipitation methods for preparation. The coprecipitation was further divided into two different zinc precursors, namely zinc oxide and zinc nitrate. Besides the influence of the different synthesis routes the effect of the calcination temperature was studied. The effects of calcination temperature were studied on the sol-gel prepared catalyst and showed best CO_2 conversion at 400 °C and best methanol selectivity at 300 °C, the best methanol productivity was achieved at a calcination temperature of 400 °C. Looking at the three different synthesis methods the best in terms of conversion was found to be coprecipitation from zinc oxide and in terms of methanol selectivity, sol gel method. While these two methods showed good behavior in different categories, the best methanol productivity was achieved by the coprecipitation method from zinc nitrate precursor. This catalyst calcined at 400 °C achieved a CO_2 conversion of 23 % and a methanol selectivity of 33 % (Angelo *et al.*, 2015).

2.2.5. Operating Conditions

By operating condition, we understand the reaction temperature, reaction pressure and the gas hourly space velocity (GHSV) of the feed. According to the stoichiometry of the main reactions and the fact that they are exothermic, the high pressures and low temperatures should be favorable. However, that is not always true in reality. Different catalysts have different optimum operating ranges for temperature; not all pressures and temperatures are feasible under industrial conditions and a lot of side reactions take place. Several researches have been done to investigate the operating conditions and their effects on the methanol synthesis from synthesis gas.

One research was made on $Cu/ZnO/Al_2O_3$ catalysts which were prepared under different conditions. The aim of the study was to reveal the dependence of CO_2 conversion, methanol selectivity and methanol space time yield on reaction temperature. The temperatures of 473 K, 493 K, 513 K and 533 K were tested. Independent of the catalyst, an increase in conversion was observed when temperature increased. The increase in conversion became smaller with higher temperatures. In total the conversion increased more than 100 % from 473 to 533 K. The selectivity exhibited the opposite effect and decreased with increasing temperature. This decrease was also nonlinear and was between 100 and 200 % from 473 to 533 K. Accordingly the space time yield was highest at 513 K and second at 493 K (Wang *et al.*, 2011a). An investigation of the influence of reaction temperature and pressure on the methanol synthesis using a $Cu/ZnO/Al_2O_3$ catalyst was done. The methanol yield increased from 4.6 to 7.7 % at 20 bar and from 6.0 to 16.0 % at 50 bar when temperature was increased from 200 to 250 °C. Therefore they concluded that higher pressure and higher temperature are favorable for a higher methanol yield and these two variables have a synergetic effect (Kleymenov *et al.*, 2012). Other experiments on $Cu/ZnO/Al_2O_3$ catalyst prepared by different methods were conducted by another team. They investigated the effect of reaction temperature and reaction GHSV. A general trend was observed that methanol productivity increased with temperature and GHSV. It was also observed that the effect was strongly depended on the method that the catalyst was prepared by. While for catalyst prepared by sol-gel method, a change in temperature from 240 to 260 °C roughly tripled the methanol productivity and a change from 5000 h^{-1} to 10000 h^{-1} yielded only a very small increase in productivity, for catalyst prepared by coprecipitation a change in temperature from 240 to

260 °C yielded an increase approx. 50 % and a change from 2500 h^{-1} to 10000 h^{-1} brought a 3-4 fold increase in productivity. A further increase in temperature to 280 °C had almost no effect at all (Angelo *et al.,* 2015).

One study experimented with different reaction temperatures, reaction pressures and GHSV on a $Cu/ZnO/ZrO_2/\gamma\text{-}Al_2O_3$ catalyst promoted by MgO. The GHSV was changed from 1500 h^{-1} to 5000 h^{-1} and the effects were observed. It showed that the GHSV had insignificant effects on CO_2 conversion and methanol selectivity. The space time yield (STY) increased like expected from 30 to 80 $g/kg_{cat}*h$. Increasing the reaction temperature from 230 to 310 °C brought an increase of CO_2 conversion from 7 to 17 % and a decrease in methanol selectivity from 33 to 13 %. The STY had its highest value at 270 °C and was 80 $g/kg_{cat}*h$. When the hydrogenation pressure was changed from 16 to 32 atm; a minimal increase of 1 % in conversion was observed but a significant increase in methanol selectivity from 23 to 47 % was achieved. The STY changed from 40 to 120 $g/kg_{cat}*h$. The team concluded that a higher reaction temperature will inhibit methanol synthesis and favor the reverse WGS and methanation reactions, whereas high reaction pressure and GHSV increase the space time yield of methanol (Ren *et al.,* 2015).

2.2.6. Different Reactor Systems

While the previously reviewed work mainly considered the use of commercial fixed bed heterogeneous reactor systems, some researches focus on novel structured reactors like coated foam reactors. Another field of research is the low temperature, low pressure methanol synthesis in slurry reactors with homogeneous and homogeneous/heterogeneous mixed catalysts within an alcoholic solvents.

One group conducted a modeling study on innovative, highly conductive, structured multi-tubular reactors, namely copper loaded honeycomb monoliths (HM) and copper loaded open-cell foams (OF). The performance of these reactors was compared to that of a commercial multi-tubular packed-bed reactor (PB). The study showed that PB reactors outperformed the novel structured reactors even at low stoichiometric numbers of the fresh feed. This was attributed to the effective convective heat transfer mechanism which is boosted by the high mass flow rate, higher radial heat transfer rates and lower hot-spot temperatures

than in the HM and OF reactors, in which the heat transfer takes place primarily by heat conduction. However, if short tubes are employed and the mass flow rate is low, PB reactors exhibit a poor heat transfer performance. The structured reactors show a heat transfer characteristic independent of these values and hence, can be operated with lower recycle ratios and more limited hot-spot temperatures. This makes the HM and OF attractive for small-scale applications of methanol units like biomass to liquid systems (Montebelli *et al.*, 2013). Later the same team of researchers conducted an experimental study to compare open cell foams with commercially used FB rectors. The results agreed with the modeling study. The FB was superior in terms of CO_X conversion and methanol productivity. In stability tests both showed similar behavior (Montebelli *et al.*, 2014).

A research on a low temperature, low pressure methanol synthesis at 423 K and 3 MPa in a gas-liquid-solid slurry system was made by Zhao *et al.* (2010). They used ethanol as a solvent. The study compared the use of heterogeneous Cu/MgO catalysts of different proportions, a homogeneous HCOOK catalyst solved in ethanol and a mix of both. The best results were obtained with an Cu:MgO atomic ratio of 3:1 if only Cu/MgO was used as catalyst. The catalyst yielded a CO_X conversion of 16.1 % and a methanol selectivity of 99.4 %. HCOOK, when used as the only catalyst, yielded very poor results with a conversion of up to five % and was therefore, not investigated further. A mix of both catalysts achieved a CO_X conversion of 48.4 % and a methanol selectivity of 98.6 %. The team concluded that the catalysts have a synergetic effect in the methanol production, which is happening by esterification and hydrogenolysis. Another team investigated a similar setup. They investigated a slurry system at 433 K and 5 MPa with ethanol as a solvent. Cu/MgO was used as solid catalyst and different Na compounds (HCOONa, NaOH and Na_2CO_3) were tried as the homogeneous part solved in ethanol. The Cu/MgO catalyst alone led to CO conversion of 7.2 % and a methanol selectivity of 57.5 %. While the addition of any Na compound increased conversion and selectivity, the best results were observed with Na_2CO_3. The CO conversion was 40.1 % and the methanol selectivity 93.3 %. This study confirmed the synergetic effect of heterogeneous and homogeneous catalysts in methanol synthesis in slurry systems (Hu and Fujimoto, 2010).

One research team investigated different calcination temperatures and times on a $Cu/ZnO/Al_2O_3$ catalyst in a liquid phase methanol synthesis. The experiments were con-

ducted in order to examine STY and stability in form of deactivation rate of the catalyst. While for all catalysts, the methanol selectivity was higher than 99 %, the STY and deactivation rate varied. Three calcination temperatures (573, 623 and 673 K) were tried with a calcination time of 4 h. The best results were achieved with a calcination temperature of 623 K. A STY of 169 $g/kg_{cat}h$ and a deactivation rate of 1.01 %/day were observed. Keeping this temperature, the calcination time was varied between 0.5 and four h. A calcination time of 2 h yielded the best results of 172 $g/kg_{cat}h$ and 0.43 %/day. They compared it with the commercial $Cu/ZnO/Al_2O_3$ catalyst which yielded 138 $g/kg_{cat}h$ and 5.47 %/day. They concluded that, with the right calcination temperature and time, the activity and stability of a commercial catalyst can be improved significantly (Zhang *et al.*, 2010). Another team experimented with Al, Zr and Ce modified Cu/ZnO catalyst, prepared by different calcination temperatures, in a low temperature liquid phase methanol synthesis at 170 °C with ethanol as a solvent. The best productivity of 68.72 $g\ kg^{-1}h^{-1}$ and selectivity of 88.06 % was achieved with the CuZnZr catalyst. That catalyst was further investigated under different calcination temperatures (no calcination, 300, 400 and 500 °C). It was observed that with increasing temperature the productivity as well as the selectivity decreased. The best results were achieved with not-calcined catalyst and reached a STY of 106.02 $g\ kg^{-1}\ h^{-1}$ and a methanol selectivity of 87.04 %. This was contributed to the effect that a lower calcination temperature results in a higher Cu dispersion and smaller Cu crystal size (Xiao-bo *et al.*, 2014).

One team conducted a kinetic study on a liquid phase methanol synthesis with Cu/ZnO catalysts using different alcohol solvents. They compared the reaction rates when ethanol or 2-propanol is used as a catalytic solvent. In all reactions, in which the alcohol is involved, the 2-propanol showed quicker reaction rates and therefore, the overall reaction rate was faster when 2-propanol was used. The researchers postulated that the reaction activity of the methanol synthesis strongly depends on the structure of the alcohol which is used as promoter (Zhang *et al.*, 2008).

2.3. Data Mining

Data mining has its roots in statistics and machine learning and started as a small discipline in the computer engineering domain. Since then it advanced rapidly and became an independent field of study. Data mining is a set of data driven methods and techniques to

analyze and explore big data sets in order to find hidden, previously unknown and important information, rules, relationships and trends. Data mining can be divided into two main categories. The first is descriptive data mining. Descriptive data mining explores the given data and extracts nontrivial information which are present but too complex to be comprehended by manual inspection. Aims of descriptive data mining can be to sort data into sub groups in which the individual data is similar to each other but different from data in other sub groups. This is called clustering. Another task is to analyze the data and find anomalies, individual data that does not fit to the data set, and exclude them or focus on them. This is called outlier detection. The second category is predictive data mining. Predictive data mining processes the given data and builds a model that explains the data set. Based on that model, predictions for future or excluding the data can be made. One use of predictive data mining is classification. Classification, based on a model, assigns a data item to one of several predefined classes. Another important use is regression. Regression assigns a real value to a data item, based on the previously constructed model from the data set. Besides the mentioned tasks, data mining is used for many other tasks, which would be too many to list here (Kantardzic, 2011).

In the beginning, data mining was used in laboratory research, clinical trials, actuarial studies and risk analysis. Nowadays it spreads into fields like: genomics, astrophysics, customer relation management, e-commerce, social media, fraud detection, mobile telephony, quality control, product management, medicine, pharmacology, food science, image recognition, audience prediction for television and a vast amount of other applications. All fields, in which a large amount of data has to be analyzed and decisions have to be made, are potential candidates for data mining. Decision assistance is an important objective and already established as in medicine, where some treatments have been developed on a statistical basis without understanding the underlying biological mechanisms of the disease, and in aviation, where pilots receive assistance in form of for example autopilot systems (Tuffery, 2011).

To apply data mining to a problem, much more than just the application of a tool to random data is necessary. Data mining is a process of a few steps, which are all important to achieve useful results. Knowledge discovery requires pre- and post-work and each of the steps is of iterative nature. One can decide to use different tools to see if they yield different results or one can decide to use different information in the modelling step to build slightly

different models. Next a brief summary of the important steps will be given, which are also illustrated in Figure 2.1.

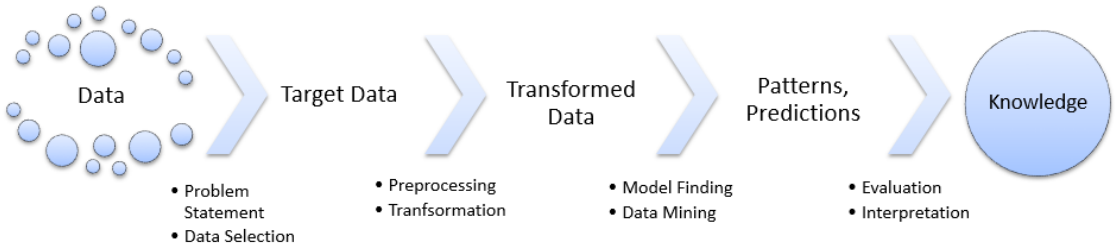

Figure 2.1. Knowledge discovery process.

First the problem has to be stated clearly and a particular application domain has to be chosen. Furthermore the researcher has to specify a set of variables for the unknown dependencies. This step requires domain specific knowledge and expertise. The second step is to select the data according to the stated problem. The data selection can be done ether by a human expert, which is called designed experiment or without the influence of the expert, which is called observational approach. Since data collection can affect the model and later the final results, an a priori knowledge can be very useful in this step (Kantardzic, 2011).

The selected data is called the target data and in most cases consists of a data base. This data base has individual data entities, which are called instances, data points or objects. They represent the rows of a data base and its size. The columns of the data base are called attributes, properties, predictors or simply variables and represent the dimensionality of the data base. A typical data base has variables of different kinds. One kind are the numeric variables, which can either be continuous like height of an object or discrete like age of an object. The other kind of variables are categorical, which are either nominal like gender or zip code of a city, or ordinal like a performance divided into “good”, “normal” and “bad” with an intrinsic order. Some other special kinds also exists and are used for particular purposes.

After the target data is collected it needs to be preprocessed and transformed. Preprocessing means to detect outliers, data values that are not consistent with most observations like measurement or recording errors or natural abnormal values, and remove them or use

them only with insensitive, robust models. Also missing values has to be paid attention to. If a data point lacks a value for a variable, the data point has to be removed or the value has to be estimated. After the target data is preprocessed, it needs to be transformed in order to be used for model building. The data can be transformed in different ways. One way is scaling. If variables have different ranges they can have different weights in the modelling step. So they have to be scaled, by different techniques. Often a scaling with respect to their range or standard deviation is applied. Another transformation is the encoding of a variable to a different kind of variable, which is easier to handle for the model. For example a continuous variable can be transformed into a discrete and vice versa or a multi valued categorical variable can be encoded to several binary categorical variables and the other way around (Kantardzic, 2011).

After the data is transformed, one or more appropriate data mining techniques have to be chosen and applied. The implementation of a data mining technique to build a model and discover dependencies takes place ether by a software with a graphic user interface or by a written algorithm in a programming environment, which allows for more detailed modifications and adjustments to the specific problem. Data mining techniques work either in a supervised or unsupervised manner. Supervised means that the model is built from known input-output samples and the error signal that is defined as the difference between the desired and actual response. Classification and regression are typical tasks of supervised learning. On the other hand, in unsupervised methods only input samples are given without predefined outputs and the model is required to evaluate itself on its own. Typical applications of unsupervised learning are clustering and discovering of relationships of the input data (Kantardzic, 2011).

After the model delivers patterns and predictions, these have to be evaluated, validated and interpreted. This is, in most cases, a semi-automatic step, where the human expert has to interfere again. Since the gained information should help in decision making, it needs to be accurate and interpretable. This two attributes are in nature contradictive. Usually simple models are more interpretable but less accurate and complex models are more accurate but less interpretable. Here a tradeoff according to the problem statement has to be done (Kantardzic, 2011).

The model can be built by several techniques like multiple linear and logistic regression, decision trees, random forest, support vector machines, artificial neural networks, k-mean and a vast of other machine learning algorithms, which all have benefits and drawbacks in their particular fields. In following sections, the techniques used in this work are explained in detail.

2.3.1. Clustering

As already mentioned previously, clustering has the aim to sub group data into sets in which the data points are similar to each other but not similar to data points in other sets. This sets are called clusters. To perform the clustering, the algorithm needs the samples, in most cases a predefined number of clusters and an objective criterion to compute the similarities between the data points.

2.3.1.1. Similarity Measure. The similarity measure is a metric that shows how close data points are to each other and therefore which cluster they should belong to. For convenience a dissimilarity measure is used instead of the similarity measure. It functions the same way but shows how far from each other or dissimilar data points are. The most common dissimilarity measure is the Euclidean distance. This is the straight line distance between two points if they are two or three dimensional. However, the concept is extendable to a multidimensional space. Equation 2.5 shows how a Euclidean distance can be calculated for two data points.

$$d_2(x_i, x_j) = \sqrt{\left(\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2\right)} \tag{2.5}$$

where $d_2(x_i,x_j)$ is the dissimilarity between data point i and j, p is the total number of variables, k is the actual variable, and x_{ik} and x_{jk} are the values of the data points i and j for the variable k. There are also distance measures like the city block distance, the Minkowski metric, the cosine-correlation and many others. Each suitable for their own purpose (Kantardzic, 2011).

However, if the data base is composed of categorical variables or of different types in general, the Euclidean distance cannot be computed for the data points. In that case a dissimilarity metric is needed that can take account of all different type of variables. This kind of dissimilarity was introduced by J. C. Gower in 1971 (Gower, 1971). The Gower distance can handle numerical and categorical variables of every kind and can deal with missing values. Equation 2.6 shows how the Gower distance is computed.

$$d_{ij} = \frac{\sum_{k=1}^{p} \omega_k \delta_{ij}^{(k)} d_{ij}^{(k)}}{\sum_{k=1}^{p} \omega_k \delta_{ij}^{(k)}} \tag{2.6}$$

where, d_{ij} is the Gower distance between data point i and j and takes values between 0 and 1. It is the weighted mean of the distances $d_{ij}^{(k)}$ between two data points in each variable. ω_k is the weight or importance of variable k, which can be assigned manually if a priori knowledge exists. The 0 or 1 weight $\delta_{ij}^{(k)}$ represents the comparability of data points i and j in variable k. It is 0 if one or both data points have a missing value in that variable or if the variable is asymmetric binary and both values are zero. Asymmetric binary means that they don't match if both are 0 but match if both are 1. In all other cases the weight is 1 and the data points can be compared in variable k. How the distance $d_{ij}^{(k)}$ for each variable is computed, depends on the type of variable. If the variable is nominal or binary and both values are equal, the contribution to the total dissimilarity is 0, otherwise 1. The contribution of other variables is the absolute difference of both values, divided by the total range of that variable, which gives a value between 0 and 1. Since all individual dissimilarities are in the range of 0 to 1, the Gower dissimilarity d_{ij} will also remain in this range (Maechler *et al.*, 2016).

Once a dissimilarity matrix including each data point is constructed it can be used as input in different clustering algorithms to cluster the data.

2.3.1.2. Partitioning Around Medoids (PAM). Partitioning Around Medoids is an objective function-based clustering, which builds clusters by minimizing a performance index. The idea of PAM is to represent the data by a collections of medoids, which are the most centrally positioned data points. To choose the median instead of the mean as the measure for the

cluster center has two essential benefits. First, the clusters are more robust and insensitive to outliers and second the cluster center is an actual data point and promotes interpretability.

PAM Algorithm

1. Select k medoids.
2. Compute distances between each data point and the medoids.
3. Assign data points to the medoid with the shortest distance (dissimilarity).
4. Compute average dissimilariy using equation 2.6.
5. For each cluster search if any of the entities of the cluster lower the average dissimilarity, if it does, select the entity that lowers the dissimilarity the most as the medoid for this cluster.
6. If at least one medoid has changed, go to step 3, else end the algorithm.

Figure 2.2. Algorithm for partitioning around medoids.

The performance index that has to be minimized, in order to form optimum clusters, is the average sum of dissimilarities, equation 2.7.

$$F(x) = minimize \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} z_{ij} \tag{2.7}$$

where F(x) is the average sum of dissimilarities, n the total number of data points, i and j the actual data points, d_{ij} the distance between i and j, and z_{ij} an indicator if the points i and j belong to the same cluster or not. z_{ij} is 1 if the two points belong to the same cluster and 0 otherwise. The PAM algorithm is given in Figure 2.2 (Kaufman and Rousseeuw, 1897).

2.3.1.3. Hierarchical Clustering. Hierarchical clustering is done by two modes. The bottom-up, also called the agglomerative mode and the top-down mode, also called the divisive mode. In the agglomerative mode each data point is treated as a single cluster and then merged with its closest one until just one cluster, which includes the whole data points, remains. The divisive approach works the opposite way around. One single cluster, including all data points, is divided until each data point represents one cluster. Hierarchical clustering

algorithms present a graphical data representation called dendogram. Figure 2.3 is an example dendogram of eight data points, divided into three clusters. The nodes located at the bottom of the graph correspond to the data points. Moving upward, the closest points are merged according to a similarity function. For instance, the distance between g and h is the smallest, and thus these two are merged. The length of clusters before being merged, shows how different they are to each other. The dotted line shows an arbitrary chosen threshold distance. If that distance is exceeded, the algorithm stops. In the case of this example the end result are three clusters, namely: {a}, {b, c, d, e} and {f, g, h}. There are several ways to calculate the distance, each suitable for its particular task. Some of the distances are: single link distance (minimal distance between two clusters), complete link distance (farthest distance between two clusters) and group average link distance (average between distances of each pair of data points from different clusters) (Cios *et al.*, 2007).

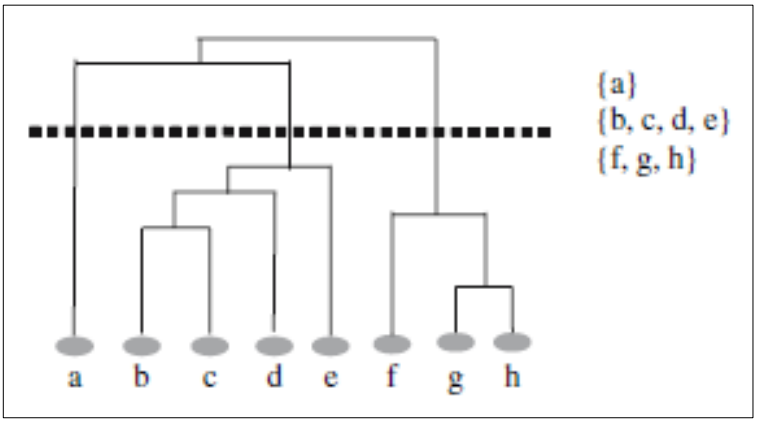

Figure 2.3. Example dendogram (Cios *et al.*, 2007).

2.3.2. Multiple Linear Regression

Multiple linear regression (MLR) is one of the most often used techniques in regression due to its simplicity and interpretability. Linear regression assumes a linear relation between the input variables (predictors) and the output variable (regressor). Although most real cases don't follow a linear dependency, they can be rewritten to fit a linear equation. Nevertheless, MLR in most cases is a good approximation to the real relationship of the variables. Equation 2.8 shows the general multiple linear regression equation.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \tag{2.8}$$

where, y is the target variable, x_i's are the input variables, p is the maximum number of variables, β_j's are partial regression coefficients and ε is the error or residual term. The error term represents hidden variables, which cannot be controlled or are missing in the model. Considering all data points, the regression equations for each data point can be written in matrix form as in equation 2.9.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{2.9}$$

This matrix equation can be written in short form as in equation 2.10.

$$Y = X\,\beta + \epsilon \tag{2.10}$$

where, Y is the vector of outputs, X is the matrix of inputs, β is the vector of partial regression coefficients and ϵ is the vector of errors. To get the regression model, β should be known. Assuming that ε_i are normally distributed and $\bar{\varepsilon} = 0$, β can be estimated using the least square estimates method shown in equation 2.11 with RSS being the residual sum of squares of errors in matrix representation and $\hat{\beta}$ the estimated regression coefficient vector.

$$RSS = \left(Y - \hat{\beta}X\right)\left(Y - \hat{\beta}X\right) \tag{2.11}$$

To find the optimum RSS, equation 2.10 has to be minimized. This is done by taking the derivative of equation 2.11 with respect to $\hat{\beta}$ and setting it to 0. This is shown in equation 2.12 and 2.13. X^t being the transpose of X.

$$\frac{\delta RSS}{\delta \hat{\beta}} = 0 \;\rightarrow (X^tX)\hat{\beta} = X^tY \tag{2.12}$$

$$\widehat{\beta} = (X^tX)^{-1}(X^tY) \tag{2.13}$$

Having this result, the multiple linear regression model can be estimated as equation 2.14 with $\hat{Y}$ being the vector of estimated (fitted) outcomes, which are different from the observed

outcomes in vector Y. The difference between the fitted and the observed values are the residuals ε_i's. Therefore the residuals vector can be computed by equation 2.15 (ReliaSoft Corporation, 2015).

$$\hat{Y} = X\hat{\beta} \tag{2.14}$$

$$\epsilon = Y - \hat{Y} \tag{2.15}$$

Sometimes, some variables are not as important for the model building process as others, or are not important at all. Including only the variables which are important and contribute to a more accurate prediction, the model can become simpler in terms of the amount of predictor variables. To achieve that, a variable selection has to be done. There are different ways to do that, namely: forward stepwise selection, backward stepwise selection and combined stepwise selection. In the forward stepwise selection, there are no variables in the initial model, than one variable is added at a time and the new model is compared with the previous one. This is done until all variables are added or until a threshold accuracy criteria is reached. The order by which variables are added depends on a significance measure, which has to be computed beforehand. Such a significance measure can be for example the p-value of the variable. This method is recommended for data sets with a high dimensionality. Backward selection works in the opposite direction. Initially all variables are included and then they are removed one by one beginning with the one that contributes least to the model. This method is recommended for data sets with low dimensionality. Combined stepwise selection is a combination of both, where each forward step is followed by one or more backward steps. The selection stops when no more variable can be added or the addition or removal results in an already evaluated model. This method is the most accurate one but demands the most computational time (Tuffery, 2011).

2.3.3. Decision Trees

The decision tree is a very popular technique due to its intuitive and simple approach, which in general yields good results. Decision trees can deal with heterogeneous data, missing values and nonlinearity. Decision trees can be used for classification and regression. If used for classification, explicit rules can be extracted and used for classification in an easy

comprehensible way. Basically the decision tree is a supervised, divisive, hierarchical method. It splits the entire data set (root node) into sub sets (nodes) by choosing the variable that provides the best separation into more pure groups of data point, containing the largest possible proportion of individual classes. The choice of the variable, which should be used for splitting, depends on a certain criterion. The value, at which the variable is separated depend on the type of variable. While binary variable offer just one way for separation, categorical and continuous variables offer n-1 possible separation values, n being the number of separate values for that variable. Usually all values or all values that fits a certain criteria are tested and the best possibility is chosen. This procedure is repeated recursively until no further separation of the data points is possible, i.e. the maximum purity of the node is reached or no further separation is desired. The terminal sub sets or nodes are called the leaves of the tree and with a reasonably high probability, belong to a certain class (Tuffery, 2011).

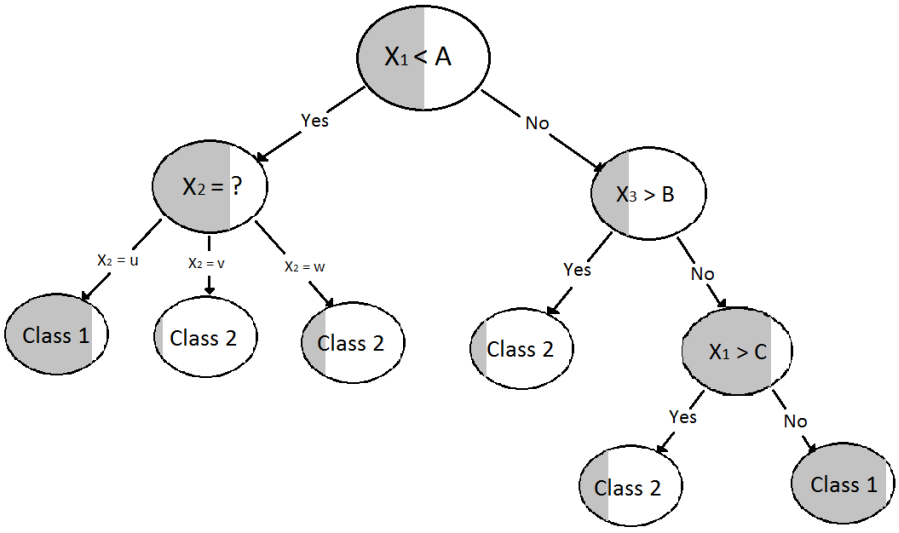

Figure 2.4. Example decision tree.

Figure 2.4 shows an example decision tree, which classifies the data into two classes. The gray region represents class one and the white region class two. As we follow the tree from the root node down to the leaf nodes, it becomes visible how the two classes becomes purer after each separation but don’t reach a 100 % probability in the end. It also can be observed that some variables can be used more than once in the tree construction and that numerical as well as categorical variables are used. If a tree makes binary or multiple splits, depends on the algorithm it is built with, but both are possible. The most common algorithms

are CART (Breiman *et al.*, 1984) and C5.0 (Pandya and Pandya, 2015) which agree in most points but have some slight differences.

The previously mentioned criterion for the best split variable differs between different algorithms. The most common ones are the Gini index (used by CART) and information gain (used by C5.0). They are not superior to each other in general but strongly depend on the conditions they are used under. The CART algorithm uses the Gini index as the default split criterion. The Gini index tells how much partitioned a node is. The Gini index of a node is at its minimum value 0 if the node is pure. If a node has equally distributed classes, the Gini index is at its maximum. The maximum value depends on the number of predefined classes. If two classes exist, the maximum Gini index is 0.5. If three classes exist, the maximum is 2/3. It measures the probability that two individuals, picked from a node, belong to two different classes. The Gini index can be calculated for a data set S (root, node or leaf) by equation 2.16.

$$Gini(S) = 1 - \sum_{i=1}^{C} f_i^2 \tag{2.16}$$

where C is the number of predefined classes, S the data set, $f_i = s_i/S$ the relative frequency of class c_i in the set and s_i the number of samples belonging to class c_i (Tuffery, 2011).

The quality of a split into k subsets S_i is calculated as the weighted sum of Gini indices of the resulting subsets from equation 2.17, with n_i being the size of the subset S_i.

$$Gini_{split} = \sum_{i=1}^{k} \frac{n_i}{n} Gini(S_i) \tag{2.17}$$

Since the greatest increase in node purity is desired, equation 2.16 has to be minimized by trying different variables. One property of the Gini index is that the impurity reduction is always positive, like shown in equation 2.18.

$$Gini(S) - Gini_{split} \geq 0 \tag{2.18}$$

The variable importance in decision trees, constructed using the Gini index, can therefore be computed by calculating the impurity reductions of the nodes which use the same variable for splitting and adding them up (Tuffery, 2011).

If decision trees are used for prediction, they are called regression trees and work in the same matter as classification trees. The differences are that the nodes predict the sample mean of the dependent variable in that node and the measure for the best split needs to be changed. Since real numbers have to be predicted, the criterion for the best split, has to be chosen accordingly. Basically it has to follow two rules. First the dependent variable must have a smaller variance in the child nodes than in the parent node and second, the dependent variable must have means which are as different as possible from one child node to the other. Considering this, the most often used criterion for node impurity is the sum of squared deviations about the mean of a node (Tuffery, 2011).

After a decision tree is build, in most cases it is pruned afterwards. Which means, subtrees are discarded and replaced by leaves. By pruning the tree, it gets simplified and branches which might lead to overfitting can be removed. Pruning is expected to lower the predicted error rate and increase the quality of classification. Pruning can take place by different criteria, some of them are: no quality improvement before and after splitting, number of samples in the leaf node are less than a minimum threshold number, a maximum of depth of the tree is reached or many others (Kantardzic, 2011).

2.3.4. Random Forest

Random forest is a method that has its roots in decision trees and a method called bagging. It builds vast amounts of de-correlated trees and averages them. The idea of bagging is to average large amounts of noisy but less biased or unbiased models and so reduce the variance. Decision trees are noisy models with low bias and therefore, ideal candidates for bagging. It can be proven that the bias of such a collection of trees is the same as the bias of each individual tree. Hence, the improvement has to take place by variance reduction. Equation 2.19 shows the variance of an average of B random variables, each with variance σ^2 and a positive pairwise correlation ρ (Hastie *et al.*, 2009).

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{2.19}$$

With increasing B, the second term goes to 0 and therefore the size of the correlation of pairs limits the benefits of the averaging. Random forests reduce the correlation between trees, without increasing the variance too much and therefore increase the variance reduction. This is done by random variable selection in the tree growing process. Before each split, a number of random variables m is selected as candidates for the split. With decreasing m, the tree correlation also decreases and the variance of the forest improves. The split criteria are the same like in decision trees. The predicted results are averages of the results, predicted by each individual tree. The quality of the forest can be adjusted by the number of individual trees and the number of random chosen variables. Also the variable importance is calculated in the same way as in decision trees with the difference that the importances are summed over all individual trees (Hastie *et al.*, 2009).

2.3.5. Model Validation

As we construct different models, it is inevitable to compare them and to measure their quality. This can be done in different ways. First of all we can evaluate the goodness of fit or in other words the capability of the model to fit values that have been used to construct the model. To do so, some quantities need to be defined. These are residual sum of squares (RSS), total sum of squares (TSS) and model sum of squares (MSS), which are calculated by equations 2.20 – 2.22. RSS represents the sum of squared differences between the real (observed) response y_i and the estimated (fitted) response $\hat{y}_i$.

$$RSS = \sum_{i=i}^{n} (y_i - \hat{y}_i)^2 \tag{2.20}$$

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2.21}$$

TSS is the total variance that the model can explain and is used as a reference quantity for standardized quality parameters. It is computed as the sum of squared differences between

the observed outcome and the average observed outcome $\bar{y}$. The last quantity MSS is defined as the difference between TSS and RSS.

$$MSS = TSS - RSS \tag{2.22}$$

Using these three quantities, all important quality parameters for the goodness of fit can be computed. The most common of those are the coefficient of determination (R^2), which is computed by equation 2.23, and the adjusted coefficient of determination ($R^2_{adj.}$), which takes into account the number of variables p and size of the data base n, and is computed by equation 2.24. Finally a direct measure of quality, the root mean squared error (RMSE), is calculated by equation 2.25 (Todeschini, 2007).

$$R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{2.23}$$

$$R^2_{adj} = 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right) \tag{2.24}$$

$$RMSE = \sqrt{\frac{RSS}{n}} \tag{2.25}$$

The previously mentioned quantities are used to express the goodness of fit of a model, and they show how well the model explains the change of dependent (response) variable with the independent variables. However, if predictions on previously unseen data have to be done, other but very similar quantities are used to quantify the prediction power of a model. For that, the database has to be divided into training and a test sets. The training set is used to build the model while the test set is used to predict previously unseen data points. It is crucial that the test set should be, in no way, used during the model construction since that would corrupt the validation. Usually 70 % of the data are used as training set and the remaining 30 % as test set. If the data base is small, and the 70 % are not sufficient to train the model adequately, a method called k-fold cross validation is applied, which includes the prediction of all data points using different training and test sets. In k-fold cross validation, the data set is divided in k equally sized subsets, from which k-1 subsets are used for training

and the remaining subset for testing. Than another subsets becomes the test set and the remaining k-1, the training sets. This is repeated until all subsets have once been the test set. The parameter k, in most cases is set to five or 10, but other values work as well. Figure 2.5 illustrates an example schematic of a 4-fold cross validation.

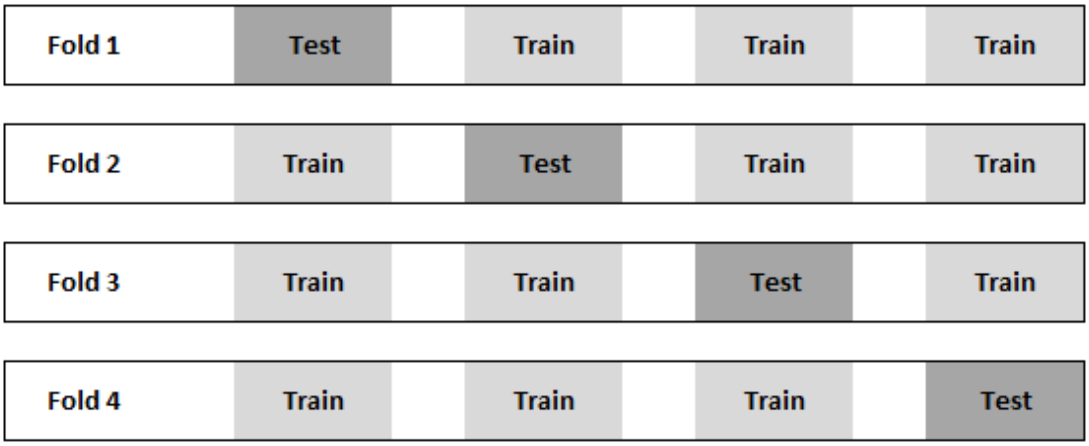

Figure 2.5. Schematic of k-fold cross validation.

To quantify the prediction power, a measure similar to RSS needs to be computed. The predicted sum of squares (PSS), which is the sum of squared differences of the observed outcome y_i and the outcome $\hat{y}_{CV,i}$, predicted by cross validation. PSS is computed by equation 2.26.

$$PSS = \sum_{i=1}^{n} \left(y_i - \hat{y}_{CV,i}\right)^2 \tag{2.26}$$

Therefore, an equivalent parameter to R^2 can be computed using PSS instead of RSS. This parameter called cross-validation R^2 and can be computed by equation 2.27. The equivalent to the RMSE is called prediction root mean squared error (PRMSE) and is computed by equation 2.28 (Todeschini, 2007).

$$R^2_{CV} = 1 - \frac{PSS}{TSS} \tag{2.27}$$

$$PRMSE = \sqrt{\frac{PSS}{n}} \tag{2.28}$$

Figure 2.6 shows an example for the trends of R^2 and R^2_{CV}. The difference of the two values is that R^2 improves the goodness of fit for a regression model with additional variables, but the prediction power, represented by R^2_{CV}, reaches a maximum and declines again. Therefore, a model should be chosen to have the number of variables, where R^2_{CV} is at its maximum, if prediction is desired. It is important to mention that for models with very poor prediction power, the R^2_{CV} can take negative values and therefore, should be used to represent and compare models with a prediction quality that yields positive values, only.

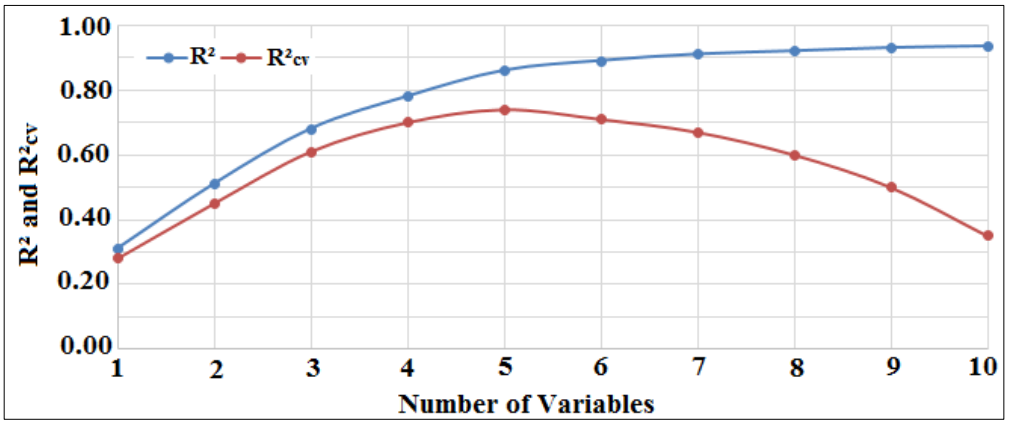

Figure 2.6. Comparison of R² and R²cv with respect to number of variables.

2.4. Data Mining Studies in Methanol Synthesis and Heterogeneous Catalysis

Data mining studies in methanol synthesis are very limited in number, but the field is growing noticeably. Also data mining research in other catalytic systems is experiencing a static growth. Although, the most published researches on data mining and catalysis implement artificial neural networks (ANN) as the method of choice, some are thinking out of the box and implement improved algorithms and other methods. ANN have the advantage of good predictive power but lack interpretability, which is very important to researchers from the chemical and engineering fields.

One of the early works on methanol synthesis used a combination of a genetic algorithm (GA) and a radial basis function network (RBFN) to find the optimum Cu-Zn-Al-Sc catalyst, which had higher activity than the conventional catalyst. The team also used a combination of ANN and GA to find the optimum composition and calcination temperature for a Cu-Zn catalyst (Umegaki *et al.*, 2003). Another team used a combination of ANN and GA to find the optimum catalyst composition and preparation conditions for low pressure, low temperature methanol synthesis from syngas. The optimum catalyst had the composition of Cu/Zn/Al/Sc/B/Zr = 43/17/23/11/0/6 prepared by using 2.2 times the equivalent of oxalic acid and calcined at 605 K. The activity was almost twice of the activity of an industrial catalyst (Watanabe *et al.*, 2004).

One team used artificial neural networks to model a methanol synthesis in a multi reactor system to find optimum parameters. On these parameters, a model predictive control scheme, which had to fulfill certain process constrains, was developed for yield maximization (Fissore *et al.*, 2004). A study with a hybrid approach, using a first principle mechanistic model and an ANN, was conducted to simulate and analyze a packed bed reactor for CO_2 hydrogenation to methanol. It was found that the hybrid model outperformed both single models and was able to predict the experimental outcomes very accurate (Zahedi *et al.*, 2005).

One of the first works on data mining in heterogeneous catalysis is the investigation of optimum reaction condition for a Fischer-Tropsch synthesis using a Co/SiO_2-Al_2O_3 catalyst. The objective was to find the condition for the optimum C_{5+} liquid hydrocarbon selectivity. Experimental data of reaction conversions and steady state concentrations of different products were used to train an ANN model. Satisfactory results were achieved and future studies on this model were suggested (Sharma *et al.*, 1998). A recently conducted study on Fischer-Tropsch synthesis with $Co(III)/Al_2O_3$ catalyst used ANN to predict product concentrations by looking at the reaction conditions and a GN to find the optimum reaction parameters once the concentrations were predicted. The ANN model predicted the outcomes with correlation coefficients of $R^2 = 0.94$ for CH_4, $R^2 = 0.93$ for CO_2 and $R^2 = 0.96$ for CO (Adib *et al.*, 2013).

One team of researchers extracted knowledge for water gas shift reaction over noble metals from literature. They used decision trees to determine rules and conditions for high catalytic performance, ANN to determine the importance of different variables and support vector machines (SVM) to predict outcomes of unstudied experiments. They deducted that each technique is successful in its particular task. The predicted outcomes lay in the range of +/- 10 % of the experimental outcomes (Odabaşı *et al.*, 2014).

An ANN analysis was done by (Günay and Yildirim, 2011) on selective CO oxidation over Co-based catalysts. Relative significances of the variables for catalyst preparation and operating conditions were found and CO conversions were predicted with an RMSE of 8.69. It was found that the reaction temperature is the most important variable with a relative importance of 54.3 %, followed by Cu weight % with 14.0 %. The same team of researchers conducted a similar study for CO oxidation over noble metal catalysts. They used decision trees to deduct rules that lead to high performance. For example, it was found that if the Pt amount is high enough in the catalyst and the operating temperature is in a suitable range, high CO conversions are reached without any help of a promoter. They also used ANN to determine the relative importance of the variables. Finally they clustered the data base by a genetic algorithm, build ANN model for each cluster and predicted unseen outcomes accordingly with a RMSE of 12.23 (Günay and Yildirim, 2013b). This team also developed a global reaction rate model for CO oxidation over Au catalyst using local reaction rate models, published in literature. They used an ANN to develop power laws for each support type, determine the reaction order with respect to each reactant and estimate the Arrhenius parameters. After validation of the models by reported information, it was deducted that they can be used for estimation of reaction rates in the absence of specific rate equations (Günay and Yildirim, 2013a).

One research group analyzed catalytic data on oxidative methane coupling by analysis of variance, correlation analysis, and decision tree. The decision tree was used for regression and it could be deducted from it that high performance catalysts are mainly based on Mg and La oxides. Alkali and alkaline-earth increased the selectivity and Mn, W and Cl anions increased the catalytic activity (Zavyalova *et al.*, 2011). One team compared different decision tree algorithms with different SVM algorithms, using them on olefin epoxidation catalysts. The aim was to optimize the catalyst and find the most important variables. Eight SVM and

six decision tree algorithms were tested and compared. While all decision trees identified the same variable to be the most important one, they were all outperformed by the SVM algorithms in terms of predictive power (Baumes *et al.*, 2006).

3. COMPUTATIONAL DETAILS

This chapter deals with the selection of the published papers to be used, the construction of the database, details of the input variables and the machine learning algorithms, which were used to extract knowledge from that data. The construction of the database is explained first, followed by some of the discussed preprocessing steps like standardization and encoding, finally the application of the data mining methods, which were presented in detail in Chapter 2.3 is explained.

3.1. Experimental Data Collection

To construct a database for methanol production from synthesis gas, scientific papers were collected from online sources like Science Direct, Wiley, American Chemical Society (ACS), and Springer. It was chosen to cover the researches from the last 10 years, from 2005 until 2015. Although, research on methanol synthesis dates far more back than 10 years, this period seems to be a suitable and sufficient in terms of essential research topics. In the course of the data collection process, initially 89 papers were collected and reviewed. The first screening was done to exclude papers which focus on kinetics, thermodynamics, catalyst surface science, simulations or homogeneous catalysis with autoclave reactors. After that screening, 53 articles were left. The detailed examination had the purpose of finding suitable papers in terms of clear defined input and output variables. The input variables needed to include the catalyst composition, preparation method, calcination conditions, reduction conditions, reaction conditions and feed composition. Two out of three output variables (CO_X conversion, methanol selectivity and methanol yield) had to be reported, so the third could be calculated. After this examination 29 articles were left. After building the database, a final screening was done to remove data points with rare attributes. If the attribute appeared less than five times or was used in just one paper, it was considered as rare. This led to the final amount of 24 articles. From these articles, 357 unique data points with a total of 28 variables could be extracted and used for modelling purposes. The database was constructed with Microsoft Excel 2013. The software WebPlotDigitizer 3.8 was used for attribute extraction from graphical representations.

Some of the input variables were related to the catalyst composition; those included the base metal, support material and promoter. The catalyst composition is an essential research topic and is known to influence the catalytic performance, therefore it was inevitable to include these variables. Table 3.1 shows all included species, the number of data points, and their ranges. Some research papers reported the catalyst composition in terms of mole fractions while some did as metal to support ratios. These were converted to weight % to have a common basis. The conversions were done, using an Excel worksheet from the Ames Laboratory U.S. Department of Energy (Ames Laboratory, 2008). These variables were encoded as continuous variables that sum up to 100 %.

Table 3.1. Number of data points with respect to their species and their ranges.

Species	**Number of Data Points**	**Range (wt. %)**
CuO	284	0 - 66.03
ZnO	281	0 - 100.00
ZrO2	147	0 - 98.00
Al2O3	95	0 - 80.00
CeO2	43	0 - 92.60
Zr	37	0 - 24.15
Cr	37	0 - 28.45
γ-Al2O3	26	0 - 90.91
Pd	23	0 - 15.00
TiO2	22	0 - 56.74
Mg	21	0 - 0.63
Au	19	0 - 2.91
SiO2	18	0 - 94.30
Ga2O3	16	0 - 98.00
SBA-15	10	0 - 94.30

The next variable was the catalyst preparation method. The way a catalyst is prepared significantly contributes to its structure, therefore, a lot of research is conducted on this topic; consequently this variable had to be included. The catalyst preparation methods, used in the database were co precipitation, citrate decomposition, incipient wetness impregnation, deposition precipitation, ion exchange, sol-gel, colloidal deposition, wet impregnation, sequen-

tial impregnation, reverse precipitation and co impregnation. Figure 3.1 shows the distribution of the data points among the preparation techniques. The last three methods named above were combined into "others". It can be observed that co-precipitation was overwhelmingly the most preferred method. The variable was encoded as a categorical variable with nine unordered levels, which were represented by integer numbers. The number of each method was written in Figure 3.1 underneath its name.

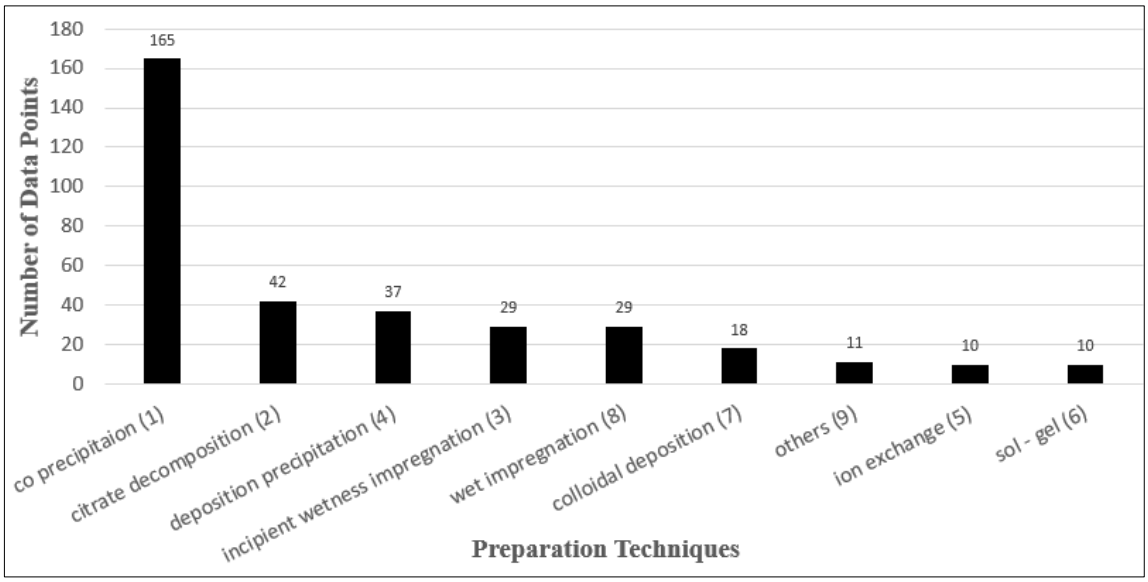

Figure 3.1. Distribution of catalyst preparation methods.

One step of the catalyst preparation is the calcination, which is done in almost all cases. Calcination can influence the base metal dispersion, support structure and other important properties of the catalyst and therefore, is subject to a lot of researches. Two calcination conditions were included into the database as input variables. These were the calcination temperature in Kelvin (K) and the calcination time in hours (h). If multiple calcination steps were conducted, the calcination with the highest temperature and its time was chosen. These two variables were encoded as continuous variables and presented in Figure 3.2 and Figure 3.3.

Figure 3.2 shows the distribution of the calcination temperature in intervals of 40 K. Most of the catalysts were calcined at 673 K, some lower and higher temperatures were tested as well, and some experiments were conducted on uncalcined catalyst; those cases

were presented as ambient temperature (298 K) in the database. The calcination time distribution has the shape of a normal distribution. The most frequent calcination time is four hours. All data points except one have integer values as hours.

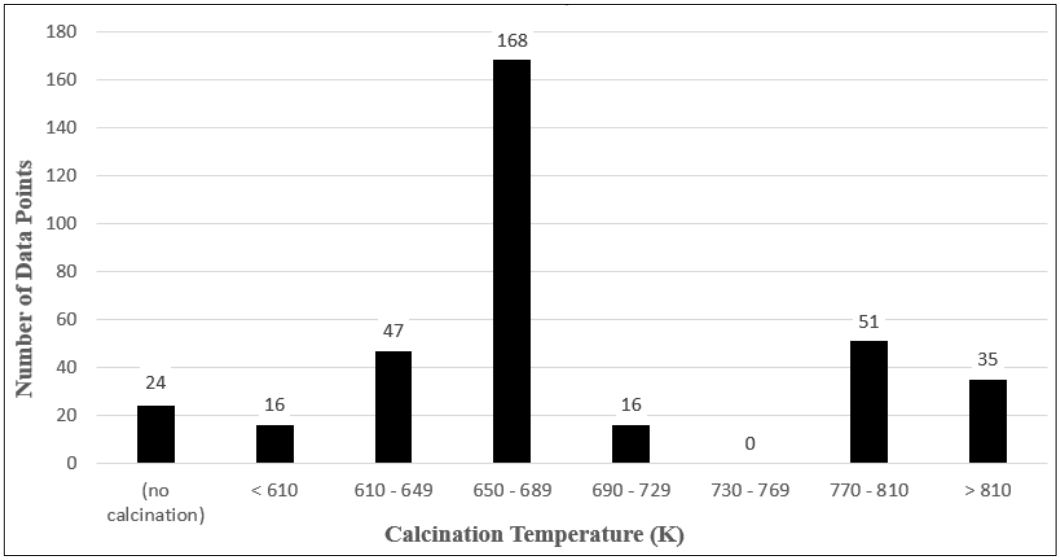

Figure 3.2. Distribution of calcination temperature

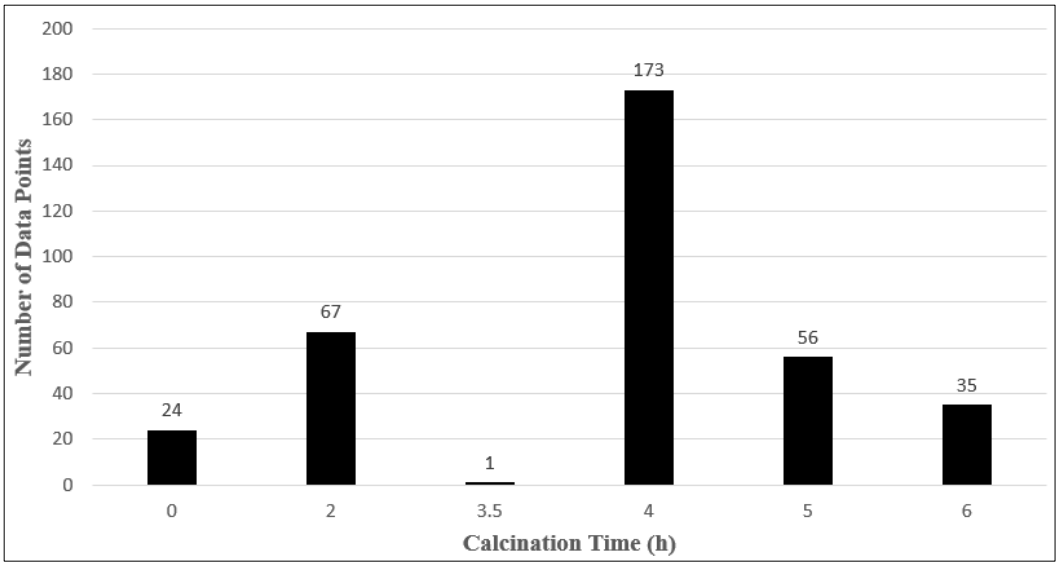

Figure 3.3. Distribution of calcination time.

Before the reaction, the catalyst undergoes a treatment called reduction, which defines the crystal structure of the active phase. The catalyst can be reduced in H_2 or H_2 diluted with an inert. The considered reduction parameters were reduction temperature in Kelvin (K),

reduction time in hours (h) and reduction H_2 content in volume %. All three variables were encoded as continuous variables. Figure 3.4 shows the distribution of the reduction temperature. The distribution is very scattered with few values with higher frequencies, therefore the values were presented in intervals of 40 K. The most used temperatures are in the range between 550 and 590 K. The distribution has a Gaussian shape.

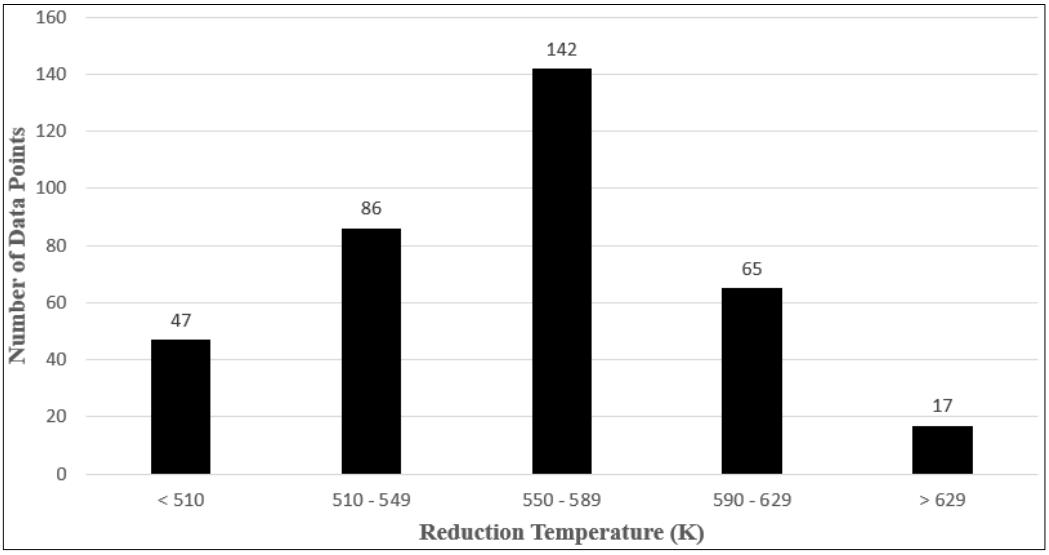

Figure 3.4. Distribution of reduction temperature.

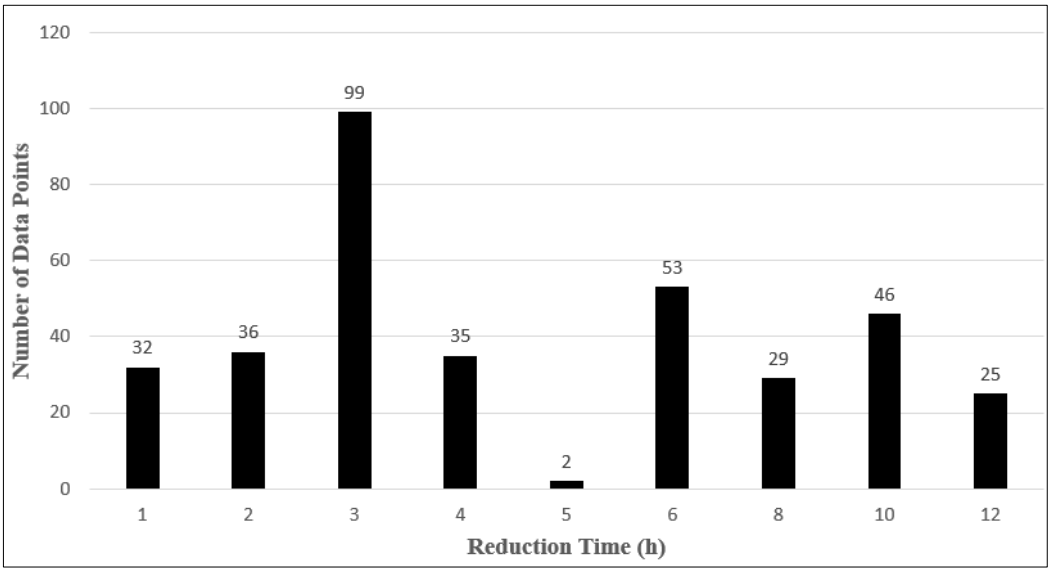

Figure 3.5. Distribution of reduction time.

Reduction time distribution is presented in Figure 3.5. Most catalyst were reduced three hours. A closer examination of the distribution, indicates two groups of reduction times: short reduction time with up to four hours and long reduction time with more than four hours. The last reduction parameter was the reduction gas composition in terms of hydrogen content. Figure 3.6 shows the distribution of the H_2 content with respect to the number of data points. It can be deducted that the majority of the catalysts is reduced under a flow of 100 % hydrogen. If hydrogen was diluted by an inert, the range of hydrogen varied between two and 10 % in the most cases. Different inert gases like N_2, He and Ar were used to dilute hydrogen. No difference were made in the database.

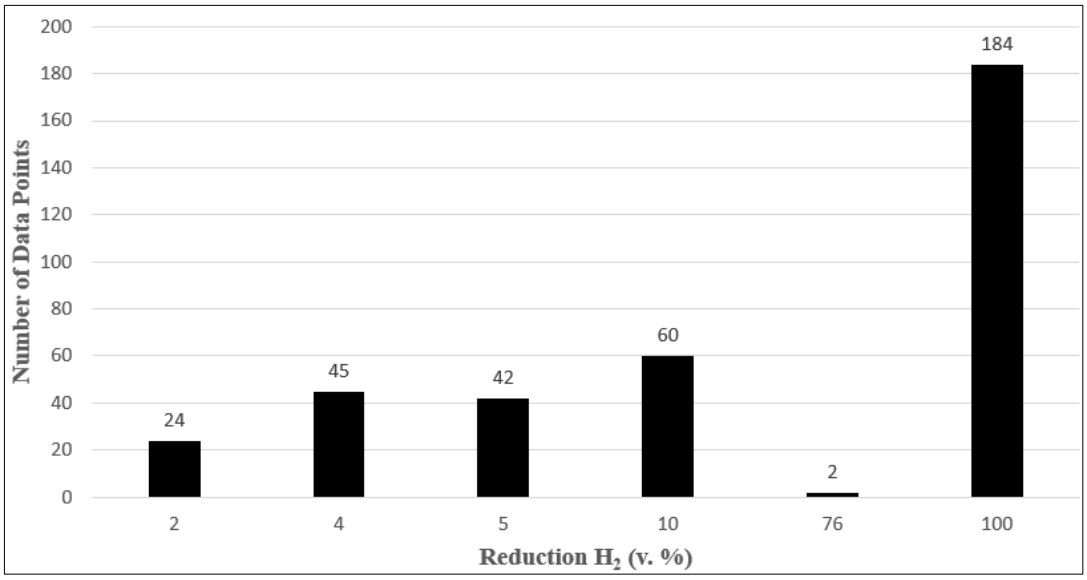

Figure 3.6. Distribution of reduction H_2 content.

Reaction temperature in Kelvin (K), reaction pressure in (MPa) and gas hourly space velocity in ($mL/g_{cat}*h$) were used as the parameters for reaction condition. They were all encoded as continuous variables. The reaction temperature has a wide range of values with a denser distribution between 490 and 550 K. The most frequent value was 523 K, which is in accordance with the industrial methanol synthesis process. The distribution of the reaction temperature is shown in Figure 3.7. Due to the scattered distribution, intervals of 20 K were presented.

The distribution of reaction pressure is presented in Figure 3.8. For an easily understandable representation the units of the pressure were chosen to be bar, although they are in

MPa in the database and intervals of 10 bar were presented. The reaction pressure distribution is more irregular than the reaction temperature distribution. The most experiments were conducted under a pressure of 30, 40 or 50 bar, which coincides with the commercial applications. Some experiments were also done under pressures as low as 7 bar, probably to explore the ability to produce methanol under less severe conditions.

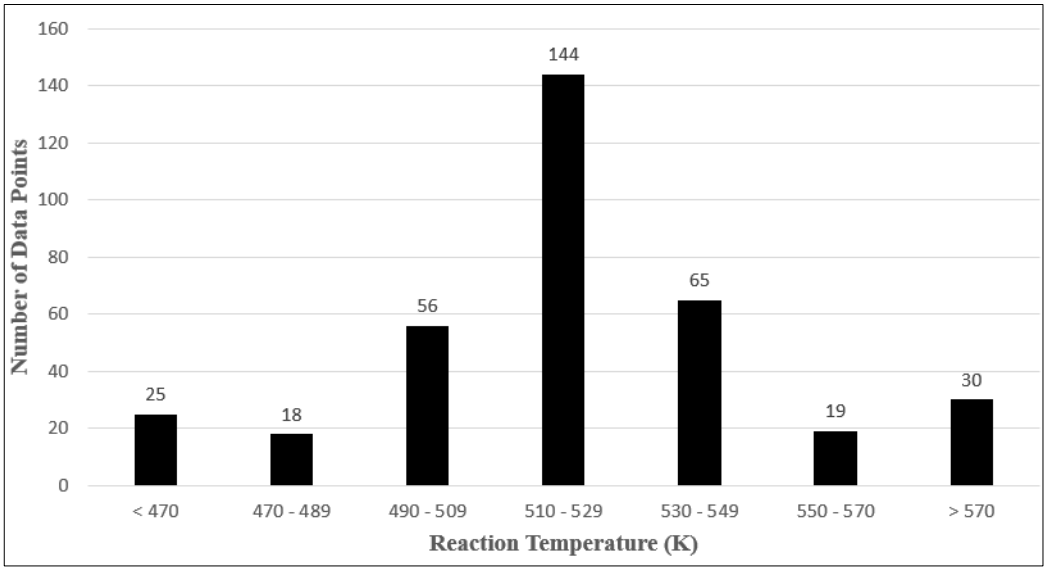

Figure 3.7. Distribution of reaction temperature.

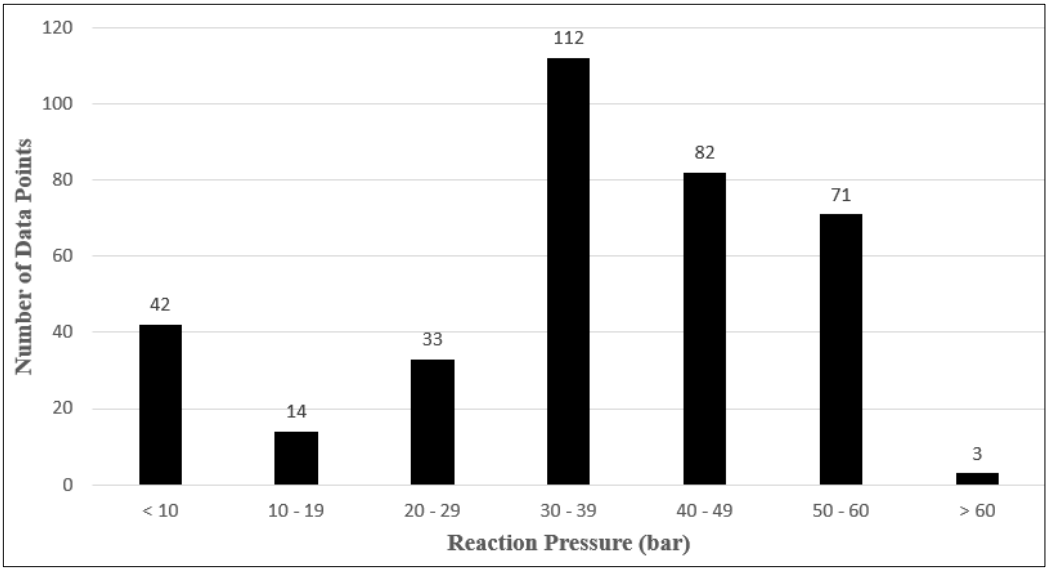

Figure 3.8. Distribution of reaction pressure.

The last variable related to reaction condition was the gas hourly space velocity (GHSV) of the feed gas in the units of ml/g_{cat}*h. A lot of articles reported the gas hourly space velocity in the units of h^{-1}, which does not provide any information about the catalyst weight. In that cases, the provided value was divided by the average density of the catalyst, which was calculated as a mixed density from the catalyst composition. If the GHSV was given with respect to moles instead of mL, the assumption of an ideal gas was made and the GHSV was multiplied by the standard molar volume V_m = 22.4 L/mol. Figure 3.9 shows the distribution of the feed GHSV in intervals of 500 ml/g_{cat}*h. Since the flowrate as well as the catalyst weight influence the value of the variable, values between 125 and 18000 mL/g_{cat}*h were possible. Although, extreme values were more seldom used, the number of experiments was not of insignificant. The most frequent GHSV values were in a range between 500 and 2500 ml/g_{cat}*h, and the range between 4500 and 5000 ml/g_{cat}*h.

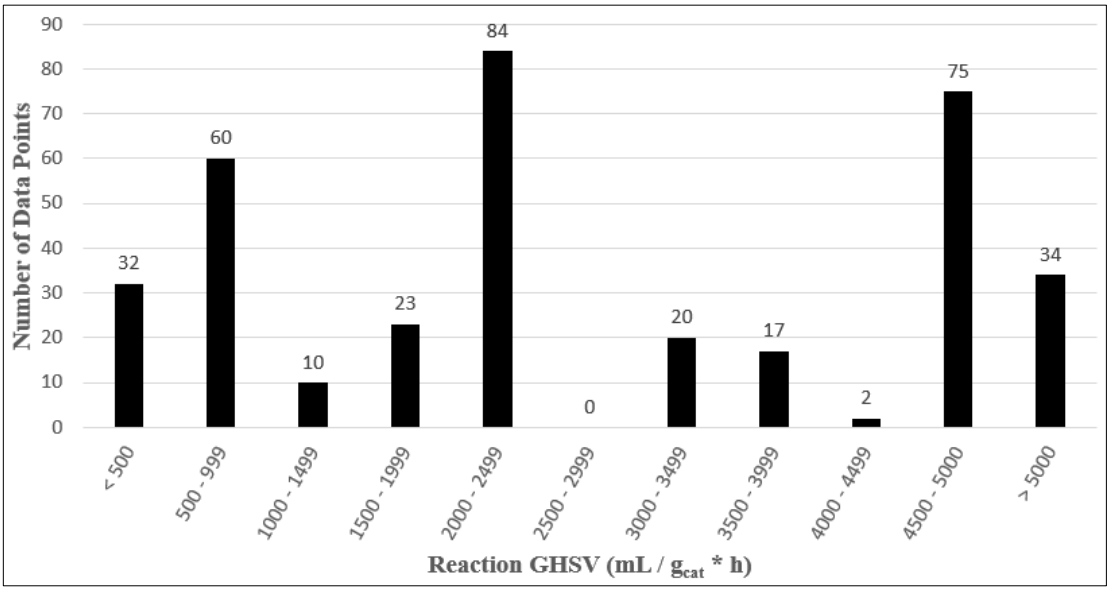

Figure 3.9. Distribution of feed gas hourly space velocity.

The last group of variables was the feed gas composition; it was composed of H_2, CO, CO_2 and inert. The inert gases were reported as different gases in different papers. Similarly to the reduction, the identity of inert gas was not taken into account; instead it was used as a single variable and named as inert. The feed composition variables were encoded as continuous variables. Table 3.2 shows the range distributions of the feed components in volume %. While all data points included hydrogen in various amounts, CO and CO_2 were not included in all data points. The amount of CO didn't exceed 33.33 % and the amount of CO_2

did not exceed 25.00 %. In some experiments an inert gas was also included, which might had the purpose of increasing the flow rate, slowing down the reaction by absorbing some of the heat or as internal standard to calculate the product distribution accurately.

Table 3.2. Range distribution of feed composition.

Feed Gas	**Range (v. %)**
H_2	59.4 - 90.00
CO	0.00 - 33.33
CO_2	0.00 - 25.00
Inert	0.00 - 12.50

As allready mentioned, some experiments were conducted only with CO, some only with CO_2 and the remainings with mixed carbon sources. According to literature, the reaction conditions for different carbon sources differ from each other. Table 3.3 shows the feed gas carbon source with respect to the amount of data points. About a half of the experiments were conducted with CO_2 under exclusion of CO. Another big part was conducted with mixed feed gas and a small portion was conducted only with CO. This might be due to the established fact that a certain addition of CO_2 increases the productivity.

Table 3.3. Number of data points with respect to carbon source.

Feed Gas Carbon Source	**Data Points**
CO	29
CO_2	216
CO + CO_2	112

The response variables (CO_X conversion, MeOH selectivity and MeOH yield) were encoded as continuous variables. The conversion was reported in all of the papers. Some papers reported a total carbon conversion and some seperated conversions of CO and CO_2. In case of seperated conversions, they were summed up. Methanol selectivities were reported directly in the most cases. In some cases the total alcohol selectivity and the methanol to higher alcohol ratio was reported. In that cases the methanol selectivity was calculated

accordingly. The methanol yield reported in % was used in the database. Methanol yield reported as space time yield (STY) in g_{MeOH}/g_{cat}*h could not be used for the database constuction. If selectivity or yield was missing, they were computed by equation 3.1.

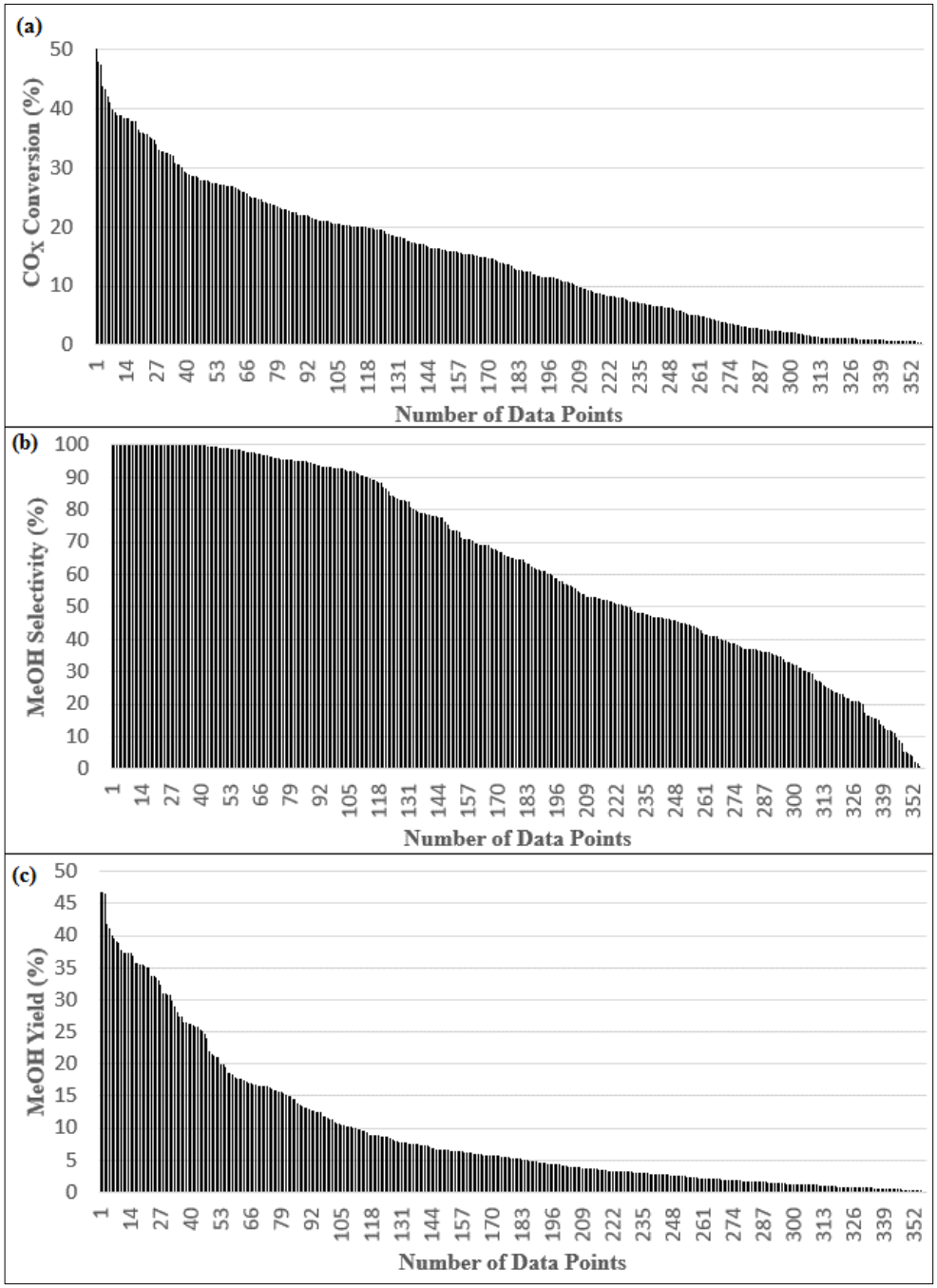

Figure 3.10. Continuous (a) CO_X conversion (b) MeOH selectivity (c) MeOH yield distribution.

In equatıon 3.1, X_{COx} is the CO_X conversion, S_{MeOH} the methanol selectivity and Y_{MeOH} the methanol yield. Figure 3.10 shows the distributions of the response variables, sorted from high to low.

$$Y_{MeOH} = X_{CO_X} S_{MeOH} \tag{3.1}$$

The conversion is equally distributed with slightly less data points in the higher region. It has the range between 0 and 50 %. The selectivity has clearly more values in the higher region, including several data points with 100 %, and very few in the lower region. The selectivity range is between 0 and 100 %. The yield has its most data points in the lower region with values smaller than 10 %, while approx. one third are above. The methanol yield ranges from 0 to 47 %.

3.2. Modeling

The computational work was conducted in Rx64 3.2.3 (R Foundation for Statistical Computing, 2008) medium within the included interface. Various additional packages, which will be named in appropriate places in the remaining part of the thesis, were used to conduct the modelling tasks. Descriptive and predictive data mining algorithms were applied to extract previously unknown information from the data base. First, the database was pre-processed by standardizing the input variables to achieve a comparable scale. Than the catalyst composition variables were encoded into one multivalued, categorical variable by partitioning around medoids and hierarchical clustering, to reduce dimensionality. Multiple linear regression, decision trees and random forest were applied on the complete and clustered databases to extract knowledge, determine the importance and the effect of variables on the outcome, and predict unseen experiments.

Preprocessing started with standardizing the input variables to comparable ranges. This is necessary to prevent domination of variables with large scales over variables with small scales. The categorical variable “preparation method” was not standardized because it is treated differently by the modelling algorithms. The scaling was done by the z-standardization technique, shown in equation 3.2.

$$z = \frac{x - \mu}{\sigma} \tag{3.2}$$

were, x is the value of the variable, μ is the mean value of the variable, σ is the standard deviation (SD) and z is the standardized value of the variable, also called z-score. The z-score is computed in a way that its mean is 0 and its standard deviation is one. Positive scores represent values above the average and negative scores represent values below the average of the variable. The response variables were transformed to fractions by division by 100. This ensured a proper representation between zero and one and better comprehension of the RMSE. Table 3.4 shows the means and standard deviations of all untransformed variables. They can be used for back transformation to improve interpretability, if desired. After that, the variables were assigned proper classes, so they could be treated by the algorithm accordingly. The classes were: numeric, integer and factor. After this transformation the database could be used as input for the data mining algorithms.

Table 3.4. Mean and standard deviation of variables.

	Mean	**SD**
Pd	0.57	2.62
Au	0.10	0.47
Mg	0.01	0.06
Zr	0.76	3.06
Cr	1.18	4.34
CuO	24.27	20.85
CeO_2	7.19	21.73
ZnO	32.99	31.72
Al_2O_3	4.02	9.53
γ-Al_2O_3	5.81	21.15
ZrO_2	15.49	25.77
SBA-15	1.99	11.90
Ga_2O_3	1.28	10.36
TiO_2	1.93	9.20
SiO_2	2.42	12.04

	Mean	**SD**
calc.temp.	673.48	126.12
calc.time	3.71	1.51
red.temp.	563.71	49.52
red.time	5.21	3.32
red.H_2	54.87	46.88
rxn.temp.	518.80	33.43
rxn.pres.	3.27	1.42
rxn.GHSV	2960.05	2259.41
comp.H_2	73.75	7.99
comp.CO	8.48	11.47
comp.CO_2	15.01	8.97
comp.inert	2.76	4.19
X_{COx}	0.15	0.11
S_{MeOH}	0.64	0.29
Y_{MeOH}	0.10	0.11

For classification tasks, the responses had to be discretized into groups or classes, which represent the performance. For a proper classification, the algorithms require approx. equal sized classes. This precondition and the distribution of the response variables, shown in Figure 3.10, led to the discretization shown in Table 3.5.

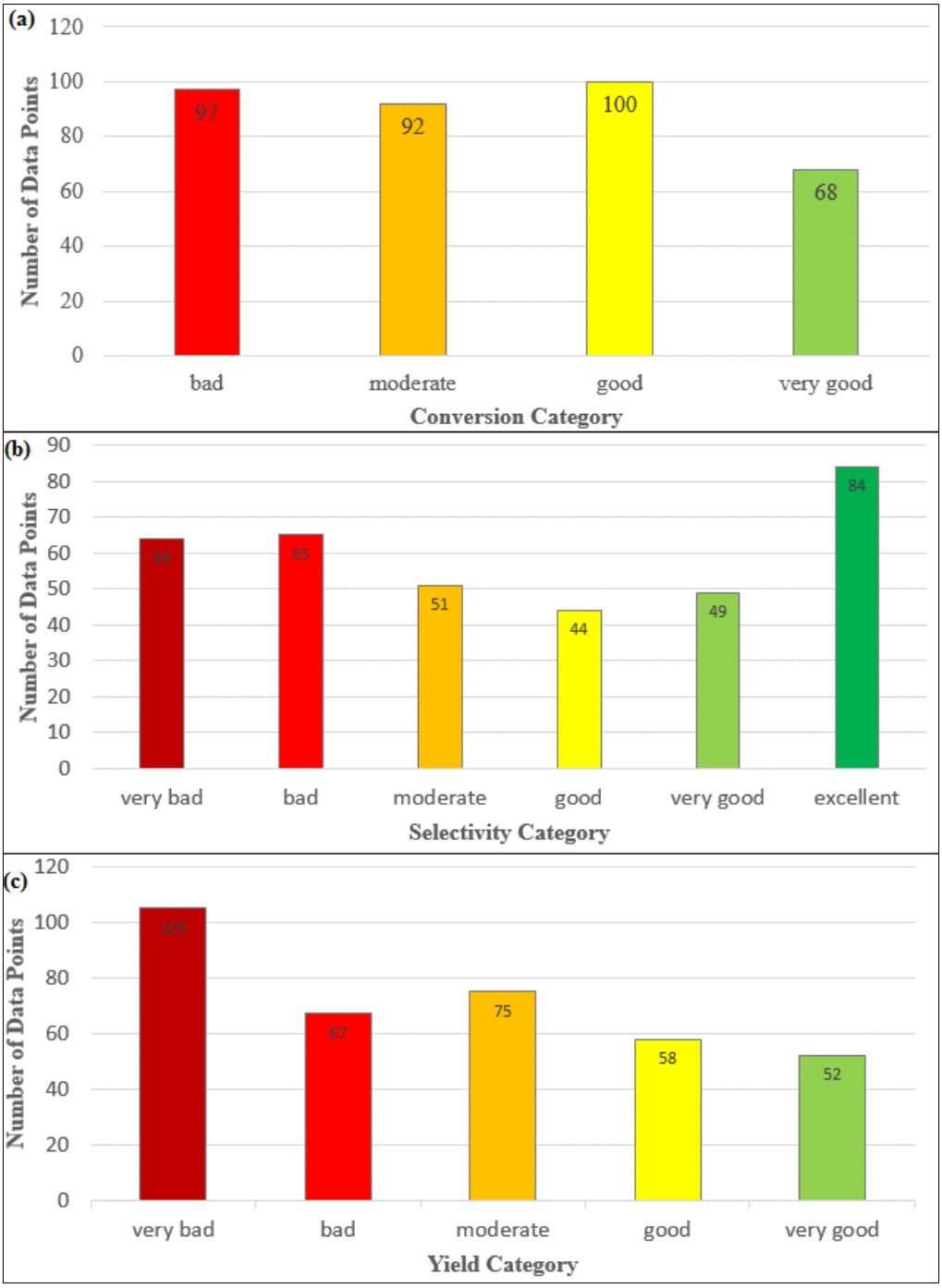

Figure 3.11. Discrete (a) conversion (b) selectivity (c) yield distribution.

The CO_X conversion was divided into four classes with sizes between 68 and 100 data points. The class "very good" had a higher range of values than the other classes but since the data points were less in number, it was the optimum choice. The methanol selectivity was separated into six classes with class sizes between 44 and 84. Here the highest class "excellent" had the most data points, due to the fact that a lot of experiments yielded a near

100 % MeOH selectivity. Yield was classified by five classes and had sizes between 52 and 105 data points. The cut at 2.5 % was an inconvenient choice, but needed to be done to keep the classes of approx. equivalent size. The three responses were discretized to variables with different amounts of individual values and therefore, were not convertible into each other. To be used properly, the response variables had to be assigned the right classes. They were treated as ordinal, categorical variables. Figure 3.11 shows a more comprehensive presentation of the response classes, their size and their relative size to each other. Slightly more equal distributed classes could be achieved, but it was decided to use convenient separation values and focus on the interpretability instead.

Table 3.5. Discretization of conversion, selectivity and yield.

		Class					
		Very Bad	**Bad**	**Moderate**	**Good**	**Very Good**	**Excellent**
CO_X Conversion	**Range**	-	0 - 5	5 - 15	15 - 25	> 25	-
	Data Points	0	97	92	100	68	0
MeOH Selectivity	**Range**	0 - 35	35 - 50	50 - 65	65 - 80	80 - 95	> 95
	Data Points	64	65	51	44	49	84
MeOH Yield	**Range**	0 - 2.5	2.5 - 5	5 - 10	10 - 20	> 20	-
	Data Points	105	67	75	58	52	0

3.2.1. Clustering

For the clustering tasks, the R packages “fpc”, “cluster” and the already included package “stats” were needed. First the dissimilarity matrix was computed with the “gower” attribute, then the catalyst composition variables and preparation method were clustered by two different unsupervised techniques into one multivalued categorical variable. This reduced dimensionality and therefore was beneficial for the small database. Cluster distributions were presented for both methods. The clustered databases were included in later modelling.

Partitioning around medoids and hierarchical clustering were applied to the catalyst preparation method and catalyst composition variables. The clustering serves the purpose of dimensionality reduction for a higher data points to variables ratio and therefore, a better performance of later applications on the database. No information extraction in form of catalyst distribution in each cluster was done.

3.2.1.1. Partitioning Around Medoids. The PAM algorithm found 13 clusters and therefore, 13 values of the new categorical variable to be optimum. This was done according to the equation 2.6. Figure 3.12 shows the silhouette plot, plotted by the PAM algorithm. It shows the 13 clusters, their number of data points and how close the data points within the clusters are, i.e. their silhouette width. The most similar points have a silhouette width of 1.00 and the least similar data points, one of 0.42. This sums up to an average silhouette width of 0.61, which was the highest among all tries with different numbers of clusters.

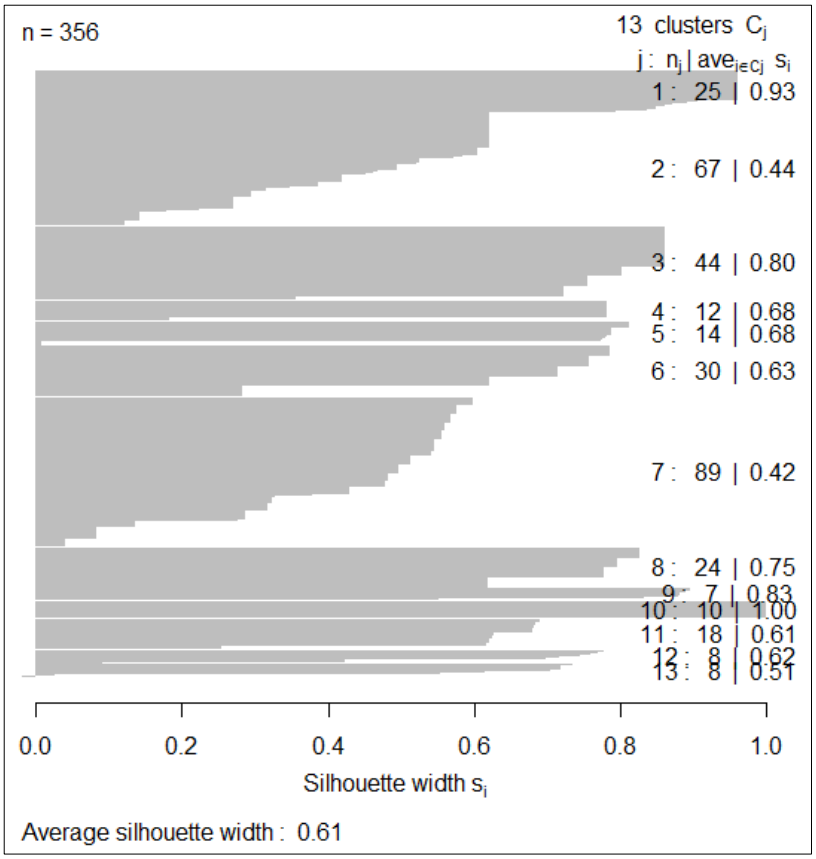

Figure 3.12. Silhouette plot of clusters.

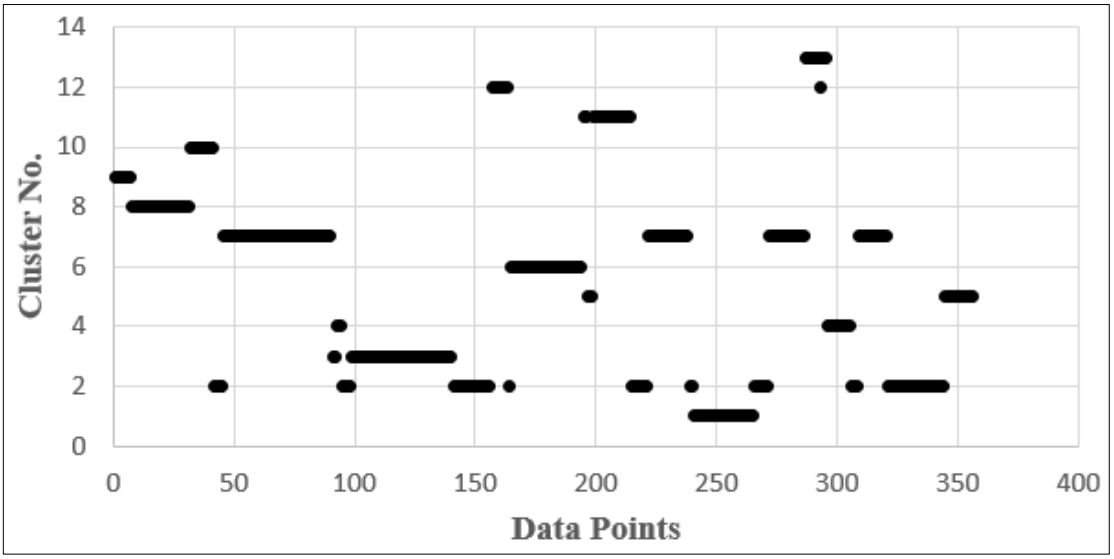

Figure 3.13. Cluster membership by PAM (13 cluster).

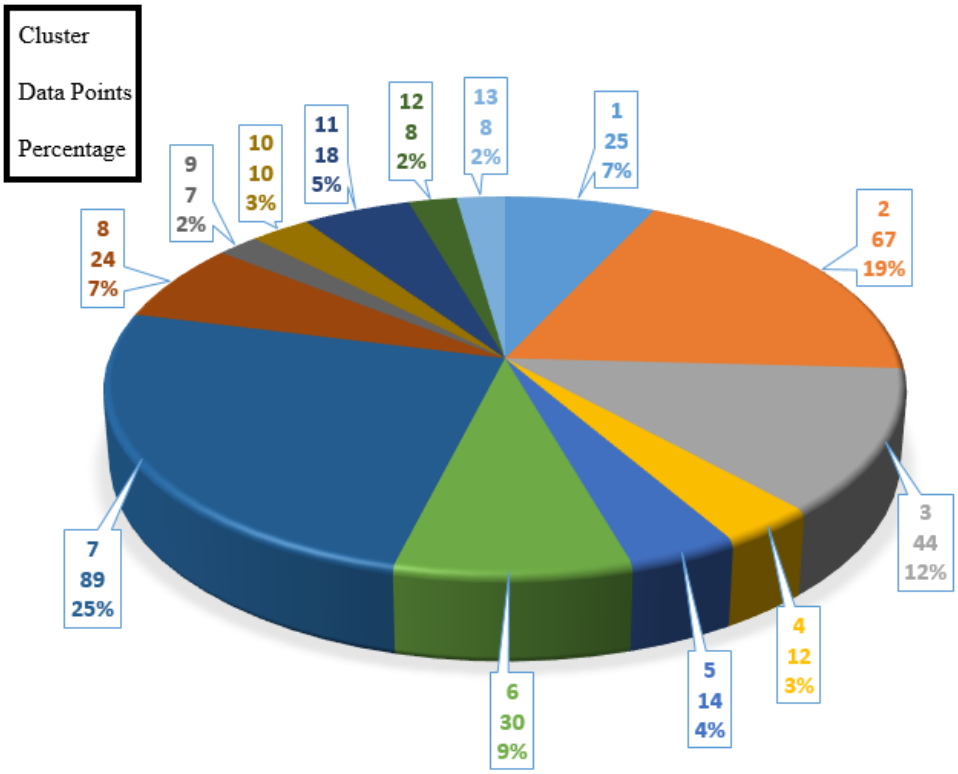

Figure 3.14. Cluster distribution by PAM.

Further visualization was done manually, using Microsoft Excel. Figure 3.13 shows the membership of each data point to one of the 13 clusters. It can be seen that data points next to each other, were assigned to the same cluster in most of the time. This indicates that

data points from the same articles were clustered together, since they follow each other in the database.

To get a more comprehensive overview of the cluster distribution, the clusters were plotted as a pie diagram in Figure 3.14. It shows the cluster name or rather the value of the categorical, unordered cluster variable, the number of data points in the clusters and the percentage of the total database. Cluster two, three and seven account for more than half of the database and are the biggest. Whereas the other clusters are of mediocre or small size, as small as 2 % of the database. Surprisingly, merging the smaller clusters, yields lower average silhouette widths, which means more inner cluster dissimilarities.

3.2.1.2. Hierarchical Clustering. Hierarchical clustering was applied according to the algorithm described in Chapter 2.3.1 and graphically presented as a dendogram in Figure 3.15. Cutting the dendogram at a height of 0.25 yielded the clusters with the highest inter cluster dissimilarities and led to six clusters. The decision was made manually by inspecting the dendogram. Further processing for graphical presentation was done with Microsoft Excel.

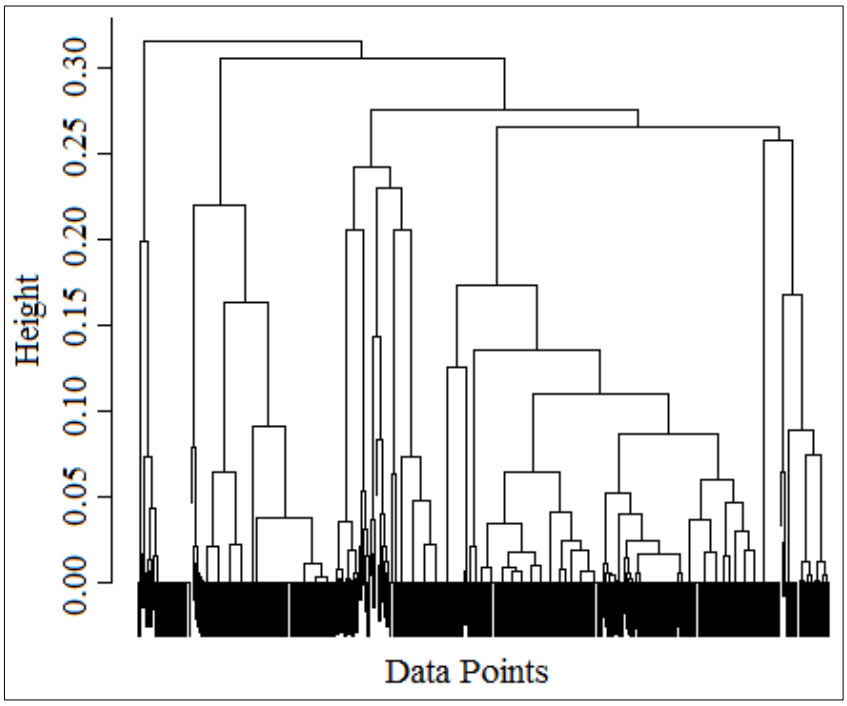

Figure 3.15. Cluster dendogram from hierarchical clustering.

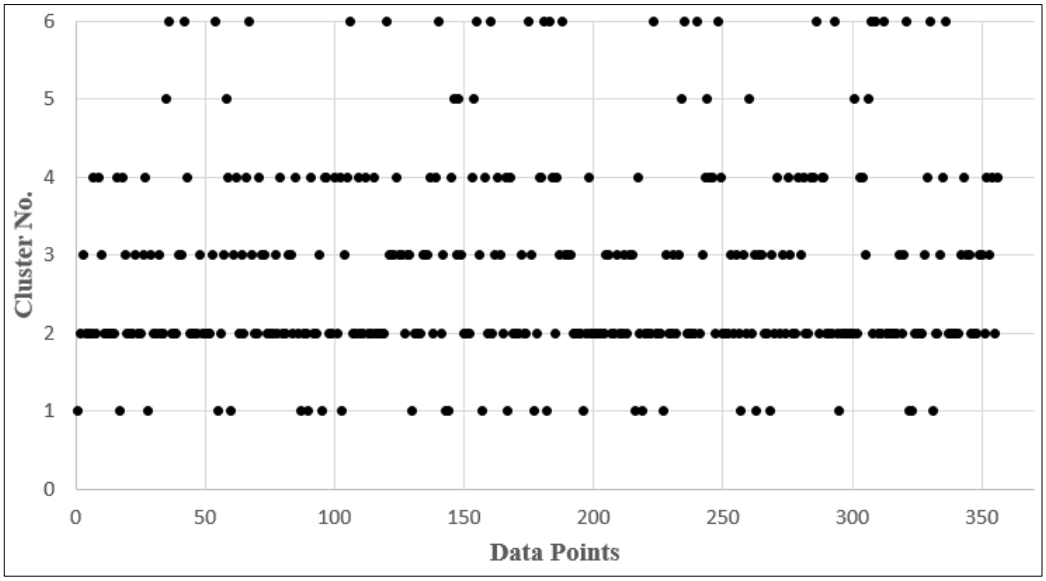

Figure 3.16. Cluster membership by hierarchical clustering (six cluster).

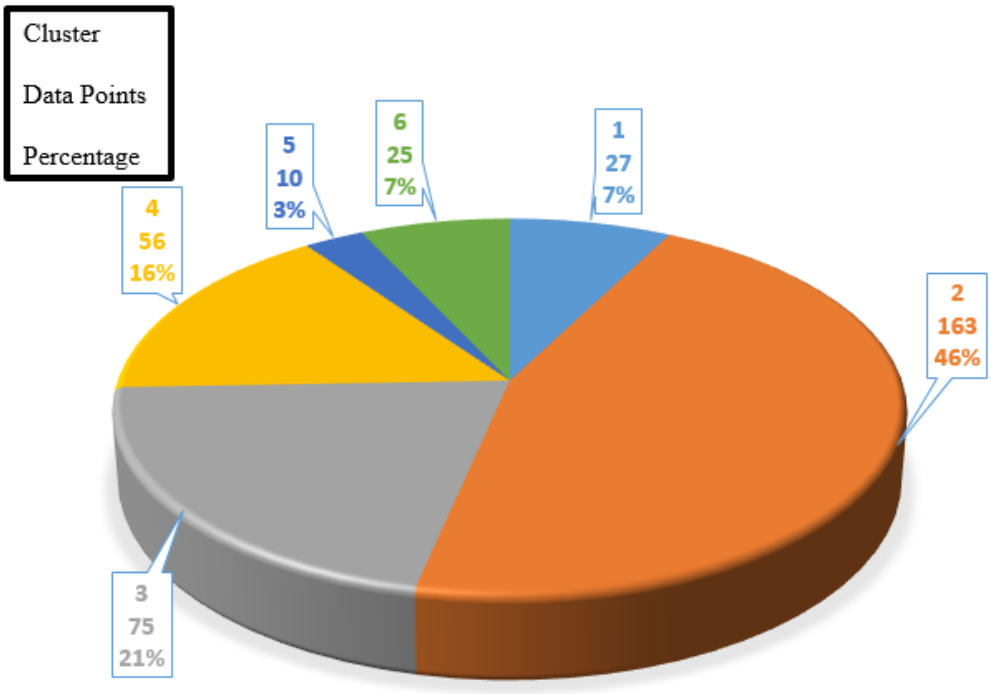

Figure 3.17. Cluster distribution by hierarchical clustering.

Figure 3.16 shows the cluster membership of the data points to one of the six clusters. Also some clusters occupy more data point than others it can be seen that the cluster sizes are more equally distributed the ones from the PAM clustering. Also adjacent data points do

not necessary belong to the same cluster and were more fragmented than the data points by PAM clustering. This indicates that data points from the same article were distributed among different clusters.

The cluster distribution is further visualized in Figure 3.17. Approximately the half of the data points were assigned to cluster two and the rest were distributed among the other clusters with just one very small cluster of 3 % of the database.

3.2.2. Multiple Linear Regression

The multiple linear regression model was built without additional packages. First the data points were divided by their article number and separated into five approx. equal sized sets with unique articles for a 5-fold cross validation. The separation by articles instead of data points, had to be done, to not spoil the prediction. This way it was ensured that the prediction was really done on previously unseen data. The model was built on the four training sets and the predictions were done on the test set. This was repeated for all folds with changing test and training sets. After that the quality parameters for fitted and predicted values were calculated manually and p-values, as indicators for variable significance, were extracted. Fitted vs. observed and predicted vs. observed responses were plotted. The whole procedure was repeated for every response variable.

After that, a combined stepwise selection of the input variables was done to build a reduced model. The whole procedure was repeated with this model and results were saved for comparison. Also the clustered models were treated the same way, the complete model was treated. The only difference is that the clustered databases were used and no stepwise selection was performed on these models. Representations of variable importance and model comparison were constructed manually with Microsoft Excel.

3.2.3. Decision Trees

To model the database with decision trees, the package “rpart” was chosen. Additionally, the packages “rpart.plot” and “rattle” were required to produce appealing tree plots. The procedure was similar to the one of the MLR algorithm. The data points were divided

by article numbers and separated into five approx. equal sized groups by unique articles for a 5-fold cross validation. The algorithm was applied on the training sets to build a regression tree model. The tuning parameter “minimum node size” and “minimum decrease of lack of fit” were set to default, because a pruning was done afterwards. During the pruning step, the algorithm automatically selects the complexity parameter associated with the smallest cross-validated error. Minimum node size was left as default because changing it to larger values, decreased the accuracy. After that, the responses of the test set were predicted. This was repeated for all test sets. Afterwards, quality parameters were calculated, the variable significance extracted and fitted vs. observed as well as predicted vs. observed values plotted. This was done for all three responses.

A reduced model, excluding all variables with zero variable importance, was built and the procedures repeated. Finally the clustered databases were modelled in the same way and the results were saved for comparison. Further interpretive graphs and overviews were computed with Microsoft Excel.

Classification trees were build the same way as the regression trees. The discretized database was used as input. The confusion matrix and the misclassification error were computed manually. The classification tree was plotted in a hierarchical way for representation.

Reduced and clustered models were modelled also with classification trees. All subsequent graphical representations and rule deductions were done manually and prepared with Microsoft Excel.

3.2.4. Random Forest

Random forest was applied from the “randonForest” R package. The optimum number of trees was determined previous to the actual modeling by a number of trees vs. error plot and was found to be 500. The best number of randomly chosen variables was found for every run individually. As well as the MLR, RT and CT algorithms, the random forest was validated by a 5-fold cross validation of sets with unique articles. First the model was built and

test responses predicted. Afterwards, quality parameters were computed, variable significance extracted and fitted vs. observed and predicted vs. observed responses plotted. According to the variable importance, a reduced model was build and tested in the same way.

Additionally clustered models were processed and partial dependencies of the most important variables were plotted. The partial dependence plots show the ranges of the input variables, which are beneficial for high response variables.

Finally the database was divided into subsets by logical decision making with respect to variable distributions in chapter 3.1 and by clustering techniques. The subsets were modeled independently by random forest to see if an increase in performance can be achieved.

4. RESULTS AND DISCUSSION

Methanol synthesis performance of various catalysts, in terms of CO_X conversion, methanol selectivity and methanol yield, were modelled by applying various techniques to the database constructed from literature (the complete, reduced, clustered or sub-setted database). The results of the best models were presented as plots of predicted vs. experimentally observed CO_X conversion, MeOH selectivity and MeOH yield. Multiple linear regression, regression tree and random forest were used for prediction; furthermore, random forest was used to predict on sub sets of the data base. Classification was also done by classification trees and the results were presented in form of confusion matrices. Empirical rules were deducted from the classification tree algorithm. Additionally importance analysis from every model was presented.

4.1. Results of Multiple Linear Regression (MLR)

Due to its simplicity MLR was applied first. Table 4.1 shows the detailed results for different MLR models with respect to CO_X conversion, MeOH selectivity and MeOH yield. The models with the best results are highlighted wıth an asterisk and presented as fitted (training) vs. experimentally observed values in Figure 4.1 and predicted (testing) vs. experimentally observed values in Figure 4.3. Since all R^2_{CV} were negative, this quantity was not included for comparison.

Table 4.1. Results of multiple linear regression models.

			Multiple Linear Regression			
			Complete	**Reduced**	**PAM**	**Hierarchical**
Fitted	**$R^2_{adj.}$**	**X_{COx}**	*0.7952	0.7924	0.7532	0.7279
		S_{MeOH}	*0.7952	0.7918	0.7787	0.7550
		Y_{MeOH}	*0.8177	0.8163	0.7952	0.7575
	RMSE	**X_{COx}**	*0.0493	0.0496	0.0532	0.0584
		S_{MeOH}	*0.1264	0.1275	0.1398	0.1421
		Y_{MeOH}	*0.0437	0.0439	0.0465	0.0518
Predicted	**PRMSE**	**X_{COx}**	1.6042	1.1019	0.2434	*0.2048
		S_{MeOH}	1.1100	0.7409	*0.4622	0.5190
		Y_{MeOH}	0.7626	0.5736	*0.1908	0.2399

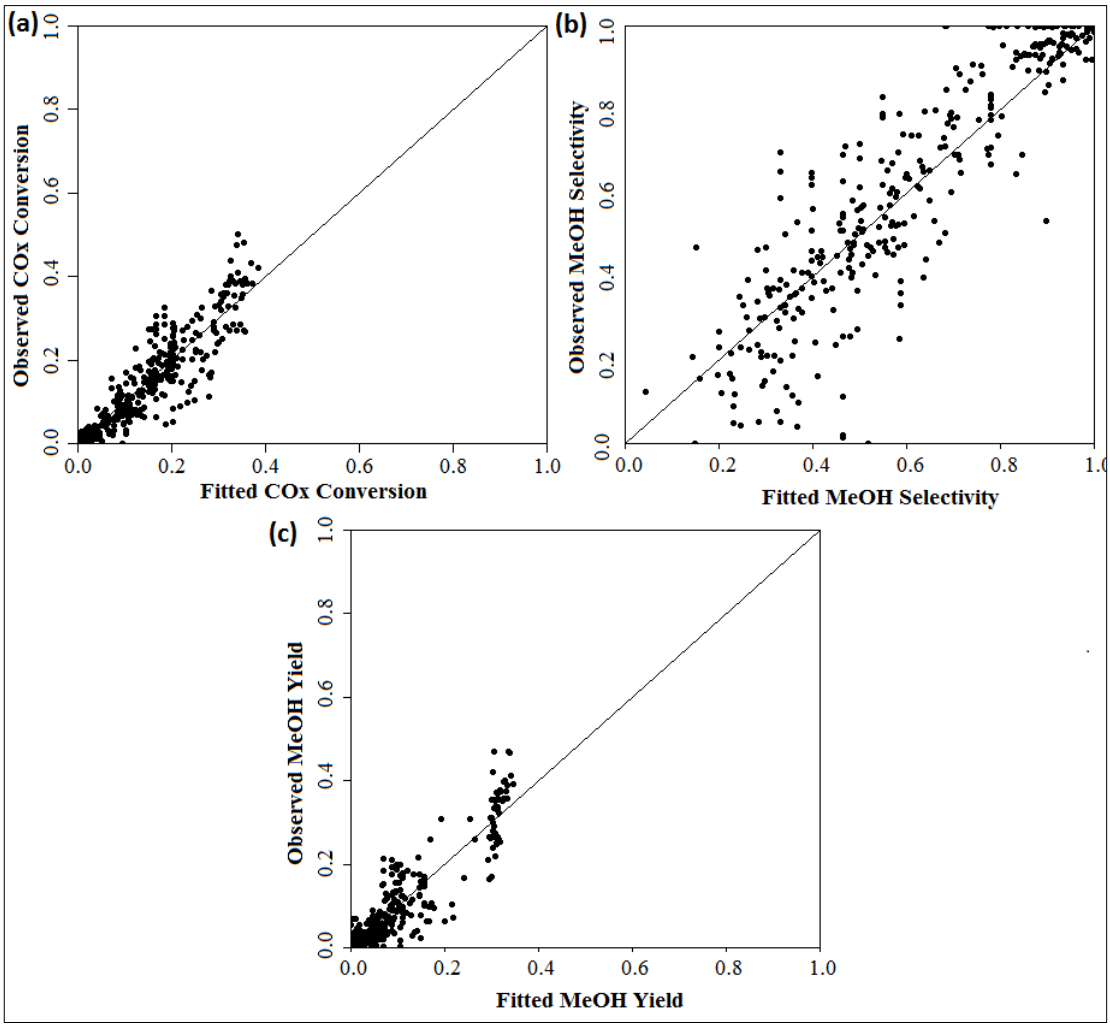

Figure 4.1. Fitted vs. observed (a) CO_X conversion (b) MeOH selectivity (c) MeOH yield for complete MLR model.

While the best results for the fitted values were found by the complete model, using all input variables, the best predicted values were found using the clustered data by PAM clustering. This agrees with the nature of linear regression, i.e. the more input variables are used to describe the model, the more accurate it gets. Since the complete model has the most variables, it produced the best results of $R^2_{adj.}$ of up to 0.8177 and RMSE as less as 0.0437 for methanol yield, which are slightly better than the results for CO_X conversion and MeOH selectivity. Despite the variable selection process, the reduced model yielded almost the same results, even slightly worse than the complete model. The clustered models also achieved lower accuracy in terms of goodness of fit.

In prediction, the clustered models were clearly superior to the unclustered ones. While the complete and reduced model achieved PRMSE's between 0.5736 and 1.6042 the clustered models achieved PRMSE's between 0.1908 and 0.5190. This can be explained by the fact that the prediction power of regression models reaches its maximum at a certain amount of variables and decreases if additional variables are added. These are called noise variables. Obviously in the case of the complete and reduced models, noise variables are contained. These are clustered away into one “catalyst” variable.

Observing the fitted values, it can be said that the goodness of fit was quite accurate for CO_X conversion and MeOH yield and less accurate for MeOH selectivity. High conversion and yield values were a little bit under fitted and high selectivity values were miss fitted. Here, it is important to understand that linear regression extrapolates values and therefore, can fit them above and below natural limits. All three responses had fitted values slightly smaller than zero and the selectivity had fitted values of up to 1.2, although that is not visible in the plots. The majority of the values were described well by the MLR models and therefore, were used for further analysis. P-values and hence, variable importance was also extracted by implementing analysis of variance. Figure 4.2 shows the p-values for each variable for each outcome.

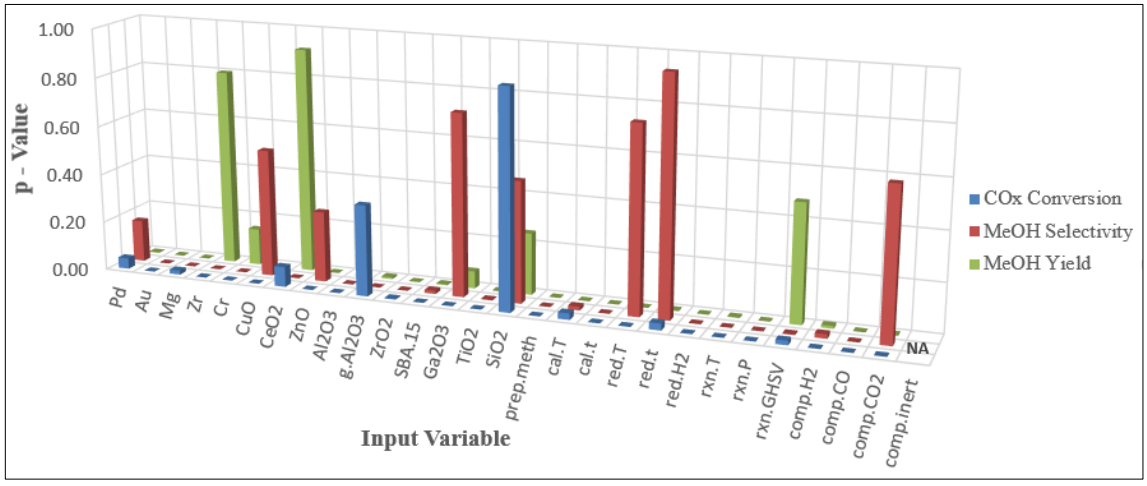

Figure 4.2. Significance (p-values) of input variables for MLR.

Taking a commonly used threshold p-value of lower than 0.05 for important variables, it can be deducted that all variables except CeO_2, γ-Al_2O_3 and SiO_2 are important for the CO_X conversion; this does not, unfortunately, allow to draw real conclusions. On the other

side, considering the fact that the unimportant variables for MeOH selectivity are Pd, CuO, ZnO, Ga_2O_3, SiO_2, reduction temperature, reduction time and composition CO_2, it can be deducted that, according to the MLR model, methanol selectivity is less sensitive to the catalyst composition, reduction conditions and CO/CO_2 ratio of the feed. It appears that the selectivity depends more on calcination and reaction conditions. The unimportant variables for the MeOH yield are found to be Zr, Cr, CeO_2, Ga_2O_3, SiO_2 and reaction GHSV. The p-value for the composition inert variable could not be calculated due to its correlation with the other feed composition variables.

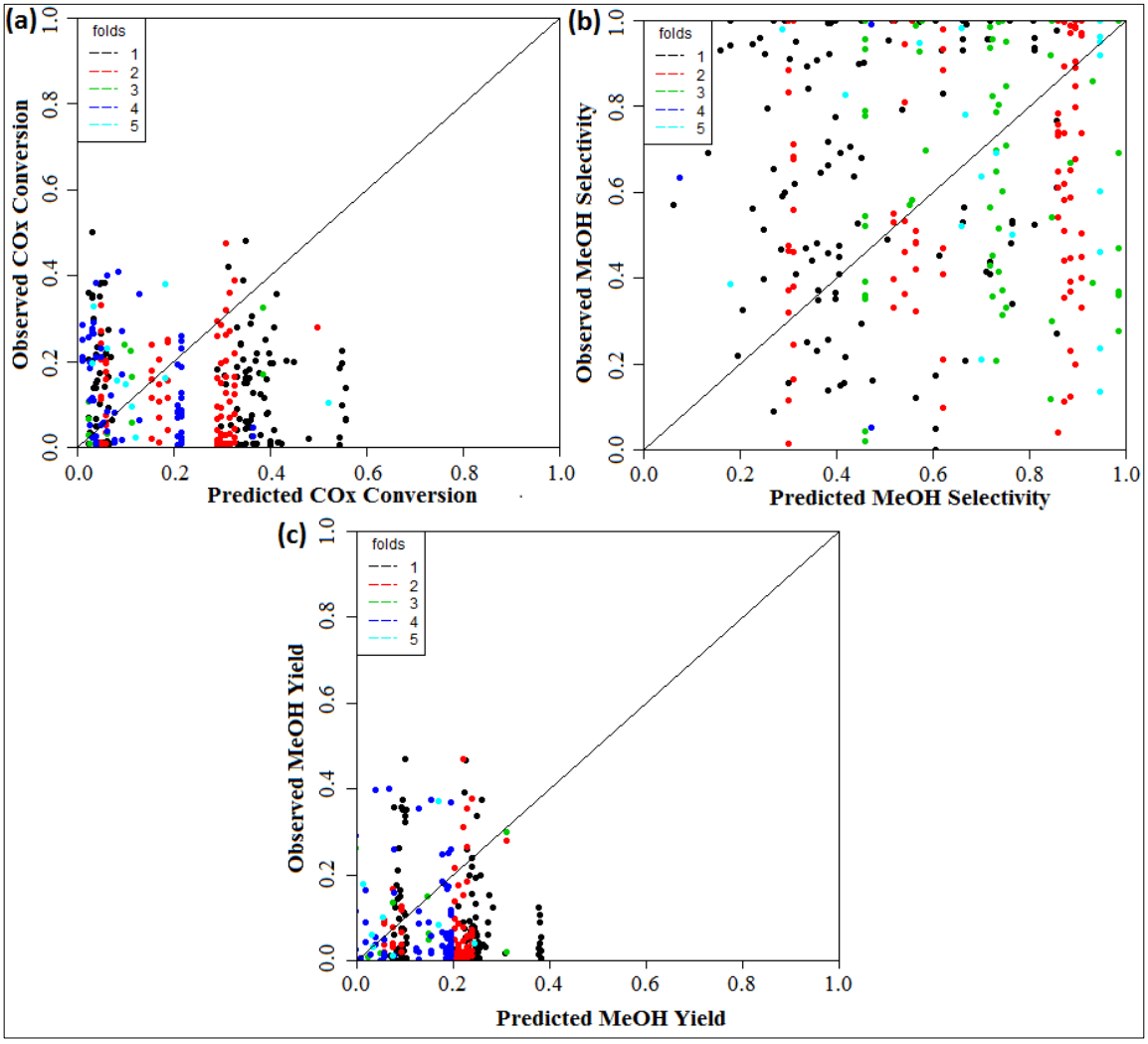

Figure 4.3. Predicted vs. observed (a) CO_X conversion (b) MeOH selectivity (c) MeOH yield for PAM clustered MLR model.

Despite the good descriptive power, the predictive power of the MLR models was very poor, with PRMSE's of larger than one. Although, the best model had PRMSE values not larger than 0.5, the predicted vs. observed plot in Figure 4.3 illustrates an almost random distribution. Different colors represent different folds of the cross validation. Negative values of up to -0.3 and values above one, of up to 1.5 were also predicted. Considering these circumstances, the MLR models cannot be used to predict future experiments. The only thing that can be deducted from this part is that clustering increased the predictive power significantly (conversion 8-fold, the selectivity approx. 2.5-fold and yield 4-fold, according to the decrease in PRMSE like shown in Table 4.1); otherwise, even the observation in the previous paragraph for the input significance is not conclusive due to the low reliability of the models.

4.2. Results of Decision Trees

Due to their good interpretability through hierarchical tree plots, empirical rules and low bias, the decision trees were used as regression trees for prediction and as classification trees for classification of the discretized database. It needs to be mentioned that tree based methods don't extrapolate, which has the positive effect that predicted responses comply with natural limits but has the drawback that the test results will be predicted within the range of the experimentally observed training set results only, even if the experimentally observed test results are above or below that range.

4.2.1. Results of Regression Tree (RT)

The regression tree algorithm was applied in the same manner as the MLR algorithm was applied, including reduced and clustered models. Table 4.2 shows the results in detail. The most successful models were highlighted with an asterisk and presented as fitted vs. observed and predicted vs. observed plots. As in the MLR models, the R^2_{CV} values were not included for comparison due to their very low or unrealistically negative values.

It can be observed that the best model describing the data is the reduced model even though it is just slightly better if compared with the complete model; to be more precise, the increase in accuracy in MeOH yield, is approx. 1 % according to the $R^2_{adj.}$ value. This is legitimate since decision tree is an empiric method and the model with the most descriptors

does not necessary has to be the one with the most accurate fitted results, like it is with MLR. The goodness of fit is slightly better for some models, if compared with MLR models. The best model for prediction differed for every response variable. The PAM clustered model was chosen as example model because it was best in predicting the methanol yield, which is considered an overall quality value because it includes the conversion and selectivity. All models were superior in prediction than the MLR models. Figure 4.4 shows the fitted vs. observed plot for the best regression tree model. Although, the quality increased in terms of $R^2_{adj.}$ and RMSE, the visual perception is contra intuitive due to the prediction of mean values instead of real values. This significantly reduces the interpretability of the regression tree, which was one of the main benefits of this method.

Table 4.2. Results of regression tree models.

			Regression Tree			
			Complete	**Reduced**	**PAM**	**Hierarchical**
Fitted	**$R^2_{adj.}$**	**X_{COx}**	0.7898	0.7898	*0.8244	0.8022
		S_{MeOH}	*0.8105	*0.8105	0.8019	0.7939
		Y_{MeOH}	0.8275	*0.8377	0.7912	0.8234
	RMSE	**X_{COx}**	0.0521	0.0521	*0.0477	0.0505
		S_{MeOH}	*0.1270	*0.1270	0.1304	0.1322
		Y_{MeOH}	0.0444	*0.0430	0.0490	0.0449
Predicted	**PRMSE**	**X_{COx}**	0.1521	0.1613	0.1470	*0.1379
		S_{MeOH}	0.3954	*0.3834	0.4100	0.3871
		Y_{MeOH}	0.1461	0.1478	*0.1401	0.1441

Figure 4.5 shows the variable importance, computed from the increase in node purity and scaled to a sum of 100 %. To build the reduced model, all variables with zero importance were excluded. In general, the plot shows that the important variables are the catalyst preparation conditions and the feed composition, whereas the catalyst composition is less important in general, albeit more important for CO_X conversion than for MeOH selectivity and yield. It can be concluded that the conversion is depending more on the catalyst composition than that is on the selectivity and yield. Especially, the preparation method and conditions for calcination and reduction seems to be important. This agrees with the literature, since these variables are all responsible for the change in the base metal size and dispersion, which in turn contributes to a higher conversion. The RT models agree with the MLR models on the issue that the reduction is less important for the selectivity, but disagrees on the issue of

the feed composition, which plays a key role according to the RT model. Looking to the overall yield, the most important variables are the reduction parameters, followed by the feed composition.

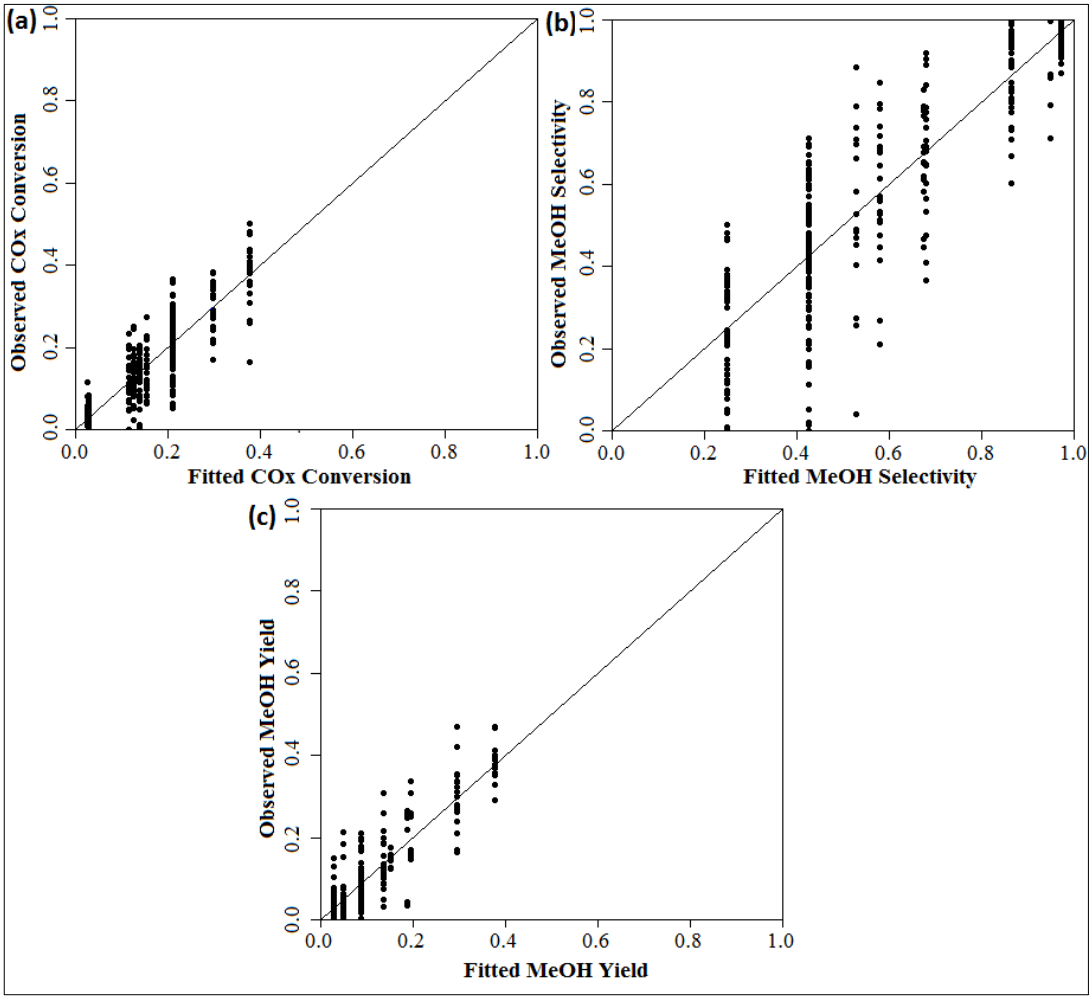

Figure 4.4. Fitted vs. observed (a) CO_X conversion (b) MeOH selectivity (c) MeOH yield for reduced RT model.

In prediction, the regression tree algorithm showed poor results. Figure 4.6 shows the predicted vs. experimentally observed values for CO_X conversion, MeOH selectivity and MeOH yield. Additional inconvenience is that the regression tree predicts only the mean values, consequently some values were strongly miss predicted. Most of the time, each fold was predicted by two or three mean values, which could mean that every paper was predicted by the same value. That in turn shows that the regression tree cannot successfully predict future experiments in methanol synthesis on the basis of this data base.

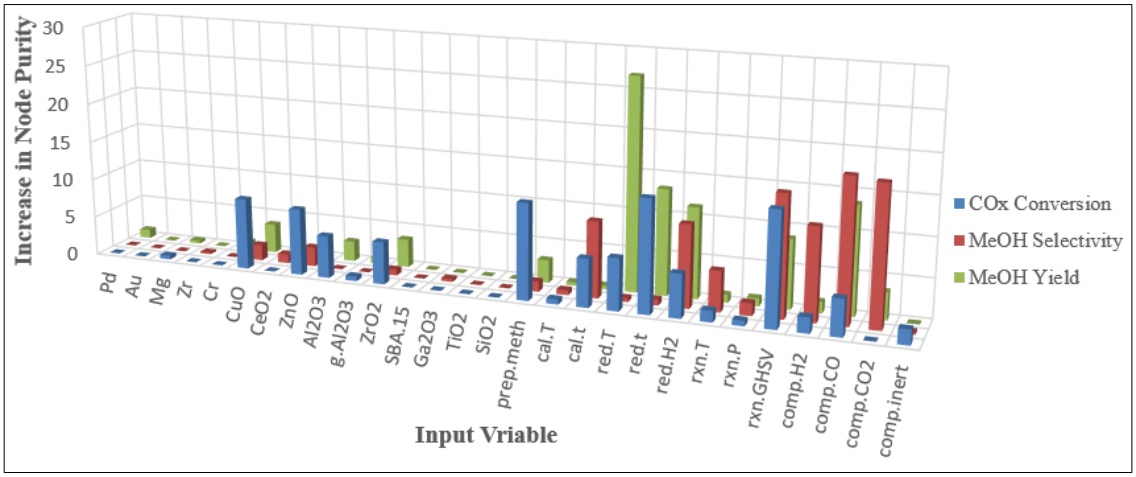

Figure 4.5. Variable importance by increase in node purity of input variables for RT (scaled to 100 %).

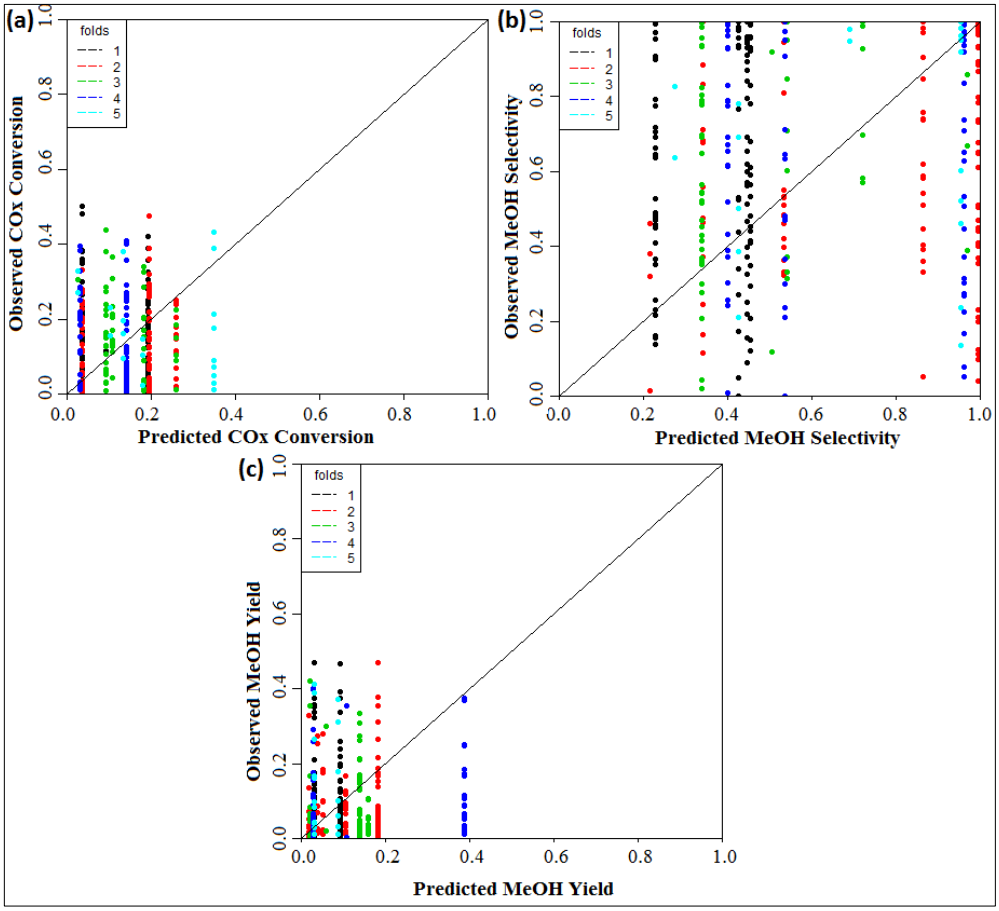

Figure 4.6. Predicted vs. observed (a) CO_X conversion (b) MeOH selectivity (c) MeOH yield for PAM clustered RT model.

4.2.2. Results of Classification Tree (CT)

A classification approach, using classification tree, was also utilized with the discretized data base. As the quality measures, the confusion matrices for fitted vs. observed and predicted vs. observed outcomes were computed, and the missqualification error was calculated. This was also done for both the reduced and clustered databases. On the basis of the fitted values, comprehensive empirical rules for classification were tried to be deducted.

Table 4.3. Results of classification tree models.

			Classification Tree			
			Complete	**Reduced**	**PAM**	**Hierarchical**
miss classification error	**Fitted**	X_{COx}	0.2297	0.2213	0.2605	*0.2045
		S_{MeOH}	*0.3053	0.3221	0.3277	0.3529
		Y_{MeOH}	0.2633	*0.2549	0.3025	0.3053
	Predicted	X_{COx}	0.7339	0.7871	*0.7255	0.7267
		S_{MeOH}	0.8319	*0.8039	0.8067	0.8067
		Y_{MeOH}	*0.7731	0.8151	0.8095	0.8088

Table 4.3 shows an overview of the classification tree results. The best models are highlighted with an asterisk. It can be observed from the fitted values that different models are better for different dependent variables. The reduced model was chosen for presentation due to the reason that it achieved best results for the yield, which is considered the overall qualitative quantity because it can be computed from the other two quantities. In prediction, on the other hand, the complete model was chosen as best model due to the same reason.

Figure 4.7 shows the confusion matrices for the reduced model. It can be observed that all three variables are fitted very accurately; the most of miss fitted instants are placed to be adjacent classes. The conversion was fitted most accurately with a misclassification error of just 22.13 % and the MeOH selectivity was fitted the worst with a misclassification error of 32.21 %, which may be also considered as acceptable. Due to the good accuracy of the fitted values, a variable significance analysis was conducted using an intrinsic function with respect to increase in node purity. Also a visual representation is presented and classification rules deducted.

(a)

CO_X Conversion		Observed			
		very good	good	moderate	bad
Fitted	very good	**63**	22	8	0
	good	2	**70**	17	3
	moderate	3	8	**57**	6
	bad	0	0	10	**88**
Missclassification Error: 0.2213					

(b)

MeOH Selectivity		Observed					
		excellent	very good	good	moderate	bad	very bad
Fitted	excellent	**72**	2	2	0	0	0
	very good	12	**42**	9	4	2	0
	good	0	1	**13**	2	0	0
	moderate	0	3	13	**32**	8	9
	bad	0	0	0	10	**36**	8
	very bad	0	1	5	3	19	**47**
Missclassification Error: 0.3221							

(c)

MeOH Yield		Observed				
		very good	good	moderate	bad	very bad
Fitted	very good	**43**	2	0	2	0
	good	5	**50**	16	3	5
	moderate	0	3	**48**	10	2
	bad	2	3	11	**41**	14
	very bad	2	0	0	11	**84**
Missclassification Error: 0.2549						

Figure 4.7. Confusion matrices for fitted vs. observed (a) CO_X conversion (b) MeOH selectivity (c) MeOH yield for reduced CT model.

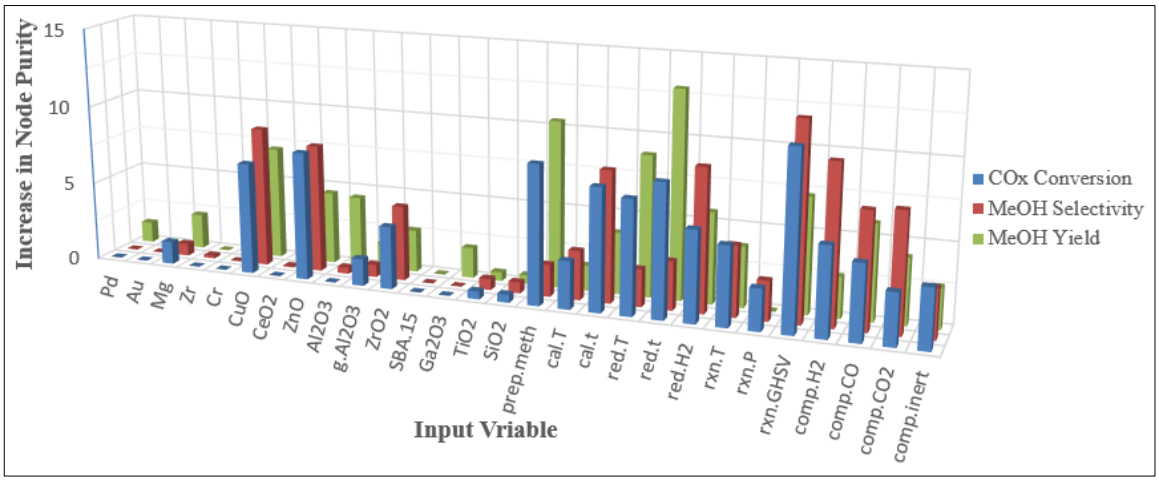

Figure 4.8. Variable importance by increase in node purity of input variables for CT (scaled to 100 %).

In Figure 4.8 the variable importance, computed by the classification tree algorithm is shown. It is comparable to the variable significance computed by the regression tree algorithm with respect to the importance focus on the preparation conditions, like preparation method, calcination and reduction. It is different with respect to the form of the distribution. The CT distributes the variable importance more flat and widely among the input variables.

It shows all the catalyst materials as important which are regarded as important in the literature, namely Cu, ZnO, Al_2O_3 and ZrO_2. Furthermore, the preparation method has one of the highest significances, especially for the conversion, which is due to the influence of the preparation method on the Cu particle size and dispersion. Furthermore, it appears that calcination and reduction is more important for CO_X conversion than MeOH selectivity and the feed composition more important for the selectivity than for the conversion. GHSV is the most important variable for conversion and selectivity but less important for the yield. This might be due to the reason that conversion and selectivity are of an anti-proportional nature.

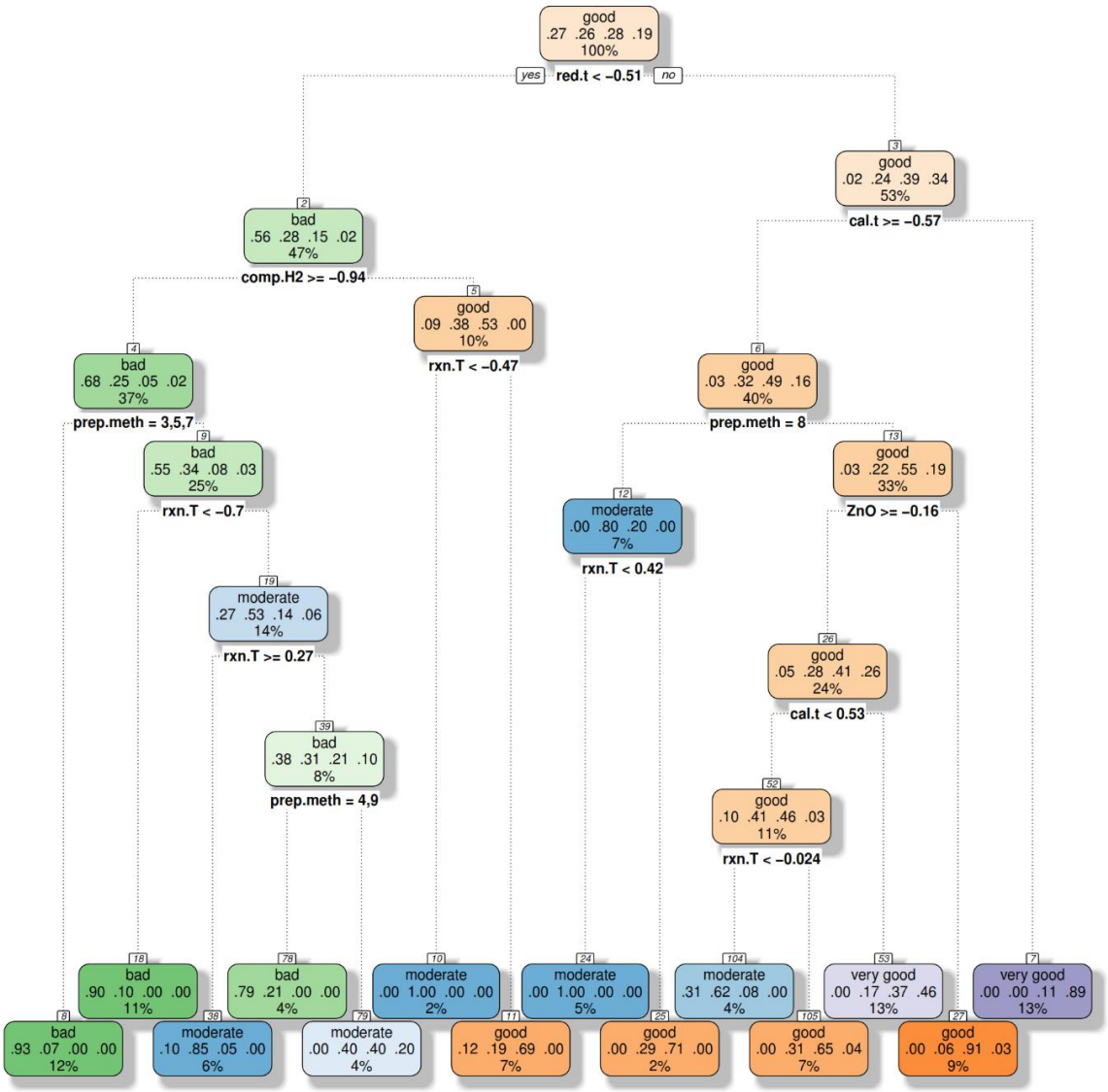

Figure 4.9. Standardized classification tree for CO_X conversion.

A graphical classification tree representation is shown in Figure 4.9. The nodes show the class which contains the most data points, the percentage distribution among the classes and the % of the total data in that node. In addition, the nodes are gradient color coded with respect to the classes and class probabilities. Under the nodes, the splitting criteria for a splitting variable is written. The variables are written as standardized quantities. If a criteria is met, the tree proceeds to the left; otherwise it continues to the right. Although, the conversion was divided into four classes, there are 14 leaf nodes. This is due to the fact that classes can be classified by following different routes.

For an easier understanding and application, decision rules were extracted from the tree and presented in Figure 4.10. The standardized values were back-transformed to the real values to gain some interpretability. Next to the class, the class probability also is presented, to see how certain each classification is. The class probabilities for the chosen class differ between 40 – 100 %. To improve this, smaller leaf nodes and more complex trees would be required. Since some leaf nodes are already as small as 2 % of the data base, the presented solution is a good trade-off. The percentage of data points belonging to each leaf is presented next to the class probabilities.

COx Conversion							Class	Class Probability (%)				Data (%)
								bad	moderate	good	very good	
reduction time < 3.5 h	composition H2 >= 66 %	preparation method = 3,5,7					bad	93	7	0	0	12
		preparation method = 1,2,4,6,8,9	reaction temperature < 495 K				bad	90	10	0	0	11
			reaction temperature >= 495 K	reaction temperature >= 528 K			moderate	10	85	5	0	6
				reaction temperature < 528 K	reaction	preparation method = 4,9	bad	79	21	0	0	4
						preparation method = 1,2,6,8	moderate	0	40	40	20	4
	composition H2 < 66 %	reaction temperature < 503 K					moderate	0	100	0	0	2
		reaction temperature >= 503 K					good	12	19	69	0	7
reduction time >= 3.5 h	calcination time >= 2.8 h	preparation method = 8	reaction temperature < 533 K				moderate	0	100	0	0	5
			reaction temperature >= 533 K				good	0	29	71	0	2
		preparation method = 1,2,3,4,5,6,7,9	ZnO >= 28 %	calcination temperature < 740 K	calcination	reaction temperature < 518 K	moderate	31	62	8	0	4
						reaction temperature >= 518 K	good	0	31	65	4	7
				calcination temperature >= 740 K			very good	0	17	37	46	13
			ZnO < 28 %				good	0	6	91	3	9
	calcination time < 2.8 h						very good	0	0	11	89	13

Figure 4.10. Tabulated decision rules for CO_X conversion.

Reduction time, H_2 % in feed gas and calcination time were chosen as the variables which affect the tree the most because a change in higher nodes variables affect the bigger portion of the data. This is in agreement with the importance analysis because these variables also achieved a high importance there. The most often used variable was reaction temperature. This variable was used to make decisions between adjacent classes and hence, can be

used to improve CO_X conversion in a certain degree even if the precedent condition were semi optimal.

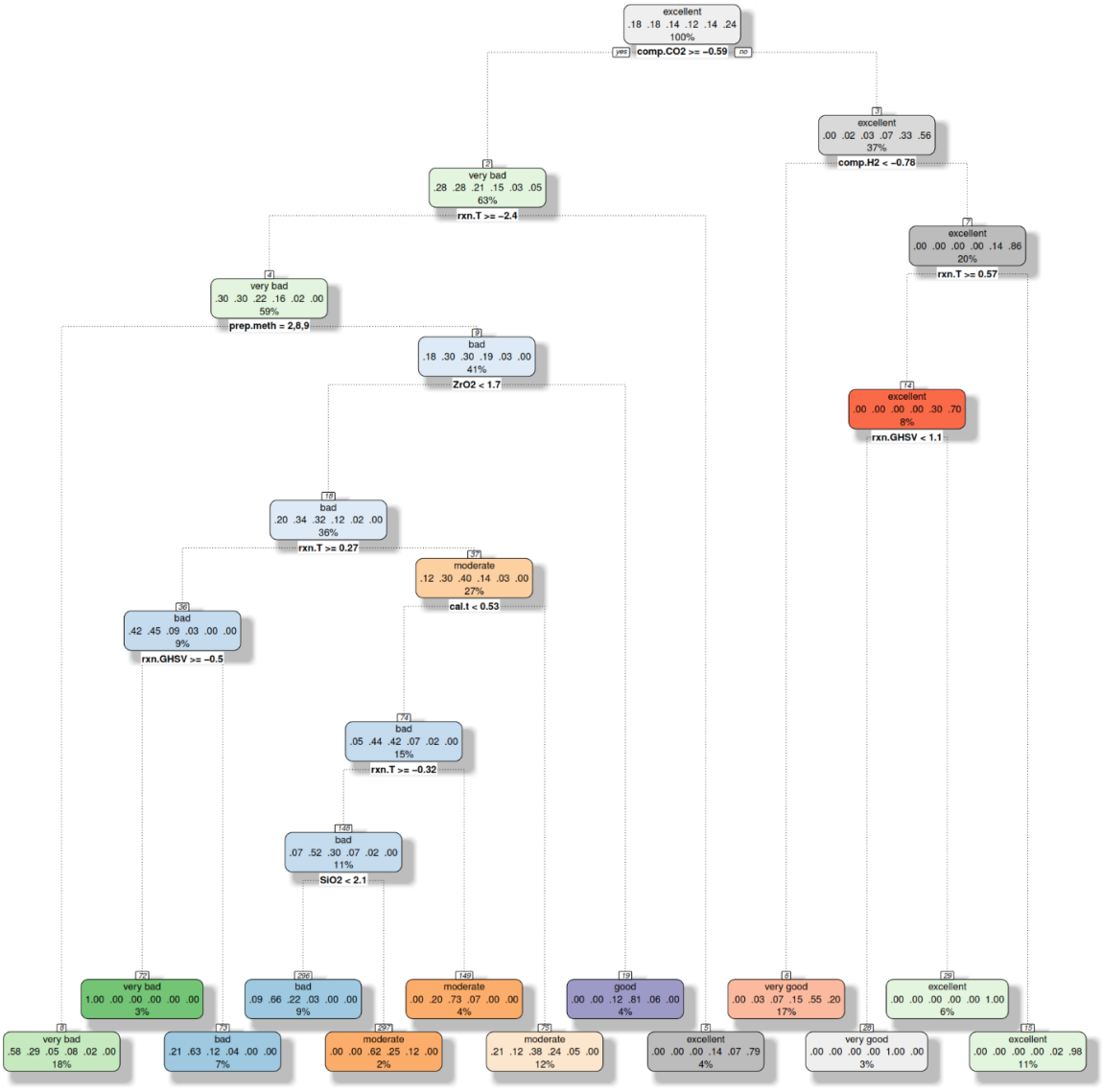

Figure 4.11. Standardized classification tree for MeOH selectivity.

Figure 4.11 shows the classification tree for methanol selectivity. It is built in the same way like the CO_X conversion tree. A gradient color code indicates the class and its probability. The selectivity range was divided into six categories and has 13 leaf nodes. For a more transparent presentation, rules were extracted and presented in Figure 4.12.

MeOH Selectivity								Class	Class probability (%)						Data (%)
									very bad	bad	moderate	good	very good	excellent	
composition CO2 >= 10 %	reaction temperature >= 438 K	preparation method = 2,8,9						very bad	58	29	5	8	2	0	18
		preparation method = 1,3,4,5,6,7	ZrO2 < 59 %	reaction temperature >= 528 K	reaction GHSV >= 1830 mL / gcat*h			very bad	100	0	0	0	0	0	3
					reaction GHSV < 1830 mL / gcat*h			bad	21	63	12	4	0	0	7
				reaction temperature < 528 K	calcination temperature < 740 K	reaction temperature >= 508 K	SiO2 < 28 %	bad	9	66	22	3	0	0	9
							SiO2 >= 28 %	moderate	0	0	62	25	12	0	2
						reaction temperature < 508 K		moderate	0	20	73	7	0	0	4
					calcination temperature >= 740 K			moderate	21	12	38	24	5	0	12
			ZrO2 >= 59 %					good	0	0	12	81	6	0	4
	reaction temperature < 438 K							excellent	0	0	0	14	7	79	4
composition CO2 < 10 %	composition H2 < 67 %							very good	0	3	7	15	55	20	17
	composition H2 >= 67 %	reaction temperature >= 538 K	reaction GHSV < 5445 mL / gat*h					very good	0	0	0	0	100	0	3
			reaction GHSV >= 5445 mL / gcat*h					excellent	0	0	0	0	0	100	6
		reaction temperature < 538 K						excellent	0	0	0	0	2	98	11

Figure 4.12. Tabulated decision rules for MeOH selectivity.

The data points were classified with certainties between 38 and 100 %. In general, the classes: good, very good and excellent were better classified than the classes of lower selectivity; this is a positive result considering that a more accurate determination of the conditions leading to high selectivity is more important. It is obvious that the CO_2 composition of the feed stream determines whether selectivity will be high or low; a content of less than 10 % CO_2 in the stream lead to higher MeOH selectivity. Decisions in the upper half of the spectrum are made by feed composition and reaction conditions, while decisions in the lower half are made by catalyst related variables like preparation, composition and calcination, together with reaction conditions. Although, the rules for MeOH selectivity classification are more complex than the ones for CO_X conversion classification, the selectivity rules present a more transparent structure.

The decision tree, shown in Figure 4.13, presents the classification of methanol yield into five classes and has twelve leaves. It follow the same structure as the previously presented trees and the values of the variables are in the standardized form. It can be observed that the first split separates the very good results from the rest, therefore reduction temperature can be regarded the main criteria for high methanol yield.

Looking at the deducted rules for methanol yield in Figure 4.14, it can be seen that the reduction time separates the very good from the others, and that the catalyst preparation method is the other important variable to be adjusted to get good results. The catalyst composition, especially the ratios of the Cu and Zn, are important to distinct the good and bad or moderate results. The classes were found with certainties between 43 and 100 %. The conditions for the optimum yield don't coincide with the conditions for the optimum CO_X conversion or methanol selectivity. This can be explained by the fact that a high yield can be

reached either by a high selectivity, a high conversion or moderate values of both of them; therefore, depending on the desired target variable, various sets of rules may be applicable.

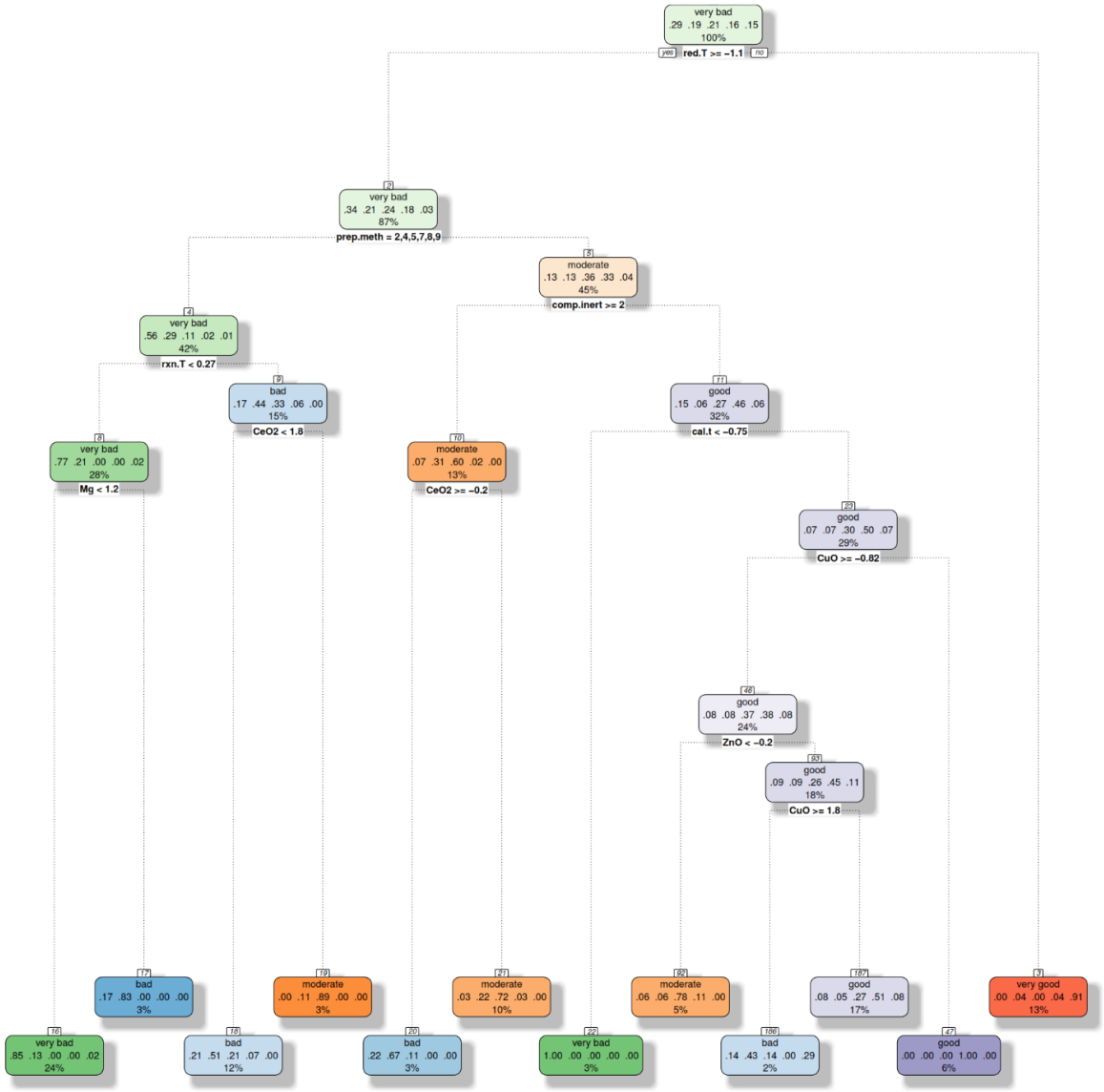

Figure 4.13. Standardized classification tree for MeOH yield.

The rules for conversion, selectivity and yield should be used as a first estimation and not as an absolute classification, since they are deducted from fitted and not from predicted values. The prediction power of the classification tree is shown in Figure 4.15. The misclassification errors are between 73 and 83 %. The misclassification for all three responses takes place between all classes and not just the adjacent ones, which could be considered with a

less weighted error. Therefore it can be concluded that the classification tree could not capture the important relations for prediction of new unseen data and should only be used for exploratory purposes.

MeOH Yield								Class	Class Probability (%)					Data (%)
									very bad	bad	moderate	good	very good	
reduction temperature >= 509 K	preparation method = 2,4,5,7,8,9	reaction temperature < 528 K	Mg < 0.08 %					very bad	85	13	0	0	2	24
			Mg >= 0.08 %					bad	17	83	0	0	0	3
		reaction temperature >= 528 K	CeO2 < 46 %					bad	21	51	21	7	0	12
			CeO2 >= 46 %					moderate	0	11	89	0	0	3
	preparation method = 1,3,6	composition inert >= 11 %	CeO2 >= 2.8 %					bad	22	67	11	0	0	3
			CeO2 < 2.8 %					moderte	3	22	72	3	0	10
		composition inert < 11 %	calcination time < 2.6 h					very bad	100	0	0	0	0	3
			calcination time >= 2.6 h	CuO >= 7.2 %	ZnO < 27 %			moderate	6	6	78	11	0	5
					ZnO >= 27	CuO >= 62 %		bad	14	43	14	0	29	2
						CuO < 62 %		good	8	5	27	51	8	17
				CuO < 7.2 %				good	0	0	0	100	0	6
reduction temperature < 509 K								very good	0	4	0	4	91	13

Figure 4.14. Tabulated decision rules for MeOH yield.

(a)

CO_X Conversion		Observed			
		very good	good	moderate	bad
Predicted	very good	**5**	13	9	9
	good	34	**56**	53	58
	moderate	11	16	**17**	13
	bad	18	15	13	**17**
		Missclassification Error:			0.7339

(b)

MeOH Selectivity		Observed					
		excellent	very good	good	moderate	bad	very bad
Predicted	excellent	**15**	7	12	10	12	12
	very good	16	**13**	8	9	13	13
	good	1	1	**0**	0	0	0
	moderate	38	12	16	**16**	21	26
	bad	9	9	5	8	**9**	6
	very bad	5	7	3	8	10	7
		Missclassification Error:					0.8319

(c)

MeOH Yield		Observed				
		very good	good	moderate	bad	very bad
Predicted	very good	**4**	5	6	2	9
	good	0	**3**	4	4	6
	moderate	28	33	**42**	36	62
	bad	12	13	14	**20**	16
	very bad	8	4	9	5	**12**
		Missclassification Error:				0.7731

Figure 4.15. Confusion matrices for predicted vs. observed (a) CO_X conversion (b) MeOH selectivity (c) MeOH yield for complete CT model.

An attempt was done to improve classification by discretizing all responses into four classes only. In general smaller number of classes improve the classification. To do the discretization, each continuous response variable was divided into four approx. equal sized classes. That had the effect that the classes became unnaturally in range. For example the yield was divided into the classes bad (0 – 2.1 %), moderate (2.2 – 5.5 %), good (5.6 – 13 %)

and very good (> 13 %). This had the effect that everything between 13 and 100 % was considered as very good and no further subgrouping was done to differentiate inside that range. It was proceeded similarly with conversion and selectivity. The results of these new discretized database increase for some responses and decreased for others in terms of goodness of fit; but the changes were insignificantly small, i.e. less than 5 %. In prediction, the misclassification errors could be improved by two, three and 5 %, depending on the response. Overall, the misclassification errors for prediction stayed between 70 and 75 %. Since the improvement was not large, but the loss of interpretability was; the tradeoff was decided to have a contra-productive effect.

4.3. Results of Random Forest (RF)

Due to their known accuracy, i.e. the same bias as a decision tree but less variance, random forest models were applied to the database. All complete, reduced and clustered datasets were used for modelling. The optimum number of random variables was chosen at every run independently, like explained in chapter 3.2.4. The number of trees was chosen to be 500. This guaranteed a minimum RMSE with relatively short computation time and enough trees for a large basis for knowledge extraction. After the acquisition of the goodness of fit and prediction power, the database was divided into subsets and these subsets were modeled independently to see if the prediction power could be improved.

4.3.1. Results of Complete Database

Table 4.4 shows the results of different random forest models. The reduced model, shows the best goodness of fit, with $R^2_{adj.}$ of up to 0.95 and RMSE of 0.0224, albeit they are only slightly better than that from the complete model. The selectivity was fitted with the least accuracy and the clustered models were inferior to the not-clustered ones. The best model for prediction, was the model clustered by PAM. The PRMSE's were the best compared with other methods and lay in the range of 0.1176 and 0.3558. Since R^2_{CV} were negative and therefore meaningless for comparison, they were excluded from the results.

The fitted vs. observed outcomes in Figure 4.16 show a very accurate goodness of fit for all three responses, although, the MeOH selectivity is again fitted weaker than the CO_X

conversion and MeOH yield. All three variables show similar grouped deviations around some lower values; this could indicate a research paper, which could not be fitted well by the model. Each plot displays the number of random variables, used to construct the forest. While the optimum number for conversion was found to be 15, the optimum number for selectivity and yield was 10. In general, the random forest fitted the conversion, selectivity and yield most accurate, compared to the other models and hence, the RF variable importance analysis should be paid the most attention to. The variable importance by increase in node purity is presented in Figure 4.17 with values scaled to a total of 100 %.

Table 4.4. Results of random forest models.

			Random Forest			
			Complete	**Reduced**	**PAM**	**Hierarchical**
Fitted	$\mathbf{R^2_{adj.}}$	$\mathbf{X_{COx}}$	0.9454	*0.9482	0.8936	0.8902
		$\mathbf{S_{MeOH}}$	*0.9437	0.9435	0.9063	*0.9437
		$\mathbf{Y_{MeOH}}$	0.9496	*0.9549	0.9022	0.8997
	RMSE	$\mathbf{X_{COx}}$	0.0257	*0.0255	0.0367	0.0373
		$\mathbf{S_{MeOH}}$	*0.0670	0.0684	0.0884	0.0972
		$\mathbf{Y_{MeOH}}$	0.0232	*0.0224	0.0331	0.0335
Predicted	**PRMSE**	$\mathbf{X_{COx}}$	0.1315	0.1424	0.1313	*0.1271
		$\mathbf{S_{MeOH}}$	0.3520	0.3558	*0.3416	0.3533
		$\mathbf{Y_{MeOH}}$	0.1304	0.1377	*0.1176	0.1211

The important variables for conversion are reduction conditions, calcination time, reaction temperature and catalyst preparation method. While the effects of preparation method, reduction and calcination conditions can be explained by the fact that they influence the catalyst structure and base metal dispersion; the reaction temperature is important due to the exothermic behavior of the hydrogenation reactions. The most important variables for the methanol selectivity are the feed composition variables and the reaction temperature. Since CO and CO_2 have different reaction paths and therefore can lead to different products, it is reasonable that the amount of them has a big influence on the methanol selectivity. Finally the methanol yield is most affected by the reduction temperature and amount of CO in the feed gas. All other catalyst preparation conditions and reaction conditions have some equally distributed influence on the MeOH yield. Looking at the catalyst composition, it is evident

that it has less importance, but the species that are considered significant for all three responses, are the ones on which the most research is done. These are: CuO, ZnO, Al_2O_3 and ZrO_2.

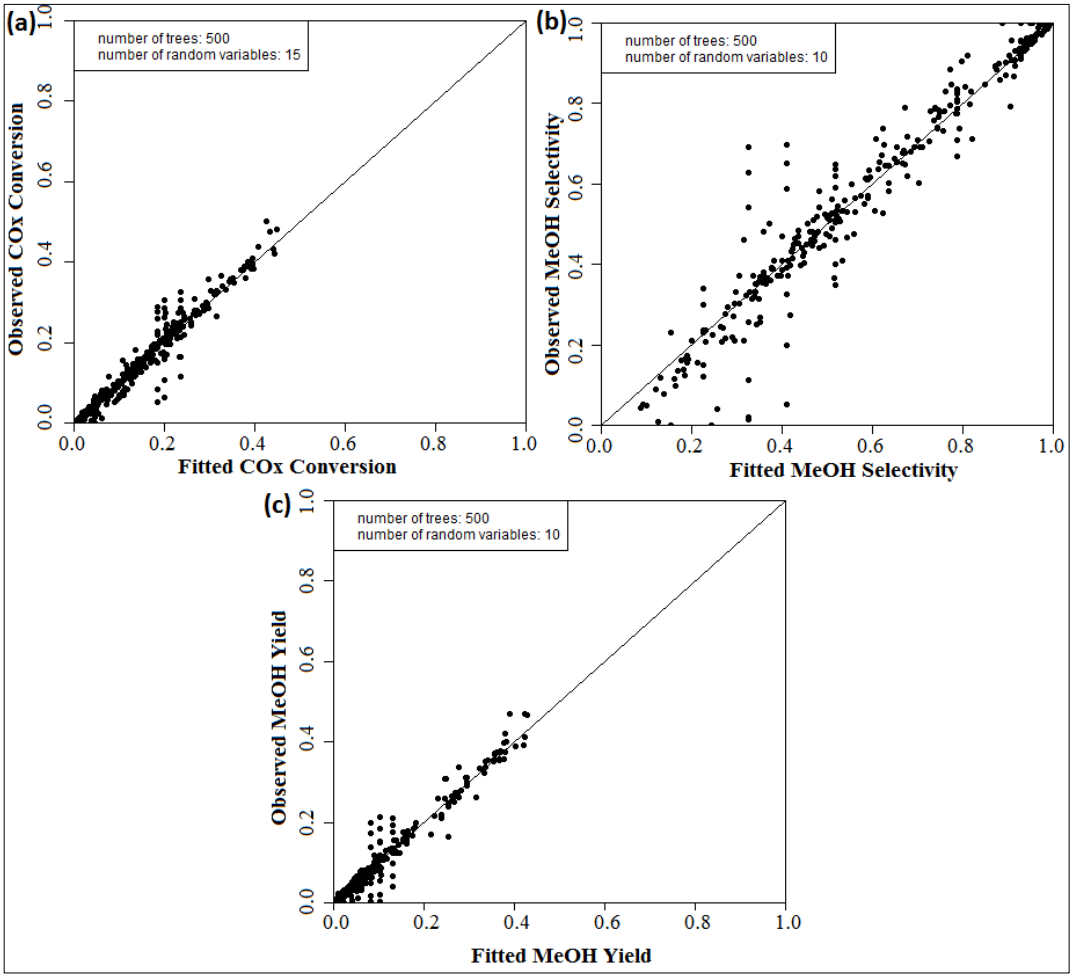

Figure 4.16. Fitted vs. observed (a) CO_X conversion (b) MeOH selectivity (c) MeOH yield for reduced RF model.

For a deeper insight, the five most important input variables for each response variable were plotted as a partial dependence plot in Figure 4.19. The abscissa represents the input variable and the ordinate the response conversion (a), selectivity (b) and yield (c). The intrinsic partial dependence function of the random forest algorithm was used to compute the partial dependencies. The plots show how much the input variables influence the outcome and in what range the best results can be expected. The dependencies include the relation between the input variable and the response variable as well as the averaged effects of all interactions of the chosen input variable with all other input variables.

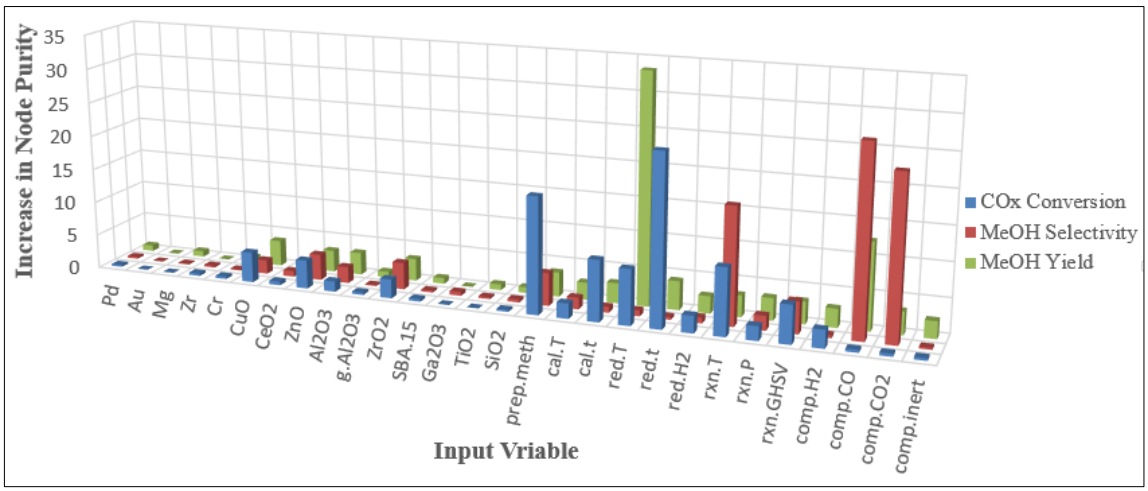

Figure 4.17. Variable importance by increase in node purity of input variables for RF (scaled to 100 %).

For example if the reduction time is larger than 3.7 h, the conversion increases by approx. 6 %, but the reduction temperature shouldn't exceed 500 K. Whereas, an increase in reaction temperature over 500 K, increases the conversion by 8 %. Catalyst preparation method one (coprecipitation) seems to produce the highest conversion. The calcination time should be as short as two hours to have an enhancing effect on the conversion.

A CO feed composition higher than 15 % results in a stable MeOH selectivity, lower values decrease the selectivity. On the other side, a CO_2 feed composition higher than 8 % decreases the selectivity dramatically. Therefore a minimum CO/CO_2 ratio of 1.875 should be considered. A high reaction temperature decreases the methanol selectivity and therefore, the MeOH selectivity is inversely proportional to the CO_X conversion requiring a trade-off. It can be observed that the selectivity is less sensitive to the catalyst preparation method than the conversion. A feed GHSV of 5000 mL/gcat*h is suggested as a minimum GHSV for high MeOH selectivity.

The methanol yield follows a very similar reduction temperature and reduction time dependence as the conversion. While a CO amount of 15 % is sufficient for high conversion, it needs to be at least 25 % for a high yield; a minimum CO/CO_2 ratio of 3.125 is also required for a high yield. It can be also inferred that a CuO amount of at least 55 % in the catalyst is very beneficial for a high methanol yield, with a slight increase at higher values

but a drastic drop at lower values. Two catalyst preparation methods, namely coprecipitation and sol-gel method, appears to be superior to the others. While the reduction conditions and preparation method dependence are acquired from the conversion, the feed composition dependence stems from the methanol selectivity.

Figure 4.18 shows the prediction power of the best random forest model, which was the PAM clustered model. The observed responses were strongly miss predicted. Especially high conversions and yields were under predicted and low selectivities over predicted. That means that the random forest models weren't able to catch the essence needed to predict the responses accurately and therefore, should only be used for exploratory analysis.

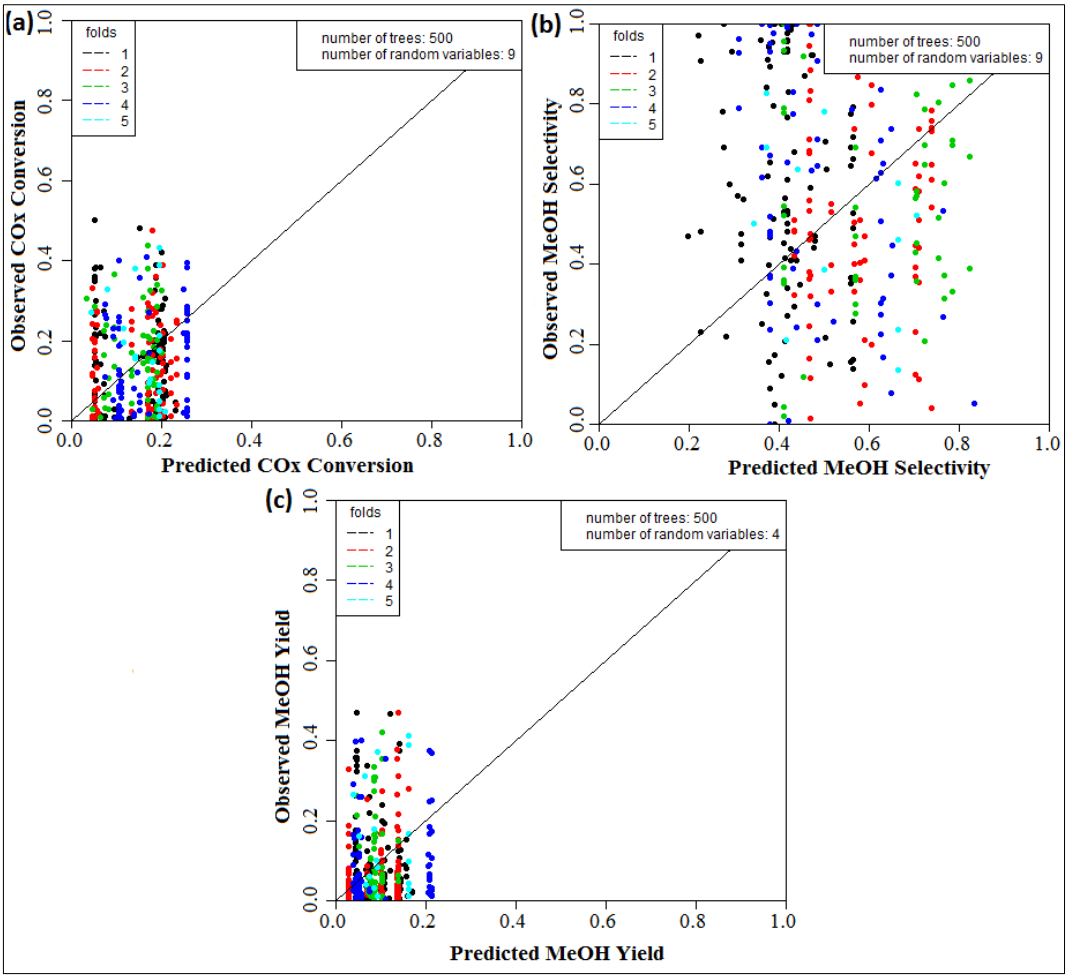

Figure 4.18. Predicted vs. observed (a) CO_X conversion (b) MeOH selectivity (c) MeOH yield for PAM clustered RF model.

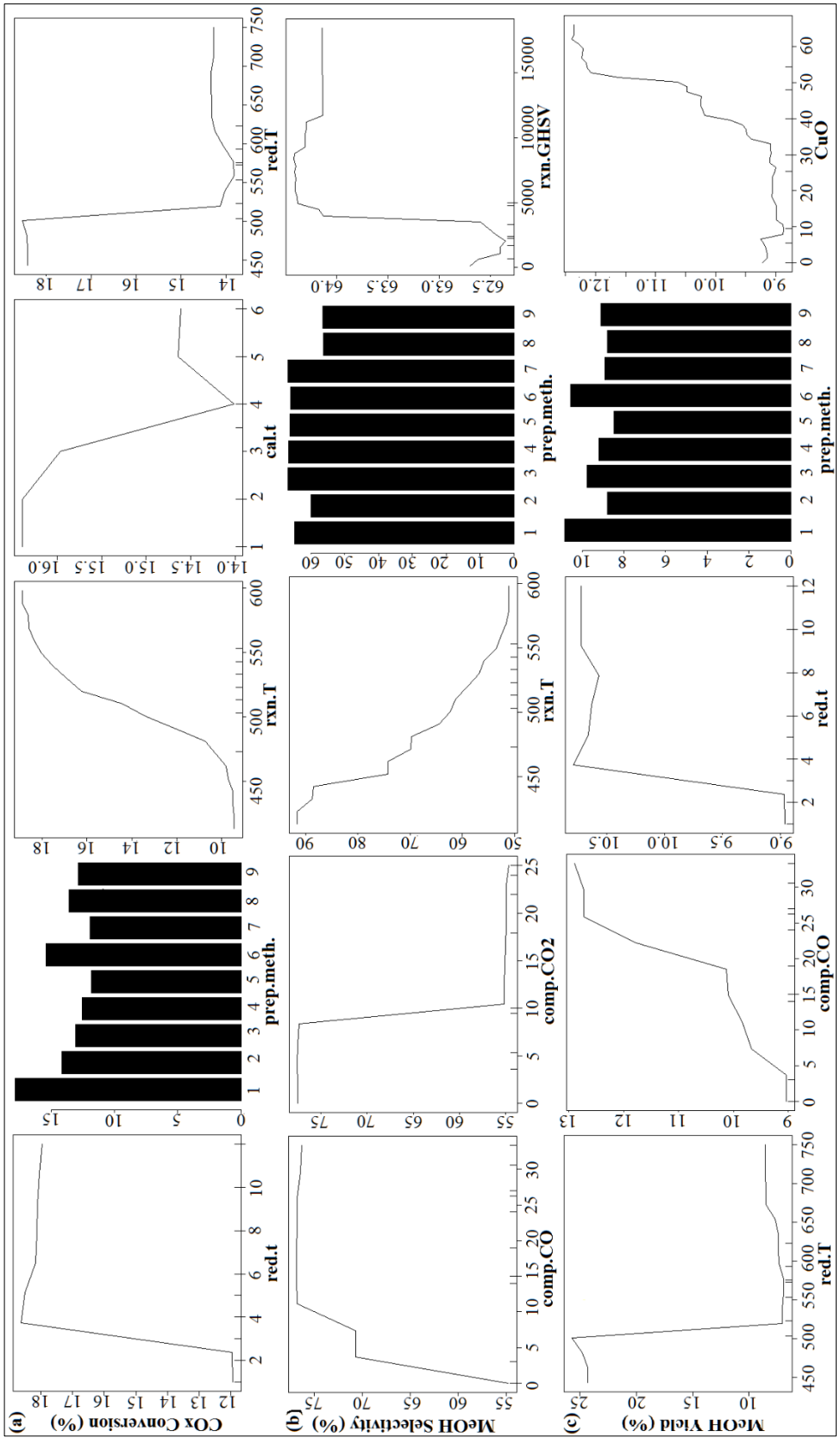

Figure 4.19. Partial dependence on the five most important input variables for (a) CO_X conversion (b) MeOH selectivity (c) MeOH yield

4.3.2. Results of Subsets of the Database

With the aim to increase the prediction power, the database was sub-setted into smaller databases and modeled with the random forest algorithm. Since the database is very small, a sub-setting into more than three groups was not permitted to keep the data point to variable ratio high enough. Sixteen different subsets were tried. Eleven subsets were created manually by inspection of the input variable distributions in chapter 3.1. If a value of a variable was extremely predominant, the data was divided into a subset with only that value for the variable and one excluding that value. If a variable had a bimodal distribution, the data was divided into two subsets with two unimodal distributions. Also if a variable had a very wide range with a lot, but different extreme values, the range was cut to reassemble a proper normal distribution. Subsets were also created by clustering the complete database into two clusters by PAM, hierarchical clustering and random forest.

Two subsets were built by separating the database by catalyst preparation method. One subset included all data points with coprecipitation as preparation method and one excluded all data points with coprecipitation method. This was chosen due to the reason that coprecipitation has a big predominance among the preparation methods like it was shown in Figure 3.1. Another two subsets were created by separating the database by reduction time. This was done due to the binominal distribution of the variable, seen in Figure 3.3. The subsets were divided into reduction times higher than four hours and lower than four hours. Division by reduction temperature contributed the next two subsets. The distribution of the reduction temperature can be divided into two approx. normal distributions and it is a variable that plays an important role in the decision tree algorithm as a split variable. Looking at the feed stream GHSV in Figure 3.9, a broad distribution can be observed among values with up to 1000-fold differences. A subset with values between 1200 and 6000 $mL/g_{cat}*h$ was chosen to include only the most often used values within a range without extreme values. Furthermore, one subset was chosen according to the catalyst composition. It had the requirement to be composed of $Cu/ZnO/Al_2O_3$ to at least 75 %. This was chosen to represent the commercial catalyst, including modifications in composition ratios and addition of promoters. Finally three subsets were build according to the carbon source. Table 3.3 shows that some data points have only CO, some only CO_2 and some a mix of both as a carbon source. Since the reaction paths differ between them, it was decided to divide the database accordingly.

One subset, including only CO_2 in the feed and no CO and one subset with only mixed carbon sources was built. Since there were too less data points with only CO and no CO_2 in the feed, the third subset was built by allowing a maximum CO_2 content of 7.5 %. This is in accordance with the previous results that showed a small amount of CO_2 does not alter the results drastically.

An unsupervised approach to divide the database, by clustering it into two subsets, was done by the clustering methods PAM and hierarchical clustering, which were also used previously to cluster the catalysts. Additionally an intrinsic random forest clustering method was applied to cluster the database. While PAM and hierarchical clustering yielded two subsets, big enough to be processed, the random forest clustering generated one big and one small cluster, of which just the big one could be used.

Table 4.5 shows the results for all subsets for fitted and predicted outcomes and the size of the subsets. The best results are highlighted with asterisks. Most of the subsets increased the goodness of fit and prediction power of the models. The best $R^2_{adj.}$ was as high as 0.9927 and the best RMSE was 0.0105. The best PRMSE was 0.0565. The best prediction for conversion and yield were done by model two, which excludes coprecipitation as a preparation method. The best selectivity prediction was done by model 14; the hierarchical cluster one model.

Figure 4.20 shows the best fitted vs. observed and predicted vs. observed plots. In terms of goodness of fit, the "PAM cluster two" subset led to the most accurate conversion and yield fits and the best selectivity was fitted by the subset having the reduction time of more than or equal to 4 h. The best conversion and yield predictions were done by the subset excluding preparation method one while the best selectivity predictions were done by the "hierarchical cluster one" subset. From the plot, it is clear that the goodness of fit improved significantly but the prediction power is still very poor and the predictions almost randomly. Hence, the sub-setting did not lead to a successful model for prediction.

Table 4.5. Random forest results for subsets.

				Random Forest		
				Fitted		Predicted
Nr.	**Subset**	**Size**	**Response**	**$R^2_{adj.}$**	**RMSE**	**PRMSE**
1	preparation method = 1 (coprecipitation)	165	X_{COx}	0.9405	0.0344	0.1141
			S_{MeOH}	0.8917	0.0839	0.3229
			Y_{MeOH}	0.9294	0.0303	0.1416
2	preparation method != 1 (all except coprecipitation)	192	X_{COx}	0.9495	0.0181	*0.0937
			S_{MeOH}	0.9631	0.0554	0.3636
			Y_{MeOH}	0.9411	0.0174	*0.0565
3	reduction time >= 4 h	155	X_{COx}	0.9614	0.0194	0.1161
			S_{MeOH}	*0.9739	*0.0415	0.3386
			Y_{MeOH}	0.9815	0.0160	0.1449
4	reduction time < 4 h	202	X_{COx}	0.8671	0.0298	0.1061
			S_{MeOH}	0.9160	0.0833	0.3706
			Y_{MeOH}	0.7599	0.0252	0.0682
5	reduction temperature >= 555 K	204	X_{COx}	0.9560	0.0164	0.0952
			S_{MeOH}	0.9737	0.0425	0.3732
			Y_{MeOH}	0.9405	0.0125	0.0635
6	reduction temperature < 555 K	153	X_{COx}	0.9226	0.0360	0.1483
			S_{MeOH}	0.8990	0.0900	0.3601
			Y_{MeOH}	0.9413	0.0321	0.1509
7	1200 <= GHSV <= 6000 $mL/g_{cat}*h$	254	X_{COx}	0.9445	0.0281	0.1401
			S_{MeOH}	0.9294	0.0780	0.3706
			Y_{MeOH}	0.9555	0.0243	0.1331
8	commercial catalyst ($Cu/ZnO/Al_2O_3$ >= 75 %)	206	X_{COx}	0.9423	0.0267	0.1389
			S_{MeOH}	0.9405	0.0693	0.3560
			Y_{MeOH}	0.9464	0.0248	0.1270
9	carbon source: CO_2	216	X_{COx}	0.8745	0.0295	0.1027
			S_{MeOH}	0.8530	0.0848	0.2444
			Y_{MeOH}	0.7828	0.0242	0.0619
10	carbon source: CO_X	112	X_{COx}	0.9776	0.0196	0.1597
			S_{MeOH}	0.9236	0.0316	0.1489
			Y_{MeOH}	0.9776	0.0192	0.1521
11	carbon source: CO + (CO_2<=7.5 %)	94	X_{COx}	0.9771	0.0194	0.1537
			S_{MeOH}	0.9173	0.0225	0.1322
			Y_{MeOH}	0.9716	0.0211	0.1551
12	PAM cluster 1	235	X_{COx}	0.8824	0.0290	0.1023
			S_{MeOH}	0.8711	0.0836	0.2645
			Y_{MeOH}	0.8173	0.0241	0.0651
13	PAM cluster 2	122	X_{COx}	*0.9806	*0.0154	0.1606
			S_{MeOH}	0.9348	0.0262	0.1296
			Y_{MeOH}	*0.9927	*0.0105	0.1685
14	hierarchical cluster 1	129	X_{COx}	0.9826	0.0169	0.1588
			S_{MeOH}	0.9361	0.0255	*0.1017
			Y_{MeOH}	0.9924	0.0172	0.1587
15	hierarchical cluster 2	228	X_{COx}	0.8807	0.0295	0.0976
			S_{MeOH}	0.8575	0.0849	0.2588
			Y_{MeOH}	0.7900	0.0252	0.0607
16	random forest cluster 1	312	X_{COx}	0.9416	0.0304	0.1074
			S_{MeOH}	0.8607	0.1039	0.3296
			Y_{MeOH}	0.8366	0.0241	0.0715

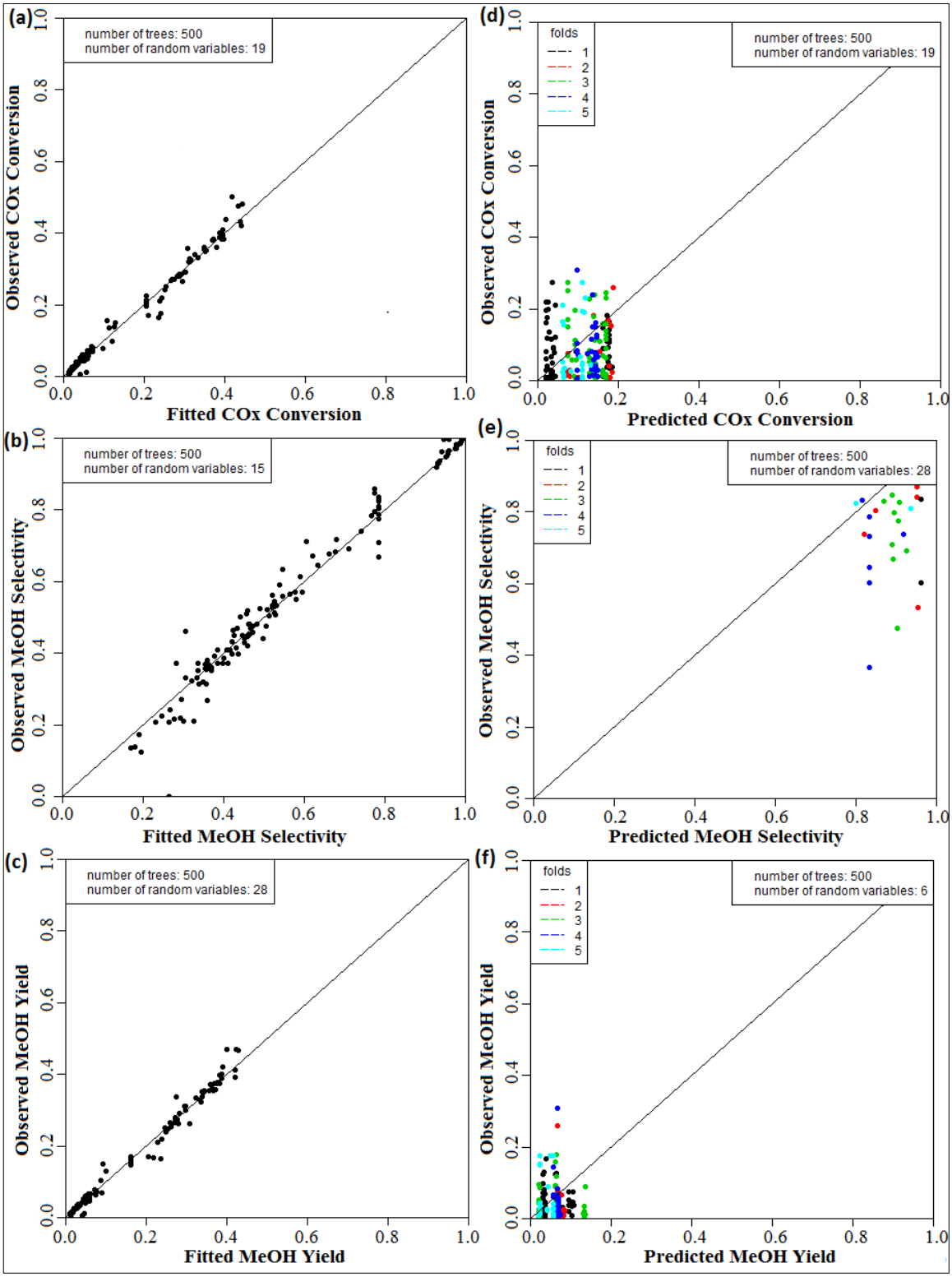

Figure 4.20. Fitted vs. observed (a) CO_X conversion of subset 13 (b) MeOH selectivity of subset 3 (c) MeOH yield of subset 13 and predicted vs. observed (d) CO_X conversion of subset 2 (e) MeOH selectivity of subset 14 (f) MeOH yield of subset 2.

4.4. Discussion

Comparing the quality of the models, it can be observed from Table 4.6 that multiple linear regression and regression tree yielded very similar results for the goodness of fit for the descriptive exploratory task. Although, the regression tree algorithm was slightly better than the MLR algorithm in terms of $R^2_{adj.}$, the RMSE was higher. This might be due to the fact that RT just fits mean values of responses. Since nonlinear relations are probable to exist in the used database, the relatively low fitting power of the MLR model was expected. It was surprising that the decision tree analysis did not improved the results either, although methods are known to capture nonlinear dependencies and generally perform better that MLR models. However, random forest increased the goodness of fit significantly. This result was expected according to literature involving random forest applications. Clustering the database did not yield any improvement for the goodness of fit either. The best models were found to be the complete or reduced models. MLR was expected to perform best with the complete model, due to the positive correlation of $R^2_{adj.}$ with the number of input variables. The MeOH selectivity was fitted with the highest errors; this might be due to the fact that selectivity values in the database ranged from zero to one although conversion and yield values of the database did not exceed 0.55.

None of the models could predict unseen data points successfully. Despite, acceptable PRSM values, all the predicted responses had an almost random pattern. This was shown by the predicted vs. observed plots of CO_X conversion, MeOH selectivity and MeOH yield. However, compared between each other, the random forest had the best prediction power; regression tree being slightly worse and MLR having the largest errors in prediction. Opposed to the fitted vs. experimentally observed (training) plots, the most successful predicted vs. observed (testing) plots were found using clustered models; with less input variables. This may be indicating that dimensionality reduction by clustering may improve the results but need further investigation. In some cases, sub setting the database according to variable distributions or clustering methods indeed led to an increase in prediction power but it was still not sufficient to be considered as successful.

Comparing the variable importance of different models, one can see that the variable importance computed by the MLR model is contra intuitive because it is represented by p-

values. Since most variables were considered important by the MLR model, only the unimportant variables with p-values larger than 0.05 were listed in Table 4.7. The other variable importances are based on increase in node purity and can be compared easier.

Table 4.6. Result comparison of best models.

		Fitted			
		MLR	**RT**	**CT**	**RF**
		Complete	**Reduced**	**Reduced**	**Reduced**
$R^2_{adj.}$	**X_{COX}**	0.7952	0.7898		0.9482
	S_{MeOH}	0.7952	0.8105		0.9435
	Y_{MeOH}	0.8177	0.8377		0.9549
RMSE	**X_{COX}**	0.0493	0.0521		0.0255
	S_{MeOH}	0.1264	0.1270		0.0684
	Y_{MeOH}	0.0437	0.0430		0.0224
Misclassification error	**X_{COX}**			0.2213	
	S_{MeOH}			0.3221	
	Y_{MeOH}			0.2549	
		Predicted			
		PAM	**PAM**	**Complete**	**PAM**
PRMSE	**X_{COX}**	0.2434	0.1470		0.1313
	S_{MeOH}	0.4622	0.4100		0.3416
	Y_{MeOH}	0.1908	0.1401		0.1176
Misclassification error	**X_{COX}**			0.7339	
	S_{MeOH}			0.8319	
	Y_{MeOH}			0.7731	

The important catalyst components, namely, CuO, ZnO and ZrO_2, were found to be the same by all methods; this coincides with the published. The catalyst related variables, including preparation method, calcination conditions and reduction conditions, are weighted differently by different models. While it is clear that preparation method and reduction temperature seems to predominate the models, the others are equally distributed with less importance. The reaction conditions like temperature, pressure, GHSV and feed composition were specified as important by the decision trees algorithms but not by the random forest. Random forest had a higher goodness of fit and identified the Cu/Zn and Cu/Zr interactions as well as the pre-reaction procedures like calcination and reduction as important. These quantities are known to have a key effect on the base metal particle size and distribution and therefore, on the conversion. Due to these reasons, this model should be regarded as the most accurate one. Since RF regards most other variables as unimportant, which might not be true,

the lesser importances can be taken from the classification tree model, which distributes the importances more equally.

By all models, the selectivity was found to be affected by the similar variables, which are the pre-reaction procedures and the reaction conditions like temperature, GHSV and feed composition. Especially reaction temperature and CO/CO_2 ratio seems to be important by every model. While the RT model excludes the catalyst composition from the list of important variables, CT and RF models include the CuO, ZnO and ZrO_2 variables. The CT model has also here a wider range for distribution of important variables while RF has a more definitive list.

The variable importance for the yield includes variables of every kind. This makes sense since the yield is a function of the conversion and the selectivity. Here, all species of the commercial catalyst (CuO, ZnO and Al_2O_3) are classified as important. Also the preparation method and especially the reduction conditions seem to play an important role. Furthermore, more attention is paid to the reaction conditions like pressure and GHSV. The RF models also consider the reduction temperature and CO amount the most important variables.

In general it can be said that the models agree on most of the variable importances with different rankings. Some models present more equally distributed weights for variable importance than other do. It can be said that the catalyst composition and treatment, including calcination and reduction are more important for the conversion and the reaction conditions and feed composition is more important for the selectivity.

		Multiple Linear Regression	Regression Tree	Classification Tree	Random Forest
CO_X Conversion	**Important**	almost all	CuO, ZnO, ZrO2, pep.meth, cal.t, red.T,red.t, red.H2, rxn.GHSV, comp.CO, comp.CO2	CuO, ZnO, ZrO2, prep.meth, cal.T, cal.t, red.T, red.t, red.H2, rxn.T, rxn.P, rxn.GHSV, comp.H2, comp.CO, comp.CO2, comp.inert	CuO, ZnO, ZrO2, prep.meth, cal.t, red.T, red.t, rxn.T, rxn.GHSV
	Unimportant	CeO2, g.Al2O3, SiO3			
MeOH Selectivity	**Important**		cal.t, red.H2, rxn.T, rxn.GHSV, comp.H2, comp.CO, comp.CO2	CuO, ZnO, ZrO2, cal.T, cal.t, red.T, red.t, red.H2, rxn.T, rxn.P, rxn.GHSV, comp.H2, comp.CO, comp.CO2, comp.inert	ZnO, ZrO2, prep.meth,rxn.T, rxn.GHSV, comp.CO, comp.CO2
	Unimportant	Pd, CuO, ZnO, Ga2O3, SiO2, red.T, red.t, comp.CO2			
MeOH Yield	**Important**		CuO,ZrO2, red.T, red.t, red.H2, rxn.GHSV, comp.CO, comp.CO2	CuO,ZnO,Al2O3, prep.meth, red.T, red.t, red.H2, rxn.T, rxn.GHSV, cop.CO, comp.CO2	CuO, ZnO, Al2O3, ZrO2, prep.meth, cal.t, red.T, red.t, rxn.T, rxn.P, rxn.GHSV, comp.H2, comp.CO, comp.CO2
	Unimportant	Zr, Cr, CeO2, Ga2O3, SiO2, rxn.GHSV			
Comments		with respect to p-values > 0.05		variable importances are more equal distributed as by random forest	<u>high significance:</u> conversion: prep.meth, red.t selectivity: rxn.T, comp. CO, comp.CO2 yield: red.T, comp.CO

Table 4.7. Comparison of variable importances.

5. CONCLUSION

5.1. Conclusions

To extract knowledge from published papers for methanol production from synthesis gas, 89 articles from the years between 2005 and 2015 were evaluated. From these articles, 24 met the requirements and were used to build a database with 357 unique data points and 28 variables. Multiple linear regression, decision trees and random forest were applied on the constructed database to analyze it, extract hidden relations, compute variable importances and predict unseen data. The computational work was conducted in R x64 3.2.3.

First the catalyst composition variables and catalyst preparation method were clustered by PAM and hierarchical clustering to reduce 16 continuous and one categorical variable to one combined categorical catalyst variable. PAM yielded a variable with 13 levels and hierarchical clustering with six. The clustered databases were constructed to investigate the effect of dimensionality reduction on the responses.

The first applied data mining method was multiple linear regression with combined stepwise variable selection. The complete, reduced and clustered models were compared. The model achieved acceptable results in terms of goodness of fit with $R^2_{adj.}$ between 0.80 and 0.82 for all three responses. The variable importance could be extracted in form of p-values only and didn't yield very meaningful results. The prediction power was very poor with almost random predicted responses, which partly were predicted out of the natural limits. The PRMSE for CO_X conversion, MeOH selectivity and MeOH yield were 0.205, 0.462 and 0.191, respectively. The MLR model was nether good in extracting variable importances nor in predicting unseen results and therefore, should not be used as the method of choice for data mining applications in the field of heterogeneous catalytic methanol synthesis.

A regression tree was built and achieved a $R^2_{adj.}$ of up to 0.84. It was slightly more successful than the multiple linear regression model in terms of $R^2_{adj.}$ and RMSE, but due to

the nature of regression trees to fit and predict mean values only, it suffered from interpretability loss. The best prediction model yielded a PRMSE of 0.140 but the responses were still almost randomly distributed. For all these reasons, regression tree cannot be suggested as a proper application for the methanol synthesis database.

Classification tree was used on a database with discretized response variables to classify CO_X conversion, MeOH selectivity and MeOH yield into different classes. Misclassification errors of 0.205, 0.305 and 0.255 were found for the fitted conversion, selectivity and yield, respectively. Classification rules for each response variable were extracted and presented in tabulated form. Since these rules are based on the fitted values they should be considered as guidelines, instead of absolute rules. Since the misclassification errors for predicted values lay between 0.723 and 0.804, the classification tree was regarded as unsuccessful for prediction purposes.

The last applied method was random forest. It was superior to all other models in every point of view. It reached a high goodness of fit with $R^2_{adj.}$ values between 0.944 and 0.955 and RMSE values between 0.022 and 0.067. Therefore, the variable importance extracted by the RF model was considered most accurate. Cu, ZnO and Al_2O_3 were found to be the most important catalyst species. While the variable importances were distributed mostly in a flat manner, some key variables stood out for each response variable. Catalyst preparation method and reduction time were found to be most important for CO_X conversion; reaction temperature, CO and CO_2 feed composition were found to be most important for MeOH selectivity; reduction temperature and CO composition were found to be most important for MeOH yield. The importances agree in most cases with the classification tree variable importances and reflect the deducted rules. Partial dependencies of the most important variables were extracted and optimum ranges for these variables were suggested. In prediction, the random forest algorithm yielded best results, compared with all other models, but still had a randomly distributed prediction pattern with PRMSE values between 0.118 and 0.342. Taking into account the real ranges of the response variables, these errors can be considered high. Therefore, random forest was unable to predict the responses properly.

As a last step, the database was divided into subsets according to different criteria like logical separation and clustering, and random forest predictions were applied on these sets.

The sub-setting led in some cases to slightly better PRMSE but nonetheless was not enough to consider the prediction a success.

It can be concluded that none of the models can be used for prediction purposes. On the other side, classification trees can be used to extract empirical rules and hierarchical tree representations. These can be used to give a first idea about the studied responses and some threshold values of variables which decide about high or low results. Random forest algorithms can be used as a black box method to extract variable importance to help directing the focus of research on the highest rated variables. Furthermore, partial dependence plots can be used to find optimum ranges for the input variables. These can further be exploited by using genetic algorithms to find the optimum catalyst for methanol synthesis.

5.2. Recommendations

Since the database was very small in size, an attempt can be done to collect scientific articles from before the year 2005 to increase the database size and include values which might reveal variable relations which could improve the prediction power. An increase in database size can also be reached if separate databases for CO_X conversion, MeOH selectivity and MeOH yield would be build, since some articles report only some of the responses. With larger database sizes, the database could be sub-setted successfully and more accurate models for each subset could be built. A prediction on these subset models could improve the results.

A dimensionality reduction by encoding the catalyst composition variables in a more sophisticated way could probably increase the prediction power, since these variables are sparse and include a lot of zero values. A dimensionality reduction by predicting the responses on physical catalyst properties like pore size, pore volume, active metal area, active metal dispersion, etc. instead of on the catalyst composition, preparation method, reduction and calcination conditions, might improve the results, since the physical properties are a direct result of the named variables. With a sufficient dimensionality reduction, nonlinear regression models could be applied in the hope of catching the nonlinear relations more precisely.

Other machine learning algorithms like artificial neural networks, support vector machines and ensemble methods can be tried as black box methods to achieve a better prediction. Genetic algorithms can be applied to find the optimum catalyst and reaction conditions which optimize the CO_X conversion, MeOH selectivity and MeOH yield.

Simultaneous multiple response prediction can be applied by artificial neural networks and decision tree algorithms to include the relations of the response variables. Empirical rules for simultaneous conversion and selectivity classification could be extracted from those trees. An attempt to first predict the CO_X conversion and include it as an input variable for MeOH selectivity prediction could improve the prediction quality.

Since a lot of researchers report the results as space time yield in g_{MeOH}/g_{cat}*h, an improvement might be done by modelling the STY directly as the response quality.

REFERENCES

Adib, H., R. Haghbakhsh, M. Saidi, M. A. Takassi, F. Sharifi, M. Koolivand, M. R. Rahimpour and S. Keshtkari, 2013, "Modeling and optimization of Fischer-Tropsch synthesis in the presence of Co (III)/Al2O3 catalyst using artificial neural networks and genetic algorithm", *Journal of Natural Gas Science and Engineering*, Vol. 10, pp. 14 - 24.

Ames Laboratory, 2008, *Atomic%Weight%Converter*, https://www.ameslab.gov/files/At_Wt_Converter20080909.xls, accessed at December 2015

Angelo, L., K. Kobl, L. M. M. Tejada, Y. Zimmermann, K. Parkhomenko and A.-C. Roger, 2015, "Study of CuZnMOx oxides (M = Al, Zr, Ce, CeZr) for the catalytic hydrogenation of CO2 into methanol", *Comptes Rendus Chimie*, Vol. 18, pp. 250 - 260.

Baltes, C., S. Vukojevic and F. Schüth, 2008, "Correlations between synthesis, precursor, and catalyst structure and activity of a large set of CuO/ZnO/Al2O3 catalysts for methanol synthesis", *Journal of Catalysis*, Vol. 258, pp. 334 - 344.

Baumes, L. A., J. M. Serra, P. Serna and A. Corma, 2006, "Support Vector Machines for Predictive Modeling in Heterogeneous Catalysis: A Comprehensive Introduction and Overfitting Investigation Based on Two Real Applications", *Journal of Combinatorial Chemistry*, Vol. 8; pp. 583 - 596.

Bertau, M., H. Offermanns, L. Plass, F. Schmidt and H.-J. Wernicke, 2014, *Methanol: The Basic Chemical and Energy Feedstock of the Future*, Springer, Berlin.

Breiman, L., J. Friedman, C. J. Stone and R. A. Olshen, 1984, *Classification and Regression Trees*, Taylor & Francis.

Cios, K. J., W. Pedrycz, R. W. Swiniarski and L. A. Kurgan, 2007, *Data Mining: A Knowledge Discovery Approach*, Springer Science+Business Media, LLC, New York, USA.

Farahani, B. V., F. H. Rajabi, M. Bahmani, M. Ghelichkhani and S. Sahebdelfar, 2014, "Influence of precipitation conditions on precursor particle size distribution and activity of Cu/ZnO methanol synthesis catalyst", *Applied Catalysis A: General*, Vol. 482, pp. 237 - 244.

Fissore, D., A. A. Barresi and D. Manca, 2004, "Modelling of methanol synthesis in a network of forced unsteady-state ring reactors by artificial neural networks for control purposes", *Chemical Engineering Science*, Vol. 59, pp. 4033 - 4041.

Fornero, E. L., P. B. Sanguineti, D. L. Chiavassa, A. L. Bonivardi and M. A. Baltanás, 2013, "Performance of ternary Cu–Ga2O3–ZrO2 catalysts in the synthesis of methanol using CO2-rich gas mixtures", *Catalysis Today*, Vol. 213, pp. 163 - 170.

Frei, E., A. Schaadt, T. Ludwig, H. Hillebrecht and I. Krossing, 2014, "The Influence of the Precipitation/Ageing Temperature on a Cu/ZnO/ZrO2 Catalyst for Methanol Synthesis from H2 and CO2", *ChemCatChem*, Vol. 6, pp. 1721 - 1730.

Gao, P., F. Li, H. Zhan, N. Zhao, F. Xiao, W. Wei, L. Zhong and Y. Sun, 2014, "Fluorine-modified Cu/Zn/Al/Zr catalysts via hydrotalcite-like precursors for CO2 hydrogenation to methanol", *Catalysis Communications*, Vol. 50, pp. 78 - 82.

Gao, P., F. Li, H. Zhan, N. Zhao, F. Xiao, W. Wei, L. Zhong, H. Wang Y. Sun, 2013, "Influence of Zr on the performance of Cu/Zn/Al/Zr catalysts via hydrotalcite-like precursors for CO2 hydrogenation to methanol", *Journal of Catalysis*, Vol. 298, pp. 51 - 56.

Gao, P., F. Li, N. Zhao, F. Xiao, W. Wei, L. Zhong and Y. Sun, 2013, "Influence of modifier (Mn, La, Ce, Zr and Y) on the performance of of Cu/Zn/Al catalysts via hydrotalcite-like precursors for CO2 hydrogenation to methanol", *Applied Catalysis A: General*, Vol. 468, pp. 442 - 452.

Gao, P., R. Xie, H. Wang, L. Zhong, L. Xia, Z. Zhang, W. Wei and Y. Sun, 2015, "Cu/Zn/Al/Zr catalysts via phase-pure hydrotalcite-like compounds for methanol synthesis from carbon dioxide", *Journal of CO2 Utilization*, Vol. 11, pp. 41 - 48.

García-Trenco, A. and A. Martínez, 2013, "A simple and efficient approach to confine Cu/ZnO methanol synthesis catalysts in the ordered mesoporous SBA-15 silica", *Catalysis Today*, Vol. 215, pp. 152 - 161.

Gower, J. C., 1971, "A General Coefficient of Similarity and Some of Its Properties", *Biometrics,* Vol. 27, No. 4, pp. 857 - 871.

Günay, M. E. and R. Yildirim, 2011, "Neural network Analysis of Selective CO Oxidation over Copper-Based Catalysts for Knowledge Extraction from Published Data in the Literature", *Industrial & Engineering Chemistry Research,* Vol. 50, pp. 12488 - 12500.

Günay, M. E. and R. Yildirim, 2013, "Developing global reaction rate model for CO oxidation over Au catalysts from past data in literature using artificial neural networks", *Applied Catalysis A: General*, Vol. 498, pp. 395 - 402.

Günay, M. E. and R. Yildirim, 2013, "Knowledge Extraction from Catalysis of the Past: A Case of Selective CO Oxidation over Noble Metal Catalysts between 2000 and 2012", *ChemCatChem,* Vol. 5, pp. 1395 - 1406.

Guo, Y., W. Meyer-Zaika, M. Muhler, S. Vukojevic and M. Epple, 2006, "Cu/Zn/Al Xerogels and Aerogels Prepared by a Sol–Gel Reaction as Catalysts for Methanol Synthesis", *European Journal of Inorganic Chemistry*, Vol. 2006, No. 23, pp. 4774 - 4781.

Hastie, T., R. Tibshirani and J. Friedman, 2009, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, USA.

Hu, B. and K. Fujimoto, 2010, "Promoting behaviors of alkali compounds in low temperature methanol synthesis over copper-based catalyst", *Applied Catalysis B: Environmental*, Vol. 95, pp. 208 - 216.

Iyer, S. S., T. Renganathan, S. Pushpavanam, M. V. Kumar and N. Kaisare, 2015, "Generalized thermodynamic analysis of methanol synthesis: Effect of feed composition", *Journal of CO2 Utilization*, Vol. 10, pp. 95 - 104.

Jiang, X., N. Koizumi, X. Guo and C. Song, 2015, "Bimetallic Pd–Cu catalysts for selective CO2 hydrogenation to methanol", *Applied Catalysis B: Environmental*, Vol. 170 – 171, pp. 173 - 185.

Kantardzic, M., 2011, *DATA MINING: Concepts, Models, Methods, and Algorithms*, John Wiley & Sons, Inc, Hoboken, New Jersey.

Karelovic, A., A. Bargibant, C. Fernández and P. Ruiz, 2012, "Effect of the structural and morphological properties of Cu/ZnO catalysts prepared by citrate method on their activity toward methanol synthesis from CO2 and H2 under mild reaction conditions", *Catalysis Today*, Vol. 197, pp. 109 - 118.

Kaufman, L. and P. Rousseeuw, 1897, "Clustering by means of Medoids", *In Statistical Data Analysis Based on the L1–Norm and Related Methods*, pp. 405 - 416.

Kleymenov, E., J. Sa, J. Abu-Dahrieh, D. Rooney, J. A. V. Bokhoven, E. Troussard, J. Szlachetko, O. V. Safonova and M. Nachtegaal, 2012, "Structure of the methanol synthesis catalyst determined by in situ HERFD XAS and EXAFS", *Catalysis Science & Technology*, Vol. 2, pp. 373 - 378.

Lee, S., 1990, *Methanol Synthesis Technology*, CRC Press Inc, Boca Raton, Florida.

Lei, H., R. Nie, G. Wu and Z. Hou, 2015, "Hydrogenation of CO2 to CH3OH over Cu/ZnO catalysts with different ZnO morphology", *Fuel*, Vol. 154, pp. 161 - 166.

Lim, H.-W., M.-J. Park, S.-H. Kang, H.-J. Chae, J. W. Bae and K.-W. Jun, 2009, "Modeling of the Kinetics for Methanol Synthesis using Cu/ZnO/Al2O3/ZrO2 Catalyst: Influence of Carbon Dioxide during Hydrogenation", *Industrial & Engineering Chemistry Research*, Vol. 48, No. 23, pp. 10448-10455.

Li, C., X.Yuan and K. Fujimoto, 2014, "Development of highly stable catalyst for methanol synthesis from carbon dioxide", *Applied Catalysis A: General*, Vol. 469, pp. 306 - 311.

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, K. Hornik, M. Studer and P. Roudier, 2016, *Finding Groups in Data: Cluster Analysis Extended Rousseeuw et al.*, https://cran.r-project.org/web/packages/cluster/cluster.pdf, accessed at March 2016

Ma, Y., Q. Ge, W. Li and H. Xu, 2009, "Methanol synthesis from sulfur-containing syngas over Pd/CeO2 catalyst", *Applied Catalysis B: Environmental*, Vol. 90, pp. 99 - 104.

Maniecki, T. P., P. Mierczynski, W. Maniukiewicz, K. Bawolak, D. Gebauer and W. K. Jozwiak, 2009, "Bimetallic Au–Cu, Ag–Cu/CrAl3O6 Catalysts for Methanol Synthesis", *Catalysis Letters*, Vol. 130, pp. 481 - 488.

Maniecki, T. P., P. Mierczynski, W. Maniukiewicz, D. Gebauer and W. K. Jozwiak, 2009, "The Effect of Spinel Type Support FeAlO, ZnAl2O4, CrAl3O6 on Physicochemical Properties of Cu, Ag, Au, Ru Supported Catalysts for Methanol Synthesis", *Kinetics and Catalysis*, Vol. 50, pp. 228 - 234.

Meshkini, F., M. Taghizadeh and M. Bahmani, 2010, "Investigating the effect of metal oxide additives on the properties of Cu/ZnO/Al2O3 catalysts in methanol synthesis from syngas using factorial experimental design", *Fuel*, Vol. 89, pp. 170 - 175.

Methanol Institute, 2011, *Applications for Methanol*, http://www.methanol.org/Methanol-Basics/Methanol-Applications.aspx, accessed at April 2016

Mierczynski, P., R. Ciesielski, A. Kedziora, M. Zaborowski, W. Maniukiewicz, M. Nowosielska, M. I. , Szynkowska and T. P. Maniecki, 2014, "Novel Pd–Cu/ZnAl2O4–ZrO2 Catalysts for Methanol Synthesis", *Catalysis Letters*, Vol. 144, pp. 723 - 735.

Mierczynski, P., P. Kaczorowski, A. Ura, W. Maniukiewicz, M. Zaborowski, R. Ciesielski, A. Kedziora and T. P. Maniecki, 2014, "Promoted ternary CuO-ZrO2-Al2O3 catalysts for methanol synthesis", *Central European Journal of Chemistry*, Vol. 12, No. 2, pp. 208 - 212.

Mierczynski, P., T. P. Maniecki, K. Chalupka, W. Maniukiewicz and W. K. Jozwiak, 2011, "Cu/ZnxAlyOz supported catalysts (ZnO: Al2O3 = 1, 2, 4) for methanol synthesis", *Catalysis Today*, Vol. 176, pp. 21 - 27.

Montebelli, A., C. G. Visconti, G. Groppi, E. Tronconi, C. Ferreira and S. Kohler, 2013, "Enabling small-scale methanol synthesis reactors through the adoption of highly conductive structured catalysts", *Catalysis Today*, Vol. 215, pp. 176 - 185.

Montebelli, A., C. G. Visconti, G. Groppi, E. Tronconi, S. Kohler, H. J. Venvik and R. Myrstad, 2014, "Washcoating and chemical testing of a commercial Cu/ZnO/Al2O3catalyst for the methanol synthesis over copper open-cell foams", *Applied Catalysis A: General*, Vol. 481, pp. 96 - 103.

Odabaşı, Ç., M. E. Günay and R. Yıldırım, 2014, "Knowledge extraction for water gas shift reaction over noble metal catalysts from publications in the literature between 2002 and 2012", *International Journal of Hydrogen Energy*, Vol. 39, pp. 5733 - 5746.

Pandya, R. and J. Pandya, 2015, "C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning", *International Journal of Computer Applications*, Vol. 117, No. 16, pp. 18 - 21.

Pasupulety, N., H. Driss, Y. A. Alhamed, A. A. Alzahrani, M. A. Daous, and L. Petrov, 2015, "Studies on Au/Cu–Zn–Al catalyst for methanol synthesis from CO2", *Applied Catalysis A: General*, Vol. 504, pp. 308 - 318.

R Foundation for Statistical Computing, 2008, *R: A language and environment for statistical computing*, https://www.r-project.org, accessed at May 2016

ReliaSoft Corporation, 2015, *Experiment Design & Analysis Reference*, http://www.synthesisplatform.net/references/Experiment_Design_and_Analysis_Reference.pdf, accessed at May 2016

Ren, H., C.-H. Xu, H.-Y. Zhao, Y.-X. Wang, J. Liu and J.Y. Liu, 2015, "Methanol synthesis from CO2 hydrogenation over Cu/g-Al2O3 catalysts modified by ZnO, ZrO2 and MgO", *Journal of Industrial and Engineering Chemistry*, Vol. 28, pp. 261 - 267.

Samei, E., M. Taghizadeh and M. Bahmani, 2012, "Enhancement of stability and activity of Cu/ZnO/Al2O3 catalysts by colloidal silica and metal oxides additives for methanol synthesis from a CO2-rich feed", *Fuel Processing Technology*, Vol. 96, pp. 128 - 133.

Santiago, M., K. Barbera, C. Ferreira, D. Curulla-Ferré, P. Kolb and J. Pérez-Ramírez, 2012, "By-product co-feeding reveals insights into the role of zinc on methanol synthesis catalysts", *Catalysis Communications*, Vol. 21, pp. 63 - 67.

Sharma, B. K., M. P. Sharma, S. K. Roy, S. Kumar, S. B. Tendulkar, S. S. Tambe and B. D. Kulkarni, 1998, "Fischer–Tropsch synthesis with Co/SiO2–Al2O3 catalyst and steady-state modeling using artificial neural networks", *Fuel*, Vol. 77, No. 15, pp. 1763 - 1768.

Simson, G., E. Prasetyo, S. Reiner and O. Hinrichsen, 2013, "Continuous precipitation of Cu/ZnO/Al2O3 catalysts for methanol synthesis in microstructured reactors with alternative precipitating agents", *Applied Catalysis A: General*, Vol. 450, pp. 1 - 12.

Siwawut, J., P. Namkhang and P. Kongkachuichay, 2015, "Co-metal catalysts (Cu, Zn, Al) on SiO2-TiO2 for methanol production from CO2: Effect of preparation methods", *Chemical Engineering & Technology*

Strunk, J., K. Kähler, X. Xia, M. Comotti, F. Schüth, T. Reinecke and T. Reinecke, 2009, "Au/ZnO as catalyst for methanol synthesis: The role of oxygen vacancies", *Applied Catalysis A: General*, Vol. 359, pp. 121 - 128.

Todeschini, R., 2007, *Useful and unuseful summaries of regression models*, http://www.moleculardescriptors.eu/tutorials/T5_moleculardescriptors_models.pdf, accessed at April 2016

Tuffery, S., 2011, *Data Mining and Statistics for Decision Making*, John Wiley & Sons, Ltd., Chichester, United Kingdom

Umegaki, T., Y. Watanabe, N. Nukui, K. Omata and M. Yamada, 2003, "Optimization of Catalyst for Methanol Synthesis by a Combinatorial Approach Using a Parallel Activity Test and Genetic Algorithm Assisted by a Neural Network", *Energy & Fuels*, Vol. 17, pp. 850 - 856.

Vesborg, P. C. K., I. Chorkendorff, I. Knudsen, O. Balmes, J. Nerlov, A. M. Molenbroek, B. S. Clausen and S. Helveg, 2009, "Transient behavior of Cu/ZnO-based methanol synthesis catalysts", *Journal of Catalysis*, Vol. 262, pp. 65 - 72.

Wang, L., L.Yang, Y. Zhang, W. Ding, S. Chen, W. Fang and Y. Yang, 2010, "Promoting effect of an aluminum emulsion on catalytic performance of Cu-based catalysts for methanol synthesis from syngas", *Fuel Processing Technology*, Vol. 91, pp. 723 - 728.

Wang, D., J. Zhao, H. Song and L. Chou, 2011, "Characterization and performance of Cu/ZnO/Al2O3 catalysts prepared via decomposition of M(Cu,Zn)-ammonia complexes under sub-atmospheric pressure for methanol synthesis from H2 and CO2", *Journal of Natural Gas Chemistry*, Vol. 20, pp. 629 - 634.

Wang, G., Y. Zuo, M. Han and J. Wang, 2011, "Cu-Zr-Zn catalysts for methanol synthesis in a fluidized bed reactor", *Applied Catalysis A: General*, Vol. 394, pp. 281 - 286.

Watanabe, Y., T. Umegaki, M. Hashimoto, K. Omata and M. Yamada, 2004, "Optimization of Cu oxide catalysts for methanol synthesis by combinatorial tools using 96 well microplates, artificial neural network and genetic algorithm", *Catalysis Today*, Vol. 89, pp. 455 - 464.

Waugh, K. C., 2012, "Methanol Synthesis", *Catalysis Letters*, Vol. 142, pp. 1153 - 1166.

Witoon, T., S. Bumrungsalee, M. Chareonpanich and J. Limtrakul, 2015, "Effect of hierarchical meso–macroporous alumina-supported copper catalyst for methanol synthesis from CO2 hydrogenation", *Energy Conversion and Management*, Vol. 103, pp. 886 - 894.

Xiao-bo, T., N. Tsubaki, X. Hong-juan, H. Yi-zhuo and T. Yi-sheng, 2014, "Effect of modifiers on the performance of Cu-ZnO-based catalysts for low-temperature methanol synthesis", *Journal of Fuel Chemistry And Technology*, Vol. 42, No. 6, pp. 704 - 709.

Xiao, J., D. Mao, X. Guo and J. Yu, 2015, "Effect of TiO2, ZrO2, and TiO2–ZrO2 on the performance of CuO–ZnO catalyst for CO2 hydrogenation to methanol", *Applied Surface Science*, Vol. 338, pp. 146 - 153.

Yoo, C.-J., D.-W. Lee, M.-S. Kim, D. J. Moon and K.-Y. Lee, 2013, "The synthesis of methanol from CO/CO2/H2gas over Cu/Ce1−xZrxO2 catalysts", *Journal of Molecular Catalysis A: Chemical*, Vol. 378, pp. 255 - 262.

Zahedi, G., A. Elkamel, A. Lohi, A. Jahanmiri and M. R. Rahimpor, 2005, "Hybrid artificial neural network—First principle model formulation for the unsteady state simulation and analysis of a packed bed reactor for CO2 hydrogenation to methanol", *Chemical Engineering Journal*, Vol. 115, pp. 113 - 120.

Zavyalova, U., M. Holena, R. Schlögl and M. Baerns, 2011, "Statistical Analysis of Past Catalytic Data on Oxidative Methane Coupling for New Insights into the Composition of High-Performance Catalysts", *ChemCatChem*, Vol. 3, pp. 1935 - 1947.

Zhang, Q., X. Li and K. Fujimoto, 2006, "Pd-promoted Cr/ZnO catalyst for synthesis of methanol from syngas", *Applied Catalysis A: General*, Vol. 309, pp. 28 - 32.

Zhang, Y., R. Yang and N. Tsubaki, 2008, "A new low-temperature methanol synthesis method: Mechanistic and kinetics study of catalytic process", *Catalysis Today*, Vol. 132, pp. 93 - 100.

Zhang, X., L. Zhong, Q. Guo, H. Fan, H. Zheng and K. Xie, 2010, "Influence of the calcination on the activity and stability of the Cu/ZnO/Al2O3 catalyst in liquid phase methanol synthesis", *Fuel*, Vol. 89, pp. 1348 - 1352.

Zhao, T.-S., K. Zhang, X. Chen, Q. Ma and N. Tsubaki, 2010, "A novel low-temperature methanol synthesis method from CO/H2/CO2 based on the synergistic effect between solid catalyst and homogeneous catalyst", *Catalysis Today*, Vol. 149, pp. 98 - 104.

APPENDIX A: ARTICLES USED IN THE METHANOL DATABASE

Table A.1 shows the articles, which were used for database construction and their number within the database.

Table A.1. Articles used in methanol database.

Article Nr.	Reference
1	(Zhang *et al*., 2006)
2	(Strunk *et al*., 2009)
3	(Ma *et al*., 2009)
4	(Lim *et al*., 2009)
5	(Wang *et al*., 2010)
6	(Wang *et al*., 2011a)
7	(Mierczynski *et al*., 2011)
8	(Wang *et al*., 2011b)
9	(Karelovic *et al*., 2012)
10	(Gao *et al*., 2013a)
11	(Gao *et al*., 2013b)
12	(García-Trenco and Martínez, 2013)
13	(Yoo *et al*., 2013)
14	(Fornero *et al*., 2013)
15	(Li *et al*., 2014)
16	(Mierczynski *et al*., 2014)
17	(Gao *et al*., 2014)
18	(Ren *et al*., 2015)
19	(Lei *et al*., 2015)
20	(Xiao *et al*., 2015)
21	(Jiang *et al*., 2015)
22	(Angelo *et al*., 2015)
23	(Gao *et al*., 2015)
24	(Siwawut *et al*., 2015)