Data Mining and Knowledge Discovery for Process identification and multivariate monitoring using spectroscopy: application to low temperature bitumen visbreaking

By

Dereje Tamiru Tefera

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

In

Process Control

Department of Chemical and Material Engineering

University of Alberta

Abstract

Data mining and knowledge discovery is a systematic process of identifying useful information from a data set where there is no or limited information about the underlying process. In this study, data mining and other learning methods are used cohesively to model a low temperature visbreaking process. Low temperature visbreaking is the process under investigation for field upgrading of oil sands bitumen. The classical visbreaker is operated at a temperature in the range of 430 to 500 °C, which would result in the formation of significant visbroken products, requiring subsequent hydrotreating. Due to this reason, several recent investigations have focused on finding an optimal operation condition that enables significant reduction of viscosity and limit the formation of olefins. These studies have indicated that the operation of a visbreaker at a temperature in the range of 150 to 400 °C could significantly decrease the viscosity of bitumen, while limiting the formation of cracked products. However, this process is at an investigation stage and there is very limited information about the underlying reaction mechanism. Spectroscopy is an ideal tool for the identification of such a complex chemical process since it provides comprehensive information about the underlying chemical changes at a given operation condition. But, the large amount of useful information contained in spectroscopic data is often difficult to extract since absorption intensities from individual chemical constituents of the sample experience a high degree of overlap, particularly for reactions involving chemically complex systems such as reactions involving heavy oils.

The notion of this thesis is to develop data-driven models that can describe the process well and can ultimately be used for real time analysis and optimization of the process of visbreaking in the temperature range of 150 to 400 °C using Fourier Transform-Infrared (FTIR) Spectroscopy data.

The first part of the research focuses on thermal kinetic modeling from the spectroscopy data acquired in the experimental analysis of the process. Obtaining mechanistic and kinetic descriptions for the chemistry involved in this process was a significant challenge, because of the compositional complexity of bitumen and the associated analytical challenges. Lumped kinetic models for heavy oil cracking can only be useful for describing the process on a preconceived reaction network, but are unsatisfactory for developing reaction networks. This study proposes a novel method to derive a reaction network of thermal cracking of oil sands bitumen from FTIR spectroscopy data using data mining and other learning methods.

The development of the kinetic network required implementation of several learning methods, including principal component analysis (PCA), data clustering and Bayesian learning. PCA is used for variable selection and a Bayesian agglomerative hierarchical cluster analysis was employed to obtain groups of pseudo-species with similar spectroscopic properties. Then, a Bayesian structure-learning algorithm was used to develop the corresponding reaction network. The reaction network derived from the model was compared to the reaction network of thermal cracking of model alkyl aromatic compounds proposed in the literature, and the agreement was encouraging. One attractive feature of the model is that it can be embedded into the process control system to predict the real-time reaction network and the process need limited or no prior description of the reaction network.

The second part attempts to design a spectroscopy-based online monitoring method for the process under consideration. The designed algorithm predicts the chemical rank of the unknown chemical mixture; resolves mixture spectra and evaluates the corresponding concentration profile of the resolved components so that the effect of different operation condition can be analyzed on a real time basis. The model includes several steps to resolve mixture spectra. In the first step, it predicts instrument noise and chemical rank of the system using PCA and Malinowski's error indicator function (IND) respectively. Once the chemical rank is determined, evolving factor analysis (EFA) is used to approximate the initial concentration profile. The final resolution of the spectra is completed using multivariate curve resolution alternating least squares (MCR-ALS). The model results agreed well with available experimental data for ^{1}H NMR characterization and other measurements such as microcarbon residue content. The model needs negligible computational effort and the only input required is the FTIR spectra and the model can be suitable for real time monitoring.

Preface

Chapter 2 and Chapter 3 of the thesis are paper manuscripts to be submitted as Dereje Tamiru Tefera, Lina Maria Yañez Jaramillo, Rajesh Ranjan, Chaoqun Li, Arno de Klerk and Vinay Prasad, "Bayesian learning to derive reaction networks for complex reacting systems from spectroscopy: Application to bitumen visbreaking"; and Dereje Tamiru Tefera, Ankit Agrawal, Lina Maria Yanez Jaramillo, Arno de Klerk and Vinay Prasad, "A Self- Modeling Multivariate Curve Resolution model for online Monitor of Bitumen Pyrolysis of using Fourier Transform-Infrared Spectroscopy (FTIR)" respectively.

Model development, computer programming, analysis and manuscript composition as well as the other chapters of the thesis was completed by myself, with the supervision of Prof. Vinay Prasad and Prof. Arno de Klerk. Prof. Vinay Prasad and Prof. Arno de Klerk are supervisory authors and were involved with concept formation and manuscript composition. Lina Maria Yañez Jaramillo, Lin Wang and Ashley Zachariah performed the visbreaking experiments. Rajesh Ranjan, Chaoqun Li and Ankit Agrawal performed preliminary analysis of the data using some of the methods implemented for analysis.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Dr. Vinay Prasad and co-supervisor Dr. Arno De Klerk for their continous supervision and valuable comments during thesis and paper writing. I consider it an honor to get the chance to work with both of them.

I also want to thank my fellows in the Dr. Vinay Prasad's group, and graduates students and postdoctoral fellows from Dr. Arno De Klerk's laboratory for their kind cooperation and support.

A special thanks to my parents for their continous encouragment.

My gratitude is also extended to my friends; Mesgana Seyoum Gizawu, Henok Asfaw, Wondemagegnehu Mulatu Teferi, Raie Taye Bekele, Berehanu Anagaw Wube, Tebikachew, Tariku,Yishak Abdulafiz Yusuf, and Ronald Ekyalimpa. Their support and personal cheering is greatly appreciated.

Table of contents

Table of Figures

List of tables

1. INTRODUCTION

1.1. BACKGROUND

Data mining and knowledge discovery refers to the systematic process of identifying potentially useful and ultimately understandable patterns in a given data set[1, 2]. Data-based knowledge discovery involves implementation of series of multivariate and/or machine learning methods [3, 4]. Data mining, a step in the knowledge discovery process, is a search for patterns of interest in a particular depictive form [5, 6]. Knowledge discovery methods have been used for a long time and in an array of disciplines, one of which is in industrial processes to extract knowledge from overwhelmingly large volumes of data generated from modern computer process control (DCS) and automatic logging systems[7-10] as well as in product development, optimization and fault detection. For example, Zheng et al., 2014[11], used data mining techniques to address critical process optimization problem in plasma display panel manufacturing. In the other study, Brudzewski *et al.*, 2006 [12, 13], used a combination of data based learning methods including principal component analysis (PCA), fuzzy C means (FCM) algorithm, hybrid neural network and support vector machines (SVM) to develop a model to predict gasoline quality using gas chromatography and Fourier transform infrared (FTIR) spectroscopy data. Some other studies focused on using knowledge discovery strategies in process monitoring, fault detection, isolation and diagnosis [14, 15]. The current study intends to use knowledge discovery technologies for the identification of a reaction network for mild thermal cracking process of oil sands bitumen and the design of a self-modeling curve resolution based multivariate online monitoring method from offline FTIR data.

The first part of this study focuses on the identification of the topology of the reaction kinetics for the mild thermal cracking of oil sands bitumen from FTIR spectroscopy data. Mild thermal cracking (visbreaking) is one of the oldest and the most commonly used heavy oil upgrading technologies, conducted at a temperatures in the range 430 to 500 °C[16] which usually results in the formation of significant cracked products. However, for visbreaking to be used for field upgrading of oils sands bitumen it is important to reduce the production of cracked products to avoid the need for subsequent hydrotreating of the visbroken product. Recent studies have indicated that the operation of a visbreaker at a temperatures in the range of 150 to 400 °C could significantly decrease the viscosity of bitumen, while limiting the formation of cracked products[17] [18]. These findings also suggested that product viscosity could be decreased by reactions other than just thermal cracking; hence, it was doubtful that kinetic models for conventional visbreaking would provide a reasonable description of the reaction network at lower visbreaking temperatures, the range of interest was 150 to 400 °C.

Lumped kinetics is one of the most commonly used methods for developing reaction network in the area of heavy oil cracking when there is enough prior knowledge about the reaction mechanism. The challenge in developing a kinetic model for visbreaking of bitumen in the temperature range 150 to 400 °C was that the reaction network was unclear. Hence, it was necessary to develop a strategy by which the reaction network could be inferred from the experimental data.

The second part of this study focuses on the design of another exploratory method, self-modeling curve resolution in order to develop spectroscopy-based online monitoring method for the mild thermal cracking process. Most multivariate online process monitoring technologies are designed based on principal component analysis (PCA) and partial least squares (PLS) methods[19]. In recent years, real-time spectroscopy based process monitoring has gained increasing acceptance as a choice for industrial chemical process control since it provides rapid and representative information such as composition and quality information directly for a complex chemical process, which otherwise are difficult/hazardous to obtain using direct measurement [20, 21]. This information is obviously valuable in process and product quality control[20, 22-24]. However, the large amount of useful information contained in spectroscopic data is often difficult to extract mainly because absorption bands from individual chemical constituents often experience a high degree of overlap, the samples are often chemically complex (typical example is reactions involving heavy oils) and susceptible to a large number of non-chemical interferences. As a result, it was important and necessary to design a valid knowledge discovery algorithm to enhance the information retrieval process.

1.2. Research Objectives

The goal of this research is to develop the reaction network and a multivariate online monitoring scheme for investigation of the field upgrading of oil sand bitumen in the temperature range of 150°C to 400°C. Hence, the main objectives are:

- To identify the most plausible reaction network for the mild thermal cracking reaction used for oil sands bitumen upgrading, from the FTIR data using PCA, data clustering and Bayesian networking learning methods.
- To design a spectroscopy based multivariate online monitoring scheme for the cracking reaction by combing PCA, Malinowski's error indicator function (IND), evolving factor analysis, and self-modeling multivariate curve resolution (SMCR).

These objectives are achieved by using FTIR spectroscopy data acquired from the pilot thermal cracker where the cracking reaction is conducted at varies reaction time and a temperature range of 150 to 400^{O}C.

This research is significant for the study of mild thermal cracking of bitumen because: (1) The reaction mechanism for thermal cracking of bitumen in the temperature range 150^{o}C to 400^{o}C is unclear and only limited domain knowledge exists to use the classical lumping methods. The model developed in this work requires very limited or no domain knowledge to identify the reaction network. The other advantage of this algorithm is that it can readily be deployed online and used in the real-time investigation of the reaction network where FTIR is used for measurement; a self-updating reaction network will be an invaluable tool in the analysis of the process under investigation. (2) Self-modeling multivariate curve resolution (SMCR)-based monitoring methods provides real-time resolution of FTIR spectra of complex chemical mixtures encountered in the thermal processing of oil sands derived bitumen; otherwise, the information contained in spectroscopic data is often difficult to extract since absorption intensities from individual chemical constituents of the sample experience a high degree of overlap. The method also predict the consecration profile of the resolved components.

1.3. THESIS OUTLINE

The thesis consists of four chapters, each of which will contribute to the main goal of the study. The second chapter focuses on the identification of a plausible chemical reaction network for mild thermal cracking of oil sands derived bitumen. The third chapter details the development of a self-modeling multivariate curve resolution-based online monitoring scheme for the thermal cracking reaction using FTIR spectroscopy. These two chapters also review the state of the art in modeling the reaction kinetics of heavy oil thermal cracking, theories of the learning algorithms implemented and spectroscopy-based monitoring. Chapter 4 presents the conclusions and summary of the finding of the study.

1.4. REFERENCES

1. Kurgan, L. A.; Musilek, P., A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review* **2006,** *21*, (01), 1-24.
2. Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R., *Advances in knowledge discovery and data mining*. AAAI press Menlo Park: 1996; Vol. 21.
3. Klösgen, W.; Zytkow, J. M., *Handbook of data mining and knowledge discovery*. Oxford University Press, Inc.: 2002.
4. Cios, K. J.; Swiniarski, R. W.; Pedrycz, W.; Kurgan, L. A. In *The knowledge discovery process*, Data Mining, 2007; Springer: 2007; pp 9-24.
5. Piatetsky-Shapiro, G.; Brachman, R. J.; Khabaza, T.; Kloesgen, W.; Simoudis, E. In *An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications*, KDD, 1996; 1996; pp 89-95.
6. Choudhary, A. K.; Harding, J. A.; Tiwari, M. K., Data mining in manufacturing: a review based on the kind of knowledge. *Journal of Intelligent Manufacturing* **2009,** *20*, (5), 501-521.
7. Köksal, G.; Batmaz, İ.; Testik, M. C., A review of data mining applications for quality improvement in manufacturing industry. *Expert systems with Applications* **2011,** *38*, (10), 13448-13467.
8. Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P. In *Knowledge discovery and data mining: towards a unifying framework*, KDD, 1996; 1996; pp 82-88.
9. Ma, C.; Ouyang, J.; Chen, H.-L.; Zhao, X.-H., An efficient diagnosis system for Parkinson's disease using kernel-based extreme learning machine with subtractive clustering features weighting approach. *Computational and mathematical methods in medicine* **2014,** *2014*.
10. Wang, X. Z.; McGreavy, C., *Data Mining and Knowledge Discovery for Process Monitoring and Control*. Springer-Verlag: 1999; p 254.
11. Zheng, L.; Zeng, C.; Li, L.; Jiang, Y.; Xue, W.; Li, J.; Shen, C.; Zhou, W.; Li, H.; Tang, L. In *Applying data mining techniques to address critical process optimization needs in advanced manufacturing*, Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014; ACM: 2014; pp 1739-1748.

12. Brudzewski, K.; Kesik, A.; Kołodziejczyk, K.; Zborowska, U.; Ulaczyk, J., Gasoline quality prediction using gas chromatography and FTIR spectroscopy: An artificial intelligence approach. *Fuel* **2006,** *85*, (4), 553-558.

13. Brudzewski, K.; Osowski, S.; Markiewicz, T.; Ulaczyk, J., Classification of gasoline with supplement of bio-products by means of an electronic nose and SVM neural network. *Sensors and Actuators B: Chemical* **2006,** *113*, (1), 135-141.

14. Pan, L., *Application of Statistical Methods for Gas Turbine Plant Operation Monitoring*. INTECH Open Access Publisher: 2011.

15. Wen, Q.; Ge, Z.; Song, Z., Data-based linear Gaussian state-space model for dynamic process monitoring. *AIChE journal* **2012,** *58*, (12), 3763-3776.

16. Gary, J. H.; Handwerk, G. E.; Kaiser, M. J., *Petroleum refining: technology and economics*. CRC press: 2007.

17. Wang, L.; Zachariah, A.; Yang, S.; Prasad, V.; de Klerk, A., Visbreaking Oilsands-Derived Bitumen in the Temperature Range of 340–400° C. *Energy & Fuels* **2014,** *28*, (8), 5014-5022.

18. Zachariah, A.; de Klerk, A., Thermal Conversion Regimes for Oilsands Bitumen. *Energy & Fuels* **2016,** *30*, (1), 239-248.

19. AlGhazzawi, A.; Lennox, B., Monitoring a complex refining process using multivariate statistics. *Control Engineering Practice* **2008,** *16*, (3), 294-307.

20. Zogg, A.; Fischer, U.; Hungerbühler, K., A new approach for a combined evaluation of calorimetric and online infrared data to identify kinetic and thermodynamic parameters of a chemical reaction. *Chemometrics and intelligent laboratory systems* **2004,** *71*, (2), 165-176.

21. Amari, T.; Ozaki, Y., Generalized Two-Dimensional Attenuated Total Reflection/Infrared and Near-Infrared Correlation Spectroscopy Studies of Real-Time Monitoring of the Initial Oligomerization of Bis(hydroxyethyl terephthalate). *Macromolecules* **2002,** *35*, (21), 8020-8028.

22. Blanco, M.; Castillo, M.; Peinado, A.; Beneyto, R., Application of multivariate curve resolution to chemical process control of an esterification reaction monitored by near-infrared spectroscopy. *Applied spectroscopy* **2006,** *60*, (6), 641-647.

23. Garrido, M.; Rius, F.; Larrechi, M., Multivariate curve resolution–alternating least squares (MCR-ALS) applied to spectroscopic data from monitoring chemical reactions processes. *Analytical and bioanalytical chemistry* **2008,** *390*, (8), 2059-2066.
24. Miller, C. E., Chemometrics for on-line spectroscopy applications—theory and practice. *Journal of Chemometrics* **2000,** *14*, (5-6), 513-528.

2. A BAYESIAN LEARNING APPROACH TO DEVELOP PSEUDO-REACTION MODELS FOR COMPLEX REACTING SYSTEMS: APPLICATION TO THE MILD VISBREAKING OF BITUMEN

A version of this chapter will be submitted as; Dereje Tamiru Tefera, Lina Maria Yañez Jaramillo, Rajesh Ranjan, Chaoqun Li, Arno de Klerk and Vinay Prasad, "Bayesian learning to derive reaction networks for complex reacting systems from spectroscopy: Application to bitumen visbreaking".

2.1. *Introduction*

Mild thermal cracking (visbreaking) is one of the oldest and the most commonly used heavy oil upgrading technologies. The cracking is typically conducted at temperatures in the range 430 to 500 °C.[1] Several studies have indicated that the operation of a visbreaker at a temperatures in the range 300 to 400 °C could significantly decrease the viscosity of bitumen, while limiting the formation of cracked products.[2, 3] These findings suggested that product viscosity could be decreased by reactions other than just thermal cracking. Limiting the formation of olefins due to cracking is advantageous for field upgrading applications, where subsequent hydrotreating of the visbroken product is impractical. It was doubtful that kinetic models that were developed for conventional visbreaking would provide a reasonable description of the reaction network at lower visbreaking temperatures. The development of a kinetic model to describe visbreaking in the temperature range 150 to 400 °C was of interest.

Modeling of the thermal cracking kinetics of heavy oil has been the subject of numerous studies. The most widely used models can be classified into two main groups: mechanistic and empirical. The compositional complexity of bitumen and its cracking products made mechanistic modeling more difficult and less practically appealing than empirical modeling. Due to their simplicity and practical application, models based on empirically lumped kinetics have been the subject of various studies. Lumping tries to represent a complex reactive system with a few pseudo-components by using chemical and/or physical properties, (e.g., boiling point range), to group many species together. This is followed by the proposal of a reaction network to relate the conversion of the pseudo-components to each other. Experimental data is then fitted to the preconceived reaction network, which forms the basis for the lumped kinetic model. However, there are many a priori assumptions made in developing lumped models, which could lead to inaccurate models for complex systems.

The challenge in developing a kinetic model for visbreaking of cold lake bitumen in the temperature range 150 to 400 °C was that the reaction network was unclear. Assuming a reaction network, which might not be representative of the actual reaction network, could render the model fundamentally unsound. It was necessary to develop a strategy by which the reaction network could be inferred from the experimental data. Furthermore, the bitumen feed and the visbroken bitumen product are both compositionally complex. These materials precluded compound-based separation and analysis.

Spectroscopy provided information about chemical identity of species in the feed and the products, and was employed to identify groups of pseudo-species to create meaningful lumps for modelling. Chemometric and machine learning methods were applied to the spectroscopic data, and provided a way to derive a plausible reaction model from the experimental data.

2.2. BACKGROUND

2.2.1 Major functional groups in oil sands-derived bitumen

A knowledge of the major functional groups present in bitumen is important for the understanding of the reaction products and expected reaction mechanism in its processing. Based on Strausz et al.,[4] the major organic functional groups in bitumen can be classified into saturates and aromatics. Though alkyl groups attached to cyclic rings represent substantially the weight of bitumen, normal alkanes and low boiling single monocyclic alkanes are negligible in bitumen, constituting only about 0.66 wt% and 4.82 wt%, respectively. The majority of the saturate class is made up of poly-cycloalkanes and/or alkyl cycloalkanes. The alkyl substitutes in the structure represent either an open chain, acyclic or aromatic hydrocarbon free radicals. Most of the ring structures including pure hydrocarbons bear at least one-alkyl substitutes.

The second major organic functional group consists of aromatics. Bitumen has an abundance of aromatics compounds ranging from alkyl benzenes to condensed polyaromatic molecules that are alkyl and naphthenic substituted. Strausz et al. [4] also state that Cold Lake bitumen contains about 8.1, 3.6 and 23.9% of monoaromatics, diaromatics and polyaromatics, respectively. Mixed naphthene and aromatic rings[5] such as perylene are other significant constituents.

Heteroatoms such as sulfur, oxygen and nitrogen are present in a variety of forms, but not in elemental form; also, hydrogen sulfide is not present.[6] Sulfur exists mainly in two forms, organic sulfides and in aromatic rings. Cold Lake bitumen contains 1.86% wt. of its sulfur in the form of aliphatic sulfides and about 3.05% wt. as thiophene[7].

2.2.2 Thermal processing of bitumen

The primary objective of thermal processing (dominated by cracking) is carbon number (molecular size) reduction and the main chemical bonds of interest are C-C, C-S, C-O and C-H, with C-S being the easiest bond to break since it has the lowest bond dissociation energy. Studies have shown that the estimated amount of highly reactive types of sulfur compounds in bitumen and asphalts ranges up to 50% of the corresponding total sulfur content and the ease of fracture of sulfide bonds has been postulated as a major mechanism for the thermal cracking of bitumen [6]. The chemistry of cracking of side chains and bridges on or between cycloalkanes and aromatics for bitumen is found to be similar to the mechanism of thermal cracking of normal alkanes, β-scission being the most important C-C bond breaking [6, 8]. For example, under favorable conditions, thermal cracking of alkylaromatics produces a mixture of alkenes, alkane, aromatics (or alkylaromatics with shorter chains) [6].

In using visbreaking as a primary upgrading technology, thermal conversion can lead to hydrogen disproportionation and eventually to the formation of coke, which is hydrogen deficient, and a hydrogen-rich group of compounds [9, 10].

2.2.3 Spectroscopic identification

Fourier transform infrared (FTIR) spectrometry is one of the most commonly used analytical tools in the identification of unknown materials[11-17]; bitumen characterization [18, 19]; and real time reaction monitoring[20-25]. A review of the spectroscopic identification of organic compounds is presented in Wong[26].

The classical method of spectroscopic identification of materials is based on the use of commercial IR libraries where a large number of reference IR spectra of pure compounds are available, or by qualitatively monitoring the change in the shape and position of the peaks [27]. FTIR library search methods treat the unknown mixtures as a linear combination of the spectra in the library. While this works well for small libraries of 10 or less spectra, it is usually unable to deconvolving the species involved in larger libraries containing spectra of similar compounds because of the overlap in their spectroscopic signatures, these difficulties are very significant for complex samples like bitumen [28].

In such cases, chemometric methods such as principal component analysis (PCA), cluster analysis (HCA) are often used [13, 14, 29]. The advantage of data mining techniques is that they need minimal expert knowledge to extract information contained in the spectra of complex mixture of samples[30-32].

2.3. DATA ACQUISITION

2.3.1 Origin of liquid samples

The data for this study was obtained from experimental investigations into the thermal cracking of Cold Lake bitumen.[2, 3, 33]The thermally processed bitumen samples were produced by performing the thermal cracking in small batch reactors with fast heat-up and cool-down (6 min) times. Each product sample was filtered through a 0.22 µm filter in order to obtain the solids yield and the liquid product that was analyzed by Fourier transform infrared (FTIR) spectroscopy. The Cold Lake bitumen feed contained suspended mineral matter: 0.9 ± 0.1 wt.%. The solids yield measured by filtration therefore contained both mineral matter and organic matter.

Experiments were conducted for various combinations of temperature and time. The range of temperatures covered was 150 to 400 °C, with reaction times varying from 0 to 8 hours. The products produced at a reaction time of 0 hours were obtained by heating the bitumen to the reaction temperature and immediately cooling it down again. The properties of these products were different to that of the Cold Lake bitumen feed.

A total of 42 thermally processed liquid samples were considered in this study.

2.3.2 Spectroscopic analysis

The infrared spectra of the liquid samples were collected using an ABB MB 3000 FTIR spectrometer. The spectra were collected at 2 cm^{-1} resolution over the spectral range 4000 to 600 cm^{-1}. Each spectrum was the average of 120 scans. The analyses were performed on the neat liquids using a Pike MIRacle™ attenuated total reflectance attachment.

2.4. METHODS

2.4.1 Data preprocessing

The collection of experimental data is not flawless. Principal component analysis (PCA) can be used to detect atypical observations or outliers. [34] The general notion of PCA is to replace a large number of correlated variables with a smaller number of uncorrelated variables, which are linear combinations of the observed variables, while capturing as much information in the original dataset as possible. These derived variables are called principal components[35] PCA was performed on the FTIR dataset to determine outliers beforehand. There are two possible ways to deal with outliers. The first option is to remove or replace with predicted value and the other method is to select a robust learning method. The second option was followed in this study.

Bayesian agglomerative hierarchical clustering, which we have employed, is effective in down-weighting the effect of outliers[36] and the clusters were also represented by the averaged intensity for Bayesian learning, which will also can reduce the effect of outliers.

The data was smoothed, scaled and centered before cluster analysis.

2.4.2 Cluster analysis: Bayesian hierarchical clustering

Cluster analysis divides observations into clusters (groups) based on the information found in the data matrix or their relationships such that observations belonging to the same group are more similar than observations belonging to different groups[37]. The matrix of data consists of samples that are characterized by multiple factors (variables). While there are different ways of classifying clustering methods, two broad categories can be identified: distance-based (also called nonparametric) and model-based (parametric) techniques [38]. In this section, we focus only on the aspects relevant to the choice of methods for the problem of interest, i.e., grouping thermal processing of bitumen.

In general, distance-based methods can often take advantage of optimization techniques but model-based methods can often find clusters of arbitrary shape and which makes them more flexible. Model-based methods also require less computational effort relative to distance-based methods. Studies have shown that as the number of variables (dimensionality) increases, the effectiveness of the distance-based methods to obtain optimum solution decreased significantly [39]; also, the notion of distance as a similarity metric becomes meaningless for high-dimensional data [40]. On the other hand, in model-based algorithms, which depend on the ratio of probability densities, the probability that the two close individuals end up in the same cluster approaches zero as number of variables greatly exceeds the number of samples[41-43].

Several methods have been suggested to overcome the curse of dimensionality[40, 44, 45]. For example, Bayesian hierarchical clustering was found to be effective in clustering high dimensional datasets such as DNA microarray data[39, 46-48]. This algorithm uses a Bayesian approach to clustering, with priors for model parameters and for the allocation of samples to groups, with the priors chosen so that the marginal posterior is analytically tractable. The marginal posterior is used as the natural measure of similarity or dissimilarity, i.e., the appropriateness of grouping. The clustering that maximizes the marginal posterior is taken to be optimal. In order to simplify the computational burden, the maximum *a posteriori* clustering over all possible partitions is estimated using the agglomerative path; this also provides a visual guide to some of the other possible data allocations through a dendrogram (Nia et al.[38]).

In the Bayesian clustering paradigm, the allocation of a variable to a given cluster is viewed as a statistical parameter; hence, with a Bayesian model for the data conditioned on the grouping structure and a prior distribution for the clusters, a search algorithm can be applied to obtain the maximum *a posteriori* grouping [49, 50].

IF a data allocation, K, clusters the individual samples into K groups of sizes $T_1, \ldots, T_k$, there is a total of $T = \sum_{k=1}^{K} T_k$ clustering individuals. Then, assuming a multinomial-Dirichlet distribution as a location prior [51],

$$f(K) \propto \frac{(K-1)!T_1!, \ . \ . \ ,T_{k!}}{T!(T+K-1)!} \tag{2-1}$$

The clustering posterior, which is the grouping criterion for clustering, is

$$f(x \mid K) = \frac{f\left(x \mid K\right)f(K)}{\beta} \tag{2-2}$$

where $f\left(x \mid K\right)$ is the marginal density of the data for the known allocation K, β is the normalization constant that plays no role in agglomerative clustering and can be omitted, and x denotes the vector of data.

In the first step in Bayesian agglomerative clustering, we start with each individual sample as a single cluster. Then, all pairwise merges are considered followed by calculation of the posterior for each pairwise merge (equation 2-2). The merge that maximizes the posterior is applied. The log posterior for the best merge having k clusters, P_k = $\log f\left(x \mid K\right)$, is used as the dendrogram height. If the best merge according to equation 2 is to join cluster k_1 to k_2 to create the new cluster k, then of course $T_k = T_{k_1} + T_{k_2}$.

The algorithm then considers all pairwise merges again, and continues until all clusters are merged and all individuals are in one cluster[39, 46]. The best grouping found using the posterior as the objective function on the agglomerative path is the one that maximizes P_k across k = 1, … , K. The groupings associated with P_k are sorted in agglomerative order with increasing c, so a dendrogram representation is possible [51].

2.4.3 Causality detection: Bayesian learning approach

Causality detection and analysis has become an important tool in process monitoring for the detection and diagnosis of plant-wide abnormalities and disturbances [52]; it has also found use in other fields such as forensic science [53]. The most commonly used methods for causality detection are Granger causality, transfer entropy and Bayesian networks[54-56]. The transfer entropy method measures the amount of directed information transferred from the cause variable to the effect, so the direction of information transfer dictates the direction of causality and the amount indicates the strength of the causal link[57]. In the Granger causality detection approach, the improvement gained in predicting the effect due to the incorporation of the cause variable as a predictor is used as indication of a causal relationship [58]. The Bayesian approach, on the other hand, measures the conditional probability of the effect given the occurrence of the cause[59, 60]. Studies indicated that Bayesian method outperforms the other two approaches when the data length is short (i.e., with a smaller number of samples);[61, 62] this better suits our case.

The first step in causality detection in a Bayesian framework is learning the topology of the causal network, commonly called the Bayesian network (BN), followed by parameter learning. The basic assumption behind Bayesian structure learning is that the data is assumed to be generated from an underlying probability distribution and this distribution in turn is induced by some Bayesian network structure[63]. The learning objective is to determine how to accurately recover the underlying causal map or BN. There are two fundamental problems in finding the correct network structure. The first issue is that there are many perfect structures for a probability distribution; thus, the best we can hope for is to get the method that enables us to recover the equivalence class for the structure[64]. The second problem is that data comes rarely without noise, which means that we need to consider trade-offs between the fit and generalizability.

A number of algorithms have been proposed in the literature for recovering structure learning from data. Despite the range of theoretical backgrounds and terminology, they fall under two major broad categories: constraint-based and score-based. Alternatively, the network structure can be built manually from the domain knowledge of a human expert and prior information available on the data.

The constraint-based approach attempts to recover a network structure that best captures the dependencies in the domain. The literature indicates that this approach can accurately capture the underlying model structure when the data has few variables, large samples and the variables are strongly dependent[63]. However, a single misleading independence test result can produce multiple errors. Besides, the algorithm is and is strongly dependent on the independence tests and can run in to trouble when the independence test results are less pronounced[63, 65].

On the other hand, score-based algorithms view structure learning as a general heuristic optimization problem or alternatively as a model selection problem. This approach hypothesizes that there is a space of possible candidate structures for a given set of data and a scoring function that measures the how well the model/structure fits the data[63, 66, 67]. Compared to the constraint-based algorithm, the score-based algorithms are less sensitive to individual failure and can also compromise between fit the data and scalability. Hence, we focus on this approach.

A detailed review of the various score based algorithms was presented in Russell and Norvig[68]. The most widely used score-based algorithms are the greedy search algorithms such as hill-climbing with random restarts or tabu-search. They explore the search space starting from a network structure, usually the empty graph, and adding, deleting, or reversing one arc at a time until the score can no longer be improved [69, 70]. The selection of the score function is the most important component in score based learning. The most commonly used score functions are the maximum log-likelihood and Bayesian score or the Bayesian information criterion (BIC). The maximum log-likelihood is often effective fitting the data but prone to overfitting[63,76], while the

Bayesian score can enable trade-offs between the likelihood of fitting the data and model complexity (equation 2-3).

$$score_{BIC} = \sum_{i=1}^{n} log f_{X_i}\left(X_i \,\middle|\, \Pi_{X_i}\right) - \frac{M}{2}\log(n) \qquad (2\text{-}3)$$

where M is the number of parameters in the network and n is the sample size.

The first term on the right hand side of the equation is the marginal likelihood and the second term controls the complexity of the structure and provides the trade-off between fit and complexity.

2.4.4 Overall procedure for modeling

Figure 2-1 describes the steps used for the developing Bayesian network structure and the information used for validation of the models. The first step in the learning process was to perform data preprocessing which includes smoothing, scaling, centering and detecting outlying samples using principal component analysis (PCA). Then Bayesian agglomerative hierarchical clustering was used for clustering preprocessed data into similar functional groups, hence the name group is used instead of cluster. Once clustering was performed, the groups were checked for within-group similarity and intergroup dissimilarity using information from standard spectrometric identification handbooks [71, 72, 73]. This was used as a reality check for the validation of the results of the cluster analysis.

Each group is a node in Bayesian learning framework, but there are different intensity values and wavenumbers in a group. A group ideally is a representative of a class of compounds with similar major functional groups.

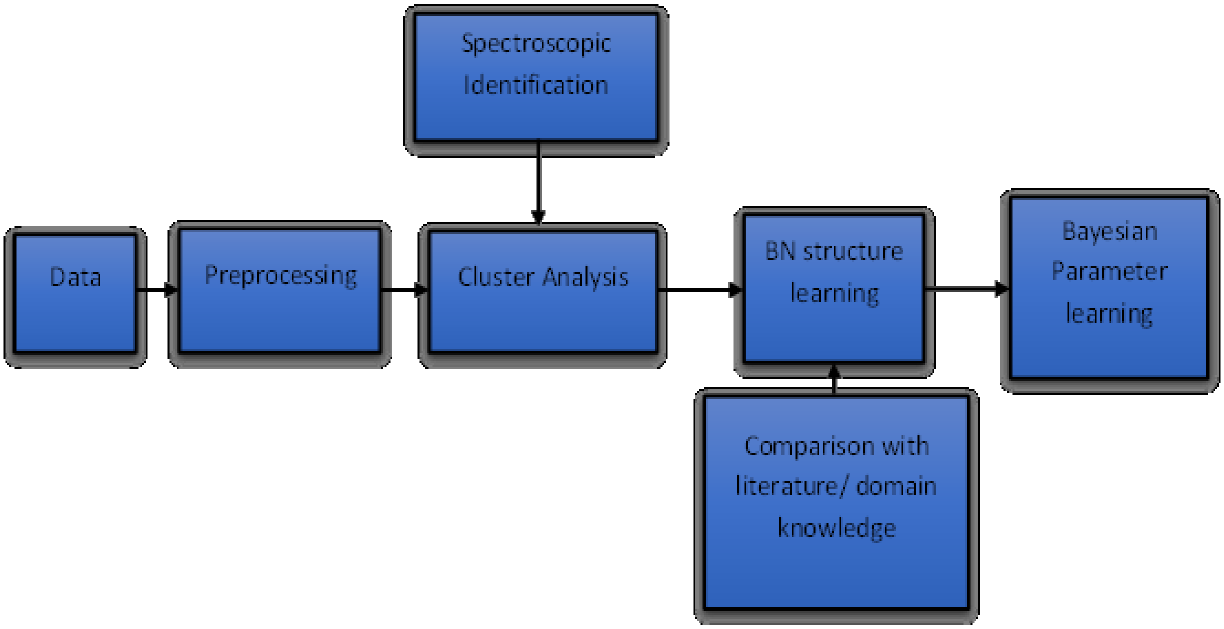

Spectroscopic Identification
Data
Preprocessing
Cluster Analysis
BN structure learning
Bayesian Parameter learning
Comparison with literature/ domain knowledge

Figure 2-1. A flow diagram for the algorithm used.

The next step is to use Bayesian learning to recover the network structure form the group data and to estimate the corresponding probability distribution/parameters for the intensity of the groups. The intensity associated with each cluster was calculated as the root- means squared average of the intensities at all wavenumbers that were present in the group. Finally, the resulting Bayesian network is compared against the possible reaction mechanisms. All the simulation/coding was done in MATLAB version 9.0.0[81] and R version 3.3.1.

2.5. RESULTS AND DISCUSSION

2.5.1 Infrared spectra

The infrared spectra that were collected are shown in relation to the solid yield for each sample in Figure 2-2. Many of the spectra are of thermally processed products that were associated with minor formation of organic deposits, having solid yields of 1–2 wt%. The spectrum with 0.9

wt% solid yield is the bitumen feed.

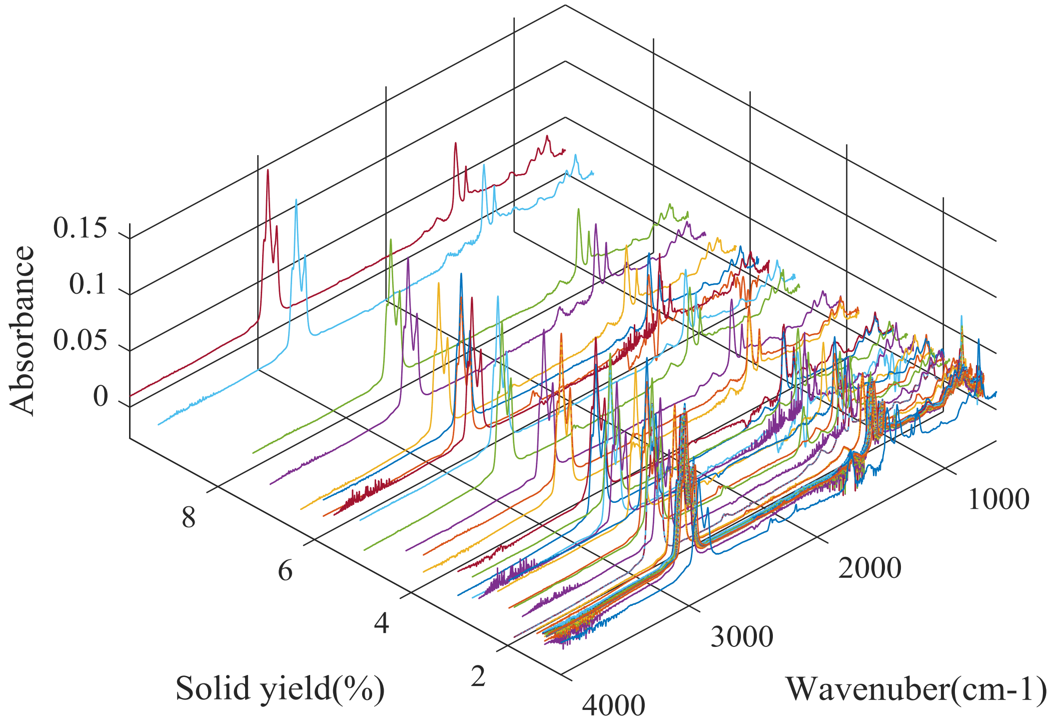


Figure 2-2. Fourier transforms infrared (FTIR) spectroscopic data of thermally processed bitumen samples.

Figure 2-3 shows the spectra of raw bitumen and a sample treated at 300ºC and for 6 hr. Clearly, the heated treated sample has much more absorbance at 2854, 2922, 2954, 1371 and 1455 cm^{-1}. This indicates that the heat-treated sample contains more saturated groups than untreated bitumen sample.

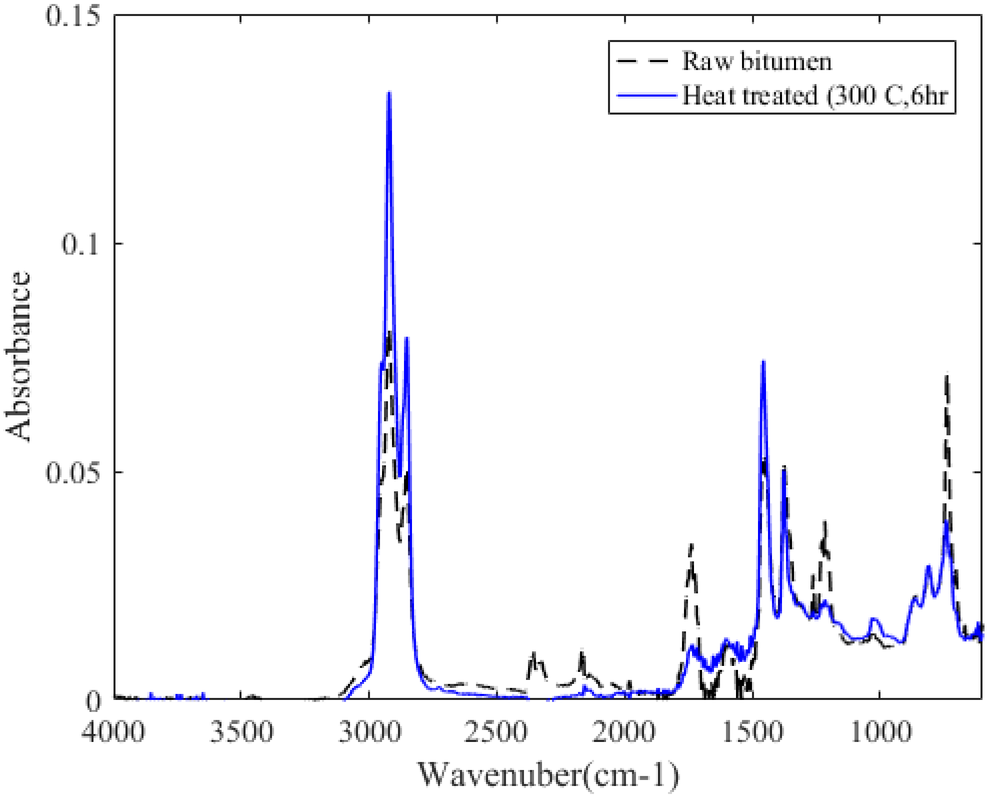


Figure 2-3. Comparison of raw bitumen spectra with sample treated at 300°C and for 6 hr.

2.5.2 Solid yield

Figure 2-4 shows the solid yield of the thermally cracked bitumen samples. The result indicates that the highest solid yield is observed for sample treated at 300 °C. In general, this result indicates that the solid yield is a none-linear function of reaction time.

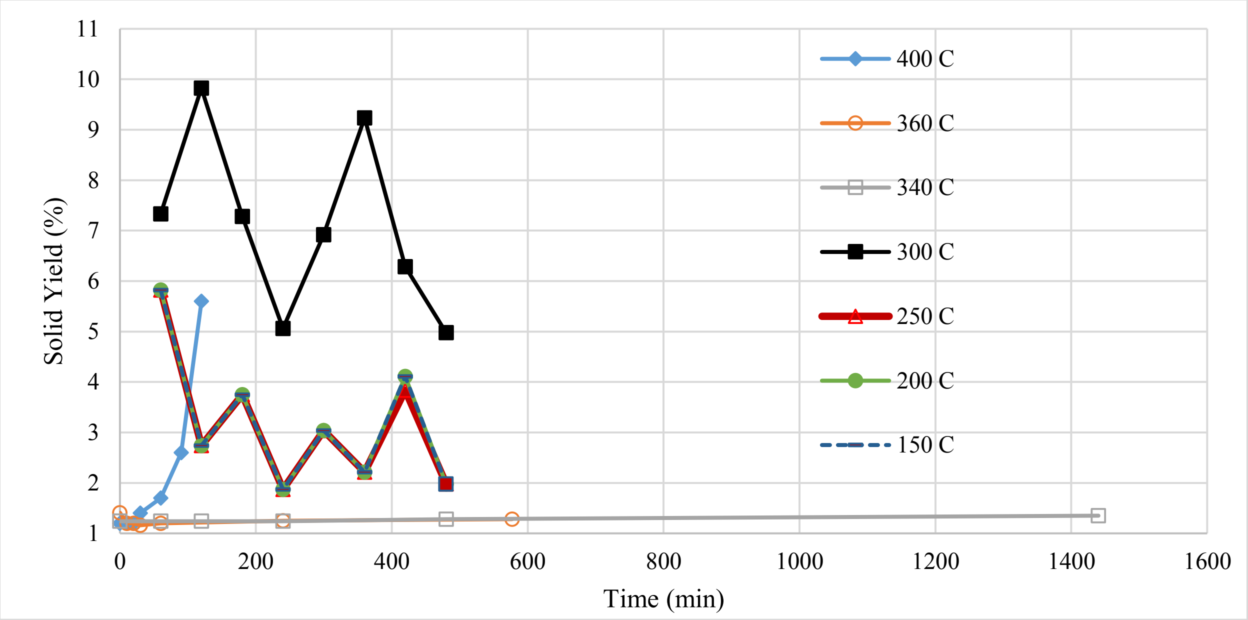


Figure 2-4. Variation of solid yield, in weight (%), with reaction time.

2.5.3 Principal component analysis

Principal component analysis (PCA) was performed on the infrared spectra. It was found that only two components were required to explain about 94% of the variation in the data. The score plot using these two components Figure 2-5 can be used to detect outlying samples.

It can be seen from Figure 3 that only two of the samples, the liquid products from thermal processing at 200 °C for 1 and 2 hours, were different from the rest of the data. These two samples were retained in the dataset, because the Bayesian agglomerative clustering employed is less sensitive to outliers, as mentioned before.

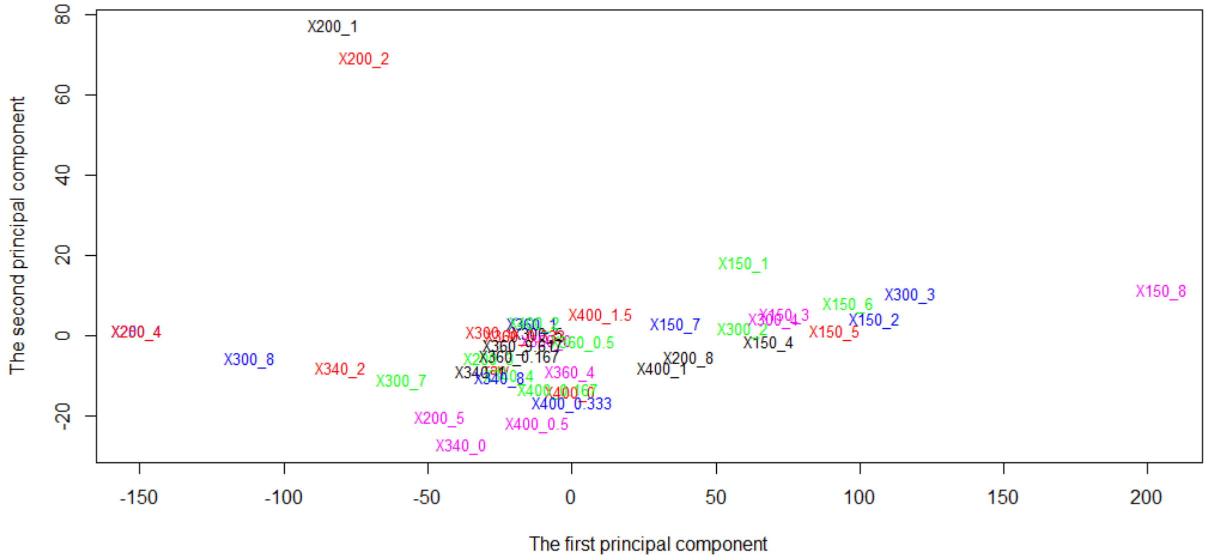


Figure 2-5. Principal component analysis score plot for the first two principal components (which explain 94% of the variance. PC1 and PC2 are represents the first and the second principal components. The legends denote the variables sample labels. For example, X200_1 and X400_0.33 represent the spectra of a samples treated at 200°C and 1 hr. and 400°C and 0.33hr respectively.

2.5.4 Clustering

Cluster analysis is used as a preprocessing method for Bayesian learning, the groups obtained were used as the variables/nodes for Bayesian learning. Clustering is similar to the concept of lumping kinetic methods where pseudo-components or lumps are generated based on some similarity criteria such as boiling point and other physicochemical properties common to the specific group[74]. In this work, functional groups and their characteristic wavenumbers are the measure of similarity of a group or pseudo-component and dissimilarity between different pseudo-components. Due to the size of the data (about 2000 variables, representing the wavenumbers), the dendrogram, Appendix A, FigureA1 has little visual information to offer; however, we choose to select five groups in total at the appropriate level of the dendrogram.

This decision on the number of groups is based on the fact that the major products of cracking heavy feedstock are mostly paraffin, alkenes, cycloalkanes, aromatics and polyaromatics, which are produced due to coking. Disregarding the heteroatoms, generally the conversion will result in the variation of those major functional groups during the thermal cracking of bitumen. Tables A1-A5 in the Appendix show the wave numbers included in each of the clusters.

In order to know the functional in each group, several sources, including handbooks of spectrometric identification of organic compounds,[75-77] were consulted. Each group consists of various wavenumbers and a functional group is assigned to the group as long as the wavenumber at which it absorbs exist. While this attribution is approximate it allows us to provide a chemical basis for the clustering and acts as a reality check for our mathematical analysis.

Table 2-1 shows the organic functional groups attributed to each group. The first group consists mainly of aromatics (including polyaromatics) and alkanes (cyclic and normal). The normal alkanes may refer to the side chains. The second group is a mixture of alkanes, aromatics and carbonyl groups. The third group includes aromatics and hetroaromatics, cycloalkanes and alkanes. The fourth group consists mainly of cycloalkanes and alkanes whereas the fifth group contains Paraffin. From the group analysis it can be seen that the aliphatic nature of the groups increase from the first to the fifth group. These groups s are then used as the variables in Bayesian network learning to generate the reaction network, which is described in the next subsection.

Table 2-1. Clusters and the constituent functional groups, the assignment is based on several sources[75-78].

Cluster/group	1	2	3	4	5
Possible functional groups	Aromatics and poly-aromatics; cycloalkanes Carboxylic: dimers, aromatic and conjugated esters, Thiocarbonyl groups and sulfoxides. Amides (tertiary)	Alkanes, Aromatics carbonyls, unsaturated alcohols and phenols, mercaptans, thiophenols, thio-acid and sulfoxides	Alkanes, Aromatics, and hetroaromatics, $\gamma -$ CH and ring bending. $\beta -$ Ring bands of: Pyridines, Furans and Thiophene. $\beta -$ Substituted naphthalene (four adjacent hydrogens). Amides (tertiary)	Alkanes; and, may have cycloalkane, normal and branched alkanes	Paraffin

2.5.5 Correlation analysis

The last subsection described the groups obtained from clustering the infrared spectra of the bitumen samples. The aim of this subsection is to get a better view of the variables obtained from clustering. Figure 2-6 describes the linear cross-correlation of the groups, qualitatively and quantitatively on the lower and upper panel, respectively. From the results, it can be seen that group4 and group5 have the most correlated variables, 97%, while group1 and group5 are the least correlated, 64%. Generally, the correlation matrix shows a nice correlation among the variables but it should be noted that correlation does not mean causality. The variables can be correlated but one may not be the cause of the other. Acyclic Bayesian learning is used to discover the causal relationship between the variables. Causal relationship is the same to the concept of reaction, which is why this method is select to uncover the reaction network between these variables.

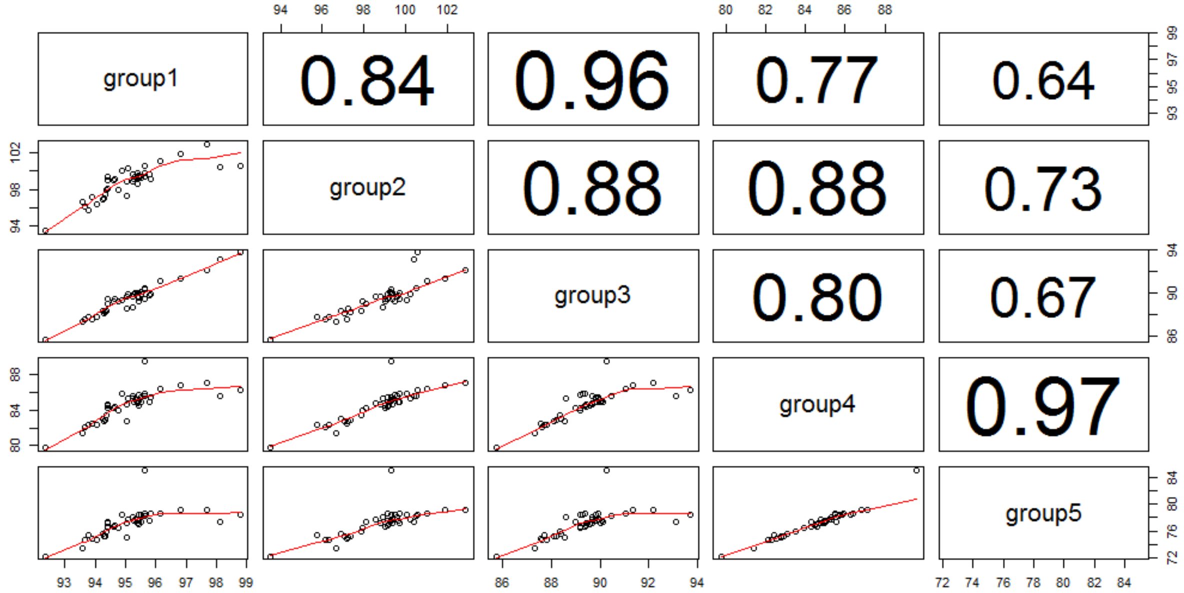


Figure 2-6. Person's cross-correlation plot of the groups, diagonal shows variable name (groups).The linear correlation of each variable with the other is shown qualitatively (graphically) and quantitatively on the lower and upper panel respectively. The horizontal and vertical axis both show intensities.

2.5.6 Bayesian network (BN) structure

In the topology of the BN, nodes that are connected together by the arcs represent the variables (groups). The arcs describe the direct causal influences. In the discussion about BN, the family metaphor is often used i.e. a node is a parent of a child if there is an arc from the former to the latter. In a directed chain of nodes, one node is an ancestor of another if it appears earlier in the chain, whereas a node is a descendant of another node if it comes later in the chain.

Most often more than one learning method are used to learn the BN topology to ensure optimality of the solution. Theoretically, if two or more different learning approaches provide the same structure, most often the solution is considered optimal. For this reason, two greedy search algorithms, Hill climbing with random restarts and Tabu search were implemented.

Hill climbing (HC) with random starts with initial structure, empty, performs an iterative HC by re-starting with a random new initial condition and if a new run of HC provides the solution that increase the score function than the previously discovered optimal structure then the new one replaces the old, otherwise the old solution will maintained and the search goes on. [79] Tabu search, on the other hand, starts with a feasible initial solution and chooses the next best move that can significantly increases the score function while introducing Tabu restrictions on possible moves to daunt the reversal and repetition of selected moves.[80]

Figure 2-7 shows graphical representation of the optimal solution when form HC with random restart is applied.

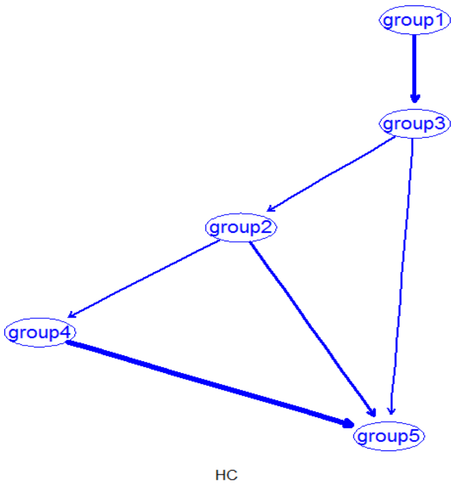


Figure 2-7. Bayesian causal network structure modeled using HC.

Table 2-2 shows the corresponding arc strength for HC graph.

Table 2-2. Arc strength for Bayesian network structure modeled using HC

From	To	arc-strength
group 4	group5	-87
group 1	group3	-53
group 2	group5	-32
group3	group2	-29
group2	group4	-29
group 3	group5	-3

Figure 2-8 describes the graphical model obtained using Tabu search.

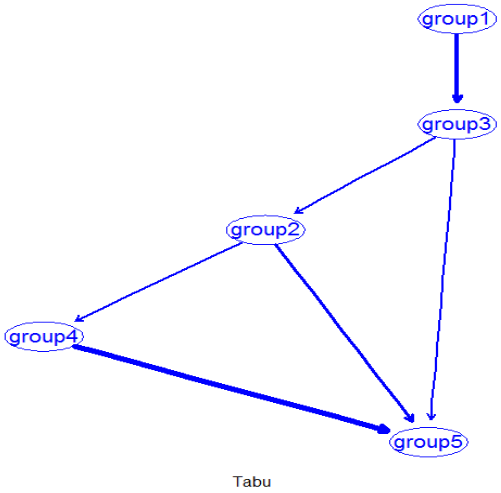


Figure 2-8. Bayesian causal network structure derived using Tabu search.

Table 2-3 illustrates the strength values of each arc in the Tabu search-based BN.

Table 2-3. Arc strength of Bayesian network developed using Tabu search.

From	To	arc-strength
group 4	group5	-87
group 1	group3	-53
group 2	group5	-32
group3	group2	-29
group2	group4	-29
group 3	group5	-3

From the above results, it can be seen that both resulted in exact the same solution for the network (Figure 2-7 and Figure 2-8), the same score, and the same arc strength, as shown in Table 2-1 and Table 2-3. This may indicate that the solution is indeed a global optimum.

The final directed acyclic graph (DAG), either of the two (Figure 2-7 or Figure 2-8), represents the causal relationships between the groups. From the DAG group1 is directly related to group3 but no direct relationship to the other groups. In the same way, group3 is the family of group2 and group5. While group4 has a single family, group2, group5 has three families, group2, group3, and group4. Hence, in this case any change in group2, group3, and group4 will directly affect the state of group5. Moreover, given group2, group3, and group4, the state of group5 is independent of group1. This applies to all family–child relationships in the graph.

In addition to the graphical representation of the causal relationship between the variables, BN also models the quantitative strength of the connections between variables (Table 2-2 and Table 2-3). Strength indicates the probabilistic beliefs about arcs to be updated automatically as new information becomes available. From this point of view, the arcs from group4 to group5 and group3 to group5 have the highest probability of updating themselves as new information become available relative to the others. In other words, the arc strength values represent the change in the score of the network that can result by an arc removal. For example, the overall score of the DAG will decrease by 87 if the arc from group4 to group5 is removed. Clearly, the strongest dependence is found between group4 and group5 while the weakest is from group3 to group5. From the reaction point of view, different arc strengths indicate the preferred reaction pathways or more probable reactions (the higher the negative number, the stronger the dependencies and vice versa).

2.5.7 Model Validation

This section presents the comparison of the developed model against the reaction network of model compound representing bitumen. Validation of this model is tricky particularly because the reaction mechanism of the process investigated; low temperature visbreaking is not well understood. Moreover, the groups used as variables for Bayesian networks are not necessarily single compounds. However, it can be seen from the graphical model representation (the DAG) that the final product, group5, is relatively saturated compared to the intermediate and the starting material. This indicates that the reaction generally leads to a hydrogen rich product, which was also observed in the experimental analysis of the reaction elsewhere.[3, 33] In the following subsection, the DAG is compared to a reaction mechanism hypothesized based on expert process knowledge.

2.5.7.1. Reaction network

The reaction network of thermal processing of bitumen is illustrated using an alkyl tricyclic naphtheno-aromatic compound (1), where R can be either hydrogen, or a more complex aliphatic and/or aromatic structure Figure 2-9 . Although no heteroatoms have been shown, the network can also describe the reactions of heteroatom containing molecules.

During thermal processing hydrogen transfer between molecules, can either lead to a net decrease in hydrogen to convert cycloalkane structures into aromatic structures, (1) to (2), or to saturation of aromatic structures, (1) to (3). Free radical addition, where stabilization does not take place by addition of hydrogen (H•) as shown, but by addition of a larger radical (R•), is not explicitly shown. Thermal cracking of weaker bonds in a cycloalkane structure, shown by intermediate (4), can lead to different reaction products. Intramolecular hydrogen transfer leads to (5) and intermolecular hydrogen transfer leads to (6). The intermediate (4) can also undergo further free radical cracking before intermolecular hydrogen transfer to stabilize the products to produce (7) and (8). The longer alkyl chain in (5) and (6) is susceptible to thermal cracking in an analogous way, as shown for the cracking of (6) to (9) and (10). Molecules present in the bitumen feed that have alkyl chains analogous to (6) will react in a similar way.

From the perspective of generalizing the reaction network, the most obvious outcome of thermal processing is the increase in lighter and more aliphatic products. This comes at the expense of generating more aromatic products, which ultimately end up as the organic deposits.

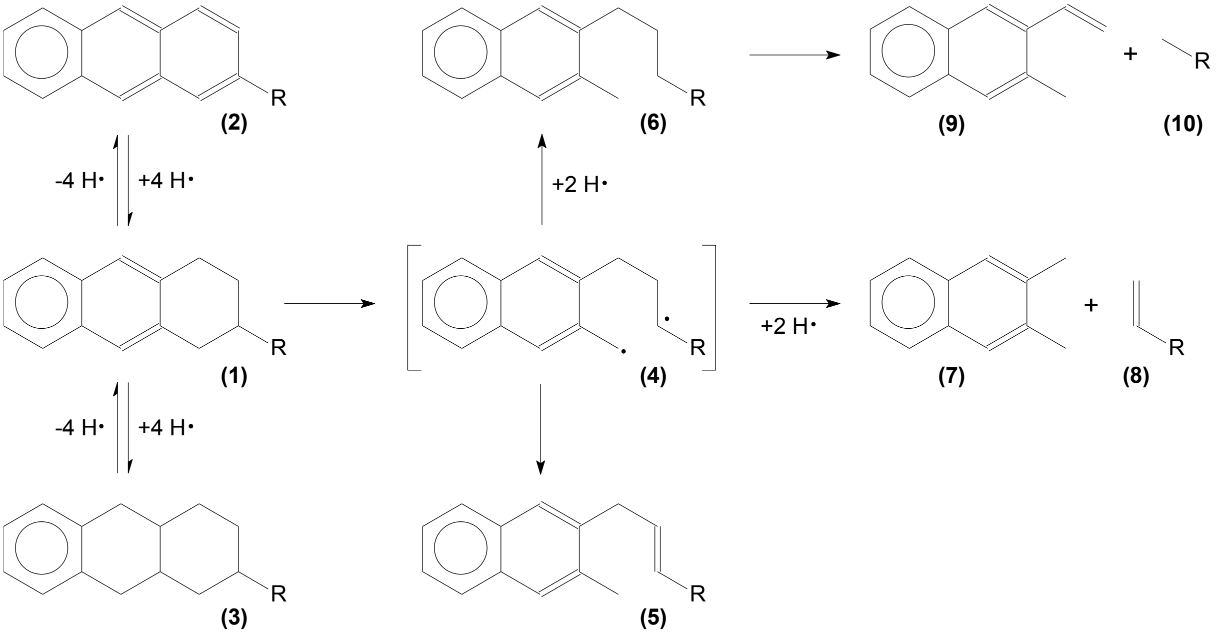


Figure 2-9. Hypothesized reaction network involving hydrogen disproportionation, hydrogen transfer and thermal cracking reactions.

Comparing the DAG (Figure 2-7) with the reaction network in Figure 2-9, one obvious similarity is that both predict the lighter and aliphatic final product. Comparing the components in Figure 2-9 with the nodes in Figure 2-8, group1 is similar to component (1) and group3 is similar to component (4). In the DAG, it was found that group1 causes group3; similarly, in Figure 2-9, the reaction of component (1) produces component (4). From the DAG, group3 also produces group5 and from the reaction network, component (4) can produce component (10) through the thermal cracking of component (5). Likewise, components (6) and (9) can represent group2, the reaction of which produces the paraffinic component (10). In the reaction network in Figure 2-7, group4 is not represented explicitly. However, component (10) can represent larger alkanes that can break down to smaller alkanes. The other possible explanation is that the intermediate in the reaction network, component (9), can undergo hydrogenation to produce a compound class similar to group4, which will finally crack to give component (10).

In summary, there is a very good agreement between the proposed reaction network and the graphical model developed through the Bayesian learning approach.

2.5.7.2. Sensitivity analysis

Inn addition to validation, sensitrivity analysis can also help to evaluate the accuracy of the model. A good model reacts correctly to changes in the process conditions. From the application point of view, this is equivalent to using the DAG as an expert system to predict the effect of different reaction conditions. In order to evaluate the response of the model to different reaction conditions, the data was divided into two sections based on the solid yield (low, <2 wt% and high, >2%). The arc strengths were calculated using each data set separately. The solid yield was selected to break the data into two sets because it represents the combined effect of resisdence time and temperature.

Figure 2-10 shows the changes in the arc strengths when the first dataset, low coke yield, is used. The result shows that less saturate will be produced under this circumstance compared to the base case (Figure 2-7).

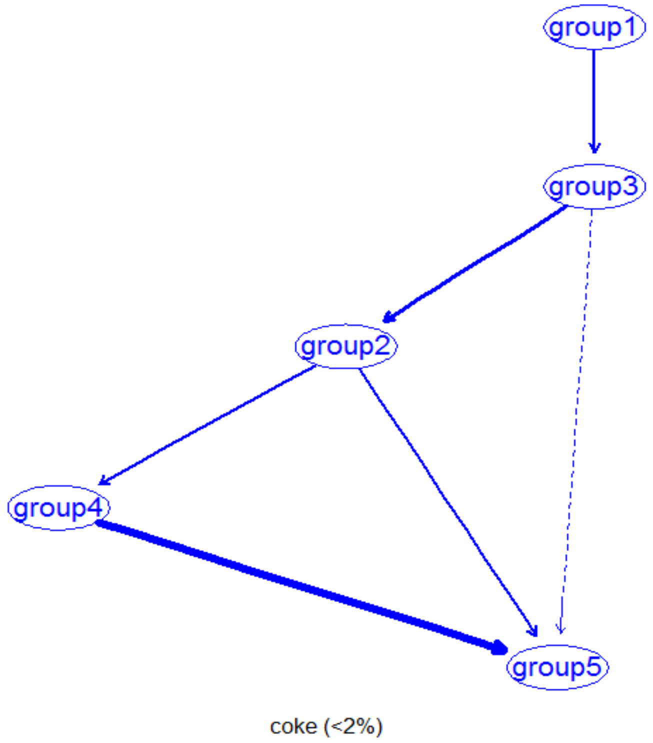


Figure 2-10. Effect of solid yield on the significance of the reaction pathways (low solid yield, <2%).

Table 2-4 shows the values of the strength of each arc for the low solid yield data. It can be seen from the strength of the arc from group3 to group5 is positive. Positive arc strength indicates that the network gains score if this branch is removed, and it is less likely to produce paraffin product from the reaction of group3.

Table 2-4. Arc strength illustrating the effect of low solid yield

From	To	arc-strength
group 4	group5	-51
group 1	group3	-22
group 2	group5	-16
group3	group2	-25
group2	group4	-10
group 3	group5	2

In the second case, data corresponding to the reactions with relatively higher coke yield were used to calculate the strength of the DAG. Figure 2-11shows the changes in the arc strengths when the second dataset, higher coke yield, is used. The result shows that more saturates will be produced under this condition compared to the first case (Figure 2-10).

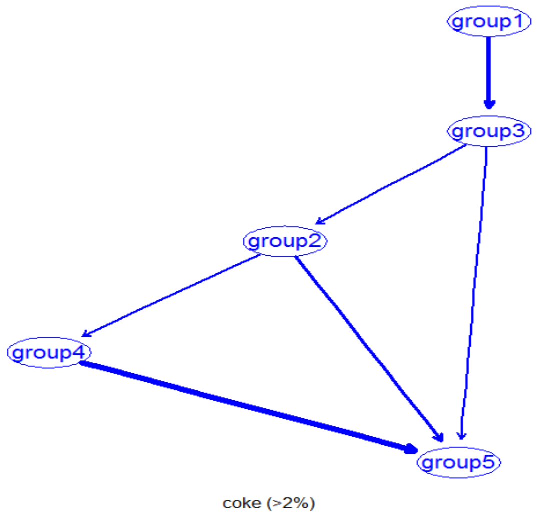


Figure 2-11. Effect of solid yield on the significance of the reaction pathways (high solid yield, >2%)

Table 2-5 describes the arc strength of the BN when high solid yield data is used. It can be seen that the strength of the arc from group3 to group5 is negative, and more than the base case scenario. This shows production of more group5 compared to the case of low solid yield.

Table 2-5. Arc strength illustrating the effect of high solid yield

From	To	arc-strength
group 4	group5	-59
group 1	group3	-32
group 2	group5	-25
group3	group2	-12
group2	group4	-9
group 3	group5	-14

2.5.8 Parameter Learning

In modeling a reactive system, once the reaction network is developed or the product and the reactants are identified, a reaction rate equation can be developed using the corresponding experimental data and the assumptions about the order. The reaction rate equation provides a quantitative description of the interaction between variables describing the reaction process. Similarly, a Bayesian network is formally defined as, $BN = \langle DAG, \Theta \rangle$, where DAG is as described earlier and Θ represents the set of parameters of the network. DAG is analogous to the reaction network and Θ is similar to the reaction rate equations and parameters. Once the topology of the BN is specified, the next step is to quantify the relationships between connected nodes and this is called parameter learning. Parameter learning is the process of specifying the conditional probability distributions and probabilistic beliefs of each nodes. The most important assumption in the Bayesian learning approach for parameter learning is that the DAG models all the direct dependencies between the variable of the system. In addition, every variable is independent of its non-descendants given the state of its parent and there are no direct dependencies.

In this study, the DAG described the directed causal map between the variables (groups) and the conditional probability distribution of each group estimates a model for the mean value of the intensity as the function of the intensity of the other groups. On top of the model for the mean, the conditional probability distribution describes the degree of uncertainty about the model used to estimate the mean. Hence, parameter learning in Bayesian framework is similar to the classical reaction rate modeling with the quantification of the uncertainty about the model.

Now, let X_j denote the intensity value of jth in the DAG (Figure 5) and μ_j be the mean value of X_j where j=1, 2, 3, 4, 5. The conditional probability distribution of each group is described in (equations 2-4 to 2-8) and the model for the mean intensity of the groups is presented in (equations2-9 to 2-13). These equations are the pseudo-kinetic equations for the graphical model, DAG.

$P\,(X_1) \sim N\,(\mu_1, 0.0056^2)$ (2-4)

$P\,(X_3|X_1) \sim N\,(\mu_3, 0.0020^2)$ (2-5)

$P\,(X_2|X_3) \sim N\,(\mu_2, 0.0037^2)$ (2-6)

$$P\,(X_4|X_2) \sim N\big(\mu_4,\, 0.0043^2\big) \tag{2-7}$$

$$P\,(X_5|X_2, X_3, X_4\,) \sim (\mu_5,0.0020\,^2) \tag{2-8}$$

$$\mu_1 = 0.0213 \tag{2-9}$$

$$\mu_3 = 1.2570\mu_1\text{-}22.91 \tag{2-10}$$

$$\mu_2 = 0.73143\mu_3 - 0.0292 \tag{2-11}$$

$$\mu_4 = 1.225\mu_2 + 0.0652 \tag{2-12}$$

$$\mu_5 = 1.70702\mu_4 - 0.4488\mu_2 - 0.3184\mu_3 - 0.0003 \tag{2-13}$$

2.5.9 Application to visbreaking

To address the application of the model to visbreaking, it is useful to restate the challenges. The main challenge in modeling the reaction mechanism of the visbreaking process at the described condition was the limited prior knowledge of the reaction mechanism to use the classical lumping kinetic modelling approach. Moreover, experimental studies indicated that the reduction in product viscosity could be by reactions other than just thermal cracking.

In order to overcome this problem and estimate the most plausible reaction model of the process a method based on data mining and Bayesian learning was designed. Lumping was performed using clustering of spectroscopic data which required limited prior knowledge about the system. The lumps (groups) obtained from the cluster analysis described the variation of the system very well. Then, the reaction network and reaction equations were developed using Bayesian learning. The network was compared against the reaction network for a model compound and the developed model described the reaction very well. Hence, one most important advantage of the proposed method was that it requires limited prior information to provide a very good approximation of the reaction model.

The developed DAG indicates that estimates the reaction of the process well compared to the model reaction and the experimental observations, the reaction network generally shows the formation of saturate material with reaction progression. Moreover, the DAG is a self-updating reaction network, which can predict the reaction network, reaction rate, for new operation condition. The model can be deployed online to monitor the real time effect of different reaction conditions. This is useful to investigate the effect of different operation conditions.

2.5.10 Conclusions

A reaction network and the rate models were modeled for low temperature visbreaking of using a Bayesian learning approach. First, Bayesian agglomerative hierarchical cluster analysis was implemented on spectroscopic data to obtain clusters or pseudo-components. Then, the Bayesian network and parameter learning approach were used to develop the kinetic model. The model described the reaction mechanism and model compound reaction very well. The kinetic model can be used for the online monitoring of the visbreaking process using FTIR spectroscopic data.

2.6. References

1. Gary, J. H.; Handwerk, G. E.; Kaiser, M. J., *Petroleum refining: technology and economics*. CRC press: 2007.
2. Wang, L.; Zachariah, A.; Yang, S.; Prasad, V.; de Klerk, A., Visbreaking oilsands-derived bitumen in the temperature range of 340–400° C. *Energy & Fuels* **2014,** *28*, (8), 5014-5022.
3. Yañez, L.; de Klerk, A., Visbreaking oilsands bitumen at 300 °C. *Energy & Fuels* **2016,** *60*, (1), 31-34.
4. Strausz, O. P.; Lown, E. M. *The chemistry of Alberta Oil Sands, bitumen and heavy oils*; Alberta Energy Research Institute: Calgary, AB, 2003; p 695.
5. H. Ali, L.; A. Al-Ghannam, K.; M. Al-Rawi, J., Chemical structure of asphaltenes in heavy crude oils investigated by n.m.r. *Fuel* **1990,** *69*, (4), 519-521.
6. Gray, M. R., *Upgrading oilsands, bitumen and heavy oil*. 2014.
7. Brons, G.; Yu, J. M., Solvent Deasphalting Effects on Whole Cold Lake Bitumen. *Energy & Fuels* **1995,** *9*, (4), 641-647.
8. Blanchard, C. M.; Gray, M. R., Free radical chain reactions of bitumen residue. *ACS Division of Fuel Chemistry, Preprints* **1997,** *42*, (1), 137-141.
9. Moschopedis, S. E.; Parkash, S.; Speight, J. G., Thermal decomposition of asphaltenes. *Fuel* **1978,** *57*, (7), 431-434.
10. Zachariah, A.; Wang, L.; Yang, S.; Prasad, V.; de Klerk, A., Suppression of Coke Formation during Bitumen Pyrolysis. *Energy & Fuels* **2013,** *27*, (6), 3061-3070.
11. Rana, M. S.; Sámano, V.; Ancheyta, J.; Diaz, J. A. I., A review of recent advances on process technologies for upgrading of heavy oils and residua. *Fuel* **2007,** *86*, (9), 1216-1231.
12. Beleites, C.; Bonifacio, A.; Codrich, D.; Krafft, C.; Sergo, V., Raman spectroscopy and imaging: Promising optical diagnostic tools in pediatrics. *Current Medicinal Chemistry* **2013,** *20*, (17), 2176-2187.
13. Ferreira, A. P.; Tobyn, M., Multivariate analysis in the pharmaceutical industry: Enabling process understanding and improvement in the PAT and QbD era. *Pharmaceutical Development and Technology* **2015,** *20*, (5), 513-527.

14. Héberger, K., Chemoinformatics-multivariate mathematical-statistical methods for data evaluation. In *Medical Applications of Mass Spectrometry*, 2008; pp 141-169.

15. Jansen, J. J.; Smit, S.; Hoefsloot, H. C. J.; Smilde, A. K., The photographer and the greenhouse: How to analyse plant metabolomics data. *Phytochemical Analysis* **2010,** *21*, (1), 48-60.

16. Paczkowska, M.; Lewandowska, K.; Bednarski, W.; Mizera, M.; Podborska, A.; Krause, A.; Cielecka-Piontek, J., Application of spectroscopic methods for identification (FT-IR, Raman spectroscopy) and determination (UV, EPR) of quercetin-3-O-rutinoside. Experimental and DFT based approach. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2015,** *140*, 132-139.

17. Li, G.; Torraca, G.; Jing, W.; Wen, Z. q., Applications of FTIR in identification of foreign materials for biopharmaceutical clinical manufacturing. *Vibrational Spectroscopy* **2009,** *50*, (1), 152-159.

18. Grewer, D. M.; Young, R. F.; Whittal, R. M.; Fedorak, P. M., Naphthenic acids and other acid-extractables in water samples from Alberta: What is being measured? *Science of the Total Environment* **2010,** *408*, (23), 5997-6010.

19. Zhao, B.; Currie, R.; Mian, H., Catalogue of Analytical Methods for Naphthenic Acids Related to Oil Sands Operations. In *Oil Sands Research and Information Network, University of Alberta, School of Energy and the Environment, Edmonton, Alberta. OSRIN Report No. TR-21*, 2012.

20. Al-Ghouti, M. A.; Al-Degs, Y. S.; Amer, M., Determination of motor gasoline adulteration using FTIR spectroscopy and multivariate calibration. *Talanta* **2008,** *76*, (5), 1105-1112.

21. Hua, H.; Dubé, M. A., Terpolymerization monitoring with ATR-FTIR spectroscopy. *Journal of Polymer Science, Part A: Polymer Chemistry* **2001,** *39*, (11), 1860-1876.

22. Darsy, G.; Bouzat, F.; Muñoz, M.; Lucas, R.; Foucaud, S.; Diogo, C. C.; Babonneau, F.; Leconte, Y.; Maître, A., Monitoring a polycycloaddition by the combination of dynamic rheology and FTIR spectroscopy. *Polymer (United Kingdom)* **2015,** *79*, 283-289.

23. Calabro, D. C.; Valyocsik, E. W.; Ryan, F. X., In situ ATR/FTIR study of mesoporous silicate syntheses. *Microporous Materials* **1996,** *7*, (5), 243-259.

24. Deng, H.; Shen, Z.; Li, L.; Yin, H.; Chen, J., Real-time monitoring of ring-opening polymerization of tetrahydrofuran via in situ Fourier Transform Infrared Spectroscopy. *Journal of Applied Polymer Science* **2014,** *131*, (15).

25. Pintar, A.; Malacea, R.; Pinel, C.; Fogassy, G.; Besson, M., In situ monitoring of catalytic three-phase enantioselective hydrogenation using FTIR/ATR spectroscopy. *Applied Catalysis A: General* **2004,** *264*, (1), 1-12.

26. Wong, K. C., Review of Spectrometric Identification of Organic Compounds, 8th Edition. *Journal of Chemical Education* **2015,** *92*, (10), 1602-1603.

27. Li, G.; Jing, W.; Wen, Z. Q., Identification of unknown mixtures of materials from biopharmaceutical manufacturing processes by microscopic-FTIR and library searching. *American Pharmaceutical Review* **2011,** *14*, (7), 60-64.

28. Nyden, M. R.; Pallister, J. E.; Sparks, D. T.; Salari, A., computer-assisted spectroscopic analysis using orthonormalized reference spectra. Part ii: application to the identification of pure compounds. *Applied Spectroscopy* **1987,** *41*, (1), 63-66.

29. Prats-Montalbán, J. M.; de Juan, A.; Ferrer, A., Multivariate image analysis: A review with applications. *Chemometrics and Intelligent Laboratory Systems* **2011,** *107*, (1), 1-23.

30. Abdolmaleki, A.; Ghasemi, J. B.; Shiri, F.; Pirhadi, S., Application of multivariate linear and nonlinear calibration and classification methods in drug design. *Combinatorial Chemistry and High Throughput Screening* **2015,** *18*, (8), 795-808.

31. Arvanitoyannis, I. S.; Katsota, M. N.; Psarra, E. P.; Soufleros, E. H.; Kallithraka, S., Application of quality control methods for assessing wine authenticity: Use of multivariate analysis (chemometrics). *Trends in Food Science and Technology* **1999,** *10*, (10), 321-336.

32. Currie, L. A., Detection: International update, and some emerging di-lemmas involving calibration, the blank, and multiple detection decisions. *Chemometrics and Intelligent Laboratory Systems* **1997,** *37*, (1), 151-181.

33. Visbreaking of Oilsands Bitumen between 150 and 300 oC. In.

34. Jackson, D. A.; Chen, Y., Robust principal component analysis and outlier detection with ecological data. *Environmetrics* **2004,** *15*, (2), 129-139.

35. James, G.; Witten, D.; Hastie, T.; Tibshirani, R., *An introduction to statistical learning*. Springer: 2013; Vol. 6.

36. Nia, V. P.; Davison, A. C., High-dimensional Bayesian clustering with variable selection: The R package bclust. *Journal of Statistical Software* **2012,** *47*, (5), 1-22.

37. Bishop, C. M., *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc.: 2006.

38. Nia, V. P.; Davison, A. C., High-dimensional Bayesian clustering with variable selection: The R package bclust. *Journal of Statistical Software* **2012,** *47*.

39. Swartz, M. D.; Mo, Q.; Murphy, M. E.; Lupton, J. R.; Turner, N. D.; Hong, M. Y.; Vannucci, M., Bayesian variable selection in clustering high-dimensional data with substructure. *Journal of Agricultural, Biological, and Environmental Statistics* **2008,** *13*, (4), 407-423.

40. Steinbach, M.; Ertöz, L.; Kumar, V., The Challenges of Clustering High Dimensional Data. In *New Directions in Statistical Physics*, Wille, L., Ed. Springer Berlin Heidelberg: 2004; pp 273-309.

41. Ahn, J.; Marron, J. S.; Muller, K. M.; Chi, Y. Y., The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika* **2007,** *94*, (3), 760-766.

42. Hall, P.; Marron, J. S.; Neeman, A., Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **2005,** *67*, (3), 427-444.

43. Ausloos, M.; Ivanova, K., Patterns, Trends and Predictions in Stock Market Indices and Foreign Currency Exchange Rates. In *New Directions in Statistical Physics*, Wille, L., Ed. Springer Berlin Heidelberg: 2004; pp 93-114.

44. Jung, S.; Marron, J. S., PCA consistency in High Dimension, Low Sample Size context. *Annals of Statistics* **2009,** *37*, (6 B), 4104-4130.

45. Yata, K.; Aoshima, M., Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis* **2012,** *105*, (1), 193-215.

46. Tadesse, M. G.; Sha, N.; Vannucci, M., Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association* **2005,** *100*, (470), 602-617.

47. Crandell, J. L.; Dunson, D. B., Posterior simulation across nonparametric models for functional clustering. *Sankhya: The Indian Journal of Statistics* **2011,** *73*, (1 B), 42-61.

48. Lian, H., Sparse Bayesian hierarchical modeling of high-dimensional clustering problems. *Journal of Multivariate Analysis* **2010,** *101*, (7), 1728-1737.

49. Darkins, R.; Cooke, E. J.; Ghahramani, Z.; Kirk, P. D. W.; Wild, D. L.; Savage, R. S., Accelerating Bayesian Hierarchical Clustering of Time Series Data with a Randomised Algorithm. *PLoS ONE* **2013,** *8*, (4).

50. Savage, R. S.; Heller, K.; Xu, Y.; Ghahramani, Z.; Truman, W. M.; Grant, M.; Denby, K. J.; Wild, D. L., R/BHC: Fast Bayesian hierarchical clustering for microarray data. *BMC Bioinformatics* **2009,** *10*.

51. Heard, N. A.; Holmes, C. C.; Stephens, D. A., A quantitative study of gene regulation involved in the immune response of Anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. *Journal of the American Statistical Association* **2006,** *101*, (473), 18-29.

52. Duan, P.; Chen, T.; Shah, S. L.; Yang, F., Methods for root cause diagnosis of plant-wide oscillations. *AIChE Journal* **2014,** *60*, (6), 2019-2034.

53. Evett, I. W.; Gill, P. D.; Jackson, G.; Whitaker, J.; Champod, C., Interpreting small quantities of DNA: The hierarchy of propositions and the use of bayesian networks. *Journal of Forensic Sciences* **2002,** *47*, (3), 520-530.

54. Marques, V. M.; Munaro, C. J.; Shah, S. L. In *Data-based causality detection from a system identification perspective*, 2013 European Control Conference, ECC 2013, 2013; 2013; pp 2453-2458.

55. Lee, S.-M.; Abbott, P. A., Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers. *Journal of Biomedical Informatics* **2003,** *36*, (4–5), 389-399.

56. Sugihara, G.; May, R.; Ye, H.; Hsieh, C. H.; Deyle, E.; Fogarty, M.; Munch, S., Detecting causality in complex ecosystems. *Science* **2012,** *338*, (6106), 496-500.

57. Schreiber, T., Measuring Information Transfer. *Physical Review Letters* **2000,** *85*, (2), 461-464.

58. Granger, C. W. J., Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **1969,** *37*, (3), 424-438.

59. Smith, V. A., Revealing structure of complex biological systems using bayesian networks. In *Network Science: Complexity in Nature and Technology*, 2010; pp 185-204.

60. Jung, A., Learning the Conditional Independence Structure of Stationary Time Series: A Multitask Learning Approach. *IEEE Transactions on Signal Processing* **2015,** *63*, (21), 5677-5690.

61. Zou, C.; Feng, J., Granger causality vs. dynamic Bayesian network inference: A comparative study. *BMC Bioinformatics* **2009,** *10*.

62. Zou, C.; Denby, K. J.; Feng, J., Granger causality vs. dynamic Bayesian network inference: A comparative study. *BMC Bioinformatics* **2009,** *10*.

63. Koller, D.; Friedman, N., *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press: 2009; p 1208.

64. Heckerman, D., A tutorial on learning with Bayesian networks. In *Studies in Computational Intelligence*, 2008; Vol. 156, pp 33-82.

65. Neapolitan, R. E., *Learning Bayesian Networks*. Prentice-Hall, Inc.: 2003.

66. De Campos, L. M., A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research* **2006,** *7*, 2149-2187.

67. Liu, J. W.; Li, H. E.; Luo, X. L., Learning technique of probabilistic graphical models: A review. *Zidonghua Xuebao/Acta Automatica Sinica* **2014,** *40*, (6), 1025-1044.

68. Russell, S. J.; Norvig, P., *Artificial Intelligence: A Modern Approach*. Pearson Education: 2003; p 1132.

69. Nagarajan, R.; Scutari, M.; Lbre, S., *Bayesian Networks in R: with Applications in Systems Biology*. Springer Publishing Company, Incorporated: 2013; p 170.

70. Taroni, F.; Aitken, C.; Garbolino, P.; Biedermann, A., *Bayesian Networks and Probabilistic Inference in Forensic Science*. 2006; p 1-354.

71. Fuchs, O., *Colthup,nb - Introduction to infrared and raman spectroscopy*. 1967; Vol. 228, p 42.

72. Ksandr, Z., Introduction to infrared and raman spectroscopy. *Chemicke Listy* **1966,** *60*, (5), 687.

73. Silverstein, R. M.; Bassler, G. C.; Morrill, T. C., *Spectrometric identification of organic compounds*. 1974; p 340-340.

74. Stagni, A.; Cuoci, A.; Frassoldati, A.; Faravelli, T.; Ranzi, E., Lumping and reduction of detailed kinetic schemes: An effective coupling. *Industrial and Engineering Chemistry Research* **2014,** *53*, (22), 9004-9016.

75. Silverstein, R. M.; Webster, F. X.; Kiemle, D.; Bryce, D. L., *Spectrometric identification of organic compounds*. John Wiley & Sons: 2014.

76. Shriner, R. L., Systematic identification of organic compounds. **1956**.

77. Colthup, N., *Introduction to infrared and Raman spectroscopy*. Elsevier: 2012.

78. Odebunmi, E.; Adeniyi, S., Infrared and ultraviolet spectrophotometric analysis of chromatographic fractions of crude oils and petroleum products. *Bulletin of the Chemical Society of Ethiopia* **2007,** *21*, (1).

79. Selman, B.; Gomes, C. P., Hill-climbing Search. In *Encyclopedia of Cognitive Science*, John Wiley & Sons, Ltd: 2006.

80. Bai, X.; Padman, R., Tabu search enhanced markov blanket classifier for high dimensional data sets. In *The Next Wave in Computing, Optimization, and Decision Technologies*, Springer: 2005; pp 337-354.

81. MATLAB Version: 9.0.0.341360. Natick, Massachusetts: The MathWorks Inc., 2016

3. A SELF- MODELING MULTIVARIATE CURVE RESOLUTION MODEL FOR ONLINE MONITORING OF BITUMEN UPGRADING USING FOURIER TRANSFORM-INFRARED SPECTROSCOPY

A version of this chapter is to be submitted as; Dereje Tamiru Tefera, Ankit Agrawal, Lina Maria Yanez Jaramillo, Arno de Klerk and Vinay Prasad, "A Self- Modeling Multivariate Curve Resolution model for online Monitor of Bitumen Pyrolysis of using Fourier Transform-Infrared Spectroscopy (FTIR)" respectively.

3.1. INTRODUCTION

Spectroscopic methods have been widely used in chemical industries for various purposes, including new product development, process performance improvement and real time monitoring.[1, 2] The main reason behind the growing interest for the use of spectroscopic methods over conventional process analysis methods is that they are generally fast, reliable, non-invasive, and cost-effective; it is also possible to use them to deduce physical parameters[3]. These methods can be used in either laboratory, pilot-plant or full-scale production experiments involving identification of new reaction chemistries, new processes and products [4]. If used in lab scale experiments, the information obtained is usually used in decision making, developing and commercializing a new product or even rejecting a new product from commercial consideration[5].

Online spectroscopy has also gained increasing acceptance as a choice for industrial chemical process control as it can provide rapid and representative information such as composition and quality information directly for a complex chemical process, which otherwise are difficult/hazardous to obtain using direct measurement [6, 7], and this information is obviously valuable in process and product quality control[5, 6, 8, 9]. However, the large amount of useful information contained in spectroscopic data is often difficult to extract mainly because absorption bands from individual chemical constituents often experience a high degree of overlap, the samples are often chemically complex (for example, reactions involving heavy oils) and susceptible to a large number of non-chemical interferences. As a result, it is often necessary to enhance the information retrieval process by using multivariate statistical methods. Chemometrics methods, and specifically self-modeling multivariate curve resolution (SMCR), are well suited for improving the effectiveness of online spectroscopic techniques by extracting useful information by deconvolving the data and facilitating the automation of online analytical techniques[10-13].

The principal advantage of SMCR over other widely used chemometrics techniques, such as partial least squares regression, principal component regression, or multiple linear regression is that it only requires a small amount of quantitative data[7, 10, 14].

The main objective in this work is to design a SMCR model for automatic prediction of the number of components or chemical rank and resolved concentration and spectral profiles of the FTIR spectroscopic data acquired from thermally processed oil sands-derived bitumen samples with a view to use the model for online spectroscopy monitoring of similar processes. The processing of bitumen offers a significant challenge for online monitoring since it is a highly complex mixture with incomplete characterization of the chemical species present. While conversion is easily defined for reactions with pure components and can be used to monitor the progress of reaction, it is not easy to define or determine for complex mixtures; often, arbitrary measures such as the liquid yield or yield of components boiling below a certain temperature are used[15, 16]. However, we show that the SMCR model can be used to develop a monitoring scheme for this case. The samples collected were analyzed offline, but the method and model developed are suitable for online monitoring.

3.2. METHODS: SELF-MODELING MULTIVARIATE CURVE RESOLUTION (SMCR)

Self-modeling curve resolution, in a broader sense, is a method similar to principal component analysis (PCA), independent component analysis (ICA), evolving factor analysis (EFA) and its derivatives, and multivariate curve resolution (MCR),[17] that helps to resolve complex mixtures into pure components where there is no or very little prior knowledge about the system [18]. SMCR is different from factor analysis techniques such as PCA which only produce an abstract decomposition of the experimental data to maximize the explained variance in the data [19], in that it forces the solution (concentration and spectral profiles) to follow chemically and physical meaningful constraints. The fundamental premise in multivariate curve resolution is that spectra of the mixtures are the linear combination of the pure spectra and concentrations of the chemical species in the mixture and that Beer's law is valid. On this ground, the spectral data of the mixture sample, D, can be modeled as

$$\mathrm{D} = \mathrm{CS}^{\mathrm{T}} + \mathrm{E} \qquad 3\text{-}1$$

where D is the NxM data matrix, C is a NxK matrix of the concentration profile of K components, S is the MxK matrix of spectral profiles of the K components and E is the NxM noise matrix.

3.2.1 Chemical Rank and Initial condition

As stated before, the first and the key step in SMCR-based analysis of process data is to determine the number of active chemical species in the sample, i.e., the chemical rank. For spectroscopic and/or chromatographic data, active chemical species refer to species that can absorb in the desired wavelength range, have distinguishable spectra and take part in the process; for instance, if a species does not change concentrations,[20] it does not contribute to the rank. In an idealized and noise free situation, the chemical rank is equal to the rank of the data matrix D[21]. In other words, the chemical rank of the systems reacting in bulk may not exceed the number of chemical reactions plus one [22]. For real data, however, the instrumental noise and other experimental errors make it difficult to identify the true chemical rank. In this context, chemical rank estimation is the process of identifying the relevant chemical information from the background noise or in the presence of species that do not take part in the process. There are an innumerable number of models and methods proposed to recover the number of linearly independent and chemically relevant factors from two-way or multi-way data from samples of chemical mixtures [23]. The major methods can be classified into two broad classes; namely, methods that require full knowledge of the experimental error and approximate methods requiring no knowledge of experimental error [23]; the details of these methods can be obtained elsewhere[20, 24]. Most of these methods are based on visual analysis of quantities such as the logarithm of the eigenvalues, etc.[20], and are usually less accurate and are difficult to automate[9, 13, 25]. An important factor is robustness, i.e., the ability of the model to provide accurate estimates in the presence of background noise, especially when the instrument noise is not constant. After a comparative analysis of various factor analysis methods, Elbergali *et Al.* [26] found out that the ratio of the second and third derivatives of the empirical function known as the indicator function (IND) derived by Malinowski [27] accurately predicted the number of relevant factors even for data without uniform background noise. Later, Wasin *et al.* [24] performed another critical analysis of rank determination techniques (applied to spectroscopic and chromatographic data) and also arrived at the same conclusion. They confirmed that the maximum of the ratio of the second and third derivatives (ROD) and the minimum of IND both accurately estimated the chemical rank of this data regardless of the noise level. For brevity, we refer the reader to Malinowski [23] and references within for details and present only the relevant quantities. Once PCA is performed, the error (the difference between the original data and the predicted data using the first L principal components) can iteratively calculated using (Eqn:3.2) by varying the number of PCA components from 1 to the number of original variables in the data . It should also be noted that the sum of square of elements of a given dataset is equal to the sum of the squares of the elements of the matrix of its corresponding principal components. Once the error function is calculated then IND and ROD are evident (Eqn:3-3 and 3-4 respectively).

$$\sum_{\mathrm{j=L+1}}^{\mathrm{N}} e_j = \sum_{\mathrm{i=1}}^{N} \sum_{i=1}^{M} \mathrm{d_{ij}}^2 - \sum_{\mathrm{i=1}}^{\mathrm{L}} \sum_{i=1}^{N} \mathrm{t_{ij}}^2 \tag{3-2}$$

$$\mathrm{IND(L)} = \frac{\sqrt{\frac{\sum_{\mathrm{j=L+1}}^{\mathrm{N}} e_j}{\mathrm{M(N-L)}}}}{\mathrm{(N-L)^2}} \tag{3-3}$$

$$\mathrm{ROD(L)} = \frac{\mathrm{IND(L-2)-IND(L-1)}}{\mathrm{IND(L-1)-IND(l)}} \tag{3-4}$$

where e_j , $\mathrm{d_{ij}}$, $\mathrm{t_{ij}}$ are error of predicting the data using the first ‘L’ PCA components, the element of data matrix, D, and elements of matrix of Principal components respectively. N and M and D are as defined earlier.

Therefore, the whole procedure of rank determination is reduced to performing PCA/SVD and evaluation of the second and third derivative of ROD(K) and locating its maxima or alternatively the minima of IND (K). One important consequence of this analysis is the estimation of the experimental error in the data, which will be very helpful in the evaluation of the performance of the curve resolution process [23].

The next step is to determine the initial condition for the least squares projection method. For this, several methods have been used, including evolving factor analysis (EFA) [28], window factor analysis (WFA) [29], orthogonal projection (OPA)[30] and simple-to-use interactive self-modeling mixture analysis (SIMPLSM) [31], etc. EFA is the most commonly used method and it exploits the inherent evolutionary structure in the data registered along the line of process change (with time, for example) to determine the local rank at each step. EFA creates a submatrix starting with the first spectrum and performs a subsequent PCA on gradually increasing submatrices, enlarged by adding one new row at a time and calculating a new set of eigenvalues. Once the forward procedure is done, EFA is also performed in the reverse order and the combined results are used for further analysis. Some of the commonly used techniques for the estimation of an initial concentration profile are combined forward and backward EFA concentration profiles (the smaller of each forward/backward pair is used); result of none iterative EFA and resolving factor analysis (RFA) [32], all with equivalent performance. The latter methods need to determine the importance level (threshold of concentration profile) which can be affected by the noise level) and in the case of the non-iterative EFA, one has to define zero concentration regions to start with (the same drawback as with the case of the importance level) and one has to perform another least squares analysis. The first method is easy to automate, and no user intervention, which makes it an ideal candidate for automation purposes [28]. For this reason and the relatively lower required computational effort, combined forward and backward EFA is used to estimate the initial concentration profile. The subsection provides a brief description of the least square based projection method.

3.2.2 Self-modeling multivariate curve resolution by alternating least squares (SMCR-ALS)

Self-modeling multivariate curve resolution by alternating least squares is by far the most popular and potent curve resolution method. Besides the simplicity and the low computational effort required, the introduction of ALS as an optimization method enabled SMCR to incorporate any data-specific constraint [1, 18, 33]. The ability to incorporate prior knowledge about the process and computational simplicity together with an automated procedure to predict the chemical rank of the data and initial condition the SMCR-ALS an ideal candidate for an online monitoring system for a complex reacting system such as is encountered in the thermal processing of bitumen. In short, SMCR-ALS is the minimization of the Frobenius norm of the residual in the search for the best fit to the data in a least squares sense[18, 34, 35], alternating between equations (3-5) and (3-6).

$$\left.\begin{array}{l}\min\limits_{C}\left(\|\mathrm{D}-\mathrm{CS}^{\mathrm{T}}\|_{2}{}^{2}\right)\\ s.t{:}\,S\geq 0\,,\;D=CS^{T}\end{array}\right\} \qquad 3\text{-}5$$

$$\left.\begin{array}{l}\min\limits_{C}\left(\|\mathrm{D}-\mathrm{CS}^{\mathrm{T}}\|_{2}{}^{2}\right)\\ s.t\ C\geq 0\ D=CS^{T}\end{array}\right\} \qquad 3\text{-}6$$

The flow chart in Figure 3-1 summarizes the algorithm used in this study for online monitoring of thermal processing of bitumen. The computation starts with singular value decomposition (SVD)/PCA and ROD determine the chemical rank. EFA estimates the initial concentration, which ALS uses to resolve the data D to active chemical species.

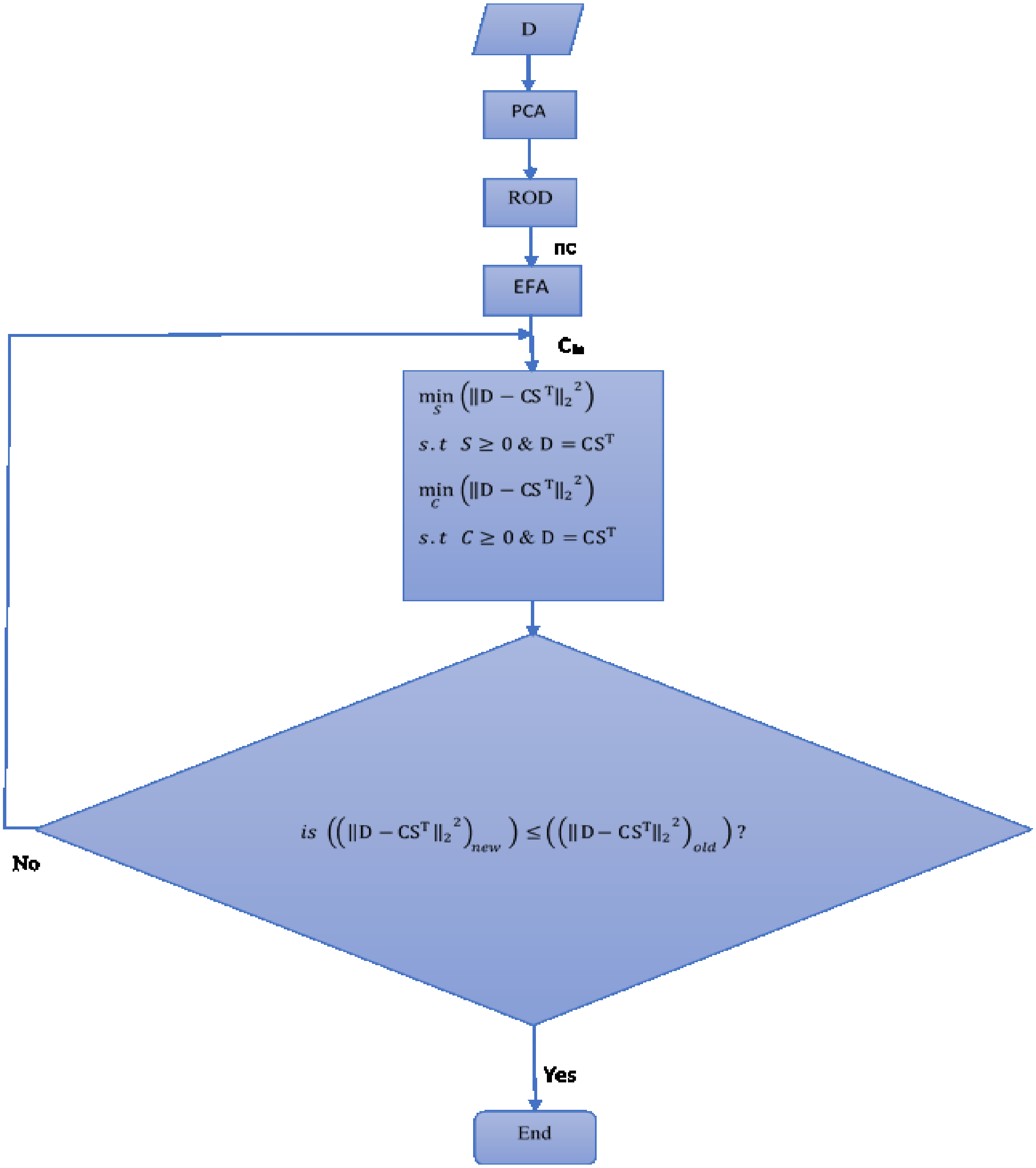

D
PCA
ROD
nc
EFA
C_{in}
$\min_S \left(\|D - CS^T\|_2^{\ 2}\right)$
$s.t\ \ S \geq 0 \ \&\ D = CS^T$
$\min_C \left(\|D - CS^T\|_2^{\ 2}\right)$
$s.t\ \ C \geq 0 \ \&\ D = CS^T$
$is\ \left(\left(\|D - CS^T\|_2^{\ 2}\right)_{new}\right) \leq \left(\left(\|D - CS^T\|_2^{\ 2}\right)_{old}\right)?$
No
Yes
End

Figure 3-1. Flow chart for SMCR-ALS algorithm; nc and C_{in} are the number of components and initial concentration profile, respectively. PCA, ROD and EFA are principal component analysis, the ratio of the second and third derivatives of the indicator function and Evolving factor analysis. D, S C are as defined earlier. C_{in} and nc represent the initial concentration and the number of components respectively.

3.2.3 Band assignment and quantitative parameters

In order to facilitate the analysis in the following section, the most commonly used IR regions for the interpretation of the FTIR spectra are summarized in Table 3-1 [36-41]. This is important and will be used as reference in the investigation of the resolved profiles.

Table 3-1. Summary of commonly used band assignments, [c] the peak centers are assigned within a spectral widow of ±10 wavenumbers.

[c] Peak center (wavenumber cm^{-1})	Assignment
700- 900	C-H out-of-plane bending
720	Aliphatic CH_2
750	Aromatic, four neighbouring C-H
820	Aromatic, two neighbouring C-H
870	Aromatic, isolated C-H
1450-1460	Aliphatic (±aromatic) CH_2 and CH_3 bending
1500-1800	Carbon-carbon and carbon-oxygen stretching
1575	Carboxyl COOH
1610	Aromatic C=C (±shited CO)
1630	Aromatic C=C(± phenol OH)
1635	Olefin C=C
1650	Carbonyl (quinone) C=O
1710	Carbonyl (ketone) C=O
1770	Carbonyl (ester) C=O and C-O
2750-3150	C-H stretching
2857	Aliphatic CH_2, symmetric
2872	Aliphatic CH_3, symmetric
2897	Aliphatic CH
2925	Aliphatic CH2, symmetric
2962	Aliphatic CH3, symmetric
3050	Aromatic C-H

In addition to the qualitative analysis of the IR spectra based on band assignment, several semi-quantitative parameters have been used to assess the effect of aging, reaction conditions and other process changes on the characteristic structure of coal, kerogen, bitumen, asphaltenes and other similar classes of materials [36-38, 42, 43]. Some of these parameters are used here to get a deeper insight into the effect of reaction conditions and to compare resolved components from the SMCR analysis.

The first parameter is the ratio of the intensity of asymmetric stretching between methylene and methyl groups (CH_2— and CH_3—). The nCH_2/nCH_3 ratio is used to measure the effect of reaction conditions on the average length of the aliphatic chain in the samples [44, 45 46] and is defined as

$$nCH_2/nCH_3 = \frac{I_{2922}}{I_{2954}} \qquad 3\text{-}7$$

or

$$nCH_2/nCH_3 = \frac{I_{2922} + I_{2954}}{I_{2954}} \qquad 3\text{-}8$$

where I_{2922} and I_{2954} are the absorbance at 2922 and 2954cm^{-1}.

A low (CH_2/CH_3) value generally indicates lower average length of the aliphatic chain in the sample while a higher value may imply either longer aliphatic chains or may also imply higher cycloalkanes content of the sample [47].

The other commonly used parameter, the degree of aromatic condensation (DOC), is used to measure the relative content of the condensed aromatic structures in the sample [37]. $A_{1550-1630}$ and $A_{700-900}$ represent the area under the peak over the range of frequencies of 1550-1630 and 700-900cm^{-1} respectively. The higher the value of DOC the higher is the condensed structure.

$$DOC = \frac{A_{1550-1630}}{A_{700-900}} \qquad 3\text{-}9$$

The third parameter, the C-factor, is used to assess the change in the oxygenated functional groups versus the aromatic ring functional group[42]. In equation 3.10 the I's, I_{1742}, I_{1603}, show the absorbance values at 1603 and 1742 cm^{-1} respectively.

$$C - factor = \frac{I_{1742}}{I_{1603} + I_{1742}} \qquad 3\text{-}10$$

Data Description and Preprocessing

Bitumen consists of different classes of compounds, namely saturates, aromatics, substituted aromatic and polyaromatic groups [48], mixed naphthene-aromatic rings[49] and other heteroatoms such as sulfur, oxygen and nitrogen. The IR spectrum shows characteristic peaks that correspond to modes of those functional groups.

FTIR data was obtained within (600 to 4000 cm^{-1}) portion of the infrared region with a resolution of 2 cm^{-1} for samples obtained from the thermal reaction of bitumen under different reaction conditions (time varying from 0.5 to 8 hrs. and temperature varying from 150^{o}C to 400^{o}C), with the details of the experimental procedure being reported elsewhere[50-52].

Figure 3-2 shows all the data together and consists of 43 FTIR spectra with 1765 spectral channels. The saturates have abundant bands at 2854, 2922, 2954, 1371 and 1455 cm^{-1}. Moreover, the band regions (1550-1630 cm^{-1}), (700-900 cm^{-1}), (3000-3150 cm^{-1}), (1630-1670 cm^{-1}), (1670-1800 cm^{-1}), (2750-3000 cm^{-1}) and (1000-1300 cm^{-1}) designate aromatic C=C ring stretching vibration; aromatic C-H out-of-plane bending vibration; aromatic C-H stretching vibration [37, 40, 42, 53, 54]; alkene group; oxygenate groups; aliphatic group and C-O functional group [55] respectively.

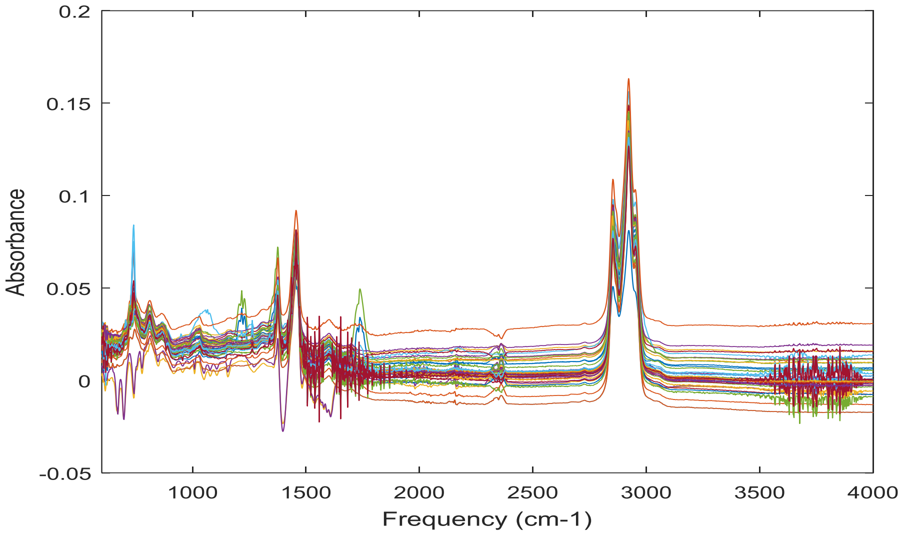


Figure 3-2. FTIR spectra of all the thermally treated bitumen samples.

Before the analysis, the data is baseline corrected and the high frequency signals were filtered out. Figure 3-3 shows one sample smoothed dataset (the first dataset at 150°C); the plots include the raw data, its smoothed counterpart and the corresponding residual. The residual consists of only high frequency signals and probably represents the instrument noise. The analysis of all the other data is done in similar fashion and is left out for brevity.

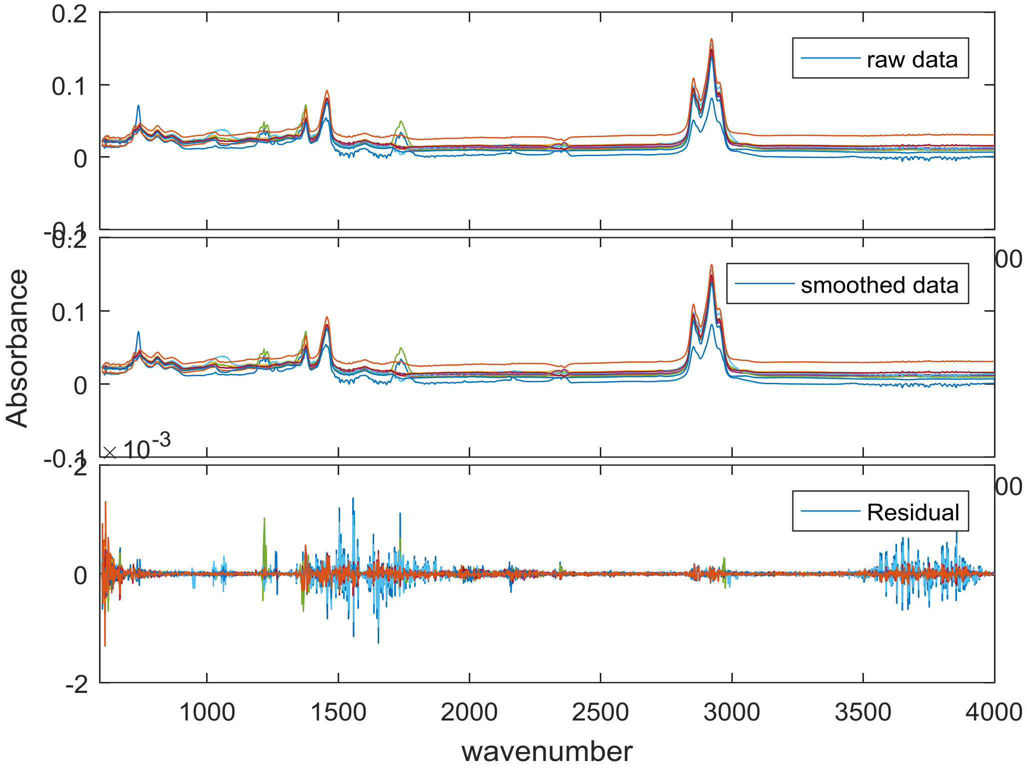
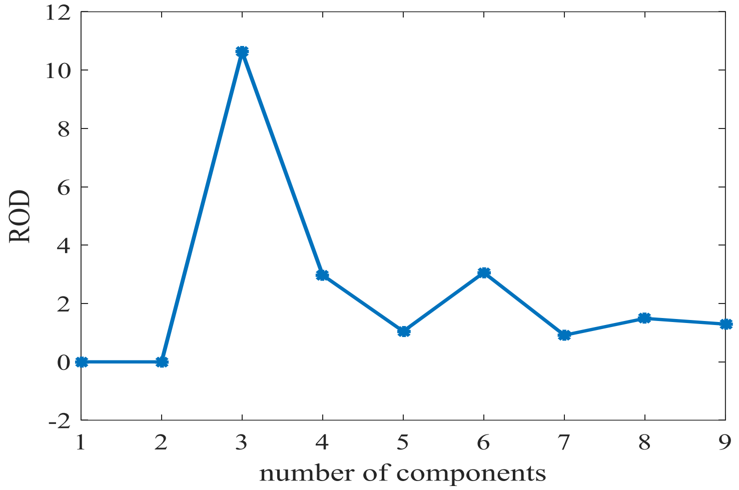


Figure 3-3. Comparison of the smoothed and raw data and the corresponding residual for the first data set (150°C).

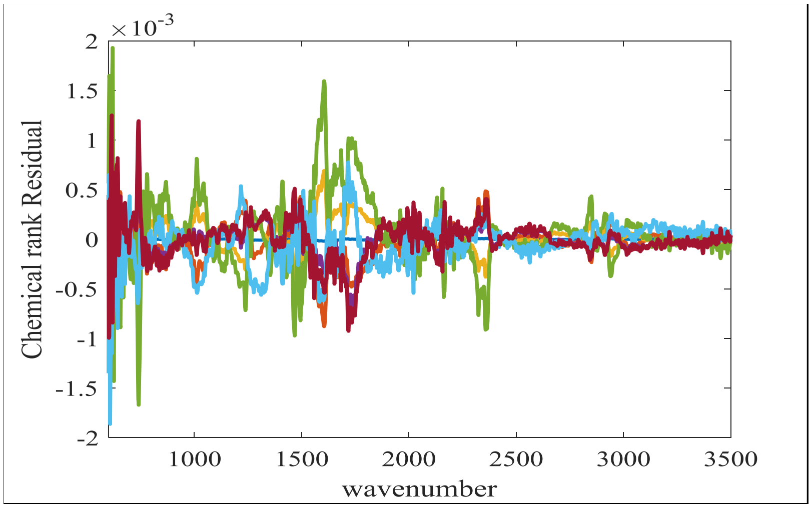
3.3. RESULTS AND DISCUSSION

3.3.1 Chemical rank

Figure 3-4a demonstrates how ROD works to estimate the chemical rank of the data, with the maximum value indicating the optimum number of components. Clearly, the number of components for this dataset is three, for the 150°C dataset, and three components are enough to explain about 99.6% of the variance in the data. Figure 3-4b presents the resulting residual, which also indicates that there is no significant information left to recover. Other datasets were analyzed in similar fashion and it was found that only three components were found significant for each dataset, explaining more than 99% of the variance in each case. Generally, it is easy to see how suitable the method is to automate compared to the graphical methods where the user has to determine the rank qualitatively based on one's understanding of the noise level.



a



b

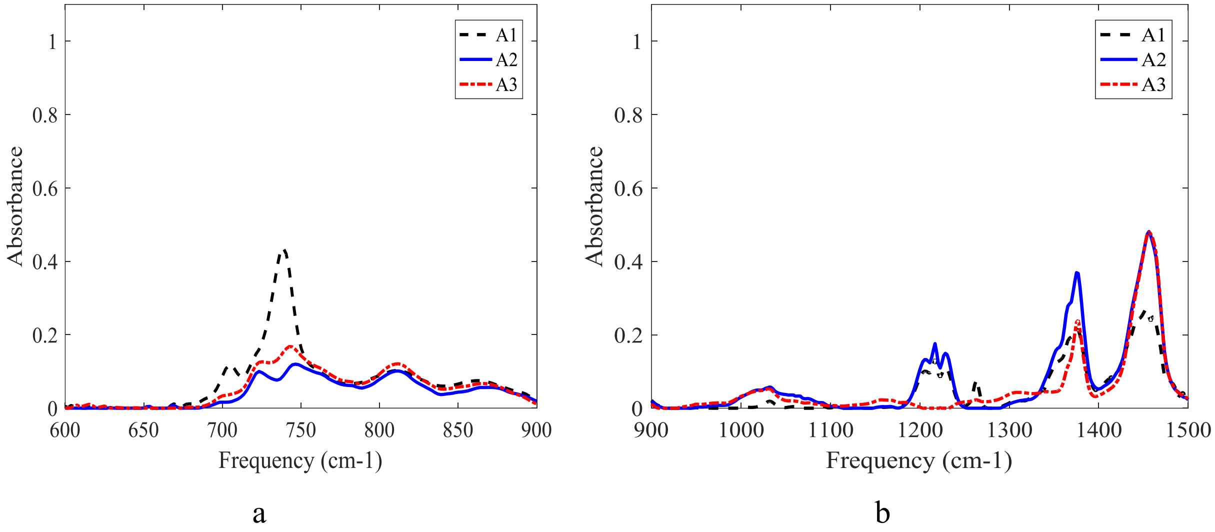
Figure 3-4. Chemical rank estimation using the ratio of the second and third derivatives of the indicator function (ROD). This is a sample result to describe how ROD works; the estimation of chemical rank is determined automatically during simulation.

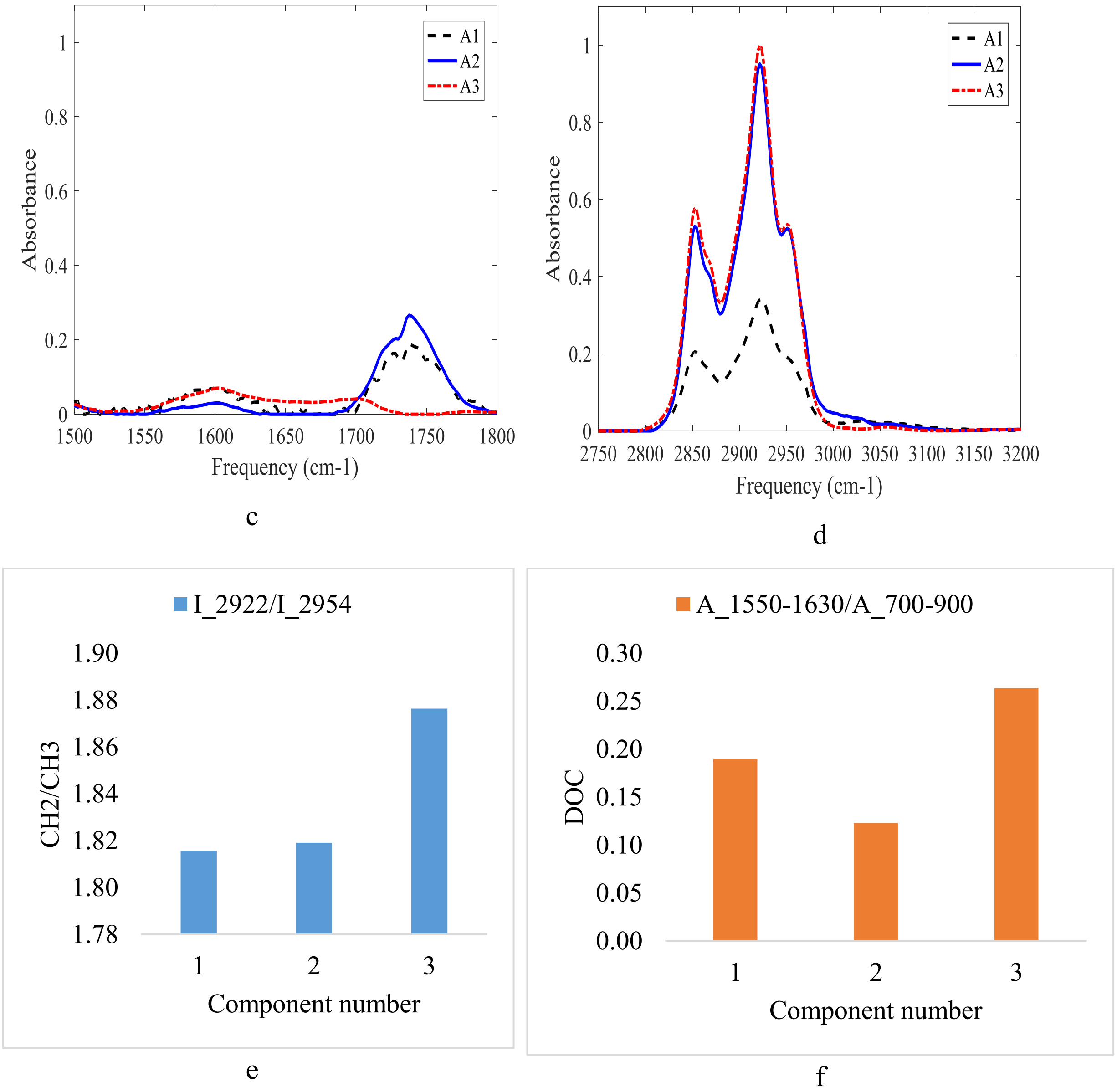
3.3.2 SMCR-ALS

This subsection presents the alternating least squares results for each dataset, obtained at reaction temperatures of 150 ºC, 200 ºC, 300ºC, 340ºC, 360ºC and 400ºC.

Figure 3-5 shows the resolved spectra, the estimated parameter values, concentration profiles, rate of convergence and ALS residual for the 150ºC dataset. The three components indicated by the analysis are designated as A_1, A_2 and A_3, respectively. For FTIR spectra of bitumen, the peak at 864 cm^{-1} indicates an aromatic structure with isolated aromatic hydrogen whereas peaks at 739 and 812 cm^{-1} designate the presence of two and four vicinal aromatics respectively[56]. Weak bands at 3050, 864, 812 and 739 cm^{-1} show the presence of polyaromatic groups and for this dataset, weaker peaks in these regions (Figure 3-5a) indicate that the second and the third components (A_2 and A_3) have higher condensed structures than the first component (A_1). Figure 3-5b compares the components in terms of the C-O band (1000-1300 cm^{-1}) and the aliphatic CH_x intensities (peaks at 1371 and 1460 cm^{-1}). From the result, it is clear that A_3 has the least C-O group and other two are equivalent. Besides the aliphatic CH_x content is least for A_1. Figure 3-5c shows two important regions, the aromatic ring C=C stretching (~1602 cm^{-1}) and the carbonyl peak ($\sim 1742\ cm^{-1}$). The aromatic ring C=C stretching intensity and carbonyl intensity are relatively the lowest for A_3 and the other two components are similar in this region. The shoulder between these two peaks is usually ascribed to the alkene group, but there is no significant absorption for this case, though A_3 has relatively better absorbance. Figure 3-5d describes the aliphatic and aromatic CH_x stretching intensities. Clearly, the intensity for A_2 and A_3 are higher in aliphatic CHx stretching intensities, indicating these compounds are more aliphatic than A_1.

Figure 3-5e-g shows the values of the semi-quantitative parameters used to evaluate the differences between the components. Figure 3-5e shows that the CH_2/CH_3 ratio is highest for A_3 followed by that for A_2. Under normal circumstances, thermal cracking usually produces shorter aliphatic chains and one may expect a decrease in the CH_2/CH_3 ratio with reaction time. In this case, the result is indicative of the fact that there is no significant thermal cracking at this reaction temperature (the difference is negligible) or it could also mean that hydrogenation is taking place, which may have produced cycloalkanes. Similar trends were observed for datasets at 150-360°C. The result also compares well with another recent study by Craddock *et al.* [37]. In their investigation of pyrolysis-based thermal maturation of bitumen samples, they observed a decreasing CH_3/CH_2 ratio with pyrolysis. Figure 3-5f shows the degree of condensation of the aromatic fraction and obviously, A_2 has the least DOC followed by A_1, with A_3 consisting of more condensed aromatic structures compared to the other two, which is reasonable. The prolonged pyrolysis period normally favors the formation of coke, which has a condensed ring structure. Finally, Figure 3-5g compares the carbonyl content of the components, which is highest for A_2, followed A_1. Figure 3-5h shows the resolved concentration profile for the same dataset, with C_1, C_2 and C_3 representing the concentrations of A_1, A_2 and A_3, respectively. From the results, one can see that the relatively aliphatic component (A_3) is dominant in terms of concentration over the majority of the reaction period. This is consistent with the results obtained through ^{1}H NMR and viscosity measurement [51] where it was found that the aliphatic content of the visbroken material is higher than the starting material and lower viscosity. Figure 3-5 i and j illustrate the speed of convergence and accuracy of the ALS optimization. The speed of convergence and residuals are similar for the other reaction conditions, and those plots are omitted for brevity.



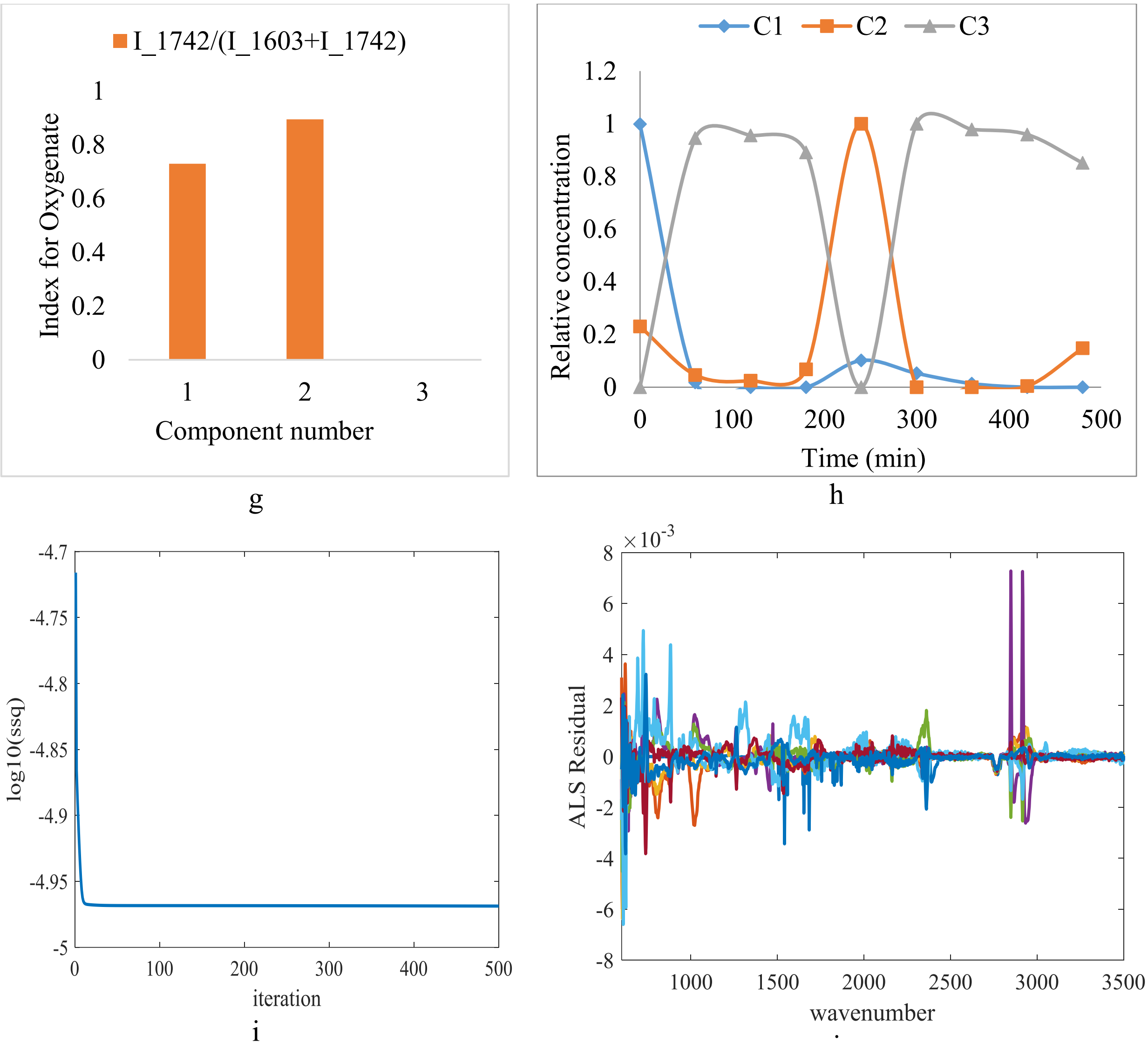

Absorbance
1
0.8
0.6
0.4
0.2
0
1500
1550
1600
1650
1700
1750
1800
Frequency (cm-1)
A1
A2
A3
Absorbance
1
0.8
0.6
0.4
0.2
0
2750
2800
2850
2900
2950
3000
3050
3100
3150
3200
Frequency (cm-1)
A1
A2
A3
I_2922/I_2954
CH2/CH3
1.90
1.88
1.86
1.84
1.82
1.80
1.78
1
2
3
Component number
A_1550-1630/A_700-900
DOC
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1
2
3
Component number

c

d

e

f



g

h

i

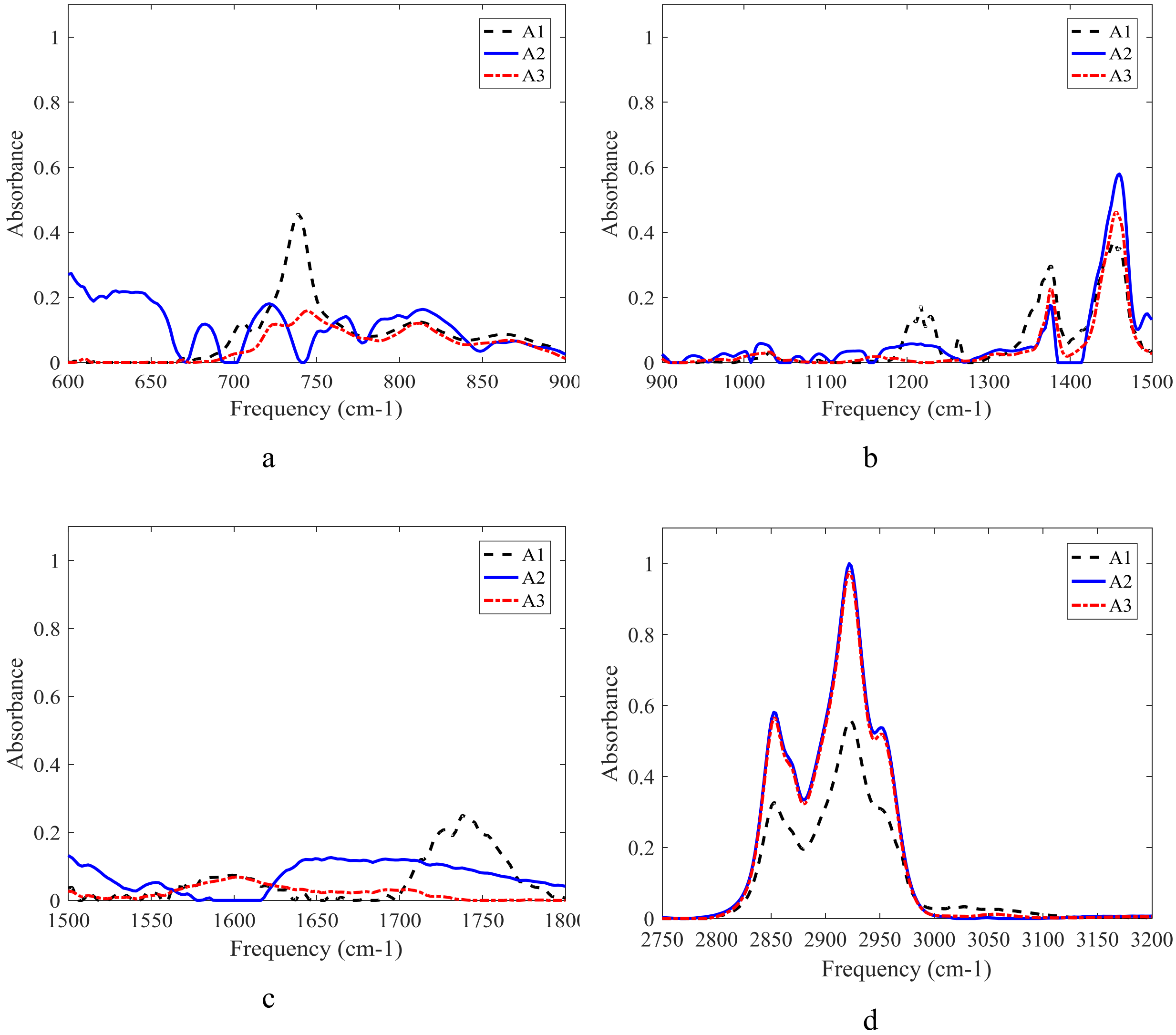
j

Figure 3-5. SMCR analysis results for 150°C dataset where (a-d) shows the resolved spectra; (e-g) the quantitative parameters; (h) represents the corresponding concentration profile; (i) shows the convergence with respect to number iteration and (j) is the ALS optimization residual. C_1, C_2 & C_3 and A_1, A_2 & A_3 represent the concentrations and the absorbance of the resolved components. DOC, I's and A's with respective subscripts stands for the degree of aromatic condensation, intensities at the wavenumber indicated as subscript and area under the peak over the range of frequency range indicated as subscript.

Figure 3-6 describes the resolved spectra of the components, the estimated parameters and the concentration profiles for the 200°C dataset. Similar to the 150°C case, it is clear that generally there is an increase in the CH_2/CH_3 ratio and increasing DOC going from A_1 to A_3, as seen in

Figure 3-6a-f. On the other hand, A_2 has the highest carbonyl fraction and A_3 the least of all (Figure 3-6g). Figure 3-6h describes the corresponding concentration profile for the resolved spectra; the product stream consists mainly of A_3 after 200 min.

In general it can be seen that the intermediate and the final product has higher aliphatic composition than the starting material, A_1, which is similar to the results obtained from ^{1}H NMR , higher aliphatic content[51].



a

b

c

d

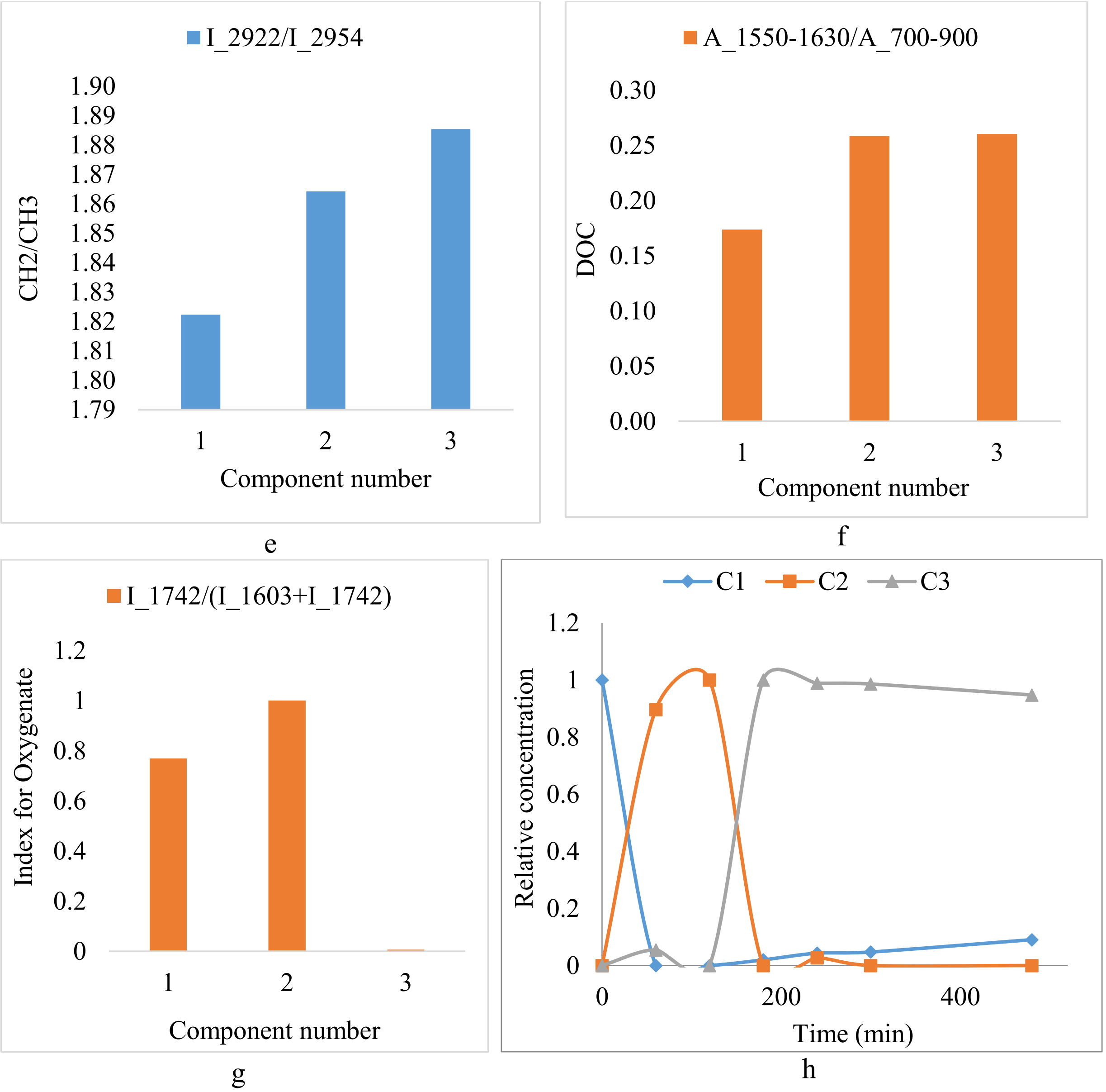
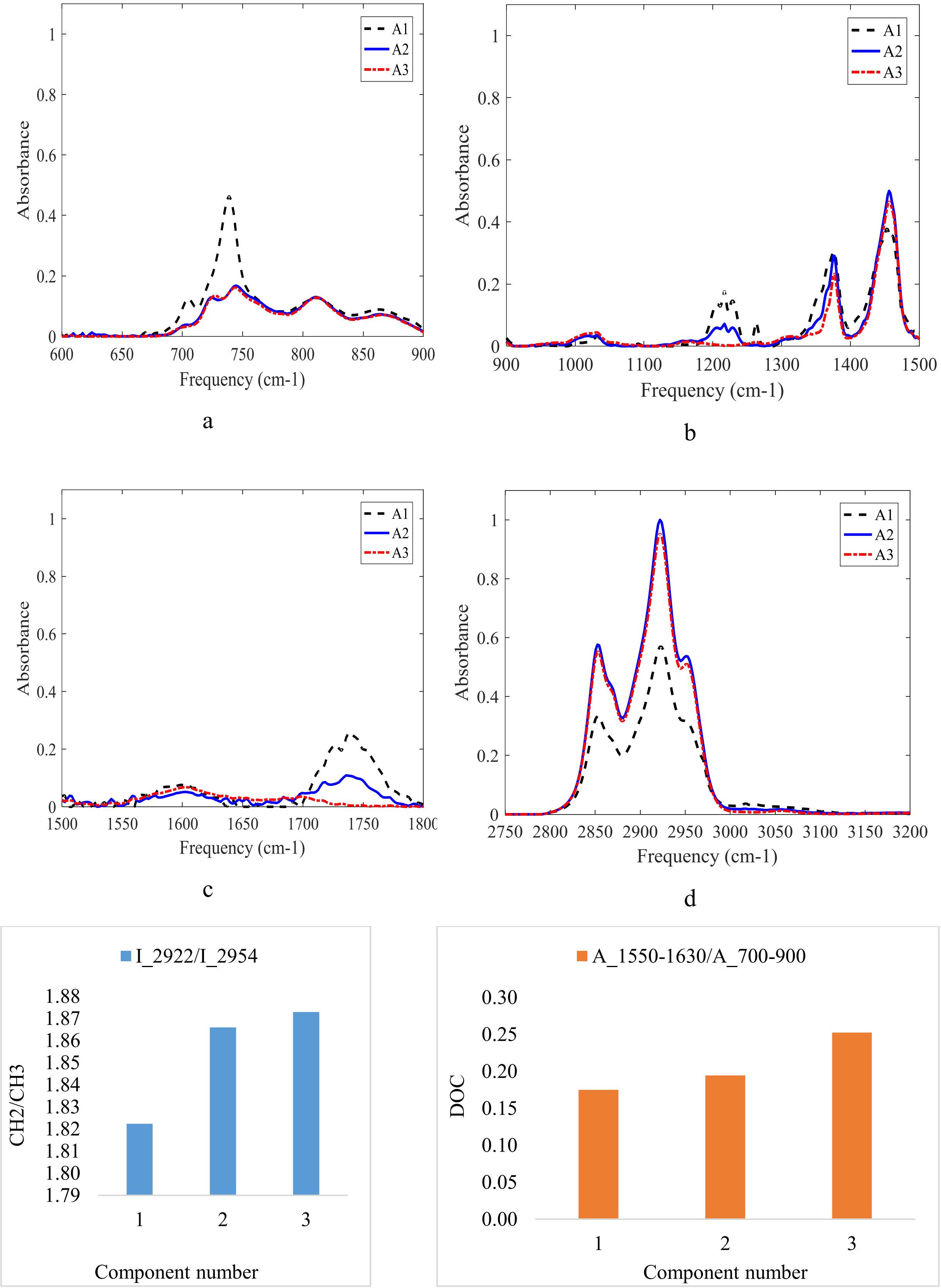


Figure 3-6. SMCR analysis results for 200°C dataset where (a-d) shows the resolved spectra, (e-g) the quantitative parameters, (h) represents the corresponding concentration profile. C_1, C_2 & C_3 and A_1, A_2 & A_3 represent the concentrations and the absorbance of the resolved components. DOC, I's and A's with respective subscripts stands for the degree of aromatic condensation, intensities at the wavenumber indicated as subscript and area under the peak over the range of frequency range indicated as subscript.

Figure 3-7 illustrates the resolved spectra of the components, the parameter values and the concentration profiles for the 300°C dataset. Similar to the previous cases, the DOC and CH_2/CH_3 ratio increase and the carbonyl fraction decreases from A_1 to A_3. Figure 3-7h presents the concentration profile of the components with respect to reaction time. The concentration of A_1 decreases quickly to zero and that of A_2 increases and stays constant after that with A_2 being the dominant species after about 300 min.

Similar to the two cases above, ^{1}H NMR was used to measure the actual content of aliphatic protons and it was found that the visbroken materials have a higher proportion of aliphatic protons than the raw bitumen. Besides, after visbreaking at 300^{o}C, the viscosity of the products decreased by two to three orders of magnitude compared with the raw bitumen[51] which may also confirm the fact that the product side has higher aliphatic content and this is consistent with the results discussed above.


Absorbance
Frequency (cm-1)
A1
A2
A3
0
0.2
0.4
0.6
0.8
1
600
650
700
750
800
850
900
1000
1100
1200
1300
1400
1500
1550
1600
1650
1700
1750
1800
2750
2800
2850
2900
2950
3000
3050
3100
3150
3200
I_2922/I_2954
CH2/CH3
1.79
1.80
1.81
1.82
1.83
1.84
1.85
1.86
1.87
1.88
Component number
1
2
3
A_1550-1630/A_700-900
DOC
0.00
0.05
0.10
0.15
0.20
0.25
0.30

a

b

c

d

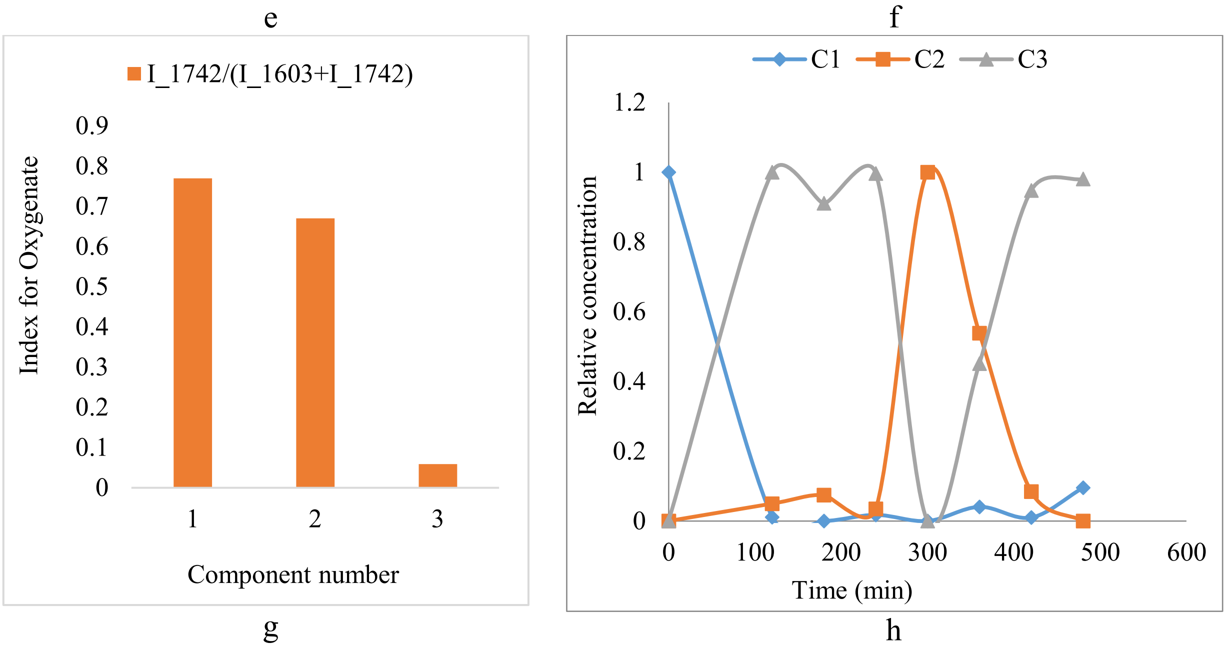
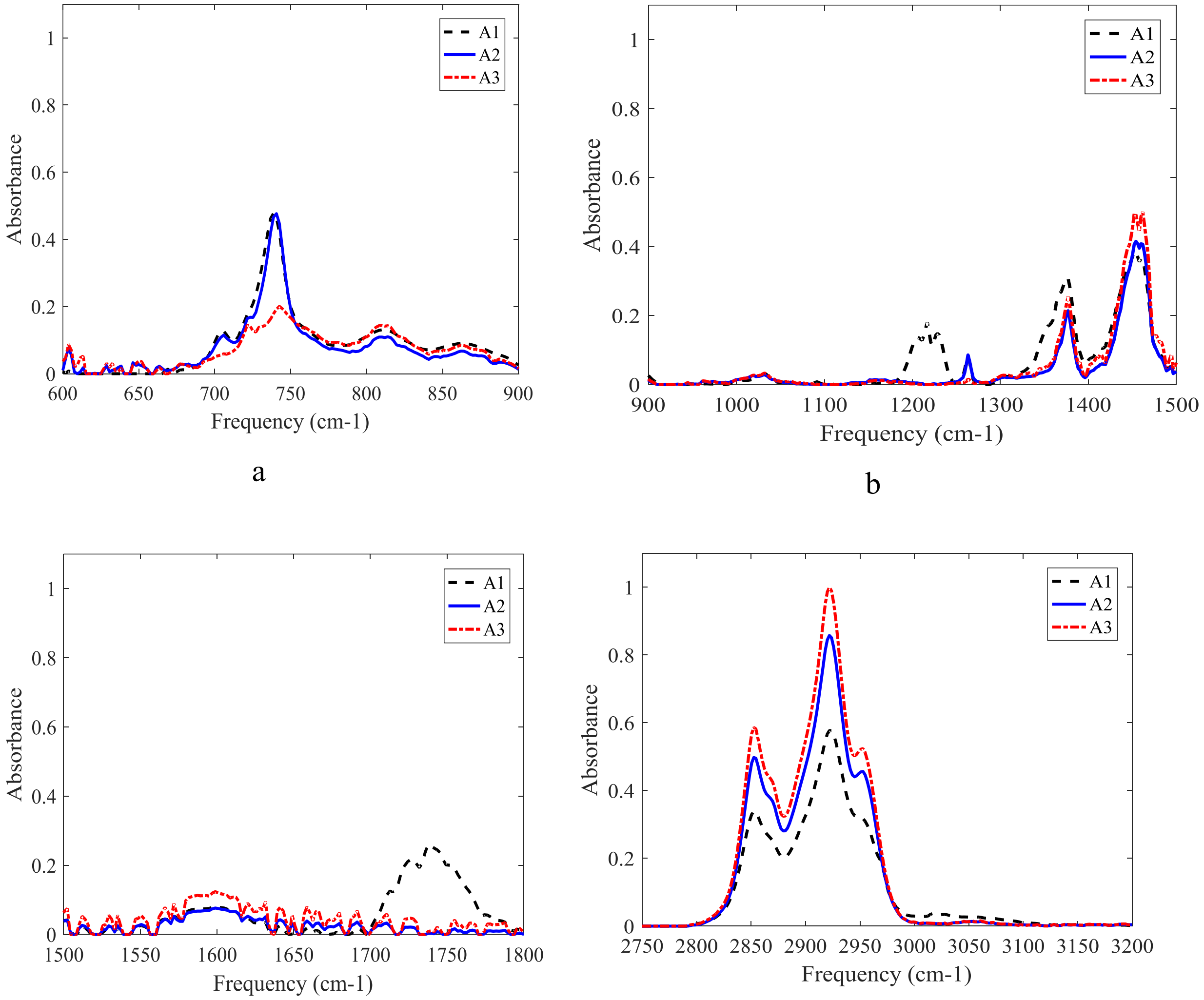


Figure 3-7. SMCR analysis results for 300°C dataset where (a-d) shows the resolved spectra, (e-g) the quantitative parameters, (h) represents the corresponding concentration profile. C_1, C_2 & C_3 and A_1, A_2 & A_3 represent the concentrations and the absorbance of the resolved components. DOC, I's and A's with respective subscripts stands for the degree of aromatic condensation, intensities at the wavenumber indicated as subscript and area under the peak over the range of frequency range indicated as subscript.

Figure 3-8 shows the SMCR analysis results, the estimated parameter values and concentration profiles for the 340°C dataset. Figure 3-8a-g show increasing DOC and CH_2/CH_3 values and decreasing carbonyl intensity from A_1 to A_3. Figure 3-8h describes the corresponding concentration profile of the components, with a fast drop in the relative concentration of A_1 and non-monotonic trends in the other two components. The fact that intermediate and the final product have more aliphatic content than the first component is consistent with observed viscosity reduction [16, 57], lower viscosity indicates higher aliphatic content.

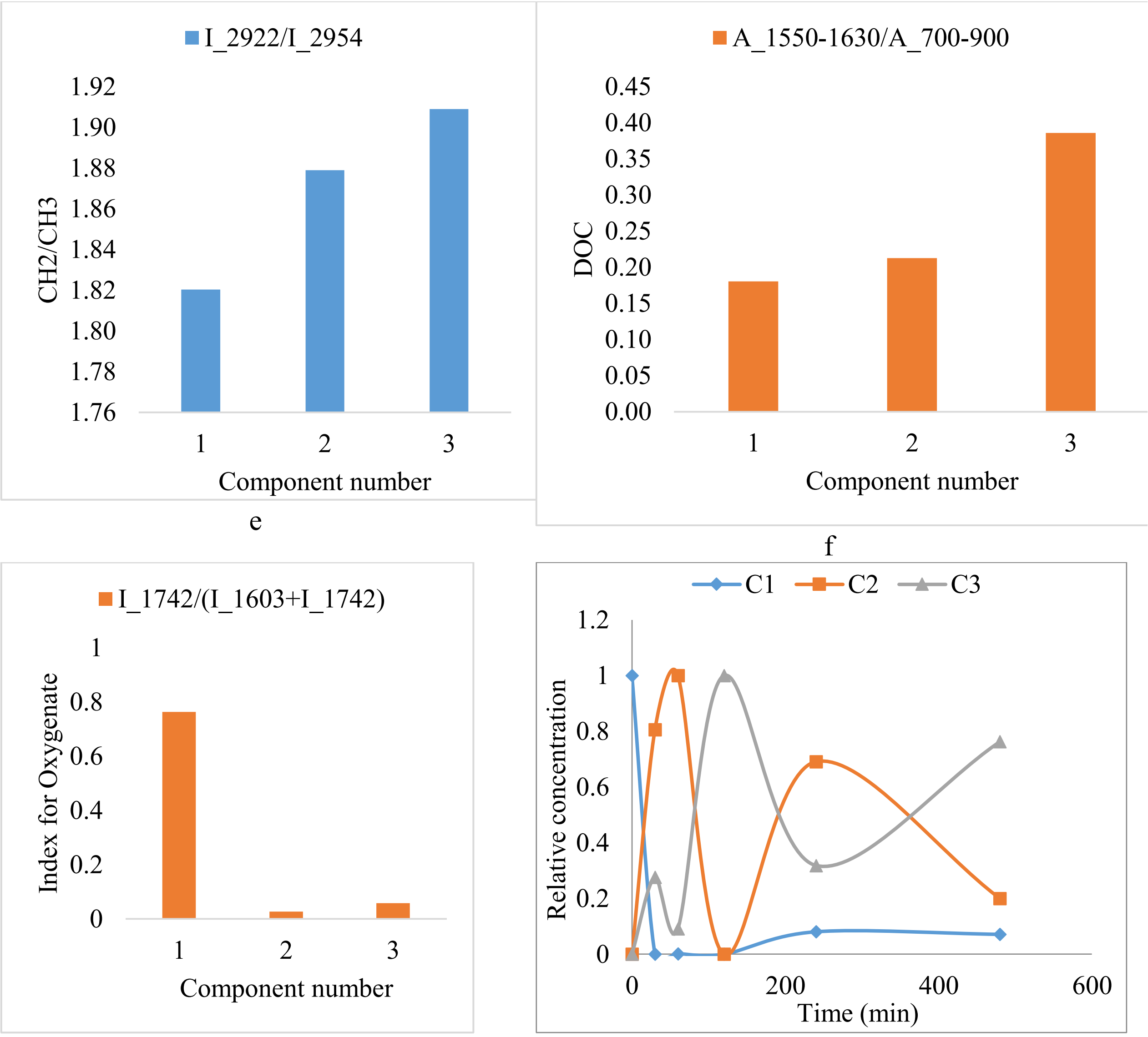


a

b

c

d



e

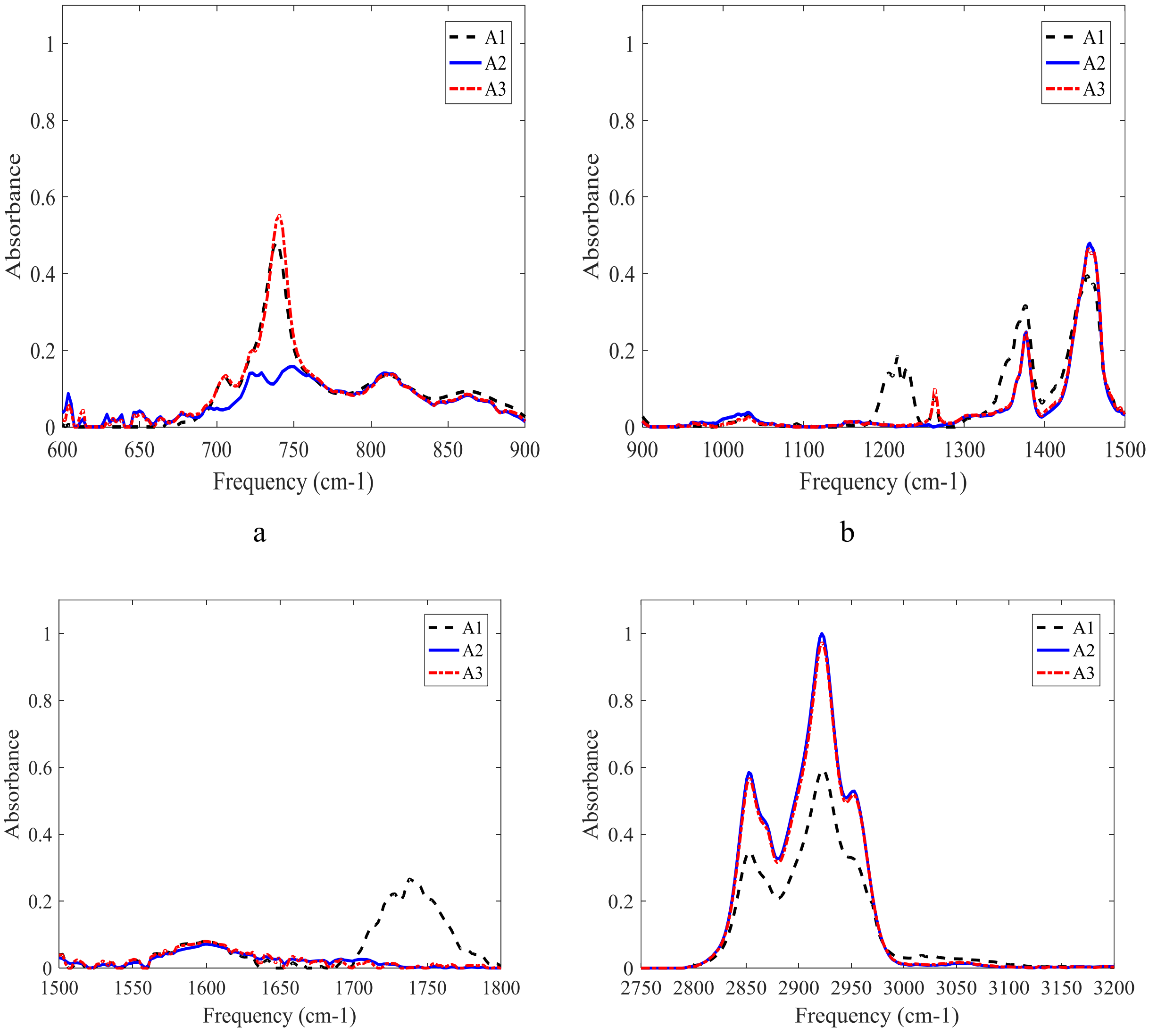
f

g

h

Figure 3-8. SMCR analysis results for 340°C dataset where (a-d) shows the resolved spectra, (e-g) the quantitative parameters, (h) represents the corresponding concentration profile. C_1, C_2 & C_3 and A_1, A_2 & A_3 represent the concentrations and the absorbance of the resolved components. DOC, I's and A's with respective subscripts stands for the degree of aromatic condensation, intensities at the wavenumber indicated as subscript and area under the peak over the range of frequency range indicated as subscript.

Figure 3-9 describes the SMCR results, the resolved spectra, the estimated parameter values and concentration profiles for the 360°C dataset, indicating a (non-monotonic) increase in the DOC and CH_2/CH_3 ratio with reaction time and decreasing carbonyl intensity from A_1 to A_3 as shown in Figure 3-9a-g. Figure 3-9h show the corresponding concentration profiles for the components. Obviously, A_1 disappears after a while and only A_2 and A_3 are available in the system. Similar to the case of 340°C visbroken material showed lower viscosity which agrees with the results of this analysis, higher aliphatic content in the product stream (A_1, A_2).[16, 57]

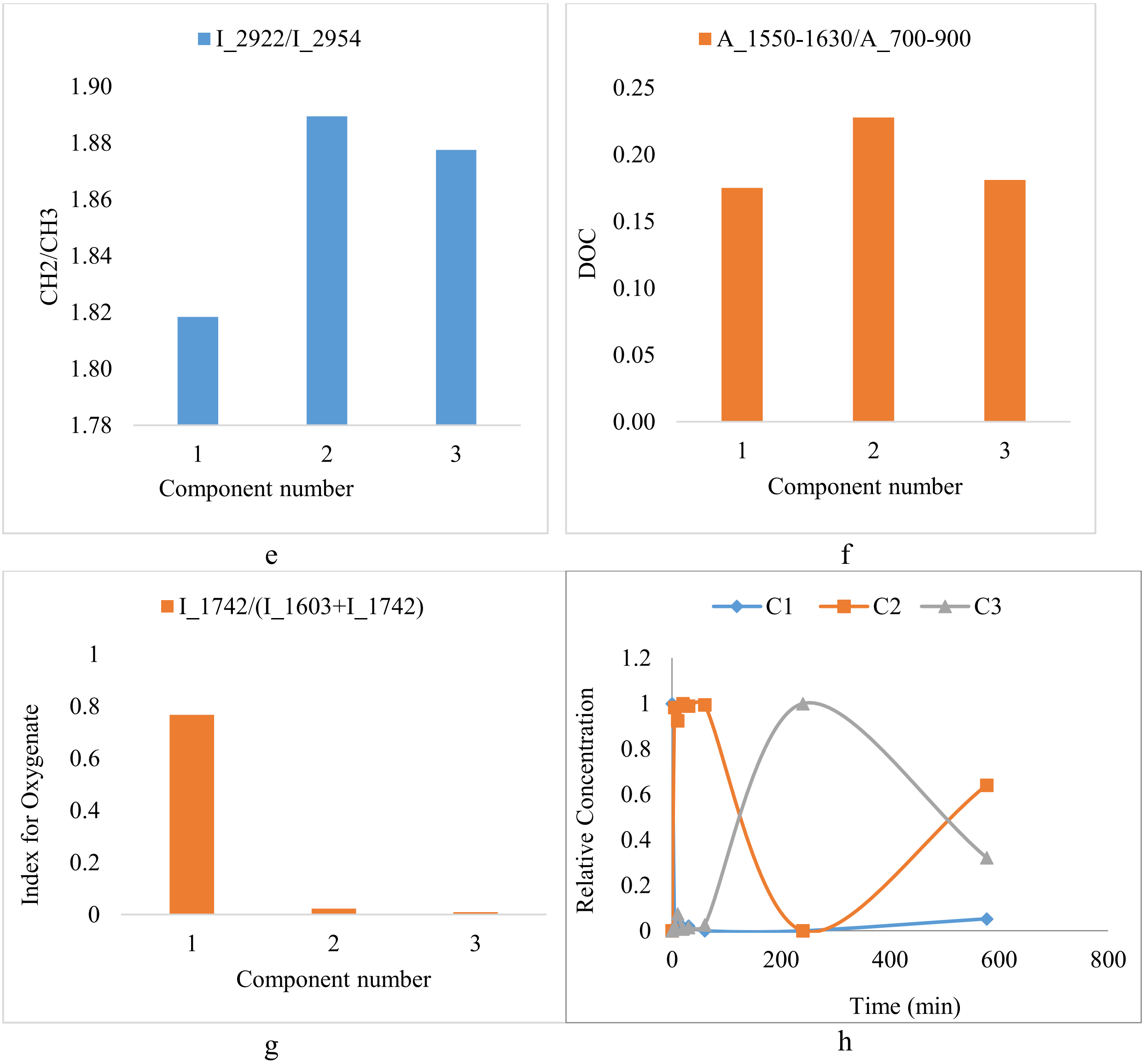


a

b

c

d



e

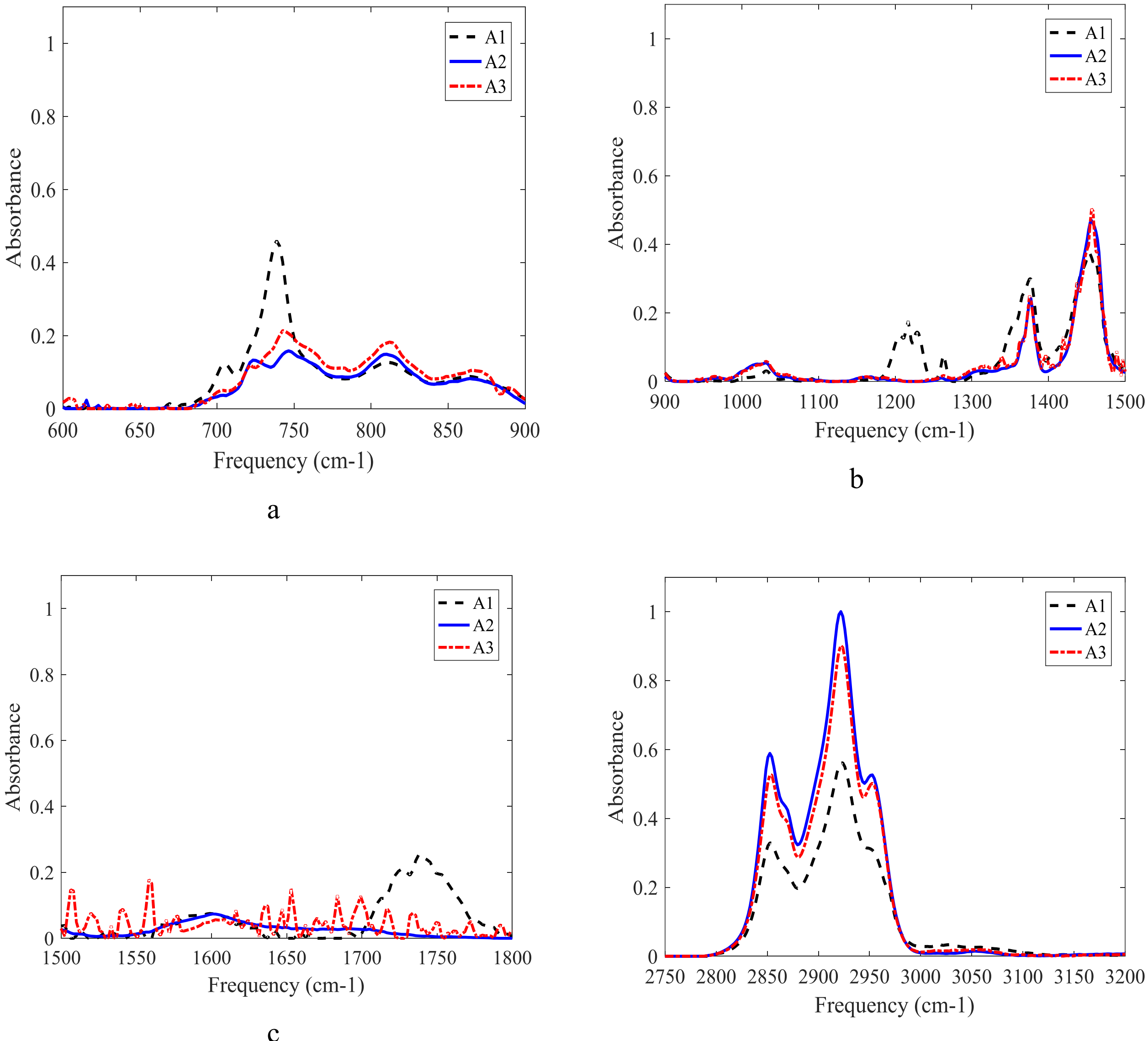
f

g

h

Figure 3-9. SMCR analysis results for 360°C dataset where (a-d) shows the resolved spectra, (e-g) the quantitative parameters, (h) represents the corresponding concentration profile. C_1, C_2 & C_3 and A_1, A_2 & A_3 represent the concentrations and the absorbance of the resolved components. DOC, I's and A's with respective subscripts, stands for the degree of aromatic condensation, intensities at the wavenumber indicated as subscript and area under the peak over the range of frequency range indicated as subscript.

Figure 3-10 shows the SMCR analysis results, the resolved spectra, the estimated parameter values and the concentration profiles for the 400°C dataset. Figure 3-10a indicates that A_3 has the least CH_2/CH_3 ratio, the relatively highest content of the shorter aliphatic chain, which is consistent with ^{1}H NMR characterization of the samples, where a decreasing and then increasing trend in the methyl group was observed [58]. This also shows that there is more thermal cracking at 400°C than the cases addressed so far. Figure 3-10f shows the increasing degree of aromatic condensation (from A_1 to A_3). In addition, A_3 is dominant in terms of its relative concentration at longer reaction times (greater than 90 min), which is in agreement with the increasing microcarbon residue content of the solid -free liquid products of the samples that was also observed in experimental analysis of these samples [58].



a

b

c

d

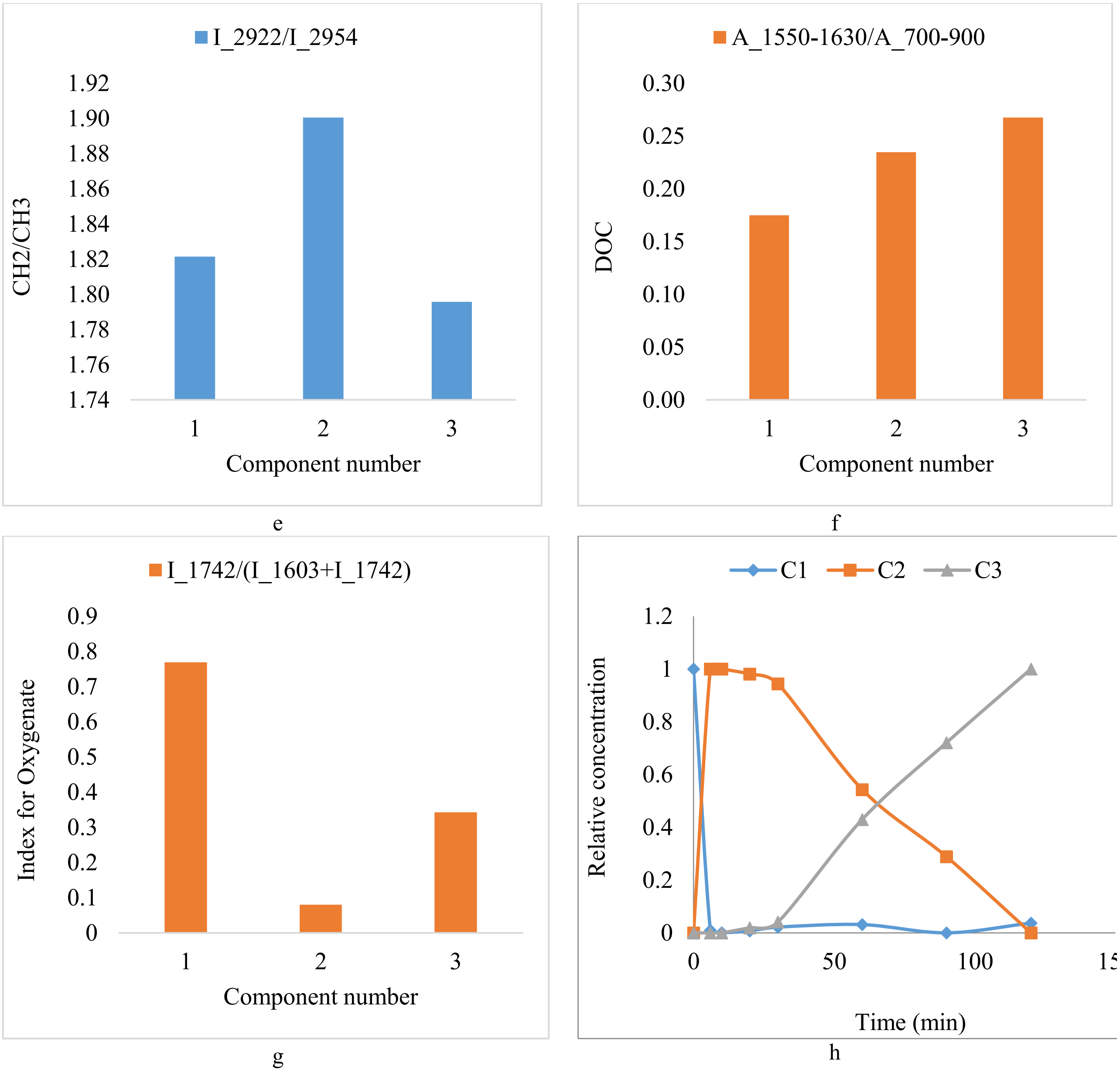


Figure 3-10. SMCR analysis results for 400°C dataset where (a-d) shows the resolved spectra, (e-g) the quantitative parameters, (h) represents the corresponding concentration profile. C_1, C_2 & C_3 and A_1, A_2 & A_3 represent the concentrations and the Absorbance of the resolved components. DOC, I's and A's with respective subscripts, stands for the degree of aromatic condensation, intensities at the wavenumber indicated as subscript and area under the peak over the range of frequency range indicated as subscript

Since this algorithm allows prediction of the most plausible components from FTIR spectra of mixtures and requires very short computation time it can be used to perform real time analysis and control of the reaction. In order to use the algorithm it can be incorporated to the existing control system and provides addition information or can be the code can be deployed online with FTIR spectroscopy. Hence one can analyze the changes and the effect of different operation condition online, identify the optimum operation condition.

3.4. CONCLUSION

A multivariate algorithm is designed to resolve the FTIR spectra of thermally cracked bitumen samples. The algorithm enables automatic estimation of the chemical rank and initial condition for FTIR spectral data. Principal component analysis is implemented to determine the error for rank estimation and evolving factor analysis to approximate the initial concentration profile. The final resolution of the spectra is performed using an alternating least squares-based constrained optimization method, SMCR-ALS, using the Frobenius norm of the residual as the cost function.

In order to assess the accuracy and the convergence speed of the algorithm six FTIR spectral of the samples of bitumen, treated at various temperature and reaction times, were investigated. The results were compared with available experimental data, ^{1}H NMR characterization and measured microcarbon residue content. The agreement with the experimental results is promising and the algorithm requires few seconds to converge, which also a good feature for online monitoring.

3.5. REFERENCE

1. De Juan, A.; Tauler, R., Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution. *Analytica Chimica Acta* **2003,** *500*, (1), 195-210.
2. de Juan, A.; Tauler, R., Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution. *Analytica Chimica Acta* **2003,** *500*, (1–2), 195-210.
3. Watari, M., Applications of near-infrared spectroscopy to process analysis using fourier transform spectrometer. *Optical review* **2010,** *17*, (3), 317-322.
4. Larrechi, M.; Rius, F., Spectra and concentration profiles throughout the reaction of curing epoxy resins from near-infrared spectroscopy and multivariate curve resolution methods. *Applied spectroscopy* **2004,** *58*, (1), 47-53.
5. Miller, C. E., Chemometrics for on-line spectroscopy applications—theory and practice. *Journal of Chemometrics* **2000,** *14*, (5-6), 513-528.
6. Zogg, A.; Fischer, U.; Hungerbühler, K., A new approach for a combined evaluation of calorimetric and online infrared data to identify kinetic and thermodynamic parameters of a chemical reaction. *Chemometrics and intelligent laboratory systems* **2004,** *71*, (2), 165-176.
7. Amari, T.; Ozaki, Y., Generalized Two-Dimensional Attenuated Total Reflection/Infrared and Near-Infrared Correlation Spectroscopy Studies of Real-Time Monitoring of the Initial Oligomerization of Bis(hydroxyethyl terephthalate). *Macromolecules* **2002,** *35*, (21), 8020-8028.
8. Blanco, M.; Castillo, M.; Peinado, A.; Beneyto, R., Application of multivariate curve resolution to chemical process control of an esterification reaction monitored by near-infrared spectroscopy. *Applied spectroscopy* **2006,** *60*, (6), 641-647.
9. Garrido, M.; Rius, F.; Larrechi, M., Multivariate curve resolution–alternating least squares (MCR-ALS) applied to spectroscopic data from monitoring chemical reactions processes. *Analytical and bioanalytical chemistry* **2008,** *390*, (8), 2059-2066.
10. Tauler, R.; Kowalski, B.; Fleming, S., Multivariate curve resolution applied to spectral data from multiple runs of an industrial process. *Analytical chemistry* **1993,** *65*, (15), 2040-2047.

11. Navea, S.; de Juan, A.; Tauler, R., Detection and resolution of intermediate species in protein folding processes using fluorescence and circular dichroism spectroscopies and multivariate curve resolution. *Analytical chemistry* **2002,** *74*, (23), 6031-6039.

12. Navea, S.; de Juan, A.; Tauler, R., Modeling Temperature-Dependent Protein Structural Transitions by Combined Near-IR and Mid-IR Spectroscopies and Multivariate Curve Resolution. *Analytical Chemistry* **2003,** *75*, (20), 5592-5601.

13. Zhang, X.; Tauler, R., Application of Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) to remote sensing hyperspectral imaging. *Analytica Chimica Acta* **2013,** *762*, 25-38.

14. Miller, C. E., Chemometrics for on-line spectroscopy applications—theory and practice. *Journal of Chemometrics* **2000,** *14*, (5-6), 513-528.

15. Nagaishi, H.; Chan, E. W.; Sanford, E. C.; Gray, M. R., Kinetics of High-Conversion Hydrocracking of Bitumen. *Energy & Fuels* **1997,** *11*, (2), 402-410.

16. Wang, L.; Zachariah, A.; Yang, S.; Prasad, V.; de Klerk, A., Visbreaking Oilsands-Derived Bitumen in the Temperature Range of 340–400° C. *Energy & Fuels* **2014,** *28*, (8), 5014-5022.

17. Jiang, J.-H.; Liang, Y.; Ozaki, Y., Principles and methodologies in self-modeling curve resolution. *Chemometrics and Intelligent Laboratory Systems* **2004,** *71*, (1), 1-12.

18. Ruckebusch, C.; Blanchet, L., Multivariate curve resolution: A review of advanced and tailored applications and challenges. *Analytica chimica acta* **2013,** *765*, 28-36.

19. Massart, D. L.; Vandeginste, B. G.; Buydens, L.; Lewi, P.; Smeyers-Verbeke, J., *Handbook of chemometrics and qualimetrics: Part A*. Elsevier Science Inc.: 1997.

20. Shen, H.; Liang, Y.; Kvalheim, O. M.; Manne, R., Determination of chemical rank of two-way data from mixtures using subspace comparisons. *Chemometrics and Intelligent Laboratory Systems* **2000,** *51*, (1), 49-59.

21. Maeder, M., Evolving factor analysis for the resolution of overlapping chromatographic peaks. *Analytical Chemistry* **1987,** *59*, (3), 527-530.

22. Amrhein, M.; Srinivasan, B.; Bonvin, D.; Schumacher, M., On the rank deficiency and rank augmentation of the spectral measurement matrix. *Chemometrics and Intelligent Laboratory Systems* **1996,** *33*, (1), 17-33.

23. Malinowski, E. R., *Factor analysis in chemistry*. Wiley: 2002.

24. Wasim, M.; Brereton, R. G., Determination of the number of significant components in liquid chromatography nuclear magnetic resonance spectroscopy. *Chemometrics and intelligent laboratory systems* **2004,** *72*, (2), 133-151.

25. Spegazzini, N.; Ruisánchez, I.; Larrechi, M. S., MCR–ALS for sequential estimation of FTIR–ATR spectra to resolve a curing process using global phase angle convergence criterion. *Analytica Chimica Acta* **2009,** *642*, (1–2), 155-162.

26. Elbergali, A.; Nygren, J.; Kubista, M., An automated procedure to predict the number of components in spectroscopic data. *Analytica Chimica Acta* **1999,** *379*, (1), 143-158.

27. Malinowski, E. R., Determination of the number of factors and the experimental error in a data matrix. *Analytical Chemistry* **1977,** *49*, (4), 612-617.

28. Gemperline, P. J.; Wade, A. P.; Keller, H. R.; Massart, D. L., Papers Presented at the Seventeenth Annual Meeting of the Federation of Analytical Chemistry and Spectroscopy SocietiesEvolving factor analysis. *Chemometrics and Intelligent Laboratory Systems* **1991,** *12*, (3), 209-224.

29. Malinowski, E. R., Window factor analysis: Theoretical derivation and application to flow injection analysis data. *Journal of Chemometrics* **1992,** *6*, (1), 29-40.

30. Sanchez, F. C.; Toft, J.; Van den Bogaert, B.; Massart, D., Orthogonal projection approach applied to peak purity assessment. *Analytical Chemistry* **1996,** *68*, (1), 79-85.

31. Windig, W., Spectral data files for self-modeling curve resolution with examples using the Simplisma approach. *Chemometrics and Intelligent Laboratory Systems* **1997,** *36*, (1), 3-16.

32. Rossmassler, S. A.; Watson, D. G., Data handling for science and technology. **2013**.

33. Parastar, H., Multivariate Curve Resolution Methods for Qualitative and Quantitative Analysis in Analytical Chemistry. *Fundamentals and Analytical Applications of Multi-way Calibration* **2015,** *29*, 293.

34. Wentzell, P. D.; Karakach, T. K.; Roy, S.; Martinez, M. J.; Allen, C. P.; Werner-Washburne, M., Multivariate curve resolution of time course microarray data. *BMC bioinformatics* **2006,** *7*, (1), 343.

35. Awa, K.; Okumura, T.; Shinzawa, H.; Otsuka, M.; Ozaki, Y., Self-modeling curve resolution (SMCR) analysis of near-infrared (NIR) imaging data of pharmaceutical tablets. *analytica chimica acta* **2008,** *619*, (1), 81-86.

36. AlHumaidan, F. S.; Hauser, A.; Rana, M. S.; Lababidi, H. M. S., Impact of Thermal Treatment on Asphaltene Functional Groups. *Energy & Fuels* **2016,** *30*, (4), 2892-2903.

37. Craddock, P. R.; Le Doan, T. V.; Bake, K.; Polyakov, M.; Charsky, A. M.; Pomerantz, A. E., Evolution of Kerogen and Bitumen during Thermal Maturation via Semi-Open Pyrolysis Investigated by Infrared Spectroscopy. *Energy & Fuels* **2015,** *29*, (4), 2197-2210.

38. Durand, B.; Espitalié, J., Geochemical studies on the organic matter from the Douala Basin (Cameroon)—II. Evolution of kerogen. *Geochimica et Cosmochimica Acta* **1976,** *40*, (7), 801-808.

39. Silverstein, R. M.; Bassler, G. C.; Morrill, T. C., *Spectrometric Identification Of Organic Compounds*. 1974; p 340-340.

40. Robin, P. L.; Rouxhet, P. G., Characterization of kerogens and study of their evolution by infrared spectroscopy: carbonyl and carboxyl groups. *Geochimica et Cosmochimica Acta* **1978,** *42*, (9), 1341-1349.

41. Painter, P. C.; Snyder, R. W.; Starsinic, M.; Coleman, M. M.; Kuehn, D. W.; Davis, A., Concerning the application of FT-IR to the study of coal: a critical assessment of band assignments and the application of spectral analysis programs. *Applied Spectroscopy* **1981,** *35*, (5), 475-485.

42. Chen, Y.; Mastalerz, M.; Schimmelmann, A., Characterization of chemical functional groups in macerals across different coal ranks via micro-FTIR spectroscopy. *International Journal of Coal Geology* **2012,** *104*, 22-33.

43. Ganz, H.; Kalkreuth, W., Application of infrared spectroscopy to the classification of kerogentypes and the evaluation of source rock and oil shale potentials. *Fuel* **1987,** *66*, (5), 708-711.

44. Lis, G. P.; Mastalerz, M.; Schimmelmann, A.; Lewan, M. D.; Stankiewicz, B. A., FTIR absorption indices for thermal maturity in comparison with vitrinite reflectance R 0 in type-II kerogens from Devonian black shales. *Organic Geochemistry* **2005,** *36*, (11), 1533-1552.

45. Petersen, H. I.; Rosenberg, P.; Nytoft, H. P., Oxygen groups in coals and alginite-rich kerogen revisited. *International Journal of Coal Geology* **2008,** *74*, (2), 93-113.

46. Coelho, R. R.; Hovell, I.; de Mello Monte, M. B.; Middea, A.; Lopes de Souza, A., Characterisation of aliphatic chains in vacuum residues (VRs) of asphaltenes and resins

using molecular modelling and FTIR techniques. *Fuel Processing Technology* **2006,** *87*, (4), 325-333.

47. Lin, R.; Patrick Ritz, G., Studying individual macerals using i.r. microspectrometry, and implications on oil versus gas/condensate proneness and “low-rank” generation. *Organic Geochemistry* **1993,** *20*, (6), 695-706.
48. Strausz, O. P.; Lown, E. M. *The chemistry of Alberta Oil Sands, bitumen and heavy oils*; Alberta Energy Research Institute: Calgary, AB, 2003; p 695.
49. H. Ali, L.; A. Al-Ghannam, K.; M. Al-Rawi, J., Chemical structure of asphaltenes in heavy crude oils investigated by n.m.r. *Fuel* **1990,** *69*, (4), 519-521.
50. Wang, L.; Zachariah, A.; Yang, S.; Prasad, V.; de Klerk, A., Visbreaking Oilsands-Derived Bitumen in the Temperature Range of 340–400 °C. *Energy & Fuels* **2014,** *28*, (8), 5014-5022.
51. Visbreaking of Oilsands Bitumen between 150 and 300 oC. In.
52. Lina Maria Yanez Jaramillo, A. d. K., Visbreaking oilsands bitumen at 300°C. . *ACS Division of Energy and Fuels 60*.
53. Ibarra, J.; Muñoz, E.; Moliner, R., FTIR study of the evolution of coal structure during the coalification process. *Organic Geochemistry* **1996,** *24*, (6–7), 725-735.
54. Li, G.; Torraca, G.; Jing, W.; Wen, Z. q., Applications of FTIR in identification of foreign materials for biopharmaceutical clinical manufacturing. *Vibrational Spectroscopy* **2009,** *50*, (1), 152-159.
55. Masson, J. F.; Collins, P., FTIR Study of the Reaction of Polyphosphoric Acid and Model Bitumen Sulfur Compounds. *Energy & Fuels* **2009,** *23*, (1), 440-442.
56. Đorđević, D. M.; Stanković, M. N.; Đorđević, M. G.; Krstić, N. S.; Pavlović, M. A.; Radivojević, A. R.; Filipović, I. M., FTIR Spectroscopic Characterization of Bituminous Limestone: Maganik Mountain (Montenegro). *Studia Universitatis Babes-Bolyai, Chemia* **2012,** *57*, (4).
57. Wang, L. Low temperature visbreaking. MSc thesis, University of Alberta, Edmonton, AB, Canada, 2013.
58. Zachariah, A.; de Klerk, A., Thermal Conversion Regimes for Oilsands Bitumen. *Energy & Fuels* **2016**.

4. CONCLUSIONS AND RECOMMENDATIONS

4.1. CONCLUSIONS

The main theme of this study is to derive the reaction network and develop a multivariate online monitoring scheme for the investigation of the mild thermal cracking of oil sand bitumen in the temperature range of 150°C to 400°C, both objectives serving the same goal, identification of the process. The objectives are achieved by analyzing FTIR spectroscopy data using a combination of a variable selection technique (PCA), data clustering, Bayesian network learning and PCA, IND, EFA and SMCR-ALS respectively.

This research is significant for the identification of the reaction network involved in the mild thermal cracking of bitumen for two basic reasons. One of the problems is that there is only limited information about the chemistry of the process in the temperature range 150°C to 400°C for the classical lumping kinetic methods to be used. Secondly, spectroscopy techniques seem to an ideal choice for the identification of such a complex chemical process since it provides comprehensive information about the underlying chemical changes at a given operation condition. However, the large amount of useful information contained in spectroscopic data is often difficult to extract since absorption intensities from individual chemical constituents of the sample experience a high degree of overlap, particularly for reactions involving chemically complex systems such as reactions involving heavy oils.

The first part of the study (Chapter 2) showed how to use combination of data mining and other learning techniques to develop the most possible reaction kinetics model from FTIR spectroscopy data alone. The required underlining assumption and domain knowledge is also minimal. Moreover, the computer codes for this algorithm can readily be deployed online for use in the real time investigation of the reaction network, a self-updating reaction network.

The second part of the study (Chapter 3) described the importance of self -modeling multivariate curve resolution (SMCR) in the analysis of the effect of various operation condition on the process. The model predicts the chemical rank of the system, resolve the spectra and calculates the resulting concentration profile with no user intervention, which is also useful for further analysis and monitoring of the system.

The major conclusions are:

- Bayesian agglomerative hierarchical cluster analysis applied to the FTIR spectroscopic data for the reaction conditions in the temperature range of 150°C to 400°C and five major clusters/ pseudo-components were identified. Then the possible compound classes in the clusters were identified using spectroscopy handbooks. The first group consists of aromatics and cycloalkanes. The second group consists of aromatics, alkanes, carbonyls, alcohols and phenols. The third cluster includes aromatics and hetroaromatics, cycloalkanes and alkanes. The fourth and the fifth clusters consists mainly of cycloalkanes & alkanes and Paraffins respectively. The clustering result shows an increase in the composition of saturates from **group1** to **group5**. The difference in composition among the groups confirms that the methods may be used to identify the varying groups in such a complex reactive system.
- A Bayesian learning method is then used to recover the possible reaction network among the pseudo (groups). The reaction network from Bayesian learning approach is compared against the representative model reaction network and the BN described the model reaction network very well. This indicates the accuracy of the approach to discover reaction mechanism while there is very limited information about the process.
- In addition to the qualitative model, the graph, Bayesian parameter learning is implemented and the reaction rate model was estimated with the uncertainties of the estimation well described.
- In general, since the proposed method is based on data mining scheme it is difficult to, completely validate the results, yet it can be concluded that a sensible reaction network can be developed from FTIR data alone using the proposed approach.
- The results of this part of the study indicates that the reaction favors formation a more saturated products.
- The second part of the study focused on self-modeling curve resolution approach to further investigate the process. The reaction at each temperature (varying reaction time) is investigated separately. The method used PCA, ROD, EFA to define IND, predict the chemical rank and approximate the initial concentration profile respectively. The final resolution of the spectra is performed using an alternating least squares-based constrained

optimization method, SMCR-ALS. The results dictate that the chemical rank of the system is three for each reaction condition.

- The findings also show that in general the reaction leads to an increase in aliphatic content of the components with reaction time.
- The agreement between the first and the second approach is an indication of the fact that both estimations are reasonably accurate. The experimental studies also showed an increase in the composition of saturates as reaction progresses.
- The results were also compared with available experimental data, ^{1}H NMR characterization and measured microcarbon residue content. The agreement with the experimental results is promising.

4.2. Recommendations and Future Work

The research highlighted some interesting techniques that can be used to recover the reaction kinetics of and resolve spectra of systems with complex chemical mixtures when there is limited information about the chemistry of the process. However, the following aspect of the topic can be investigated further.

- Even though the acyclic Bayesian network is provides a very good estimation of the reaction mechanism, thermal cracking reaction is commonly reversible and this model can be extended to cyclic Bayesian graphs for the complete description of the phenomena.
- A further parametric study can be done using the model either using offline or online data.
- The models in this work are developed using the offline data and they need to be tested on real time basis for actual implementation.
- The methods designed in this reach can be used for the analysis of similar systems such as thermal processing of biogas.
- In the presence of enough data points, concentration profiles from self-modeling curve resolution scheme can be used with Bayesian learning to identify reaction parameters.

BIBLIOGRAPHY

1. Kurgan, L. A.; Musilek, P., A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review* **2006,** *21*, (01), 1-24.
2. Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R., *Advances in knowledge discovery and data mining*. AAAI press Menlo Park: 1996; Vol. 21.
3. Klösgen, W.; Zytkow, J. M., *Handbook of data mining and knowledge discovery*. Oxford University Press, Inc.: 2002.
4. Cios, K. J.; Swiniarski, R. W.; Pedrycz, W.; Kurgan, L. A. In *The knowledge discovery process*, Data Mining, 2007; Springer: 2007; pp 9-24.
5. Piatetsky-Shapiro, G.; Brachman, R. J.; Khabaza, T.; Kloesgen, W.; Simoudis, E. In *An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications*, KDD, 1996; 1996; pp 89-95.
6. Choudhary, A. K.; Harding, J. A.; Tiwari, M. K., Data mining in manufacturing: a review based on the kind of knowledge. *Journal of Intelligent Manufacturing* **2009,** *20*, (5), 501-521.
7. Köksal, G.; Batmaz, İ.; Testik, M. C., A review of data mining applications for quality improvement in manufacturing industry. *Expert systems with Applications* **2011,** *38*, (10), 13448-13467.
8. Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P. In *Knowledge discovery and data mining: towards a unifying framework*, KDD, 1996; 1996; pp 82-88.
9. Ma, C.; Ouyang, J.; Chen, H.-L.; Zhao, X.-H., An efficient diagnosis system for Parkinson's disease using kernel-based extreme learning machine with subtractive clustering features weighting approach. *Computational and mathematical methods in medicine* **2014,** *2014*.
10. Wang, X. Z.; McGreavy, C., *Data Mining and Knowledge Discovery for Process Monitoring and Control*. Springer-Verlag: 1999; p 254.
11. Zheng, L.; Zeng, C.; Li, L.; Jiang, Y.; Xue, W.; Li, J.; Shen, C.; Zhou, W.; Li, H.; Tang, L. In *Applying data mining techniques to address critical process optimization needs in advanced manufacturing*, Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014; ACM: 2014; pp 1739-1748.

12. Brudzewski, K.; Kesik, A.; Kołodziejczyk, K.; Zborowska, U.; Ulaczyk, J., Gasoline quality prediction using gas chromatography and FTIR spectroscopy: An artificial intelligence approach. *Fuel* **2006,** *85*, (4), 553-558.

13. Brudzewski, K.; Osowski, S.; Markiewicz, T.; Ulaczyk, J., Classification of gasoline with supplement of bio-products by means of an electronic nose and SVM neural network. *Sensors and Actuators B: Chemical* **2006,** *113*, (1), 135-141.

14. Pan, L., *Application of Statistical Methods for Gas Turbine Plant Operation Monitoring*. INTECH Open Access Publisher: 2011.

15. Wen, Q.; Ge, Z.; Song, Z., Data-based linear Gaussian state-space model for dynamic process monitoring. *AIChE journal* **2012,** *58*, (12), 3763-3776.

16. Gary, J. H.; Handwerk, G. E.; Kaiser, M. J., *Petroleum refining: technology and economics*. CRC press: 2007.

17. Wang, L.; Zachariah, A.; Yang, S.; Prasad, V.; de Klerk, A., Visbreaking Oilsands-Derived Bitumen in the Temperature Range of 340–400° C. *Energy & Fuels* **2014,** *28*, (8), 5014-5022.

18. Zachariah, A.; de Klerk, A., Thermal Conversion Regimes for Oilsands Bitumen. *Energy & Fuels* **2016,** *30*, (1), 239-248.

19. AlGhazzawi, A.; Lennox, B., Monitoring a complex refining process using multivariate statistics. *Control Engineering Practice* **2008,** *16*, (3), 294-307.

20. Zogg, A.; Fischer, U.; Hungerbühler, K., A new approach for a combined evaluation of calorimetric and online infrared data to identify kinetic and thermodynamic parameters of a chemical reaction. *Chemometrics and intelligent laboratory systems* **2004,** *71*, (2), 165-176.

21. Amari, T.; Ozaki, Y., Generalized Two-Dimensional Attenuated Total Reflection/Infrared and Near-Infrared Correlation Spectroscopy Studies of Real-Time Monitoring of the Initial Oligomerization of Bis(hydroxyethyl terephthalate). *Macromolecules* **2002,** *35*, (21), 8020-8028.

22. Blanco, M.; Castillo, M.; Peinado, A.; Beneyto, R., Application of multivariate curve resolution to chemical process control of an esterification reaction monitored by near-infrared spectroscopy. *Applied spectroscopy* **2006,** *60*, (6), 641-647.

23. Garrido, M.; Rius, F.; Larrechi, M., Multivariate curve resolution–alternating least squares (MCR-ALS) applied to spectroscopic data from monitoring chemical reactions processes. *Analytical and bioanalytical chemistry* **2008,** *390*, (8), 2059-2066.

24. Miller, C. E., Chemometrics for on-line spectroscopy applications—theory and practice. *Journal of Chemometrics* **2000,** *14*, (5-6), 513-528.

25. Gary, J. H.; Handwerk, G. E.; Kaiser, M. J., *Petroleum refining: technology and economics*. CRC press: 2007.

26. Wang, L.; Zachariah, A.; Yang, S.; Prasad, V.; de Klerk, A., Visbreaking oilsands-derived bitumen in the temperature range of 340–400° C. *Energy & Fuels* **2014,** *28*, (8), 5014-5022.

27. Yañez, L.; de Klerk, A., Visbreaking oilsands bitumen at 300 °C. *Energy & Fuels* **2016,** *60*, (1), 31-34.

28. Strausz, O. P.; Lown, E. M. *The chemistry of Alberta Oil Sands, bitumen and heavy oils*; Alberta Energy Research Institute: Calgary, AB, 2003; p 695.

29. H. Ali, L.; A. Al-Ghannam, K.; M. Al-Rawi, J., Chemical structure of asphaltenes in heavy crude oils investigated by n.m.r. *Fuel* **1990,** *69*, (4), 519-521.

30. Gray, M. R., *Upgrading oilsands, bitumen and heavy oil*. 2014.

31. Brons, G.; Yu, J. M., Solvent Deasphalting Effects on Whole Cold Lake Bitumen. *Energy & Fuels* **1995,** *9*, (4), 641-647.

32. Blanchard, C. M.; Gray, M. R., Free radical chain reactions of bitumen residue. *ACS Division of Fuel Chemistry, Preprints* **1997,** *42*, (1), 137-141.

34. Moschopedis, S. E.; Parkash, S.; Speight, J. G., Thermal decomposition of asphaltenes. *Fuel* **1978,** *57*, (7), 431-434.

35. Zachariah, A.; Wang, L.; Yang, S.; Prasad, V.; de Klerk, A., Suppression of Coke Formation during Bitumen Pyrolysis. *Energy & Fuels* **2013,** *27*, (6), 3061-3070.

36. Rana, M. S.; Sámano, V.; Ancheyta, J.; Diaz, J. A. I., A review of recent advances on process technologies for upgrading of heavy oils and residua. *Fuel* **2007,** *86*, (9), 1216-1231.

37. Beleites, C.; Bonifacio, A.; Codrich, D.; Krafft, C.; Sergo, V., Raman spectroscopy and imaging: Promising optical diagnostic tools in pediatrics. *Current Medicinal Chemistry* **2013,** *20*, (17), 2176-2187.

38. Ferreira, A. P.; Tobyn, M., Multivariate analysis in the pharmaceutical industry: Enabling process understanding and improvement in the PAT and QbD era. *Pharmaceutical Development and Technology* **2015,** *20*, (5), 513-527.

39. Héberger, K., Chemoinformatics-multivariate mathematical-statistical methods for data evaluation. In *Medical Applications of Mass Spectrometry*, 2008; pp 141-169.

40. Jansen, J. J.; Smit, S.; Hoefsloot, H. C. J.; Smilde, A. K., The photographer and the greenhouse: How to analyse plant metabolomics data. *Phytochemical Analysis* **2010,** *21*, (1), 48-60.

41. Paczkowska, M.; Lewandowska, K.; Bednarski, W.; Mizera, M.; Podborska, A.; Krause, A.; Cielecka-Piontek, J., Application of spectroscopic methods for identification (FT-IR, Raman spectroscopy) and determination (UV, EPR) of quercetin-3-O-rutinoside. Experimental and DFT based approach. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **2015,** *140*, 132-139.

42. Li, G.; Torraca, G.; Jing, W.; Wen, Z. q., Applications of FTIR in identification of foreign materials for biopharmaceutical clinical manufacturing. *Vibrational Spectroscopy* **2009,** *50*, (1), 152-159.

43. Grewer, D. M.; Young, R. F.; Whittal, R. M.; Fedorak, P. M., Naphthenic acids and other acid-extractables in water samples from Alberta: What is being measured? *Science of the Total Environment* **2010,** *408*, (23), 5997-6010.

44. Zhao, B.; Currie, R.; Mian, H., Catalogue of Analytical Methods for Naphthenic Acids Related to Oil Sands Operations. In *Oil Sands Research and Information Network, University of Alberta, School of Energy and the Environment, Edmonton, Alberta. OSRIN Report No. TR-21*, 2012.

45. Al-Ghouti, M. A.; Al-Degs, Y. S.; Amer, M., Determination of motor gasoline adulteration using FTIR spectroscopy and multivariate calibration. *Talanta* **2008,** *76*, (5), 1105-1112.

46. Hua, H.; Dubé, M. A., Terpolymerization monitoring with ATR-FTIR spectroscopy. *Journal of Polymer Science, Part A: Polymer Chemistry* **2001,** *39*, (11), 1860-1876.

47. Darsy, G.; Bouzat, F.; Muñoz, M.; Lucas, R.; Foucaud, S.; Diogo, C. C.; Babonneau, F.; Leconte, Y.; Maître, A., Monitoring a polycycloaddition by the combination of dynamic rheology and FTIR spectroscopy. *Polymer (United Kingdom)* **2015,** *79*, 283-289.

48. Calabro, D. C.; Valyocsik, E. W.; Ryan, F. X., In situ ATR/FTIR study of mesoporous silicate syntheses. *Microporous Materials* **1996,** *7*, (5), 243-259.

49. Deng, H.; Shen, Z.; Li, L.; Yin, H.; Chen, J., Real-time monitoring of ring-opening polymerization of tetrahydrofuran via in situ Fourier Transform Infrared Spectroscopy. *Journal of Applied Polymer Science* **2014,** *131*, (15).

50. Pintar, A.; Malacea, R.; Pinel, C.; Fogassy, G.; Besson, M., In situ monitoring of catalytic three-phase enantioselective hydrogenation using FTIR/ATR spectroscopy. *Applied Catalysis A: General* **2004,** *264*, (1), 1-12.

51. Wong, K. C., Review of Spectrometric Identification of Organic Compounds, 8th Edition. *Journal of Chemical Education* **2015,** *92*, (10), 1602-1603.

52. Li, G.; Jing, W.; Wen, Z. Q., Identification of unknown mixtures of materials from biopharmaceutical manufacturing processes by microscopic-FTIR and library searching. *American Pharmaceutical Review* **2011,** *14*, (7), 60-64.

53. Nyden, M. R.; Pallister, J. E.; Sparks, D. T.; Salari, A., computer-assisted spectroscopic analysis using orthonormalized reference spectra. Part ii: application to the identification of pure compounds. *Applied Spectroscopy* **1987,** *41*, (1), 63-66.

54. Prats-Montalbán, J. M.; de Juan, A.; Ferrer, A., Multivariate image analysis: A review with applications. *Chemometrics and Intelligent Laboratory Systems* **2011,** *107*, (1), 1-23.

55. Abdolmaleki, A.; Ghasemi, J. B.; Shiri, F.; Pirhadi, S., Application of multivariate linear and nonlinear calibration and classification methods in drug design. *Combinatorial Chemistry and High Throughput Screening* **2015,** *18*, (8), 795-808.

58. Arvanitoyannis, I. S.; Katsota, M. N.; Psarra, E. P.; Soufleros, E. H.; Kallithraka, S., Application of quality control methods for assessing wine authenticity: Use of multivariate analysis (chemometrics). *Trends in Food Science and Technology* **1999,** *10*, (10), 321-336.

59. Currie, L. A., Detection: International update, and some emerging di-lemmas involving calibration, the blank, and multiple detection decisions. *Chemometrics and Intelligent Laboratory Systems* **1997,** *37*, (1), 151-181.

60. Visbreaking of Oilsands Bitumen between 150 and 300 oC. In.

61. Jackson, D. A.; Chen, Y., Robust principal component analysis and outlier detection with ecological data. *Environmetrics* **2004,** *15*, (2), 129-139.

62. James, G.; Witten, D.; Hastie, T.; Tibshirani, R., *An introduction to statistical learning*. Springer: 2013; Vol. 6.

63. Nia, V. P.; Davison, A. C., High-dimensional Bayesian clustering with variable selection: The R package bclust. *Journal of Statistical Software* **2012,** *47*, (5), 1-22.

64. Bishop, C. M., *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc.: 2006.

65. Nia, V. P.; Davison, A. C., High-dimensional Bayesian clustering with variable selection: The R package bclust. *Journal of Statistical Software* **2012,** *47*.

66. Swartz, M. D.; Mo, Q.; Murphy, M. E.; Lupton, J. R.; Turner, N. D.; Hong, M. Y.; Vannucci, M., Bayesian variable selection in clustering high-dimensional data with substructure. *Journal of Agricultural, Biological, and Environmental Statistics* **2008,** *13*, (4), 407-423.

67. Steinbach, M.; Ertöz, L.; Kumar, V., The Challenges of Clustering High Dimensional Data. In *New Directions in Statistical Physics*, Wille, L., Ed. Springer Berlin Heidelberg: 2004; pp 273-309.

68. Ahn, J.; Marron, J. S.; Muller, K. M.; Chi, Y. Y., The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika* **2007,** *94*, (3), 760-766.

69. Hall, P.; Marron, J. S.; Neeman, A., Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **2005,** *67*, (3), 427-444.

70. Ausloos, M.; Ivanova, K., Patterns, Trends and Predictions in Stock Market Indices and Foreign Currency Exchange Rates. In *New Directions in Statistical Physics*, Wille, L., Ed. Springer Berlin Heidelberg: 2004; pp 93-114.

71. Jung, S.; Marron, J. S., PCA consistency in High Dimension, Low Sample Size context. *Annals of Statistics* **2009,** *37*, (6 B), 4104-4130.

72. Yata, K.; Aoshima, M., Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis* **2012,** *105*, (1), 193-215.

73. Tadesse, M. G.; Sha, N.; Vannucci, M., Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association* **2005,** *100*, (470), 602-617.

74. Crandell, J. L.; Dunson, D. B., Posterior simulation across nonparametric models for functional clustering. *Sankhya: The Indian Journal of Statistics* **2011,** *73*, (1 B), 42-61.

75. Lian, H., Sparse Bayesian hierarchical modeling of high-dimensional clustering problems. *Journal of Multivariate Analysis* **2010,** *101*, (7), 1728-1737.

76. Darkins, R.; Cooke, E. J.; Ghahramani, Z.; Kirk, P. D. W.; Wild, D. L.; Savage, R. S., Accelerating Bayesian Hierarchical Clustering of Time Series Data with a Randomised Algorithm. *PLoS ONE* **2013,** *8*, (4).

77. Savage, R. S.; Heller, K.; Xu, Y.; Ghahramani, Z.; Truman, W. M.; Grant, M.; Denby, K. J.; Wild, D. L., R/BHC: Fast Bayesian hierarchical clustering for microarray data. *BMC Bioinformatics* **2009,** *10*.

78. Heard, N. A.; Holmes, C. C.; Stephens, D. A., A quantitative study of gene regulation involved in the immune response of Anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. *Journal of the American Statistical Association* **2006,** *101*, (473), 18-29.

79. Duan, P.; Chen, T.; Shah, S. L.; Yang, F., Methods for root cause diagnosis of plant-wide oscillations. *AIChE Journal* **2014,** *60*, (6), 2019-2034.

80. Evett, I. W.; Gill, P. D.; Jackson, G.; Whitaker, J.; Champod, C., Interpreting small quantities of DNA: The hierarchy of propositions and the use of bayesian networks. *Journal of Forensic Sciences* **2002,** *47*, (3), 520-530.

81. Marques, V. M.; Munaro, C. J.; Shah, S. L. In *Data-based causality detection from a system identification perspective*, 2013 European Control Conference, ECC 2013, 2013; 2013; pp 2453-2458.

82. Lee, S.-M.; Abbott, P. A., Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers. *Journal of Biomedical Informatics* **2003,** *36*, (4–5), 389-399.

83. Sugihara, G.; May, R.; Ye, H.; Hsieh, C. H.; Deyle, E.; Fogarty, M.; Munch, S., Detecting causality in complex ecosystems. *Science* **2012,** *338*, (6106), 496-500.

84. Schreiber, T., Measuring Information Transfer. *Physical Review Letters* **2000,** *85*, (2), 461-464.

85. Granger, C. W. J., Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* **1969,** *37*, (3), 424-438.

86. Smith, V. A., Revealing structure of complex biological systems using bayesian networks. In *Network Science: Complexity in Nature and Technology*, 2010; pp 185-204.

87. Jung, A., Learning the Conditional Independence Structure of Stationary Time Series: A Multitask Learning Approach. *IEEE Transactions on Signal Processing* **2015,** *63*, (21), 5677-5690.

88. Zou, C.; Feng, J., Granger causality vs. dynamic Bayesian network inference: A comparative study. *BMC Bioinformatics* **2009,** *10*.

89. Zou, C.; Denby, K. J.; Feng, J., Granger causality vs. dynamic Bayesian network inference: A comparative study. *BMC Bioinformatics* **2009,** *10*.

90. Koller, D.; Friedman, N., *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press: 2009; p 1208.

91. Heckerman, D., A tutorial on learning with Bayesian networks. In *Studies in Computational Intelligence*, 2008; Vol. 156, pp 33-82.

92. Neapolitan, R. E., *Learning Bayesian Networks*. Prentice-Hall, Inc.: 2003.

93. De Campos, L. M., A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research* **2006,** *7*, 2149-2187.

94. Liu, J. W.; Li, H. E.; Luo, X. L., Learning technique of probabilistic graphical models: A review. *Zidonghua Xuebao/Acta Automatica Sinica* **2014,** *40*, (6), 1025-1044.

95. Russell, S. J.; Norvig, P., *Artificial Intelligence: A Modern Approach*. Pearson Education: 2003; p 1132.

96. Nagarajan, R.; Scutari, M.; Lbre, S., *Bayesian Networks in R: with Applications in Systems Biology*. Springer Publishing Company, Incorporated: 2013; p 170.

97. Taroni, F.; Aitken, C.; Garbolino, P.; Biedermann, A., *Bayesian Networks and Probabilistic Inference in Forensic Science*. 2006; p 1-354.

98. Fuchs, O., *Colthup,nb - Introduction to infrared and raman spectroscopy*. 1967; Vol. 228, p 42.

99. Ksandr, Z., Introduction to infrared and raman spectroscopy. *Chemicke Listy* **1966,** *60*, (5), 687.

100. Silverstein, R. M.; Bassler, G. C.; Morrill, T. C., *Spectrometric identification of organic compounds*. 1974; p 340-340.

101. Stagni, A.; Cuoci, A.; Frassoldati, A.; Faravelli, T.; Ranzi, E., Lumping and reduction of detailed kinetic schemes: An effective coupling. *Industrial and Engineering Chemistry Research* **2014,** *53*, (22), 9004-9016.

102. Silverstein, R. M.; Webster, F. X.; Kiemle, D.; Bryce, D. L., *Spectrometric identification of organic compounds*. John Wiley & Sons: 2014.

103. Shriner, R. L., Systematic identification of organic compounds. **1956**.

104. Colthup, N., *Introduction to infrared and Raman spectroscopy*. Elsevier: 2012.

105. Odebunmi, E.; Adeniyi, S., Infrared and ultraviolet spectrophotometric analysis of chromatographic fractions of crude oils and petroleum products. *Bulletin of the Chemical Society of Ethiopia* **2007,** *21*, (1).

106. Selman, B.; Gomes, C. P., Hill-climbing Search. In *Encyclopedia of Cognitive Science*, John Wiley & Sons, Ltd: 2006.

107. Bai, X.; Padman, R., Tabu search enhanced markov blanket classifier for high dimensional data sets. In *The Next Wave in Computing, Optimization, and Decision Technologies*, Springer: 2005; pp 337-354.

108. MATLAB Version: 9.0.0.341360. Natick, Massachusetts: The MathWorks Inc., 2016

109. De Juan, A.; Tauler, R., Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution. *Analytica Chimica Acta* **2003,** *500*, (1), 195-210.

110. de Juan, A.; Tauler, R., Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution. *Analytica Chimica Acta* **2003,** *500*, (1–2), 195-210.

111. Watari, M., Applications of near-infrared spectroscopy to process analysis using fourier transform spectrometer. *Optical review* **2010,** *17*, (3), 317-322.

112. Larrechi, M.; Rius, F., Spectra and concentration profiles throughout the reaction of curing epoxy resins from near-infrared spectroscopy and multivariate curve resolution methods. *Applied spectroscopy* **2004,** *58*, (1), 47-53.

113. Miller, C. E., Chemometrics for on-line spectroscopy applications—theory and practice. *Journal of Chemometrics* **2000,** *14*, (5-6), 513-528.

114. Zogg, A.; Fischer, U.; Hungerbühler, K., A new approach for a combined evaluation of calorimetric and online infrared data to identify kinetic and thermodynamic parameters of a chemical reaction. *Chemometrics and intelligent laboratory systems* **2004,** *71*, (2), 165-176.

115. Amari, T.; Ozaki, Y., Generalized Two-Dimensional Attenuated Total Reflection/Infrared and Near-Infrared Correlation Spectroscopy Studies of Real-Time Monitoring of the Initial Oligomerization of Bis(hydroxyethyl terephthalate). *Macromolecules* **2002,** *35*, (21), 8020-8028.

116. Blanco, M.; Castillo, M.; Peinado, A.; Beneyto, R., Application of multivariate curve resolution to chemical process control of an esterification reaction monitored by near-infrared spectroscopy. *Applied spectroscopy* **2006,** *60*, (6), 641-647.

117. Garrido, M.; Rius, F.; Larrechi, M., Multivariate curve resolution–alternating least squares (MCR-ALS) applied to spectroscopic data from monitoring chemical reactions processes. *Analytical and bioanalytical chemistry* **2008,** *390*, (8), 2059-2066.

118. Tauler, R.; Kowalski, B.; Fleming, S., Multivariate curve resolution applied to spectral data from multiple runs of an industrial process. *Analytical chemistry* **1993,** *65*, (15), 2040-2047.

119. Navea, S.; de Juan, A.; Tauler, R., Detection and resolution of intermediate species in protein folding processes using fluorescence and circular dichroism spectroscopies and multivariate curve resolution. *Analytical chemistry* **2002,** *74*, (23), 6031-6039.

120. Navea, S.; de Juan, A.; Tauler, R., Modeling Temperature-Dependent Protein Structural Transitions by Combined Near-IR and Mid-IR Spectroscopies and Multivariate Curve Resolution. *Analytical Chemistry* **2003,** *75*, (20), 5592-5601.

121. Zhang, X.; Tauler, R., Application of Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) to remote sensing hyperspectral imaging. *Analytica Chimica Acta* **2013,** *762*, 25-38.

122. Miller, C. E., Chemometrics for on-line spectroscopy applications—theory and practice. *Journal of Chemometrics* **2000,** *14*, (5-6), 513-528.

123. Nagaishi, H.; Chan, E. W.; Sanford, E. C.; Gray, M. R., Kinetics of High-Conversion Hydrocracking of Bitumen. *Energy & Fuels* **1997,** *11*, (2), 402-410.

124. Wang, L.; Zachariah, A.; Yang, S.; Prasad, V.; de Klerk, A., Visbreaking Oilsands-Derived Bitumen in the Temperature Range of 340–400° C. *Energy & Fuels* **2014,** *28*, (8), 5014-5022.

125. Jiang, J.-H.; Liang, Y.; Ozaki, Y., Principles and methodologies in self-modeling curve resolution. *Chemometrics and Intelligent Laboratory Systems* **2004,** *71*, (1), 1-12.

126. Ruckebusch, C.; Blanchet, L., Multivariate curve resolution: A review of advanced and tailored applications and challenges. *Analytica chimica acta* **2013,** *765*, 28-36.

127. Massart, D. L.; Vandeginste, B. G.; Buydens, L.; Lewi, P.; Smeyers-Verbeke, J., *Handbook of chemometrics and qualimetrics: Part A*. Elsevier Science Inc.: 1997.

128. Shen, H.; Liang, Y.; Kvalheim, O. M.; Manne, R., Determination of chemical rank of two-way data from mixtures using subspace comparisons. *Chemometrics and Intelligent Laboratory Systems* **2000,** *51*, (1), 49-59.

129. Maeder, M., Evolving factor analysis for the resolution of overlapping chromatographic peaks. *Analytical Chemistry* **1987,** *59*, (3), 527-530.

130. Amrhein, M.; Srinivasan, B.; Bonvin, D.; Schumacher, M., On the rank deficiency and rank augmentation of the spectral measurement matrix. *Chemometrics and Intelligent Laboratory Systems* **1996,** *33*, (1), 17-33.

131. Malinowski, E. R., *Factor analysis in chemistry*. Wiley: 2002.

132. Wasim, M.; Brereton, R. G., Determination of the number of significant components in liquid chromatography nuclear magnetic resonance spectroscopy. *Chemometrics and intelligent laboratory systems* **2004,** *72*, (2), 133-151.

133. Spegazzini, N.; Ruisánchez, I.; Larrechi, M. S., MCR–ALS for sequential estimation of FTIR–ATR spectra to resolve a curing process using global phase angle convergence criterion. *Analytica Chimica Acta* **2009,** *642*, (1–2), 155-162.

134. Elbergali, A.; Nygren, J.; Kubista, M., An automated procedure to predict the number of components in spectroscopic data. *Analytica Chimica Acta* **1999,** *379*, (1), 143-158.

135. Malinowski, E. R., Determination of the number of factors and the experimental error in a data matrix. *Analytical Chemistry* **1977,** *49*, (4), 612-617.

136. Gemperline, P. J.; Wade, A. P.; Keller, H. R.; Massart, D. L., Papers Presented at the Seventeenth Annual Meeting of the Federation of Analytical Chemistry and Spectroscopy

SocietiesEvolving factor analysis. *Chemometrics and Intelligent Laboratory Systems* **1991,** *12*, (3), 209-224.

137. Malinowski, E. R., Window factor analysis: Theoretical derivation and application to flow injection analysis data. *Journal of Chemometrics* **1992,** *6*, (1), 29-40.

138. Sanchez, F. C.; Toft, J.; Van den Bogaert, B.; Massart, D., Orthogonal projection approach applied to peak purity assessment. *Analytical Chemistry* **1996,** *68*, (1), 79-85.

139. Windig, W., Spectral data files for self-modeling curve resolution with examples using the Simplisma approach. *Chemometrics and Intelligent Laboratory Systems* **1997,** *36*, (1), 3-16.

140. Rossmassler, S. A.; Watson, D. G., Data handling for science and technology. **2013**.

141. Parastar, H., Multivariate Curve Resolution Methods for Qualitative and Quantitative Analysis in Analytical Chemistry. *Fundamentals and Analytical Applications of Multi-way Calibration* **2015,** *29*, 293.

142. Wentzell, P. D.; Karakach, T. K.; Roy, S.; Martinez, M. J.; Allen, C. P.; Werner-Washburne, M., Multivariate curve resolution of time course microarray data. *BMC bioinformatics* **2006,** *7*, (1), 343.

143. Awa, K.; Okumura, T.; Shinzawa, H.; Otsuka, M.; Ozaki, Y., Self-modeling curve resolution (SMCR) analysis of near-infrared (NIR) imaging data of pharmaceutical tablets. *analytica chimica acta* **2008,** *619*, (1), 81-86.

144. AlHumaidan, F. S.; Hauser, A.; Rana, M. S.; Lababidi, H. M. S., Impact of Thermal Treatment on Asphaltene Functional Groups. *Energy & Fuels* **2016,** *30*, (4), 2892-2903.

145. Craddock, P. R.; Le Doan, T. V.; Bake, K.; Polyakov, M.; Charsky, A. M.; Pomerantz, A. E., Evolution of Kerogen and Bitumen during Thermal Maturation via Semi-Open Pyrolysis Investigated by Infrared Spectroscopy. *Energy & Fuels* **2015,** *29*, (4), 2197-2210.

146. Durand, B.; Espitalié, J., Geochemical studies on the organic matter from the Douala Basin (Cameroon)—II. Evolution of kerogen. *Geochimica et Cosmochimica Acta* **1976,** *40*, (7), 801-808.

147. Silverstein, R. M.; Bassler, G. C.; Morrill, T. C., *Spectrometric Identification Of Organic Compounds*. 1974; p 340-340.

148. Robin, P. L.; Rouxhet, P. G., Characterization of kerogens and study of their evolution by infrared spectroscopy: carbonyl and carboxyl groups. *Geochimica et Cosmochimica Acta* **1978,** *42*, (9), 1341-1349.

149. Painter, P. C.; Snyder, R. W.; Starsinic, M.; Coleman, M. M.; Kuehn, D. W.; Davis, A., Concerning the application of FT-IR to the study of coal: a critical assessment of band assignments and the application of spectral analysis programs. *Applied Spectroscopy* **1981,** *35*, (5), 475-485.

150. Chen, Y.; Mastalerz, M.; Schimmelmann, A., Characterization of chemical functional groups in macerals across different coal ranks via micro-FTIR spectroscopy. *International Journal of Coal Geology* **2012,** *104*, 22-33.

150. Ganz, H.; Kalkreuth, W., Application of infrared spectroscopy to the classification of kerogentypes and the evaluation of source rock and oil shale potentials. *Fuel* **1987,** *66*, (5), 708-711.

151. Lis, G. P.; Mastalerz, M.; Schimmelmann, A.; Lewan, M. D.; Stankiewicz, B. A., FTIR absorption indices for thermal maturity in comparison with vitrinite reflectance R 0 in type-II kerogens from Devonian black shales. *Organic Geochemistry* **2005,** *36*, (11), 1533-1552.

152. Petersen, H. I.; Rosenberg, P.; Nytoft, H. P., Oxygen groups in coals and alginite-rich kerogen revisited. *International Journal of Coal Geology* **2008,** *74*, (2), 93-113.

153. Coelho, R. R.; Hovell, I.; de Mello Monte, M. B.; Middea, A.; Lopes de Souza, A., Characterisation of aliphatic chains in vacuum residues (VRs) of asphaltenes and resins using molecular modelling and FTIR techniques. *Fuel Processing Technology* **2006,** *87*, (4), 325-333.

154. Lin, R.; Patrick Ritz, G., Studying individual macerals using i.r. microspectrometry, and implications on oil versus gas/condensate proneness and "low-rank" generation. *Organic Geochemistry* **1993,** *20*, (6), 695-706.

155. Strausz, O. P.; Lown, E. M. *The chemistry of Alberta Oil Sands, bitumen and heavy oils*; Alberta Energy Research Institute: Calgary, AB, 2003; p 695.

156. H. Ali, L.; A. Al-Ghannam, K.; M. Al-Rawi, J., Chemical structure of asphaltenes in heavy crude oils investigated by n.m.r. *Fuel* **1990,** *69*, (4), 519-521.

157. Wang, L.; Zachariah, A.; Yang, S.; Prasad, V.; de Klerk, A., Visbreaking Oilsands-Derived Bitumen in the Temperature Range of 340–400 °C. *Energy & Fuels* **2014,** *28*, (8), 5014-5022.

158. Visbreaking of Oilsands Bitumen between 150 and 300 oC. In.

159. Lina Maria Yanez Jaramillo, A. d. K., Visbreaking oilsands bitumen at 300°C. . *ACS Division of Energy and Fuels 60*.

160. Ibarra, J.; Muñoz, E.; Moliner, R., FTIR study of the evolution of coal structure during the coalification process. *Organic Geochemistry* **1996,** *24*, (6–7), 725-735.

161. Li, G.; Torraca, G.; Jing, W.; Wen, Z. q., Applications of FTIR in identification of foreign materials for biopharmaceutical clinical manufacturing. *Vibrational Spectroscopy* **2009,** *50*, (1), 152-159.

162. Masson, J. F.; Collins, P., FTIR Study of the Reaction of Polyphosphoric Acid and Model Bitumen Sulfur Compounds. *Energy & Fuels* **2009,** *23*, (1), 440-442.

163. Đorđević, D. M.; Stanković, M. N.; Đorđević, M. G.; Krstić, N. S.; Pavlović, M. A.; Radivojević, A. R.; Filipović, I. M., FTIR Spectroscopic Characterization of Bituminous Limestone: Maganik Mountain (Montenegro). *Studia Universitatis Babes-Bolyai, Chemia* **2012,** *57*, (4).

164. Wang, L. Low temperature visbreaking. MSc thesis, University of Alberta, Edmonton, AB, Canada, 2013.

165. Zachariah, A.; de Klerk, A., Thermal Conversion Regimes for Oilsands Bitumen. *Energy & Fuels* **2016**.

APPENDIX A.

Figure A1 show the dendrogram where the height is the log of posterior of merging and the horizontal axis represent the variables, wavenumbers.

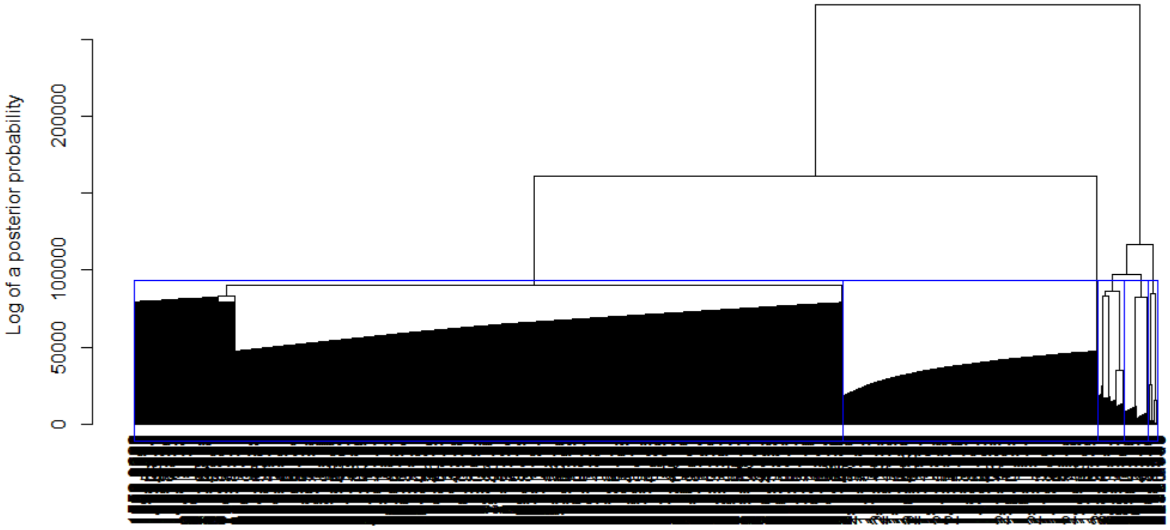


Figure A1. The dendrogram where the height (y-axis) is the log of posterior of merging and the horizontal axis represent the variables, wavenumbers.

Tables A 1 to 5 describes the wave numbers in the each cluster described in (chapter 2) and Figure A1 of the thesis.

Table A-0-1: Wave numbers (cm^{-1}) in cluster I

2982	1408	1325	1254	1182	1111	1040	968	897	825	754	665
2980	1406	1323	1252	1180	1109	1038	966	895	824	752	663
2978	1404	1321	1250	1178	1107	1036	964	893	822	750	662
2976	1402	1319	1248	1176	1105	1034	962	891	820	748	660
2974	1400	1317	1246	1175	1103	1032	960	889	818	729	658
2972	1398	1315	1244	1173	1101	1030	959	887	816	727	656
2970	1396	1313	1242	1171	1099	1028	957	885	814	725	654
2837	1394	1311	1240	1169	1097	1026	955	883	812	723	652
2835	1393	1310	1238	1167	1095	1024	953	881	810	721	650
2833	1391	1308	1236	1165	1094	1022	951	879	808	719	648
2831	1389	1306	1234	1163	1092	1020	949	878	806	717	646
2829	1387	1304	1232	1161	1090	1018	947	876	804	716	644
2827	1385	1302	1230	1159	1088	1016	945	874	802	714	642
2826	1383	1300	1229	1157	1086	1014	943	872	800	712	640
2824	1369	1298	1227	1155	1084	1013	941	870	798	710	638
1489	1367	1296	1225	1153	1082	1011	939	868	797	708	636
1487	1366	1294	1223	1151	1080	1009	937	866	795	706	635
1485	1364	1292	1221	1149	1078	1007	935	864	793	704	633
1483	1362	1290	1219	1148	1076	1005	933	862	791	702	631
1481	1360	1288	1217	1146	1074	1003	932	860	789	700	629
1479	1358	1286	1215	1144	1072	1001	930	858	787	698	627
1477	1356	1284	1213	1142	1070	999	928	856	785	696	623
1475	1354	1283	1211	1140	1068	997	926	854	783	694	621
1474	1352	1281	1209	1138	1067	995	924	852	781	692	619
1472	1350	1279	1207	1136	1065	993	922	851	779	690	615
1431	1348	1277	1205	1134	1063	991	920	849	777	689	613
1429	1346	1275	1203	1132	1061	989	918	847	775	687	611
1427	1344	1273	1202	1130	1059	987	916	845	773	685	608
1425	1342	1271	1200	1128	1057	986	914	843	771	683	606
1423	1340	1269	1198	1126	1055	984	912	841	770	681	604
1421	1339	1267	1196	1124	1053	982	910	839	768	679	602
1420	1337	1265	1194	1122	1051	980	908	837	766	677	600
1418	1335	1263	1192	1121	1049	978	906	835	764	675	
1416	1333	1261	1190	1119	1047	976	905	833	762	673	
1414	1331	1259	1188	1117	1045	974	903	831	760	671	
1412	1329	1257	1186	1115	1043	972	901	829	758	669	
1410	1327	1256	1184	1113	1041	970	899	827	756	667	

Table A -0-2: Wave numbers (cm^{-1}) in cluster II

4000	3929	3857	3786	3715	3643	3572	3501	3429	3358	3286	3215	3144	3072	3001	2770
3998	3927	3855	3784	3713	3641	3570	3499	3427	3356	3285	3213	3142	3070	2999	2768
3996	3925	3853	3782	3711	3639	3568	3497	3425	3354	3283	3211	3140	3069	2997	2766
3994	3923	3852	3780	3709	3637	3566	3495	3423	3352	3281	3209	3138	3067	2995	2764
3992	3921	3850	3778	3707	3636	3564	3493	3421	3350	3279	3207	3136	3065	2993	2762
3990	3919	3848	3776	3705	3634	3562	3491	3420	3348	3277	3205	3134	3063	2991	2760
3988	3917	3846	3774	3703	3632	3560	3489	3418	3346	3275	3204	3132	3061	2989	2758
3987	3915	3844	3772	3701	3630	3558	3487	3416	3344	3273	3202	3130	3059	2988	2756
3985	3913	3842	3771	3699	3628	3556	3485	3414	3342	3271	3200	3128	3057	2986	2754
3983	3911	3840	3769	3697	3626	3555	3483	3412	3340	3269	3198	3126	3055	2984	2752
3981	3909	3838	3767	3695	3624	3553	3481	3410	3339	3267	3196	3124	3053	2822	2750
3979	3907	3836	3765	3693	3622	3551	3479	3408	3337	3265	3194	3123	3051	2820	2748
3977	3906	3834	3763	3691	3620	3549	3477	3406	3335	3263	3192	3121	3049	2818	2746
3975	3904	3832	3761	3690	3618	3547	3475	3404	3333	3261	3190	3119	3047	2816	2745
3973	3902	3830	3759	3688	3616	3545	3474	3402	3331	3259	3188	3117	3045	2814	2743
3971	3900	3828	3757	3686	3614	3543	3472	3400	3329	3258	3186	3115	3043	2812	2741
3969	3898	3826	3755	3684	3612	3541	3470	3398	3327	3256	3184	3113	3042	2810	2739
3967	3896	3825	3753	3682	3610	3539	3468	3396	3325	3254	3182	3111	3040	2808	2737
3965	3894	3823	3751	3680	3609	3537	3466	3394	3323	3252	3180	3109	3038	2806	2735
3963	3892	3821	3749	3678	3607	3535	3464	3393	3321	3250	3178	3107	3036	2804	2733
3961	3890	3819	3747	3676	3605	3533	3462	3391	3319	3248	3177	3105	3034	2802	2731
3960	3888	3817	3745	3674	3603	3531	3460	3389	3317	3246	3175	3103	3032	2800	2729
3958	3886	3815	3744	3672	3601	3529	3458	3387	3315	3244	3173	3101	3030	2799	2727
3956	3884	3813	3742	3670	3599	3528	3456	3385	3313	3242	3171	3099	3028	2797	2725
3954	3882	3811	3740	3668	3597	3526	3454	3383	3312	3240	3169	3097	3026	2795	2723
3952	3880	3809	3738	3666	3595	3524	3452	3381	3310	3238	3167	3096	3024	2793	2721
3950	3879	3807	3736	3664	3593	3522	3450	3379	3308	3236	3165	3094	3022	2791	2719
3948	3877	3805	3734	3663	3591	3520	3448	3377	3306	3234	3163	3092	3020	2789	2718
3946	3875	3803	3732	3661	3589	3518	3447	3375	3304	3232	3161	3090	3018	2787	2716
3944	3873	3801	3730	3659	3587	3516	3445	3373	3302	3231	3159	3088	3016	2785	2714
3942	3871	3799	3728	3657	3585	3514	3443	3371	3300	3229	3157	3086	3015	2783	2712
3940	3869	3798	3726	3655	3583	3512	3441	3369	3298	3227	3155	3084	3013	2781	2710
3938	3867	3796	3724	3653	3582	3510	3439	3367	3296	3225	3153	3082	3011	2779	2708
3936	3865	3794	3722	3651	3580	3508	3437	3366	3294	3223	3151	3080	3009	2777	2706
3934	3863	3792	3720	3649	3578	3506	3435	3364	3292	3221	3150	3078	3007	2775	2704
3933	3861	3790	3718	3647	3576	3504	3433	3362	3290	3219	3148	3076	3005	2773	2702
3931	3859	3788	3717	3645	3574	3502	3431	3360	3288	3217	3146	3074	3003	2772	2700
2627	2555	2484	2413	2341	2270	2199	2127	2056	1985	1913	1842	1771	1699	1628	1556
2625	2554	2482	2411	2339	2268	2197	2125	2054	1983	1911	1840	1769	1697	1626	1555
2623	2552	2480	2409	2338	2266	2195	2123	2052	1981	1909	1838	1767	1695	1624	1553
2621	2550	2478	2407	2336	2264	2193	2122	2050	1979	1907	1836	1765	1693	1622	1551
2619	2548	2476	2405	2334	2262	2191	2120	2048	1977	1906	1834	1763	1691	1620	1549
2617	2546	2474	2403	2332	2260	2189	2118	2046	1975	1904	1832	1761	1690	1618	1547
2615	2544	2473	2401	2330	2258	2187	2116	2044	1973	1902	1830	1759	1688	1616	1545
2613	2542	2471	2399	2328	2257	2185	2114	2042	1971	1900	1828	1757	1686	1614	1543
2611	2540	2469	2397	2326	2255	2183	2112	2041	1969	1898	1826	1755	1684	1612	1541
2609	2538	2467	2395	2324	2253	2181	2110	2039	1967	1896	1825	1753	1682	1610	1539
2608	2536	2465	2393	2322	2251	2179	2108	2037	1965	1894	1823	1751	1680	1609	1537
2606	2534	2463	2392	2320	2249	2177	2106	2035	1963	1892	1821	1749	1678	1607	1535
2604	2532	2461	2390	2318	2247	2176	2104	2033	1961	1890	1819	1747	1676	1605	1533
2602	2530	2459	2388	2316	2245	2174	2102	2031	1960	1888	1817	1745	1674	1603	1531
2600	2528	2457	2386	2314	2243	2172	2100	2029	1958	1886	1815	1744	1672	1601	1529
2598	2527	2455	2384	2312	2241	2170	2098	2027	1956	1884	1813	1742	1670	1599	1528
2596	2525	2453	2382	2311	2239	2168	2096	2025	1954	1882	1811	1740	1668	1597	1526
2594	2523	2451	2380	2309	2237	2166	2095	2023	1952	1880	1809	1738	1666	1595	1524
2592	2521	2449	2378	2307	2235	2164	2093	2021	1950	1879	1807	1736	1664	1593	1522
2590	2519	2447	2376	2305	2233	2162	2091	2019	1948	1877	1805	1734	1663	1591	1520
2588	2517	2446	2374	2303	2231	2160	2089	2017	1946	1875	1803	1732	1661	1589	1518
2586	2515	2444	2372	2301	2230	2158	2087	2015	1944	1873	1801	1730	1659	1587	1516
2584	2513	2442	2370	2299	2228	2156	2085	2014	1942	1871	1799	1728	1657	1585	1514
2582	2511	2440	2368	2297	2226	2154	2083	2012	1940	1869	1798	1726	1655	1583	1512
2581	2509	2438	2366	2295	2224	2152	2081	2010	1938	1867	1796	1724	1653	1582	1510
2579	2507	2436	2365	2293	2222	2150	2079	2008	1936	1865	1794	1722	1651	1580	1508
2577	2505	2434	2363	2291	2220	2149	2077	2006	1934	1863	1792	1720	1649	1578	1506
2575	2503	2432	2361	2289	2218	2147	2075	2004	1933	1861	1790	1718	1647	1576	1504
2573	2501	2430	2359	2287	2216	2145	2073	2002	1931	1859	1788	1717	1645	1574	1502
2571	2500	2428	2357	2285	2214	2143	2071	2000	1929	1857	1786	1715	1643	1572	1501
2569	2498	2426	2355	2284	2212	2141	2069	1998	1927	1855	1784	1713	1641	1570	1499
2567	2496	2424	2353	2282	2210	2139	2068	1996	1925	1853	1782	1711	1639	1568	1497
2565	2494	2422	2351	2280	2208	2137	2066	1994	1923	1852	1780	1709	1637	1566	1495
2563	2492	2420	2349	2278	2206	2135	2064	1992	1921	1850	1778	1707	1636	1564	1493
2561	2490	2419	2347	2276	2204	2133	2062	1990	1919	1848	1776	1705	1634	1562	1491
2559	2488	2417	2345	2274	2203	2131	2060	1988	1917	1846	1774	1703	1632	1560	625
2557	2486	2415	2343	2272	2201	2129	2058	1987	1915	1844	1772	1701	1630	1558	617

Table A0-3: Wave numbers (cm^{-1}) in cluster III

2968	2966	2964	2962	2889	2887	2885	2883	2881	2880	2878	2876	2874	2872	2845	2843	2841	2839	1470	1468	1466	1448	1447
1445	1443	1441	1439	1437	1435	1433	1381	1379	1377	1375	1373	1371	746	744	743	741	739	737	735	733	731	

Table A0-4: Wave numbers (cm^{-1}) in cluster IV

2961	2959	2957	2955	2953	2951	2949	2947	2945	2943	2941	2939	2937	2903	2901	2899	2897	2895	2893	2891
2870	2868	2866	2864	2862	2860	2858	2856	2854	2853	2851	2849	2847	1460	1458	1456	1454	1452	1450	

Table A0-5: Wave numbers (cm^{-1}) in cluster V

2935	2934	2932	2930	2928	2926	2924	2922	2920	2918	2916	2914	2912	2910	2908	2907	2905