DOCTORAL (PHD) THESIS

Omar Péter Hamadi

University of Pannonia

2024

University of Pannonia

Department of Process Engineering

Doctoral School of Chemical Engineering and Material Sciences

Algorithm Development for Reaction and Composition Characterization in Multicomponent Mixtures

Doctoral (PhD) Thesis

Omar Péter Hamadi

DOI:10.18136/PE.2025.927

Supervisors:

Tamás Varga, PhD

Alex Kummer, PhD

2024

Algorithm Development for Reaction and Composition Characterization in Multicomponent Mixtures

Thesis for obtaining a PhD degree in the Doctoral School of Chemical Engineering and Material Sciences of the University of Pannonia

in the branch of Bio-, Environmental and Chemical Engineering Sciences

Written by Omar Péter Hamadi

Supervisors: Tamás Varga, Alex Kummer

propose acceptance (yes / no) …….
(supervisor)

As reviewer, I propose acceptance of the thesis:

Name of Reviewer: …..... …..... yes / no
…….
(reviewer)

Name of Reviewer: …..... …..... yes / no
…….
(reviewer)

The PhD-candidate has achieved ….....% at the public discussion.

Veszprém, …….

…….
(Chairman of the Committee)

The grade of the PhD Diploma …..... (……. %)
Veszprém, …….

…….
(Chairman of the UDHC)

Abstract

Models are an essential part of every type of industrial activity. Based on their complexity, models can rely on a few equations or be expressed in complex mathematical terms. In the case of chemical reactors, the mathematical model should contain the proper thermodynamic and kinetic equations. The most detailed kinetic model possible, which describes the reaction system includes all the reactions that every single component in the system undergoes. When dealing with a multi-component mixture, it is a difficult task to consider all reactions of every single component, and the proper validation of this kind of complex model is simply not feasible. Even, if it is possible to measure the change in the concentration of each component, (in a complex reaction system) the isolation of the effects of the different reactions on one component concentration is possible only with a high level of uncertainty. To solve these tasks, in the past few decade researchers have been developed two main approaches (lumped and molecular models), in which model simplification and order reduction became central problems.

In this work, various approaches to modelling complex chemical processes and phenomena are investigated, drawing on mathematical models. Key aspects include the development of lumped models to describe catalytic hydrocracking experiments, providing insights into the phenomena of catalyst fouling and the development of yield distribution functions. Additionally, the study delves into the selection and optimization of thermodynamic models for hydrogen solubility in hydrocarbons, emphasizing the importance of proper model selection based on operating conditions. Furthermore, a method is proposed to address retention time drifts in gas chromatography analyses of complex mixtures, enabling easier comparability and analysis of chemical changes during processes like plastic waste pyrolysis. Moreover, computational methods integrating molecular similarities and Kovats retention index are explored to refine qualitative analysis in catalytic pyrolysis processes. Despite computational challenges, the study demonstrates significant insights into molecular composition estimation, underscoring the need for further refinement and validation of computational techniques using high-level measurement properties.

Tartalmi összefoglaló

A modern vegyipari tevékenység alapvető részét képzik a különböző berendezéseket leíró matematikai modellek. Összetettségétől függően, egy rendszer leírható néhány egyenlettel, de sok esetben a leírás csak komplex matematikai kifejezésekkel lehetséges. Kémiai reaktorok esetén, a leíró egyenleteknek tartalmaznia kell a megfelelő termodinamikai és kinetikai összefüggéseket is. A legösszetettebb kinetikai modell tartalmazza az összes komponens minden reakcióját. Sok-komponensű rendszerek esetében egy ilyen összetett modell felírása rendkívül nehéz feladat, a modell validálása pedig szinte lehetetlen. Még ha mérhető is az egyes komponensek koncentrációjának változása a reakciórendszerben, a különböző reakciók hatásainak elkülönítése egy adott komponens koncentrációjára csak nagy bizonytalansággal lehetséges. Ennek a problémának a megoldására az elmúlt évtizedekben két fő megközelítést dolgoztak ki: az "*összevont" (lumped)* és a molekuláris modelleket, ahol a modellek egyszerűsítése, a komplexitásuk csökkentése vált központi kérdéssé.

Ezt a központi kérdést alapul véve, munkámban olyan modelleket mutatok be melyek alkalmasak sok-komponensű rendszerek leírására. A vizsgált módszerek közül az első az úgynevezett "*összevont*" megközelítés, melynek diszkrét és folytonos változatát alapul véve meghatároztam a hidrokrakkolás folyamatát leíró egyenleteket. Diszkrét megközelítés esetén a modellegyenleteket a katalizátor dezaktiválódással, a folytonos változatot pedig újfajta hozameloszlási függvényekkel bővítettem. A következőkben javaslatot tettem a hidrogén oldhatóságának becslésére használt termodinamikai módszer kiválasztására, valamint a különböző módszerek becslésének optimálására. Mindezek mellett kidolgoztam egy módszert mely lehetővé teszi a kromatográfiás elemzésekben fellépő retenciósidő-eltolódás csökkentését. A kidolgozott módszert alapul véve egy, a műanyaghulladék katalitikus pirolízését vizsgáló kísérletsorozat termékelegyeiben megjelenő komponenseket azonosítottam, a számolt átlagos Kováts retenciós indexek és különböző molekuláris hasonlósági mértékek felhasználásával.

Table of contents

1 Preface

This thesis is aimed at assessing and further developing the descriptive capabilities of different lumping methodologies. Moreover, it prepares a new modelling approach based on gas chromatographic analysis. The appropriateness of the developed methodologies is investigated by studying three selected problems, namely:

Special Hydrocracking of Sunflower Oil and Kerosene Mixture

- Investigate existing lumped models suitable for describing experimental data.
- Identify and validate model parameters while maintaining possible reaction pathways.
- Analyse measurement data to establish the occurrence of catalyst fouling, particularly at lower liquid loads/higher residence times.
- Develop and integrate three catalyst deactivation models into the kinetic model: Levenspiel Deactivation Kinetic Model, a simplified Eley-Rideal mechanism, and a model based on competitive adsorption.

Pilot-plant Data of Hydrocracking of Kuwait VGO Mixture:

- Develop and validate a novel modelling approach to understand hydrocracking processes, refining the continuous lumping approach to accommodate the complexities of hydrocracking chemistry.
- Utilize three different yield distribution functions to estimate the yield of each component in various cracking reactions.
- Investigate both steady-state and dynamic behaviour of the hydrocracking process.
- Compare simulation results with experimental data to validate the model's accuracy.

GC (Gas Chromatographic) Based Analysis of Waste Plastic Pyrolysis Products

- Enhance the accuracy of qualitative analysis in catalytic pyrolysis processes.

- Develop a fast and simple method to eliminate time drifts between gas chromatography (GC) chromatograms for calculating a proper Kovats index. Propose modifying the k-means algorithm to handle time shifts in chromatograms, emphasizing its potential despite its original purpose of minimizing variance.
- Apply computational methods integrating molecular similarities and Kovats retention index to refine estimation of molecular compositions, especially in scenarios with uncertain retention index database accuracy.

Structure of the thesis

The thesis consists of a total of 7 chapters.

Chapter 1 is a brief introduction to the project, providing an overview of the objectives, motivation, and structure of this work.

Chapter 2. is a literature review covering processes, theories, system-reduction techniques, and modelling approaches considered in the thesis.

Chapter 3. aims to enhance understanding of catalyst deactivation mechanisms and improve the accuracy of modelling hydrocracking processes, particularly in the context of producing high-quality aviation fuel from alternative feedstocks.

Chapter 4. aims to enhance understanding of the dynamic behaviour of a hydrocracker reactor, emphasizing temporal and spatial chemical changes in the reactor.

Chapter 5. presents an algorithm to estimate hydrogen solubility in hydrocarbons, which is a cornerstone of implementing the pressure dependency in models developed in Chapters 3 and 4.

Chapter 6. presents an algorithm to eliminate retention time drifts in Gas Chromatographic data.

Chapter 7. presents a computational approach for better estimation of molecular composition of pyrolysis product based on GC analysis. This approach can be the basis for a new, single-event type modelling approach, in which the estimated molecular composition would be the cornerstone of reaction pathway identification.

At the end of the thesis, one can find the thesis summary and formulation of possible improvements, the thesis points, my publications related to the thesis, and the references.

2 Modeling in multicomponent mixtures

Industrial processes, particularly chemical reactors, rely on accurate modeling to ensure efficient operation and product optimization. In this chapter, we delve into the complexities of system reduction in multicomponent mixtures, introducing various approaches to simplify kinetic models while maintaining predictive accuracy. Additionally, we introduce the applied concepts for computational qualitative analysis, detailing the Kovats retention index and different molecular similarity measures.

First, an overview of the selected processes (hydrocracking and pyrolysis) which we used for the analysis of multicomponent mixtures is provided. Following this, the challenges posed by multicomponent mixtures in developing detailed kinetic models, which lead to the development of lumped kinetic approaches, are introduced. Emphasis is placed on hydrocracking and modeling methodologies such as discrete and continuous lumping, tracing their evolution and applications from historical contexts to modern-day studies.

Finally, the indexes for characterizing pyrolysis products, such as Kovats retention index and molecular similarity indexes, are introduced. Through this chapter, we aim to provide insights into the world of system reduction and qualitative analysis in multicomponent mixtures, offering a comprehensive picture necessary for understanding the upcoming chapters.

2.1 Hydrocracking and Pyrolysis: Selected Processes for Analyzing Multicomponent Mixtures

Hydrocracking is a key process that transforms heavy petroleum fractions into lighter, more valuable products by applying hydrogen and a suitable catalyst. In simple terms, hydrocarbon cracking is a process to break a long chain hydrocarbon into short ones, in the absence of oxygen. Compared to thermal cracking, hydrocracking operates at lower temperatures [1]. As a result of its better catalytic activity, the product fuels have higher quality, with a high hydrogen-to-carbon ratio and lower impurity content [2].

Usually, hydrocrackers are fixed-bed reactors through which the fluid feedstock (oil) and hydrogen are passed downward through the catalyst beds. They typically operate within pressure ranges of 80 to 200 bar and temperature ranges of 300 to 450°C. The two basic and widely applied hydrocracking schemes are two-stage and single-stage hydrocracking. In the case of two-stage hydrocracking, separate hydrotreating (HDS, HDN, and HDO) is applied in stages to separate different products such as H2S and NH3. Organic nitrogen and ammonium are considered poisons to the acidic catalyst. With this setup, their levels can be kept low, which was necessary before the application of zeolitic hydrocracking catalysts, as amorphous silica-alumina catalysts required low levels of organic nitrogen and ammonium. The single-stage procedure contains no interstage separation since the applied zeolite catalyst is remarkably less sensitive to ammonia than the silica-alumina catalyst [3].

Traditionally, in petrochemical plants, hydrocracking reactors are used to convert heavy fractions like vacuum gas oil, tar, etc., into lighter fuels [4] or to convert by-products into high-value products [5], such as the conversion of polyaromatics into toluene, benzene, and xylene [6]. As the demand for renewable energy sources continues to rise, biofuels have become more popular in recent years, emerging as viable alternatives to fossil fuels [7]. However, 95% of the world’s biodiesel is derived from edible oils [8].

The type of the feed significantly influences the hydrocracking chemistry. Hence, a detailed description of possible reactions is not provided in the thesis.

However, in Case study (I), the kinetics of hydrocracking is investigated focusing on a special hydrocracking of sunflower-oil (triglycerides) and kerosene mixture: the reaction pathway for the conversion of triglycerides into alkanes is shown in Figure 2.1-1. The first step is the hydrogenation of the triglyceride, breaking it down into various intermediates presumed to be monoglycerides, diglycerides, and carboxylic acids. These intermediates are then converted into alkanes through three different pathways: decarboxylation, decarboxylation, and hydrodeoxygenation (or dehydration/hydrogenation). The alkanes can further be converted into different lighter components and/or isomers [9].

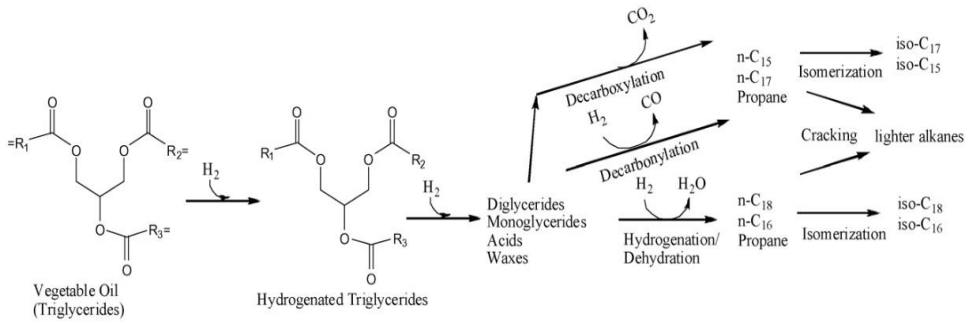

Figure 2.1-1 Reaction pathway for conversion of triglycerides to alkanes (reprinted from [10])

Pyrolysis is a high-temperature recycling process used for the thermal decomposition of organic polymers, converting them into liquid oil, char, and gases [10]. Pyrolysis is similar to cracking, as it involves breaking long-chain hydrocarbons into shorter ones in the absence of oxygen. Both processes can be performed with or without catalysts, thus distinguishing between catalytic and thermal cracking/pyrolysis. The main differences between the two processes are the operating temperature, feedstock versatility, and product yield. The temperature range for thermal cracking is typically higher (above 500°C [11]), while the typical temperature range for pyrolysis can vary from a few hundred to over 1000 °C. [12]. Moreover, the aim of thermal cracking is to maximize the yield of lighter hydrocarbon products, while pyrolysis is used to produce a range of products including liquids, solids, and gases. [13]. The feedstock for thermal cracking primarily contains hydrocarbons, such as crude oil, while various feedstocks can be utilized for pyrolysis, such as biomass or plastic waste [14].

2.2 System reduction techniques

Ideally, the most detailed kinetic model should capture every reaction that each component in the system undergoes [15]. However, when one is modelling multi-component mixture, considering all reactions for each component becomes quite a task, and validating such a complex model proves challenging. Even if we can measure the change in concentration of each component in a complex reaction setup, isolating the effects of different reactions on one component's concentration remains uncertain. To tackle these challenges, researchers have developed different lumped approaches over the past few decades.

2.2.1 Discrete lumping

Discrete lumping approach is performed using characteristics such as TBP (true boiling point), or molecular weight. The reactions between these pseudo-components are determined in advance, and the reactivity of these pseudo-components (which is an overall reactivity of every component in the lump) must be identified. The predictive performance of this kind of model is sufficient in many applications, and the easy integration into the reactor model causes its widespread usage, but the selection of the pseudo-components can be cumbersome, it is usually based on experience or on the aim of model development procedure [16]. This traditional lumping model is used since the 1960's [17], [18]. The first experimental based kinetic study of hydrocraking of real industrial raw material was reported by Qader and Hill in 1969 [19]. They reported that the gas oil hydrocracking, desulfurization, and denitrogenation are all first-order reactions. Orochko and Khimiya (1970) also reported a model to describe the hydrocracking of vacuum distillates in a fixed-bed reactor using a first-order kinetic scheme with four lumps [20]. A first-order lumping approach was developed by Mosby et al. to predict the performance of a residue hydrotreater [21], and Ayasse et al. used this reaction scheme to describe the catalytic cracking of Athabasca bitumen [22]. Stangeland assumed a first order kinetics for a model to predict the yields of a hydrocracker with four parameters: k0 and A describes the reaction rates, C quantifies the butane yield, and B is a constant varies based on the feed and the catalytic process [23].

Weekman and Nace proposed to use second order rate for oil gas cracking and first order for gasoline based on their 3-lump model [24]. This model was further developed by Lee et al, the model was extended from 3 to 10 lumps and from 3 to 20 reactions [25]. The development of discrete lumping models was continuous during the 1990's and 2000's and the application and development of these models are still popular. A kinetic study of hydrotreating of a mixture of Fluid Catalytic Cracking feedstock and light gas oil was reported by Morales-Blancas et al. [26]. In the single-lump kinetic model of hydrodesulfuration and in the five lumps kinetic model of hydrocracking the reaction orders were found to be 2.75 and 1 respectively. Forghani et al. were investigated a non-isothermal, heterogeneous model of a triglyceride hydrocracking reactor. The developed fived lumps model showed a good agreement to the experimental data [27]. Till et al. applied five different global sensitivity analysis methods to prove that a high amount of model uncertainty can be eliminated from lumped kinetic models by reducing the size of the reaction network [28]. Lechleitner et al. developed a four-lumps kinetic model to investigate the physical and chemical changes in a tubular reactor during in co-pyrolysis of PP, HDPE and LDPE [29].

In general, three conditions have to be met to use pseudo-components [30]:

- Measurability: The ability to measure all pseudo-components.
- Adequacy: Pseudo-components must contain sufficient detail to determine all desired product properties.
- Accuracy: Producing fuels from different feedstocks described with the same pseudo-components, results products, which can be characterized using the same pseudo-components.

Assuming a plug flow reactor model, where the longitudinal changes are identified with the residence time, the concentration change of the i^{th} pseudo-component caused by non-reversible reactions can be formulated as:

$$\frac{dc_i}{dt} = \sum_{l=1}^{n} k_l \cdot \prod_{k=i}^{m} c_{k,l} \qquad 2.2\text{-}1$$

Where *k* is the reaction-rate constant, and *c* denotes concentration.

2.2.2 Continuous lumping

Continuous kinetic lumping is an alternative method to predict the composition of the hydrocracking reactor, assuming that the reaction mixture contains an infinite number of compounds [31]. The continuous lumping approach assumes, that the reactivity of each individual component can be described with a continuous function of the boiling point or the molecular weight distribution. Hence, the application of the continuous approach is used for describing processes where all components are undergoing similar types of reactions and have a big advantage in that the component lumping procedure can be omitted.

The continuous-lumping theory was proposed in 1931 by De Donder [32]. Chuo and Ho [33] further developed this theory for nonlinear reactions by applying a reaction type distribution function. The yield distribution function was introduced by Laxminaraasimhan et al. [34]. They assumed that the rate of hydrocracking is a monotonic function of the boiling point (Eq. 2.2-2 - Eq. 2.2-3), and the distribution function determines the amount of the formed species in each reaction, and it is represented by a skewed Gaussian-type distribution function.

$$\theta_i = \frac{TBP_i - TBP(l)}{TBP(h) - TBP(l)} \qquad 2.2\text{-}2$$

$$k_i = k_{max} \cdot \theta_i^{\frac{1}{\alpha}} \qquad 2.2\text{-}3$$

In Eq. 2.2-2 and Eq. 2.2-3, *TBP* denotes the true boiling point, TBP(L) and TBP(h) indicates the possible lowest and highest boiling points in the mixture. θ is the normalised boiling point, *k* is the reaction rate constant, k_{max} is the reaction rate constant for the heaviest component and α is an adjustable parameter. In this case, k_i determines the rate of decomposition of the i^{th} component at a given temperature but does not provide any information about the yield distribution. According to

Laxminarasimhan et al. [34] the continuous lumping approach can be formulated as:

$$\frac{dc(k,t)}{dt} = -k \cdot c(k,t) + \int_{k}^{k_{max}} p(k,K) \cdot K \cdot c(K,t) \cdot D(K) \cdot dK \qquad 2.2\text{-}4$$

Where *p(k,K)* is the yield distribution function, describes the yield of components with reactivity *k* from cracking of species with reactivity *K*, *D(K)* is the species type distribution function, provides information about the concentrations of the chemical species. The yield and species type distribution functions are calculated according to Eq. 2.2-5 - Eq. 2.2-9.

$$p(k,K) = \frac{1}{S_0\sqrt{2\pi}} \cdot \left[e^{-[\{(k/K)^{\alpha_0} - 0.5/\alpha_1\}]^2} - A + B\right] \qquad 2.2\text{-}5$$

$$S_0 = \int_0^K \left[\frac{1}{\sqrt{2\pi}} \cdot \left[e^{-[\{(k/K)^{\alpha_0} - 0.5/\alpha_1\}]^2} - A + B\right]\right] \cdot D(k) \cdot dk \qquad 2.2\text{-}6$$

$$D(k) = \frac{n \cdot \alpha}{k_{max}^{\alpha}} \cdot k^{\alpha-1} \qquad 2.2\text{-}7$$

$$B = \delta \cdot \left[1 - \left(\frac{k}{K}\right)\right] \qquad 2.2\text{-}8$$

$$A = e^{-(0.5/\alpha_1)^2} \qquad 2.2\text{-}9$$

Where, α_0, α_1, δ and S_0 are parameters of the yield distribution function, *n* is the total number of the species represented in the reaction mixture. A detailed solution of the continuous kinetic lumping modelling can be found in [35].

Different approaches have been used to further improve the continuous lumping method. Elizalde et al. applied the continuous lumping approach for modeling the hydrocracking of heavy crude at different severity levels and temperatures. They

found that four of the model parameters are the linear function of the temperature, and there is only one parameter that is almost independent from temperature changes [36]. The next development was not only considering the effect of LHSV and temperature but the partial pressure of hydrogen on hydrocracking kinetics. The parameters of the model were correlated with temperature and pressure using a modified Arrhenius type of equation [37]. The continuous lumping model of hydrocracking was successfully extended with hydrodesulfurization reactions and applied to experimental data. The number of parameters was increased by three, and the results showed a good agreement with both the product distillation and sulphur curves [38]. Modeling the catalyst deactivation was also implemented in a continuous lumping approach. The effect of temperature and time in the stream was studied. The assumed deactivation phenomena were the deactivation by coverage and pore mouth constrictions [39].

2.3 Concepts for computational qualitative analysis

This section of the literature review offers an introduction to the computational methods applied in the qualitative analysis of multicomponent mixtures. It covers key concepts such as the Kovats retention index and various molecular similarity measures. In the upcoming chapters, these concepts will be integrated and utilized to develop a methodology aimed at improving the computation-based estimation of molecular composition.

2.3.1 Kovats retention index

The pyrolysis product contains a huge number of different components, which is usually characterized using gas-chromatography mass spectrometry (GC-MS). However, to identify the fractions in the product and to give a quick response about the performance of the reactor at the operating conditions usually only gas-chromatography is applied. Kovats retention index is a concept introduced by E. Kovats in 1958. To eliminate most of the effect of the instrument (in gas chromatography), the retention index of a compound is calculated relatively to two standard compounds [40]. Kovats retention is calculated according to Eq. 2.3-1.

$$KRI_i = 100\left[n + (N-n)\frac{log(t_i - t_0) - log(t_n - t_0)}{log(t_N - t_0) - log(t_n - t_0)}\right] \quad 2.3\text{-}1$$

Where:

- KRI_i is the Kovats retention index of peak *i*
- *n* the carbon number of the heading n-alkane peak
- t_i is the retention time of the compound *i*
- t_0 is the void time (air peak)

The Kovats retention index serves multiple purposes in gas chromatography. Analysts employ it for compound identification, where the experimental retention index of an unknown substance is compared with a reference database. Additionally, Kovats retention indices are widely used for characterizing and comparing stationary phases across various GC columns, aiding in the selection of the most suitable column for specific separations or analytical tasks. Moreover, in cheminformatics and computational chemistry, Kovats indexes are used to predict the gas chromatographic behaviour of compounds based on their molecular structures [41], [42]. Despite the improved precision and reproducibility offered by the Kovats retention index, achieving absolute system independence remains challenging. Fluctuations in an analyte's retention index on a particular column stationary phase can occur due to operational conditions such as gas flow rate, film thickness, and linear temperature programming. Consequently, variations in reported Kovats retention indexes may stem from inconsistencies in operational conditions across reporting laboratories [43].

The correction of misalignments is important in every field where samples are characterized with any kind of chromatographic data. For example, methods were developed and tested for correction of retention time shifts in case of HPLC analysis of herbal medicines [44], GC x GC data [45], diesel fuel GC profiles [46], drug metabolites LC/MS data [47], and metabonomic GC/MS data [48]. The most commonly used methods to eliminate the time drifts are the wrapping algorithms and principal component analysis (PCA). A clear summary of warping methods for chromatographic signal alignment available in [49]. PARAFAC2 is a generalization

of PCA, which is a powerful and popular tool for handling retention time shifts [50]. However, wrapping method requires the selection of a target chromatogram, which can be difficult or computationally expensive, and the segmentation during the application of PARAFAC2 method is influenced by user chosen parameters [51].

2.3.2 Molecular similarity

Polyethylene pyrolysis produces a variety of hydrocarbon compounds, and homologous series within the pyrolysis product can include alkanes, alkenes, aromatics, and hydrogen gas. The specific distribution of these compounds in the pyrolysis product can vary based on factors such as the temperature of pyrolysis, residence time, pressure, and the presence of catalysts. Additionally, secondary reactions and side reactions may lead to the formation of other compounds. [52]

In organic chemistry, homologous series are groups of molecules that have the same basic structure. The similarity between molecules can be measured by several, molecular fingerprint-based similarity indexes. Some examples are collected in Table 2.3-1, where: Variable *a* represents the count of features where both molecules exhibit a value of 1 (indicating positive matches). Variables *b* and *c* represent the count of features where one molecule has a value of 0 while the other has a value of 1. Variable *d* represents the count of features where both molecules have a value of 0. Consequently, the sum of *a* and *d* signifies the total number of matches, while the sum of *b* and *c* indicates the total number of mismatches.

Table 2.3-1 Similarity indices used for investigation in this thesis.

Similarity index	Formula	Description
Tanimoto Similarity Index [53], [54]	$\frac{a+d}{a+2\cdot(b+c)+d}$	Measures the similarity between two molecules by dividing the size of their intersection by the size of their union.
Dice Similarity Coefficient [55], [56]*:*	$\frac{a}{2\cdot a+b+c}$	It is like the Tanimoto coefficient but assigns greater significance to shared features.
Cosine Similarity [57], [58]*:*	$\frac{a}{\sqrt{(a+b)}\cdot(a+c)}$	Measures the cosine of the angle between two binary vectors.
Sokal Similarity Coefficient [56], [59]*:*	$\frac{a}{a+2\cdot(b+c)}$	It considers both matches and mismatches and is commonly used in biology to compare presence/absence data.
Russell Similarity Index [60], [61]*:*	$\frac{a}{a+b+c+d}$	Russell similarity index, like Sokal, is used for binary data. It considers both concordant and discordant pairs of elements in the sets.
Kulczynski Similarity Index [57], [62]*:*	$\frac{a}{b+c}$	Measures the similarity between two molecules based on the average of the proportions of shared elements in each molecule.
McConnaughey Similarity Coefficient [54], [62]*:*	$\frac{a^2-b\cdot c}{(a+b)\cdot(a+c)}$	Binary similarity measure that considers both matches and mismatches. It is commonly used in ecological and biological studies.
Tversky Similarity Index [53], [63]*:*	$\frac{a}{a+b+c}$	Generalization of the Tanimoto Dice coefficient.

As it was mentioned, these similarities are molecular fingerprint-based similarity indexes. The process of calculating molecular fingerprints involves rigorous steps aimed to encode structural information in a format conducive to various applications such as similarity analysis. Initially, the molecular structure undergoes conversion into a suitable 2D or 3D representation. Subsequently, specific types are assigned to atoms based on their chemical attributes (such as hydrogen, carbon, nitrogen, etc.), accompanied by the incorporation of information concerning the local environment of each atom. Identification of neighbouring atoms and their respective types follows, to capture details regarding to atomic connectivity within the molecule. This information is then utilized to generate a binary or integer vector, representing the presence or absence of specific substructures. Several algorithms are commonly employed for fingerprint generation, including the Extended Connectivity Fingerprint (ECFP) [64], MACCS keys, and Daylight algorithm [65]. Subsequently, a hashing function is applied to reduce the size of the fingerprint vector, which is a crucial step in managing its complexity for storage and comparison purposes. Finally, the fingerprint vector undergoes normalization to ensure comparability across different molecules, thereby addressing variations in molecular size and complexity.

3 Discrete Lumped Model Based Investigation of Hydrocracking of Sunflower Oil and Kerosene Mixture

In this chapter, catalyst deactivation phenomena in the case of the special hydrocracking of a sunflower oil and kerosene mixture are analyzed based on experiments and models. Alternative (bio/waste-originated) fuels are becoming more important to reduce the full life-cycle environmental pollution of transportation. One of the production possibilities for these alternative fuels is co-processing (i.e., catalytic quality improvement) of fossil and bio-based feedstocks. The huge number of individual chemical components in the system increases the complexity of the investigation. Moreover, from a chemical analysis viewpoint, only some of these can be followed during the experiments. Hence, the models to describe the processes in this system (hydrocracker) are usually based on lumps (i.e., component groups). Experimental data on the special catalytic hydrocracking of a sunflower oil and kerosene mixture for the production of high-quality aviation fuel (JET) were used for model development. Experiments were carried out in a fixed-bed tubular reactor over temperatures ranging from 533 to 613 K, pressures ranging from 30 to 70 bar, and LHSV of 1.0 h^-1 and 2.0 h^-1, employing a Pt/H-mordenite catalyst. The objective of this chapter is to investigate the existing lumped models that are suitable for describing the experimental data; moreover, while maintaining the possible reaction pathways, model parameters were identified and validated against the measurements. As a result of the analysis of the measurement data, it has been established that in the case of lower liquid load/higher residence time, a deactivation phenomenon, the so-called catalyst fouling, takes place on the applied catalyst. Three catalyst deactivation models were developed and integrated into the kinetic model: Levenspiel Deactivation Kinetic Model, a simplified Eley-Rideal mechanism, and the last one based on competitive adsorption.

3.1 Introduction

Based on our review of the literature, most of the models developed for estimating the behavior of the investigated system contain three, four, five, and six pseudo-components in the case of modeling vegetable oil hydrocracking, as shown in Table 3.1-1.

Table 3.1-1 Possible pseudo-components

3 pseudo-components	**4 pseudo-components**	**5 pseudo-components**	**6 pseudo-components**
Triglycerides (TG)	Triglycerides (TG)	Triglycerides (TG)	Triglycerides (TG)
Organic liquid product (OLP)	Organic liquid product (OLP)	Light (C_5-C_8)	Gas (G)
Gas and Coke (GC)	Gas (G)	Middle (C_9-C_{14})	Gasoline (GO)
	Coke (C)	Heavy (C_{15}-$C1_8$)	Kerosene (K)
		Oligomerized (>C_{18})	Diesel (D)
			Coke (C)
[66], [67], [68], [69],[70]	[66], [67], [68], [70], [71], [72],	[73], [74], [75]	[66], [67], [68]

A total of thirteen mechanisms were found in the literature. Ten out of these thirteen mechanisms involve the formation of coke or some oligomerized by-product as a pseudo-component; two contain oligomerization reactions, while all mechanisms contain only first-order and non-reversible reactions. The models with 3, 4 and 6 pseudo-components were applied to model the conversion of palm oil into biofuels. The model with 6 pseudo-components predicted the yield and conversion of gasoline fraction with an error of 10% [68]. The model with 4 pseudo-components was able to predict the thermal cracking of waste cooking oil with a maximum error of 23% [71]. The hydrocracking process of triglycerides was described by models containing 5 pseudo-components (with different possible

reaction pathways) with an error of 23% and 43% [73]. Three models were developed to determine the kinetics of a continuous bio-oil catalytic reactor, concentrations calculated with the best model at 450 and 500 °C resulted in a difference of 6.71% and 6.18% compared to the experimental data [70].

The main goal of this chapter is to develop a mathematical model to establish the occurring physical (fouling) and chemical changes. To achieve this goal, four models were applied. The first is a simple lumped model without any deactivation mechanism. The second one is an empirical model called Levenspiel's Deactivation Kinetic Model (LDKM); the parameters of this model do not have any physical meaning. The other two are based on catalyst fouling (deposits). The first one of these is a simple decomposition reaction model on the surface, while the second one is based on the strength of the adhesive bonds. The unknown parameters of all the proposed fouling models were identified based on measurements, and finally, the increase in model complexity is evaluated.

Here, I would like to highlight that the experiments were not performed by the author, but the analysis of the measurements was. The experiments were published in [76].

3.2 Experimental Setup and Results

A fixed-bed tubular reactor (diameter: 29 mm, length: 700 mm) was applied to perform the special hydrocracking reactions of the investigated oil mixture. The feedstock was the mixture of kerosene and sunflower oil in a mass ratio of 3:7. In the experimental system the temperature, the pressure and the LHSV can be controlled, so the hydrocracking experiments were performed at 533-613 K, 30-70 bar, 1.0-2.0 h-1 employing Pt/H-mordenite catalyst (Pt content: 0.45%, specific area, BET: 451 m2/g, micropore volume: 0.18 ml/g, Si/Al ratio from XRF: 19 mol/mol, acid sites by ammonia TPD: 0.82 mmol/g, particle diameter: 1.4 mm). The measurements were performed after reaching a steady state condition.

The experimental results are shown in Figure 3.2-1 and. Figure 3.2-2. In general, it can be observed that the total product yield is directly proportional to the increase in temperature, meaning that the total product yield increases with temperature (in the case of 30 bar and LHSV 1.0 h^{-1}, the total product yields are: 17.6% at 533 K, 18% at 553 K, 20.8% at 573 K, 32.9% at 593 K, and 44.6% at 613 K). However, by increasing the pressure, mainly the proportions of gas and gasoline fractions increase, while the proportion of the diesel fraction decreases.

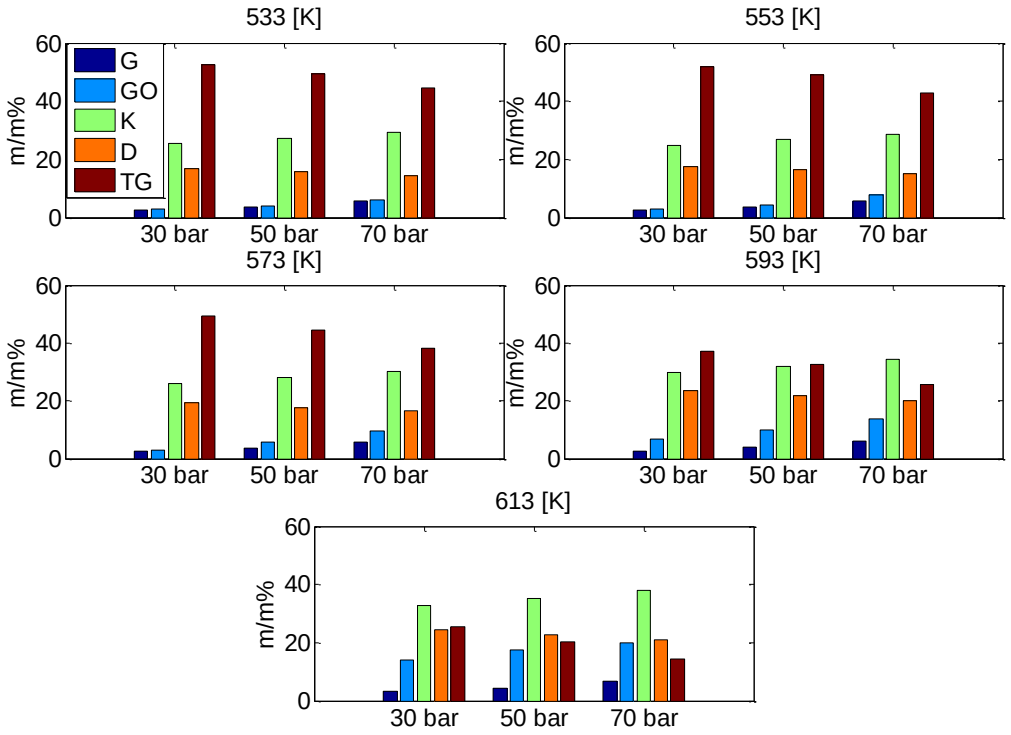

Figure 3.2-1 Experimental data, LHSV = 1.0 h-1

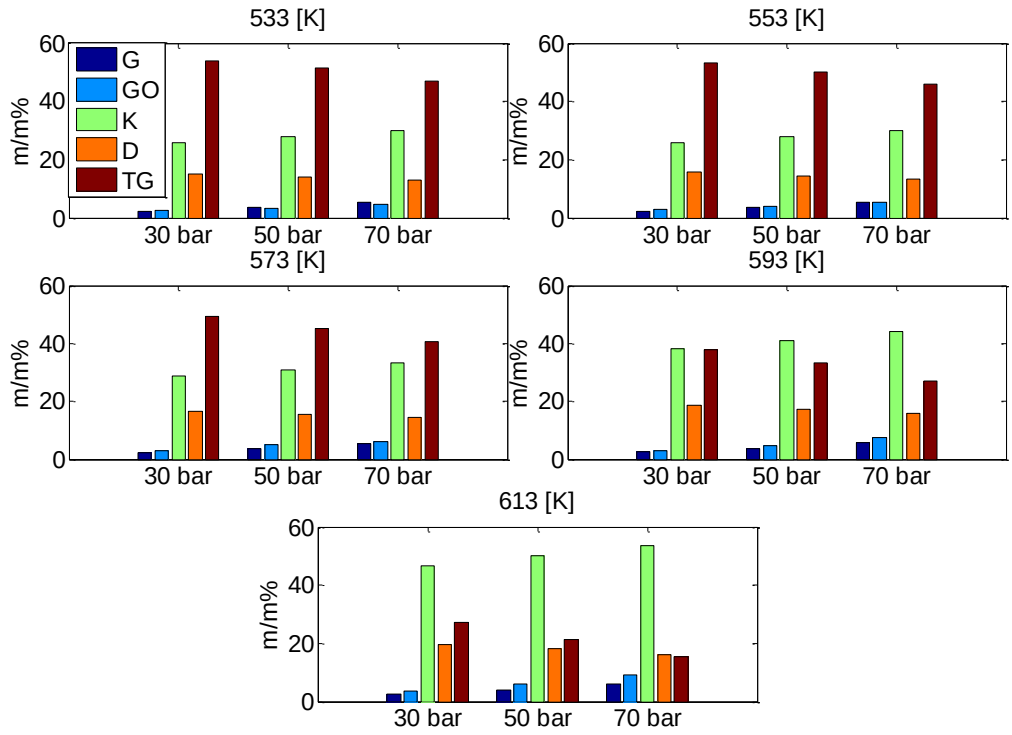

Figure 3.2-2 Experimental data, LHSV = 2.0 h-1

The proportions of conversions measured in different residence times are shown in Figure 3.2-3. In this figure, every point represents a result of the following division and shows how the residence time affects the conversion (Eq. 3.2-1).

$$conversion_{prop.} = \frac{conversion_{LHSV=1\ h^{-1}}}{coversion_{LHSV=2\ h^{-1}}} \qquad 3.2\text{-}1$$

It can be seen that the effect of residence time is not considerable. Despite our expectation that doubling the residence time should have a major effect on the conversion, the two times higher residence time only increases the conversion by 1% to 14% (e.g., at 30 bar and 533 K, the conversion increased by 9.61%). Reaching chemical equilibrium could cause such an effect; however, in this case, this is not an adequate explanation. Here, the mass transfer of hydrogen from the gas phase, the mass transfer of all the other substances from the liquid phase, and the hydrocracking reactions on active sites of the catalyst take place in the system.

Among these, the hydrocracking reactions were considered as the rate-dependent steps in the overall process based on some corresponding works [77], [78]. The chemical equilibrium of these reactions is extremely shifted towards the production of smaller, but hydrogen-saturated molecules if the temperature and the hydrogen concentration are high enough. In addition, the maximal product yield is lower than 55.8%, which means that chemical equilibrium in the case of these reactions cannot have a significant effect on the reaction rates.

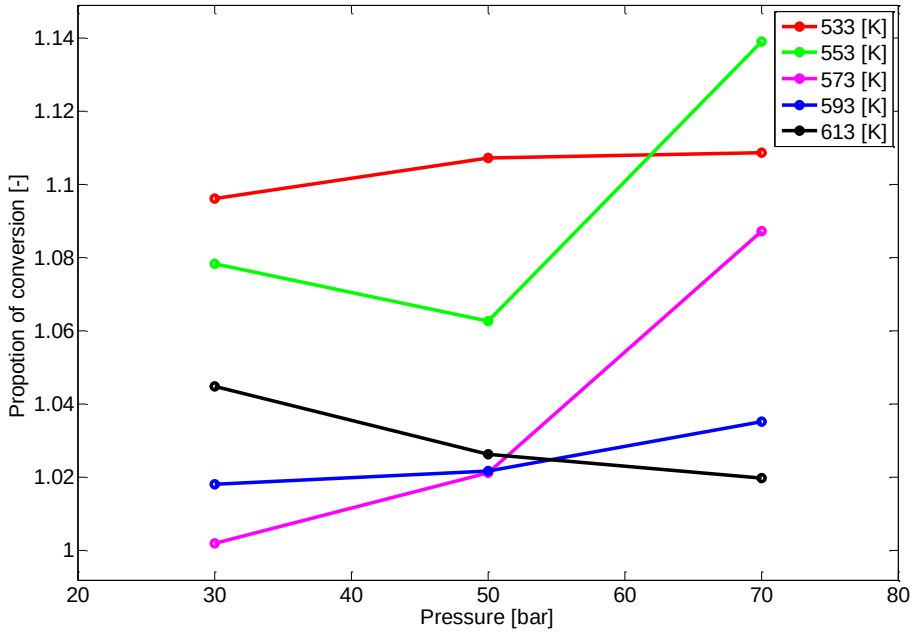

Figure 3.2-3 Proportion of conversions at LHSV = 1.0 h -1 divided with conversion at LHSV = 2.0 h-1

The minor effect of the residence time on conversion is more likely the result of some catalyst deactivation process, in which the large oil molecules cover the macro-pores of the catalyst due to the low fluid velocity. The maximum total product yield is at 613 K, 70 bar, and LHSV = 1.0 h^{-1} (55.8%); the maximum kerosene yield is at 613 °C, 70 bar, and LHSV = 2.0 h^{-1} (53.5%). It should be taken into consideration that, as mentioned earlier, the feedstock contains a considerable amount of kerosene (30%).

3.3 The developed model

3.3.1 Reaction kinetic model – M1

Since the experimental data contain the same pseudo-components, the reaction mechanism based on 6 pseudo-components was investigated during the modeling of the system [66]. During the experiments, the rate of coke formation was not monitored (assuming that there was no coke formation), so in the model, the residue is the remaining triglycerides and the coke. The reaction mechanism shown in Figure 4, and the following differential equations describe changes in the concentration of pseudo-components along the reactor:

$$\frac{dC_{TG}}{d\tau} = -k_0 C_{TG} \tag{3.3-1}$$

$$\frac{dC_D}{d\tau} = k_{1,1} C_{TG} - k_5 C_D \tag{3.3-2}$$

$$\frac{dC_K}{d\tau} = k_{1,2} C_{TG} + k_{5,1} C_D - k_6 C_K \tag{3.3-3}$$

$$\frac{dC_{GO}}{d\tau} = k_{1,3} C_{TG} + k_{5,2} C_D + k_{6,1} C_K - k_7 C_{GO} \tag{3.3-4}$$

$$\frac{dC_G}{d\tau} = k_{2,1} C_{TG} + k_{5,3} C_D + k_{6,2} C_K + k_{7,1} C_{GO} - k_4 C_G \tag{3.3-5}$$

$$\frac{dC_C}{d\tau} = k_{2,2} C_{TG} + k_{5,4} C_D + k_{6,3} C_K + k_{7,2} C_{GO} + k_4 C_G \tag{3.3-6}$$

$$k_0 = k_1 + k_2 \tag{3.3-7}$$

$$k_2 = k_{2,1} + k_{2,2} \tag{3.3-8}$$

$$k_1 = k_{1,1} + k_{1,2} + k_{1,3} \tag{3.3-9}$$

$$k_5 = k_{5,1} + k_{5,2} + k_{5,3} + k_{5,4} \tag{3.3-10}$$

$$k_6 = k_{6,1} + k_{6,2} + k_{6,3} \tag{3.3-11}$$

$$k_7 = k_{7,1} + k_{7,2} \tag{3.3-12}$$

The reactor model was defined as a plug flow reactor; the longitudinal changes were identified with the residence time. Although the experiments were carried out at varying residence times, temperatures, and pressures, the identification process resulted in a distinct set of unknown kinetic parameters being determined for each pressure level individually. The pressure dependence can only be investigated if hydrogen is also considered as a separate component in the kinetic model. The identified reaction rates are higher for lower liquid load and lower for higher liquid load, as can be expected.

3.3.2 Deactivation models

There are several reasons behind catalyst deactivation, which significantly impacts the efficiency of the production process. The physical and chemical mechanisms of deactivation can be divided into four groups: poisoning, coking or fouling, sintering, and phase transportation [79]. Catalyst fouling is a physical process in which certain species from the reaction mixture adsorbs onto the catalyst surface, blocking the active sites or pores reversibly [80].

In fundamental studies, deactivation is approached with an empirical function of time [81]. In further studies, deactivation is assumed to be a function of kinetic variables such as temperature and concentration of reactants, for the same reason that the main reaction rate depends on these kinetic variables [82], [83]. The relationship between the empirical and fundamental aspects has been established [81], [84]. Some authors developed models in which the analysis of fouling are restricted to power law model relations [85], [86], [87]. Expecting non-selective deactivation, a correlation for estimating the activity coefficients has been established [88]. Nonintegral order rate expression, the Langmuir-Hinshelwood mechanism was used to model the fouling process with typified rate expression [89], and this model was also extended with the case of combined series/parallel fouling [90].

As it was mentioned in Chapter 3.2, the minimal impact of residence time on conversion is likely due to a catalyst deactivation process known as fouling. In this phenomenon, large oil molecules, driven by low fluid velocity, accumulate and

block the catalyst's macro-pores, leading to a reduction in its effectiveness. In this thesis, three deactivation models are investigated, namely: Levenspiel's Deactivation Kinetic Model (LDKM), and two based on the Eley-Rideal mechanism.

3.3.2.1 Levenspiel's Deactivation Kinetic Model (LDKM)

LDKM is an empirical model (Eq. 3.3-13) originally developed to describe the long-term deactivation of the catalyst [91]. In Eq. 3.3-14 and Eq. 3.3-15, the elapsed time is considered as the residence time in the reactor. With this approach, the catalyst fouling can be estimated by the addition of the catalytic activity (*a*) to the kinetic model, increasing the number of unknown parameters by two (the deactivation function, Ψ_d* and residual activity, a_s), *d* is the deactivation order.

$$-\frac{da}{dt} = \Psi_d \cdot a^d \qquad 3.3\text{-}13$$

$$d(\tau) = 1 + \left(a_s + \frac{1-a_s}{1+(1-a_s)\cdot\Psi_d^*\cdot\tau}\right)\cdot\left[\Psi_d^*\cdot\tau\cdot\left(\frac{1-a_s}{1+(1-a_s)\cdot\Psi_d^*\cdot\tau}\right)^2\right]^{-1} \qquad 3.3\text{-}14$$

$$\Psi_d(\tau) = \Psi_d^*\cdot\left(\frac{1-a_s}{1+(1-a_s)\cdot\Psi_d^*\cdot\tau}\right)^2\cdot\left(a_s + \frac{1-a_s}{1+(1-a_s)\cdot\Psi_d^*\cdot\tau}\right)^{-d(\tau)} \qquad 3.3\text{-}15$$

3.3.2.2 Eley-Rideal mechanism

Eley-Rideal mechanism, a concept is used in surface-chemistry to describe reaction mechanisms where at least one reactant is absorbed on the surface, reacts with a gas-phase reactant [92].

A typical Eley-Rideal mechanism can be formulated as:

$$A(liquid) + s \leftrightarrow As \qquad 3.3\text{-}16$$

$$As + B(gas) \leftrightarrow Product \qquad 3.3\text{-}17$$

Where *A* is the reactant from the liquid phase which interacts with catalyst active site *s*, and B is the reactant from the gaseous phase. In the hydrocracking process, a gas-phase hydrogen molecule reacts with an adsorbed hydrocarbon molecule on the catalyst surface, leading to the breaking of carbon-carbon bonds and the formation of lighter hydrocarbon products. As catalyst fouling is a physical phenomenon in which a component blocks the catalyst's active sites, the Eley-Rideal mechanism can be used to model this kind of catalyst deactivation.

To describe the effect of catalyst fouling on the overall reaction rate, three models (M2 – M4) were developed and integrated into the kinetic model and all the unknown parameters of these models are identified based on the measurements. The deactivation phenomena, in every case, was considered through the catalyst activity *a* by extending the kinetic equations according to Eq. 3.3-18 which is the general form of Eq. 3.3-1 – Eq. 3.3-6.

$$\frac{dc_i}{d\tau} = a \cdot \sum_{l=1}^{n} k_l \cdot \prod_{k=i}^{m} c_{k,l} \qquad 3.3\text{-}18$$

In **M2** (Levenspiel's Deactivation Kinetic Model), the catalyst activity was calculated according to Eq. 3.3-13- Eq. 3.3-15.

The second deactivation model **M3**, (Simplified Eley-Rideal mechanism of catalyst deactivation) is based on the catalyst fouling caused by one of the considered pseudo-components, the TG (triglyceride) fraction. In this model, the catalyst is also considered as a chemical component and interacts with only this one

pseudo-component (i.e. the catalyst concentration influences all the reaction rates through the activity, but only one component can reduce it, according to Eq. 3.3-20). Therefore, the adhesively bonded component must be assumed as a pseudo-component. The component-mass balance is defined with Eq. 3.3-18 and with the following equations:

$$a = \frac{c_{cat.}}{c_{cat}^{\tau=0}} \tag{3.3-19}$$

$$\frac{dc_{cat.}}{d\tau} = -k_{TG+cat.} \cdot c_{cat.} + k_{TG-cat.} \cdot c_{TG\&cat.} \tag{3.3-20}$$

$$\frac{dc_{TG\&cat.}}{d\tau} = k_{TG+cat.} \cdot c_{cat.} - k_{TG-cat.} \cdot c_{TG\&cat.} \tag{3.3-21}$$

In **M4 (**Fouling based on the strength of adhesive bonds and the number of active sites), all the components are able to bind to active sites of the catalyst at the same time. However, the extents of the adhesions of pseudo-components are different. The number of occupied active sites by each pseudo-component is based on the considered molecular size of components. The size of pseudo-components (A_i) is estimated with Eq. 3.3-22 - Eq. 3.3-24 - where V is the volume of a single molecule, M is the molar mass, N_A is the Avogadro number, ρ is the density and r is the radius -, assuming that the molecules have (quasi) spherical shapes. Table 3.3-1 contains the material data at 298 K. The temperature dependence of the density is neglected in the developed models.

$$V_i = \frac{M_i}{\rho_i \cdot N_A} \tag{3.3-22}$$

$$r_i = \sqrt[3]{\frac{3 \cdot V_i}{4 \cdot \pi}} \tag{3.3-23}$$

$$A_i = \frac{4 \cdot {r_i}^2 \cdot \pi}{M_i} \cdot N_A \tag{3.3-24}$$

Table 3.3-1 The density and the molecular weight of pseudo-components at 298 K.

	TG	D	K	GO	G
M_i [g/mol]	839.1	222.7	178.1	106.1	33.48
ρ [g/cm^3]	0.9182	0.8355	0.7934	0.7369	0.0814

The strength of adhesive bonds and reaction rate constants are the unknown parameters in M4 that should be identified based on the experimental data. The catalyst activity in this model is proportional to the surface concentration and the size of the pseudo-components, so the higher the surface concentration and the bigger the size of the pseudo-component, the more active sites are occupied by the pseudo-component. The surface concentration (c_i) and the activity (a) can be calculated based on Eq. 3.3-25 and Eq. 3.3-26. Where Ci denotes the concentration in the liquid phase and K_i values are the equilibrium constants for the distribution of adsorbates between the surface and the liquid phase. The component mass balance is the same as in the case of M2 (Eq. 3.3-18). The temperature dependence of K_i values was also approached with the van't Hoff equation.

$$c_i = K_i \cdot C_i \tag{3.3-25}$$

$$a = \sum_i \frac{A_{cat} - A_i \cdot c_i}{A_{cat}} \tag{3.3-26}$$

All the introduced models were solved in MATLAB 2017b using the Runge-Kutta method. The reaction rate constants were identified simultaneously at all temperatures (the temperature dependence was estimated with the Arrhenius equation). Thus, the objective function was the square error between the calculated and measurement data for different components, liquid loads at all temperatures:

$$f(\underline{x}^n) = \sum_{T} \sum_{comp} \sum_{LHSV} \left(\frac{y_{exp} - y_{model}}{y_{exp}^{max}} \right)^2 \quad 3.3\text{-}27$$

To determine the minimum of the objective functions in different cases NOMAD algorithm [93] and MATLAB was applied.

3.4 Identification strategy

If the temperature dependence of the reactions is estimated with the Arrhenius equation, the 15 reactions result in 30 unknown parameters. In general, the greater the number of unknown parameters, the more uncertain the model becomes. Hence, a straightforward identification strategy was developed. The algorithm consists of sequential fitting steps. The essence of the algorithm is that the reaction rate constants are identified separately at all temperatures. After the identification, the results are ranked by error; the result (meaning the ten identified reaction rate constants) with the smallest error is fixed. Then, except for the fixed result, the reaction rate constants are identified separately at all temperatures but considering the limitation of the fixed result. The limitation is that if the fixed result temperature is higher than the just identified temperature, the reaction rate constant must be smaller than the fixed one; and if the fixed result temperature is lower than the just identified temperature, the reaction rate constant must be higher than the fixed one. The algorithm is shown in Figure 3.4-1.

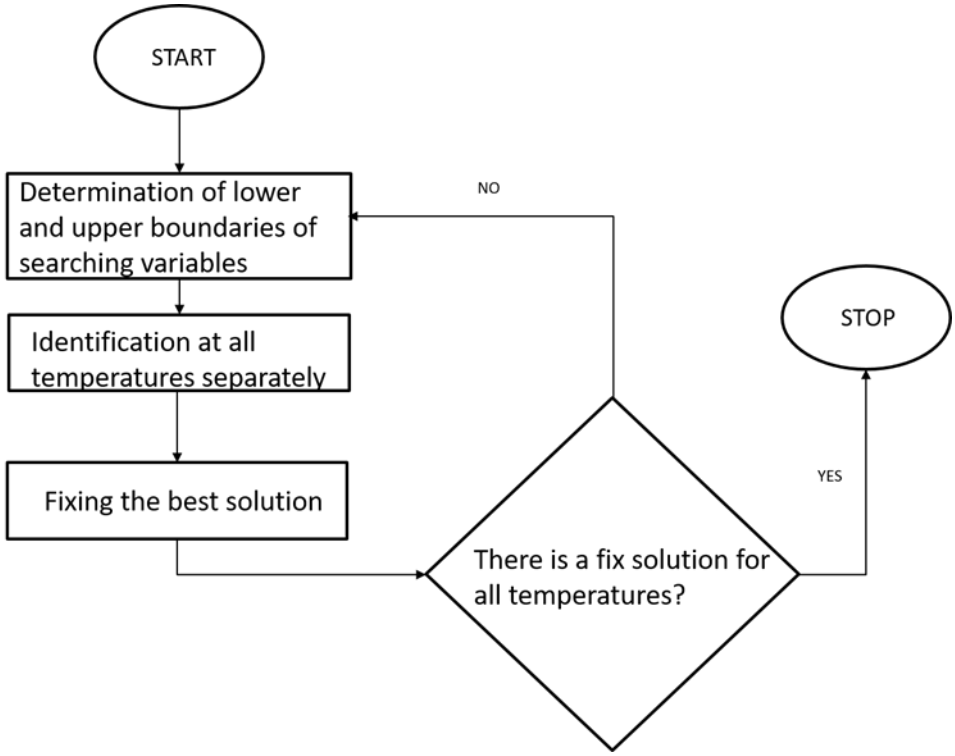

Figure 3.4-1 The proposed identification strategy.

In every optimization task, it is necessary to define the search space with the possible lowest and highest values of every unknown parameter. The range of each dimension of the search space affects the computing time and accuracy. Computing time can be reduced, and accuracy increased if the range of the search space is between 0 and 1. The parameters obtained during the searching process are converted according to the desired lower and upper limits, and then passed to the objective function (Eq. 3.4-1).

$$p = p_2 \cdot (UB - LB) + LB \qquad 3.4\text{-}1$$

Where p is the parameter set, p_2 is the parameter set with range of 0 and 1, UB is the upper bounds and LB is the lower bounds. The first step of the identification is to create two n-by-m matrices, where n denotes the number of different temperatures, and m denotes the number of parameters to be determined. One of the two matrices contains ones, the other contains zeros, later refer to them as LB_0 and UB_1. The rows of the matrices are replaced one by one during the operation of the algorithm, and the lower and upper bounds are selected from these matrices.

The algorithm consists of two embedded cycles, the internal cycle being responsible for simultaneous determination of parameters at different temperatures and for selection of the lower and upper limits, while the external cycle for comparing all the errors of the parameters identified at different temperatures and for fixing the parameters with smallest error to a suitable location in the LB_0 and UB_1 matrices. The steps of the algorithm:

1. In the first step of the external cycle, the parameters are between the range of 0 and 1.
2. The internal cycle computing the parameter values at all temperature and records the model error.
3. The external cycle compares the errors and record the parameters with the smallest error in the LB_0 and UB_1 matrices.
4. The internal cycle adjusts the lower and upper limits at all temperature according to the new LB_0 and UB_1 matrices. Selection criteria:

- Lower bound: The highest of the lower temperatures in LB_0
- Upper bound: The lowest of the higher temperatures in UB_1

5. The 3-4. steps repeated until all the elements of LB_0 and UB_1 have been replaced.
6. Stop

After the identification, with a simple linear regression the parameters of Arrhenius equation can be easily calculated.

3.5 Results

In this section the obtained results are presented in case of all models. The results are mostly shown in a correlation charts. The essence of the correlation charts is that the experimental data are depicted as a function of calculated data that means the better the result is the closer the point to the 45° line.

3.5.1 Reaction Kinetic model (M1)

The reaction kinetic model with the published reaction rate constants was tested. The structure of the model is based on the cracking order. This means that the heavier fractions can be converted into lighter fractions, and coke formation is possible from all fractions and any other can be converted from triglycerides.

In the first step, because not the parameters of the Arrhenius equation were published, it was necessary to calculate the pre-exponential factor and the activation energy from the published reaction rate constants. To determine the parameters of Arrhenius equation, a linear regression method, the linear least squares was used. The kinetic parameters (published in [66]) were identified in a completely different catalyst system. The results accordingly shows that these constants are not suitable for describing the processes that take place in the investigated system (Figure 3.5-1).

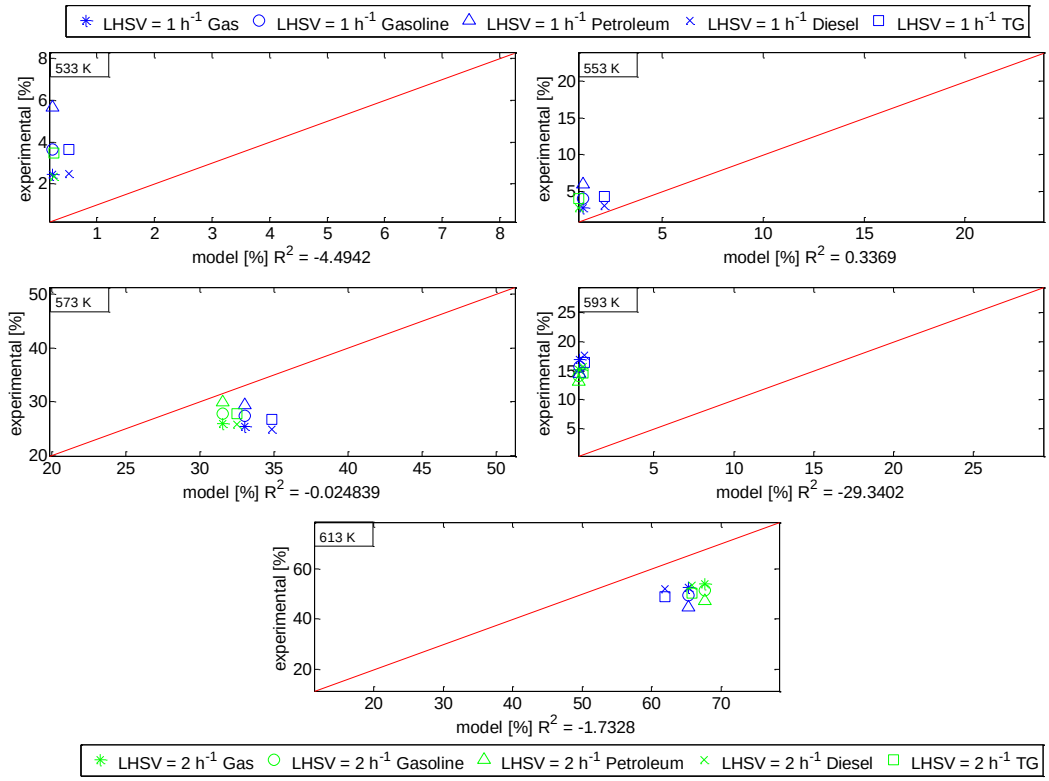

Figure 3.5-1 Correlation charts in case of published reaction rate constants

Since the reaction rate constants in the literature are not suitable for the catalyst used in our system based on some preliminary simulations, the identification of the kinetic constants is necessary. In the first step - because no coke formation is assumed - the coking reaction steps were eliminated from the considered reaction mechanism, therefore the number of the possible reaction steps are decreased to 10. The comparison of the two reaction pathways and the reaction rate constants are shown in Figure 3.5-2.

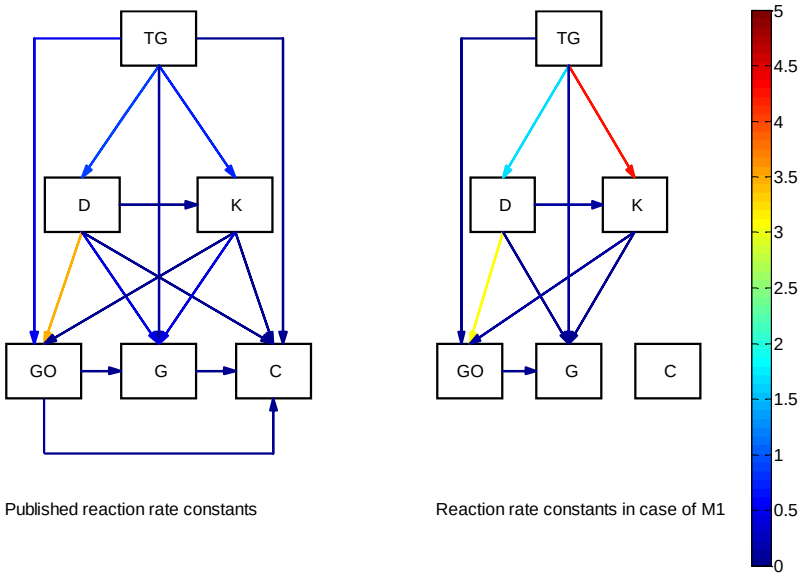

Figure 3.5-2 The considered kinetic schemes with published and identified reaction rate constants (1/min) at 673 K

With the new kinetic model, the objective function value has improved, and the fitting of the correlation charts are also much better. The results at 30 bar are shown in Figure 3.5-3. The objective function value at 30, 50 and 70 bar are 2.50, 2.27 and 1.86 so it can be established that at higher pressure the model is more accurate.

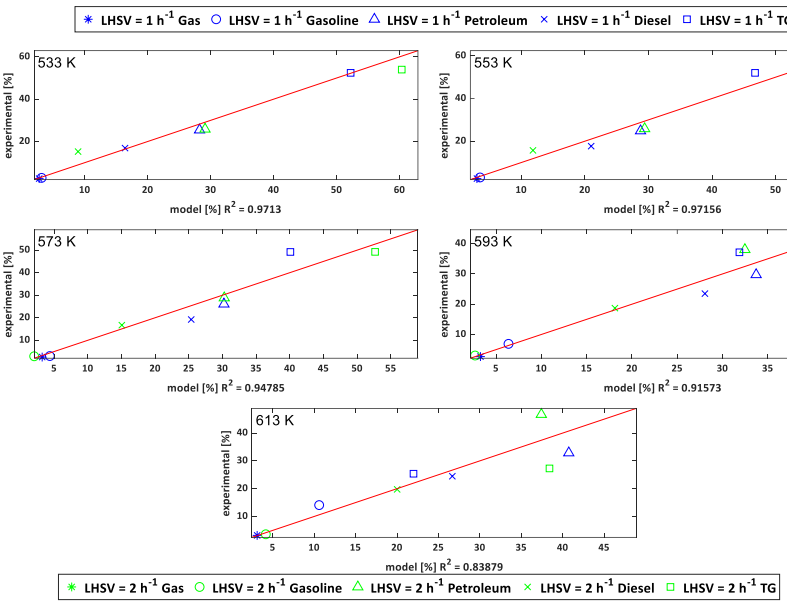

Figure 3.5-3 Correlation charts in case of M1, at 30 bar

The identified parameters are collected in Table 3. In case of all pressures the model is more inaccurate at higher temperatures. Hence, it worth to investigate, how the accuracy is changing with the integration of different catalyst deactivation models.

Table 3.5-1 The identified kinetic parameters for M1 at different pressures

	TG → D	TG → K	TG → GO	TG → G	D → K	D → GO	D → G	K→ GO	K→ G	GO→ G
30 bar										
A0 [1/min]	$1.725 \cdot 10^{3}$	$9.336 \cdot 10^{11}$	$2.604 \cdot 10^{-1}$	2.881	$4.616 \cdot 10^{5}$	$7.732 \cdot 10^{9}$	0	$1.781 \cdot 10^{-1}$	0	0
Ea [kJ/kg]	$3.868 \cdot 10^{4}$	$1.461 \cdot 10^{5}$	$1.217 \cdot 10^{4}$	$1.827 \cdot 10^{4}$	$8.349 \cdot 10^{4}$	$1.210 \cdot 10^{5}$	0	$4.008 \cdot 10^{3}$	0	0
50 bar										
A0 *[1/min]*	$1.106 \cdot 10^{4}$	$1.406 \cdot 10^{17}$	$6.950 \cdot 10^{2}$	$1.400 \cdot 10^{1}$	$4.288 \cdot 10^{2}$	$2.511 \cdot 10^{25}$	0	$1.511 \cdot 10^{1}$	0	0
Ea *[kJ/kg]*	$4.683 \cdot 10^{4}$	$2.044 \cdot 10^{5}$	$4.836 \cdot 10^{4}$	$2.364 \cdot 10^{4}$	$3.303 \cdot 10^{4}$	$3.058 \cdot 10^{5}$	0	$2.110 \cdot 10^{4}$	0	0
70 bar										
A0 [1/min]	$5.557 \cdot 10^{3}$	$1.013 \cdot 10^{15}$	$3.627 \cdot 10^{-1}$	$1.421 \cdot 10^{2}$	$1.014 \cdot 10^{3}$	0	0	$6.531 \cdot 10^{6}$	0	0
Ea [kJ/kg]	$4.362 \cdot 10^{4}$	$1.743 \cdot 10^{5}$	$4.881 \cdot 10^{3}$	$3.189 \cdot 10^{4}$	$3.910 \cdot 10^{4}$	0	0	$8.477 \cdot 10^{4}$	0	0

To prove that optimal parameters were calculated, and to investigate the weights of each reaction pathways, a sensitivity analysis was performed. A small perturbation was made in identified the reaction rate constants; the results are

shown in Figure 3.5-4. The analysis indicates that optimal parameters were found, and it can be established that the greater the change in the value of the objective function as a result of the perturbation, the bigger the weight of the reaction pathway. For example, the TG to D and the K to G can be the most dominating reactions based on the reaction kinetic constants.

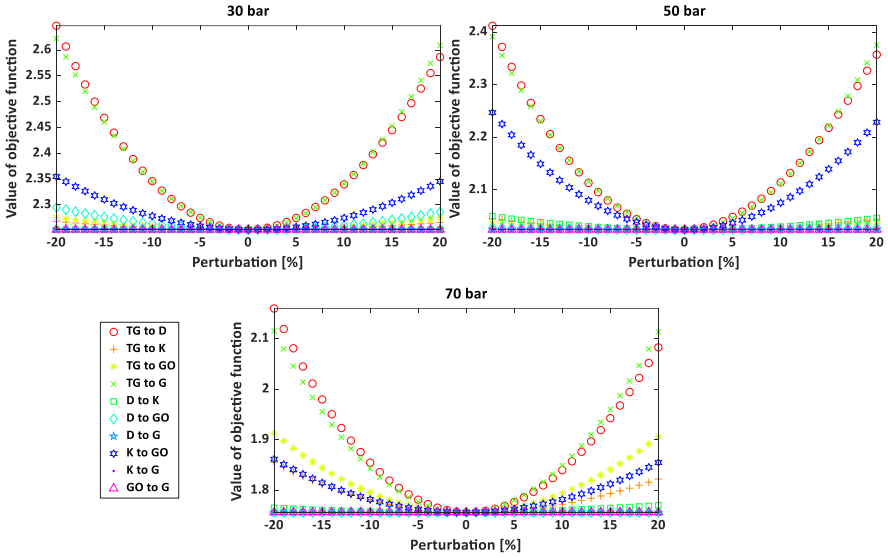

Figure 3.5-4 Sensitivity analysis of identified parameters of M1 model

3.5.2 Deactivation modes (M2-M4)

In the case of **M2**, the results appear to be in line with the assumption that deactivation is present on the catalyst. The value of the coefficient of determination (R^2) increased at almost all temperature and pressure. The previously identified parameters were used to calculate the reaction rate constants and the temperature dependence of the parameters of LDKM model was also estimated with the Arrhenius equation. The deactivation model was used to identify the deactivation on the catalyst continuously at all times, but in the model, the catalyst activity changed only when the LHSV value changed from 1.0 h^{-1} to 2.0 h^{-1}. That means the catalyst bed can be separated into two parts related to the activity and in each part, a constant activity was considered to characterize the effect of catalyst fouling. The identified parameters for M2 are summarized in Table 3.5-2.

Table 3.5-2 he identified parameters of M2.

	a_s	Ψ_d*
30 bar		
A0 [1/min]	$8.746 \cdot 10^{-2}$	$6.144 \cdot 10^{1}$
Ea [kJ/kg]	$1.566 \cdot 10^{4}$	$1.499 \cdot 10^{4}$
50 bar		
A0 [1/min]	$3.103 \cdot 10^{-2}$	$5.614 \cdot 10^{1}$
Ea [kJ/kg]	$1.102 \cdot 10^{4}$	$1.444 \cdot 10^{4}$
70 bar		
A0 [1/min]	$4.417 \cdot 10^{-2}$	$2.206 \cdot 10^{1}$
Ea [kJ/kg]	$1.495 \cdot 10^{4}$	$1.075 \cdot 10^{4}$

As for **M3** and **M4**, the initial activity can be more than 1. When the kinetics parameters were identified (M1), the best solution for both residence time was found (metaheuristic), so the activity above 1 can be expected as the correction of the pre-exponential factors (β) at higher liquid velocity. Thus Eq. 3.3-19 and 3.3-26 have been changed as the following:

$$a = \frac{c_{cat.}}{c_{cat}^{\tau=0}} \cdot \beta \qquad 3.5\text{-}1$$

$$a = \sum_i \frac{A_{kat} - A_i \cdot c_i}{A_{kat}} \cdot \beta \qquad 3.5\text{-}2$$

The previously mentioned β parameter was added to the unknown parameters at 30 bar, and the identified value was set as a constant at 50 and 70 bar. In case M3 $\beta = 3.16$, as for M4 $\beta = 1.27$. The identified parameters for M3 and M4 are collected in Table 3.5-3 and in Table 3.5-4. The conclusion can be made from the parameters of Table 3.5-4, that the temperature dependence of the *K* value is not significant at the investigated temperature and pressure range, and the dependency decreases at higher pressure. Moreover, the assumption made in M3 (the TG fraction can bind the active sites mostly) is proven since the maximum value of the equilibrium constants is at TG in case off all temperature and pressure.

Table 3.5-3 The identified parameters for M3

	$k_{TG+cat.}$	$k_{TG-cat.}$
	30 bar	
A0 [1/min]	$8.876 \cdot 10^{1}$	$1.949 \cdot 10^{2}$
Ea [kJ/kg]	$3.161 \cdot 10^{4}$	$4.475 \cdot 10^{4}$
	50 bar	
A0 [1/min]	$4.009 \cdot 10^{1}$	$4.812 \cdot 10^{1}$
Ea [kJ/kg]	$3.276 \cdot 10^{4}$	$4.287 \cdot 10^{4}$
	70 bar	
A0 [1/min]	$1.212 \cdot 10^{0}$	$1.936 \cdot 10^{-1}$
Ea [kJ/kg]	$3.338 \cdot 10^{4}$	$9998 \cdot 10^{4}$

Table 3.5-4 The identified parameters for M4

	K_{TG}	K_D	K_K	K_{GO}	K_G
			30 bar		
A0 [1/min]	$9.990 \cdot 10^{-1}$	$9.274 \cdot 10^{-1}$	$9.993 \cdot 10^{-1}$	$1.000 \cdot 10^{0}$	$1.061 \cdot 10^{-1}$
Ea [kJ/kg]	$1.343 \cdot 10^{3}$	$9.792 \cdot 10^{3}$	$6.342 \cdot 10^{3}$	$6.761 \cdot 10^{3}$	$2.584 \cdot 10^{3}$
			50 bar		
A0 [1/min]	$9.983 \cdot 10^{-1}$	$2.702 \cdot 10^{-1}$	$5.665 \cdot 10^{-1}$	$4.689 \cdot 10^{-1}$	$2.026 \cdot 10^{-1}$
Ea [kJ/kg]	$1.713 \cdot 10^{3}$	$1.028 \cdot 10^{3}$	$2.492 \cdot 10^{3}$	$2.049 \cdot 10^{3}$	$-4.848 \cdot 10^{0}$
			70 bar		
A0 [1/min]	$6.948 \cdot 10^{-1}$	$8.382 \cdot 10^{-2}$	$7.718 \cdot 10^{-1}$	$3.195 \cdot 10^{-1}$	$1.737 \cdot 10^{-1}$
Ea [kJ/kg]	$-4.975 \cdot 10^{-1}$	$7.126 \cdot 10^{0}$	$3.832 \cdot 10^{3}$	$-4.948 \cdot 10^{0}$	$-4.907 \cdot 10^{0}$

3.5.3 Comparison of models

Four models were tested in this chapter. The first one is a kinetic model without any deactivation phenomenon. The second one is an empirical model (LDKM) which was developed to determine the long-term catalyst deactivation. The third and fourth models are based on the catalyst fouling. The former one is a simplified Eley - Rideal mechanism where only one pseudo-component is able to interact with catalyst. The last model is based on the adhesive bonds. Each model has different number of parameters, which number is increasing in order of the models, numerically M1: 20; M2: 22; M3: 24; M4: 30. Since M2, M3 and M4 models using the parameters obtained in M1, these models carry the uncertainty of M1 which still increases with the number of parameters. In Figure 3.5-5 the coefficients of determination in the case of all models, temperatures and pressures are compared.

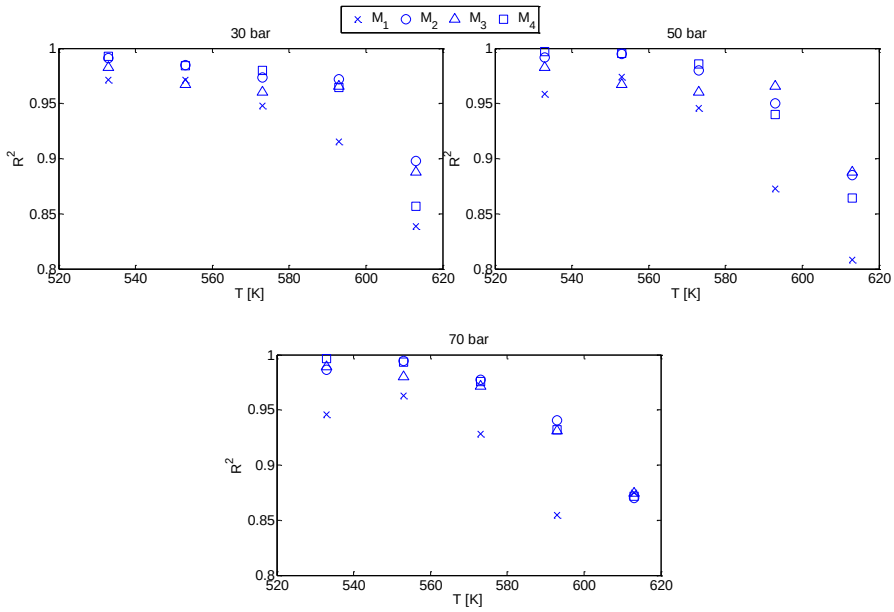

Figure 3.5-5 Coefficients of determination (R2) in case of M1-M4 at all temperatures and pressures.

The objective function values are shown in Figure 3.5-6. As can be seen, the accuracy of models is proportional to their complexity.

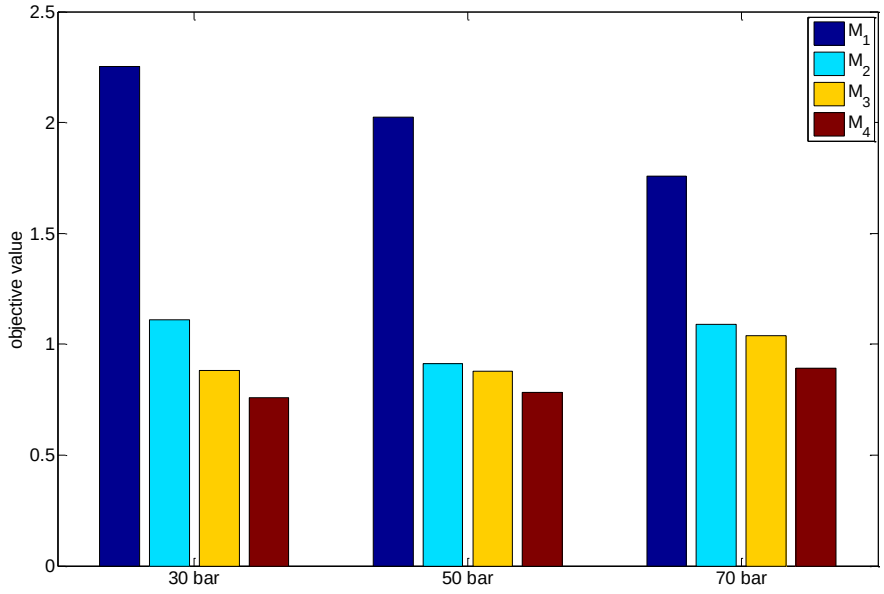

Figure 3.5-6 Objective values in case of M1-M4

It can be established that the objective and R^2 values are more dependent on temperature than the pressure. The estimates provided by M3 and M4 tend to be reliable, but the best result was achieved with M4 (Figure 3.5-7.). The results obtained with M2 are also remarkable, considering the lower level of complexity.

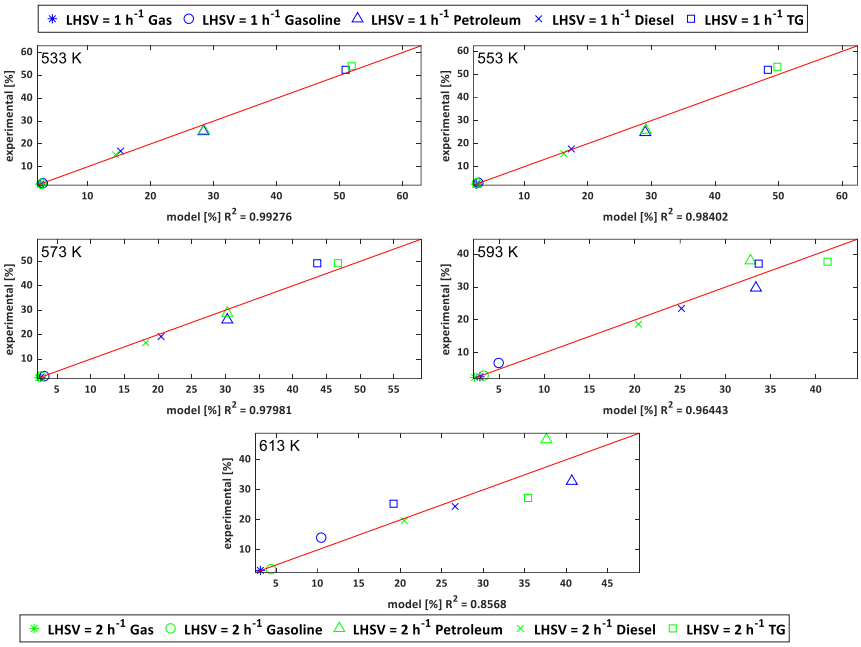

Figure 3.5-7 Correlation charts in case of M4, at 30 bar

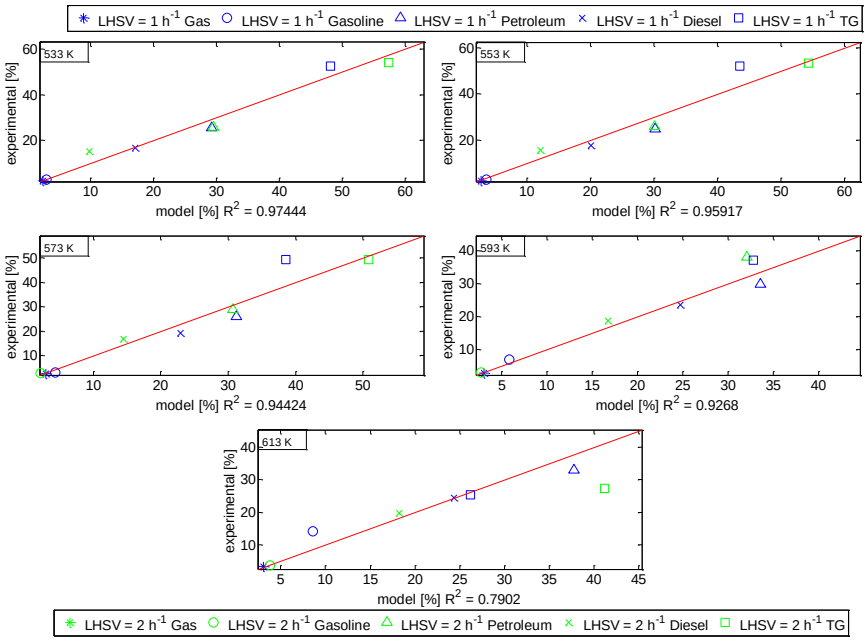

Figure 3.5-8 Correlation charts in case of M2, at 30 bar

The concentration profiles along the reactor are shown in Figure 3.5-9. It can be established that the deactivation effect is much more intensive in case of M3 and M4. The concentrations change faster, and when the activity reaches a critical value, the changes suddenly slow down.

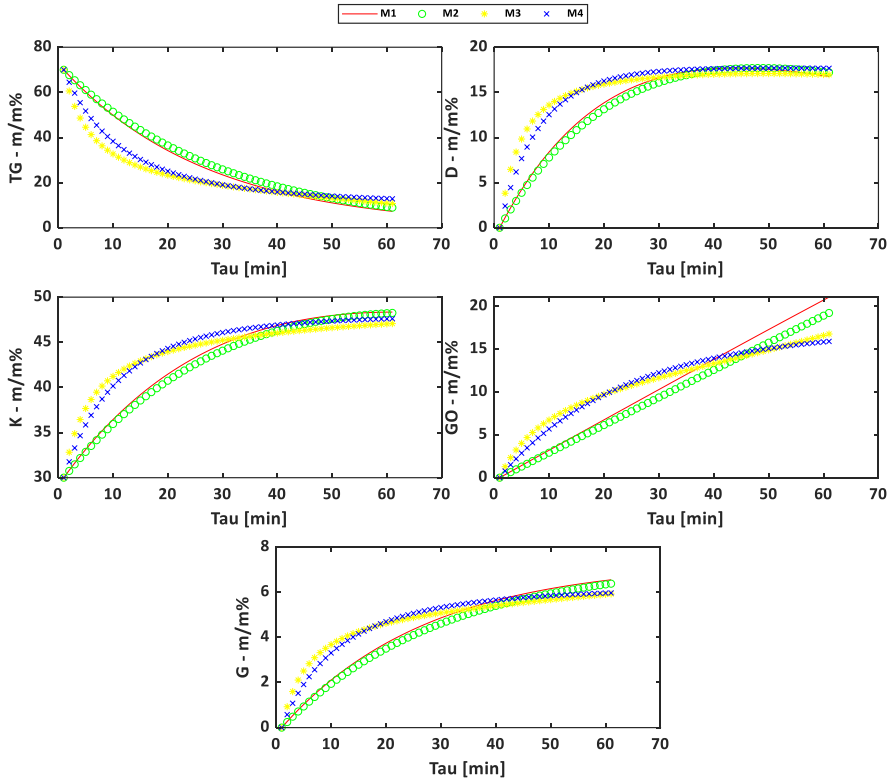

Figure 3.5-9 Concentration profiles in case of all models, at 70 bar and 613 K

3.6 Conclusion

In this work, heterocatalytic reactor models were developed and tested to establish the occurring physical (fouling) and chemical changes. In the kinetic model, the components with similar nature were lumped into pseudo-components, and the process that leads to the chemical transformation was approached with first order non-reversible reactions. The performance of the (kinetic) model was further improved by integrating three different catalyst deactivation models. The first one is an empirical model (LDKM), where the catalyst activity is the function of the residence time. The second one is a simplified ER mechanism, where only one component can reduce the catalytic activity. The last model is based on the strength of the adhesive bonds and the number of active sites. The activity in this model means the number of the unblocked active sites which depends on the surface concentration and the size of the pseudo-components. It has been established that the accuracy of models is proportional to their complexity, so the best results were provided by M4. Besides, based on the improvement in the accuracy of the developed models it can be stated, that due to the low flow rate catalyst fouling has a significant effect on conversion. Moreover, it has been stated that the *TG* fraction can mostly bind the active sites, and the adsorption ability of the molecules in all fractions is not depending on temperature significant, and the dependency decreases at high pressure.

4 Distributed parameter model-based continuous lumping approach: an application to a pilot-plant hydrocracking reactor

A novel modelling approach that captures the complex chemistry of the hydrocracking process and its detailed solution is described in this chapter. The fundamentals of the developed model are coming from the well-known continuous lumping approach. The further development of the continuous lumping approach is that yield of each component in the different cracking reactions were estimated with three different yield distribution functions, moreover the originally applied skewed-Gaussian type yield distribution was reformulated and all three approaches were solved as partial differential equations. With the developed method not only the steady state but the dynamic behaviour can be investigated. The unknown parameters of the models were estimated, and the model predictions were validated using previously published experimental data. The experimental data is pilot-plant data of VGO-hydrocracking of Kuwait VGO at four severity levels. Based on the model error and predicted hydrocracking trends the skewed-Gaussian type yield distribution was the best for prediction.

4.1 Introduction

The continuous and discrete kinetic lumping models provide a continuous description of the component concentrations as a function of the reaction time, but there is no information about the component concentrations in the reactor tube at any axial position.

This work aims to develop a lumped model which can describe the changes in time and space, using the same data (pilot-plant data of VGO-hydrocracking of Kuwait VGO at four severity levels) [94] as Laxminarasimhan et al. [19], as it seems to be that there were no attempts to develop this approach. Three different model approaches were investigated:

- The first one assumes that the cracking rate of every component is the same.
- The second approach assumes that the C-C bond strengths are the highest for the centre C-C bond and decrease monotonically toward the end bonds.
- The third approach assumes that the amount of the formed species in each cracking reaction is represented by the same skewed Gaussian-type distribution function.

As per our knowledge, among these three approaches only the Gaussian-type distribution was applied in previous publications to describe the selectivity distribution in case of hydrocracking reactions. With the proposed model the reactor volume/residence time can be investigated and optimized.

4.2 Model

In this chapter, the fundamentals which were taken over from the continuous-lumping approach and the three different yield distribution functions are introduced.

4.2.1 Fundamentals

The mathematical description of the hydrocracker reactor is based on the material balance on a plug flow reactor, which takes into account the temporal and spatial changes (4.2-1).

$$\frac{\partial c_i}{\partial t} = -\frac{\partial}{\partial x}(v_x \cdot w_i) + R_i \qquad 4.2\text{-}1$$

Where w_i is the weight fraction of the i^{th} component, t is time, x is the axial position in the reactor tube, v_x is flow velocity and R is the rate of generation. The hydrocracker is assumed to be a single-phase isothermal reactor and the effect of the density changes on space velocity is neglected in this study. The decomposition is assumed to be first-order chemical reactions based on beta-scission.

The idea of using the normalised true boiling point (*See in: Eq. 2.2-2 or* $\theta_i = (TBP_i - TBP(l)) / (TBP(h) - TBP(l))$) as the key property to identify the components comes from the continuous lumping theory, moreover the idea that the rate of hydrocracking is a monotonic function of the boiling point was also applied (*See in Eq. 2.2-3 or* $k_i = k_{max} \cdot \theta_i^{1/\alpha}$) [19]. The effect of α on the distribution of reaction (cracking) rate constant is shown in Figure 4.2-1, where the rate of cracking of the heaviest component was set to one.

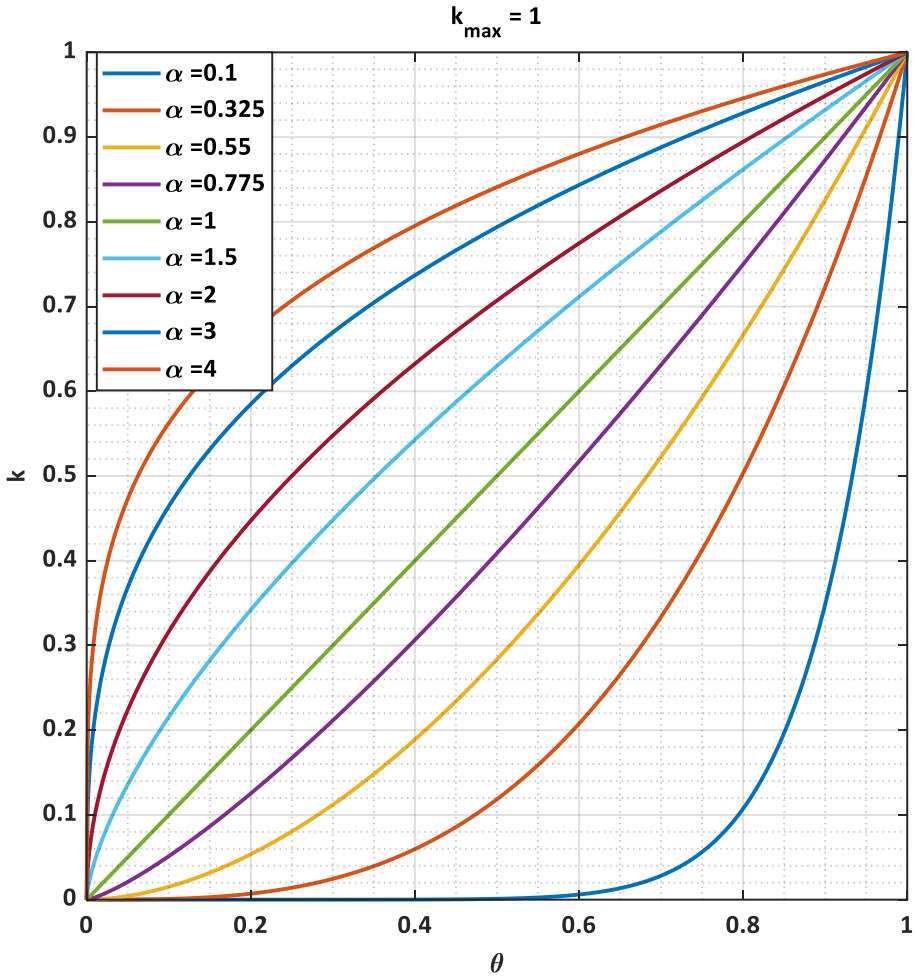

Figure 4.2-1 The effect of α on the reaction rate constant distribution

4.2.2 Selectivity distribution

This section provides three different approaches to estimate the distribution of the cracking rate constants, so the yield of each component in the different cracking reactions.

The simplest approach is the **equidistant division** of the cracking rate constants. The rate constant of the cracking of l^{th} component (k_l) is divided by the number of the lighter components and the resulted number will be the rate of reaction for each lighter component (4.2-2). This means that the weight fraction of each lighter components would be increased in the same extent caused by a single cracking reaction.

$$k_{l,i} = \frac{k_l}{l-1} \qquad 4.2\text{-}2$$

The second approach assumes of **decreasing C-C bond strength**. It means, that the C-C bond strengths are highest for the centre C-C bond and decrease monotonically toward the end bonds in the molecules. This means that the probability of formation of methane is the highest and the probability of breakdown the molecule into two equal parts is the lowest, this probability is denoted as $a_{l,i}$. The calculation of the reaction rate constant is divided into three steps:

1. $a_{l,I}$ can be calculated for chain length j from $\frac{l}{2}$ to $l-1$ with equation 6 and can be reflected over the centre (Eq. 4.2-3, Figure 4.2-2 (a)) .
2. Because the reaction rates are calculated from the weight fractions (so the stoichiometric coefficient is not considered) the rate constants have to be normalised with the weights of the product molecules. The molecular weights are unknown (the products are pseudo components), so we assumed that the product with n=50 is fifty times heavier than the product with n = 1 (Eq. 4.2-4, Figure 4.2-2(b)).
3. Equation 4.2-5 is to integrate the $k_{l,i}$ curve equal to k_l (Figure 4.2-2(c)) .

$$a_{l,i} = k_l \cdot \theta_i^{\frac{1}{\alpha_2}} \qquad 4.2\text{-}3$$

$$b_{l,i} = \frac{a_{l,i}}{n_{max} - n_i} \qquad 4.2\text{-}4$$

$$k_{l,i} = \frac{b_{l,i}}{\sum_{i=1}^{l-1} b_{l,i}} \cdot k_l \qquad 4.2\text{-}5$$

α_2 has the same effect on the $a_{l,i}$ distribution curve as α on k (Figure 4.2-1). However, the range of α_2 is between 0 and 1000. If α_2 =1000, the $a_{l,i}$ curve is nearly linear (Figure 4.2-2),meaning the only difference between M1 and M2 is step 2 (Eq. 4.2-4).

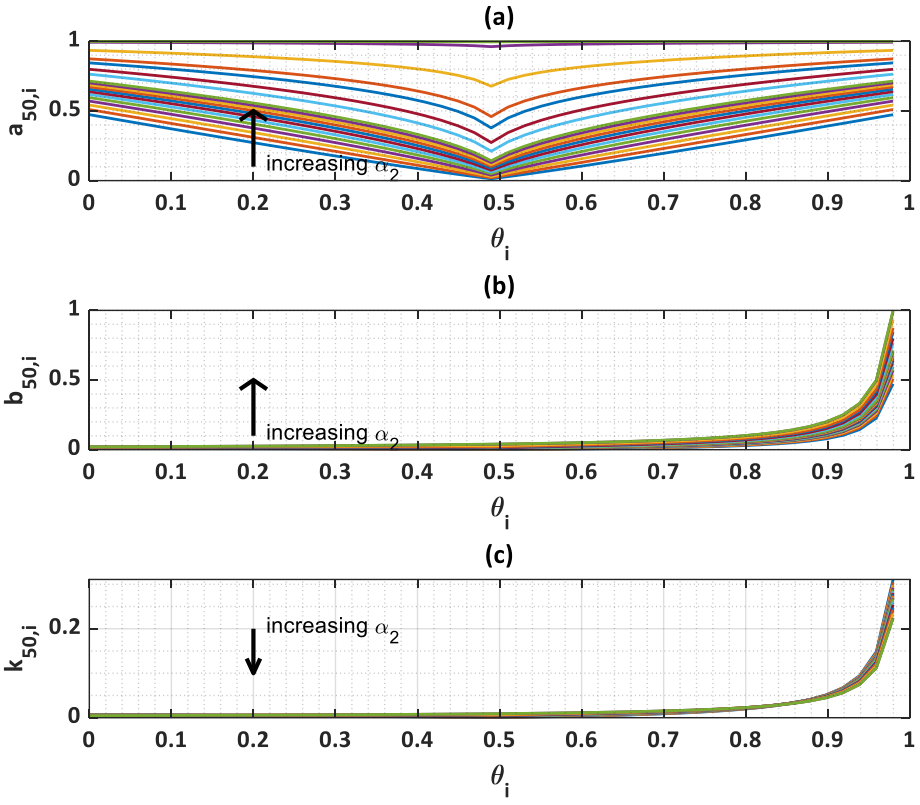

Figure 4.2-2 The effect of α_2 on Equations 4.2-3 - 4.2-5 in the range of: 0.9 <α_2 1000

The third approach assumes that the amount of the formed species in each cracking reaction is represented by the same **skewed Gaussian-type distribution function** (4.2-6).

$$f_{\alpha_3}(x) = \frac{\phi(x) \cdot \Phi(\alpha_3 \cdot x)}{\Phi(0)} \qquad 4.2\text{-}6$$

Where $\phi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2}$ is the probability density function of normal distribution, $\Phi(x) = \int_{-\infty}^{x} \phi(t)dt$ is the corresponding cumulative distribution function and α_3 determines the direction and measure of skewness. $k_{l,i}$ is normalized, so the integral of the function is equal to k_l:

$$k_{l,i} = \frac{\frac{1}{s} f_{\alpha_3}\left(\frac{x_{i-u}}{s}\right)}{\sum_{i=1}^{l-1} \frac{1}{s} f_{\alpha_3}\left(\frac{x_{i-u}}{s}\right)} \cdot k_l \ \{x \in R \mid a \le x \le b\} \qquad 4.2\text{-}7$$

Where a=-3 and b=3 are the endpoints of the interval of x, u and s are for adjusting the mean and the variance of the distribution curve respectively. x is divided into the same number of point as θ, so the coherent points can be easily matched with each other ($k_{l,i}(x_i) = k_{l,i}(\theta_i)$). Figure 4.2-3 shows the effect of parameters on $k_{l,i}$ (Eq. 4.2-7). As was mentioned, u is used to adjusting the mean of the distribution curve, which means that u determines the dominant pathways for product formation. Through the variance, s determines the ratio of different product formations.

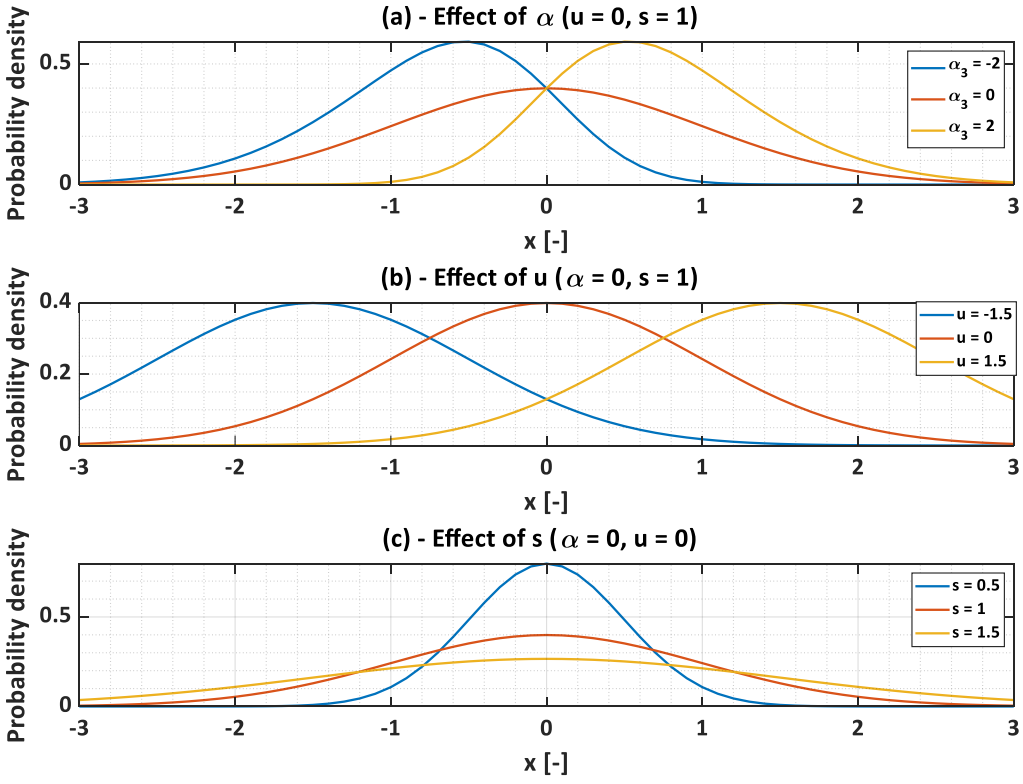

Figure 4.2-3 Effect of parameters on the skewed Gaussian distribution function.

4.3 Model Solution

All the introduced models were successfully applied to experimental data previously published by Bennett and Bourne [94]. The experimental data contains product yields from the hydrocracking of Kuwait vacuum gas oil at four residence times: 0.383, 0.952, 1.724, and 2.5 h. The models were implemented in MATLAB 2019a and the partial different equations were solved using method-of-lines technique [95]. In the method of lines procedure, the partial differential equations are approximated by ordinary differential equations based on finite differences. In this case, it means, that the spatial changes are discretized in the way, that the output of one discrete part is the input of the next one. In this form, the spatial changes are calculated by the Euler method, and the changes in time can be calculated by for example Runge-Kutta method.

Due to the nonlinear nature of the identification task, the usage of a global nonlinear optimization algorithm is needed. In this case, the so-called NOMAD (Nonlinear Optimization by Mesh Adaptive Direct Search) algorithm was chosen which is intended to use for time-consuming black box simulations with a small number of variables. One of the main advantages of the algorithm is that independently of the starting point, it can globally converge to a point that satisfies local optimality conditions based on local properties of the functions defining the problem [96].

The unknown parameters of the models were estimated using experimental data, and the abovementioned NOMAD algorithm [96] was applied to determine the minimum of the objective function (4.3-1). where y_{exp} is the experimental and y_{model} is the predicted values.

$$q = \sum_{i} \sum_{LHSV} \left(y_{exp} - y_{model}\right)^2 \qquad 4.3\text{-}1$$

Table 4.3-1 contains the unknown parameters for the different models. The number of points to discretize the reactor volume on was set to N = 10 and the

number of the components was set to $n = 50$. In this case the calculation time is not considerable, but the fluid velocity profile is nearly a plug flow. The simulation time was set a huge number (1000 hours), but the simulation was automatically terminated when the reactor reached the steady state operation (an event function was applied to detect steady state condition). The volume of the reactor is assumed to be 1 m^3 so the volumetric flowrate is: $v_x = 1/\tau$.

Table 4.3-1 Unknown parameters of the different models

	M1	**M2**	**M3**
k_{max}	✓	✓	✓
α	✓	✓	✓
α_2	×	✓	×
α_3	×	×	✓
u	×	×	✓
s	×	×	✓

4.4 Results

All the identified parameter values, the lower and upper bounds (*LB* and *UB*) of the searching domains and the model errors can be found in Table 4.4-1. The following criteria were followed to set the lower and upper limit for parameter optimization:

- The identified value for k_{max} and α by Laxminaraasimhan et al. was 1.35 (for both), so the boundaries for these two parameters were set near to this value with some tolerance due to the different selectivity distribution functions applied.
- Based on Figure 4.2-2, the $a_{l,i}$ curve is nearly linear if $\alpha_2 = 1000$, meaning there is no further effect on the result if the α_2 value exceeds 1000.
- α_3, u and s are the parameters of the skewed normal distribution function. α_3, u and s are the parameters of the skewed normal distribution function. The limits of these parameters were set to be able to estimate any kind of shape which can be achieved by a skewed normal distribution.

Results show that the accuracy of models is proportional to their complexity: the prediction capability of M3 is the best, and M1, with the linear distribution of the rate constant, performs the worst. In case of M2, α_2 reached the upper limit so the $a_{l,i}$ curve is nearly linear (Figure 4.2-2). This means that by applying Eq. 4.2-4 on $k_{l,I}$ calculated by Eq. 4.2-2 and using k_{max} and α parameters of M2, same results can be obtained with M1.

Table 4.4-1 Identified parameter values, calculated model errors and searching domains

Model	**M1: Equidistant division**			**M2: Decreasing C-C bond strength**			**M3: Skewed Gaussian-type**		
Error (q)	0.1156			0.0474			0.028		
Parameter	***Value***	***LB***	***UB***	***Value***	***LB***	***UB***	***Value***	***LB***	***UB***
k_{max}	2.2473	0	5	2.5165	0	5	1.429	0	5
α	0.2584	0	2	0.5545	0	2	1.502	0	2
α_2	-	-	-	1000	0	1000	-	-	-
α_3	-	-	-	-	-	-	-2.938	-10	41
u	-	-	-	-	-	-	-3.925	-5	5
s	-	-	-	-	-	-	8.4032	1e-5	15

Figure 4.4-1 shows the sum of the identified reaction rates for all individual pseudo components. According to M1, this selectivity function is monotonic and decreases with the increased boiling point. This result fits to the assumption of equidistant division: the lighter the component, the more reactions it takes part in with the same reaction rate constant as the heavier ones, meaning that the resulted total rate of formation has to be larger for lighter components. M2 and M3 show a skewed Gaussian distribution for selectivity, but the direction and amount of skewness is different. M2 assumes that the formation of heavier components is more pronounced, but according to M3 the formation of lighter components are slightly faster reactions.

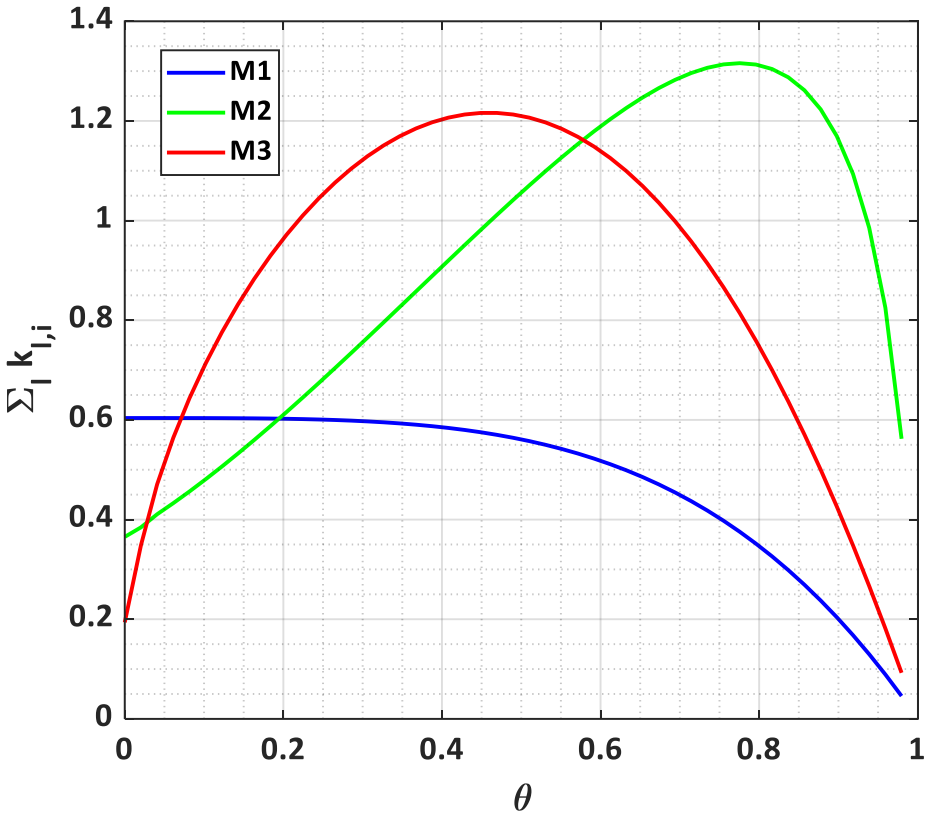

Figure 4.4-1 Selectivity function: sum of the identified reaction rate constants.

Comparison of the model predictions with experimental data (where ω is the weight fractions) is shown in Figure 4.4-2. With M2 and M3, the normal hydrocracking trends are predicted well with respect to the residence time, but M1 is not suitable to predict the hydrocracking process with tolerable precision, the distribution getting wider and more dominant toward lighter components. The prediction of M3 (dashed line) is the most accurate in the case of higher residence (red and magenta dashed lines) times and lighter components, but the error starts to increase at lower residence times (green dashed line) and heavier components. The figure also shows the calculated weight fractions obtained with continuous lumping approach from [97]. A comparison between M3 and the continuous lumping data reveals that M3 performs better at higher residence times and lower TBP. However, at higher TBP and lower residence times, the continuous lumping approach appears to deliver better results.

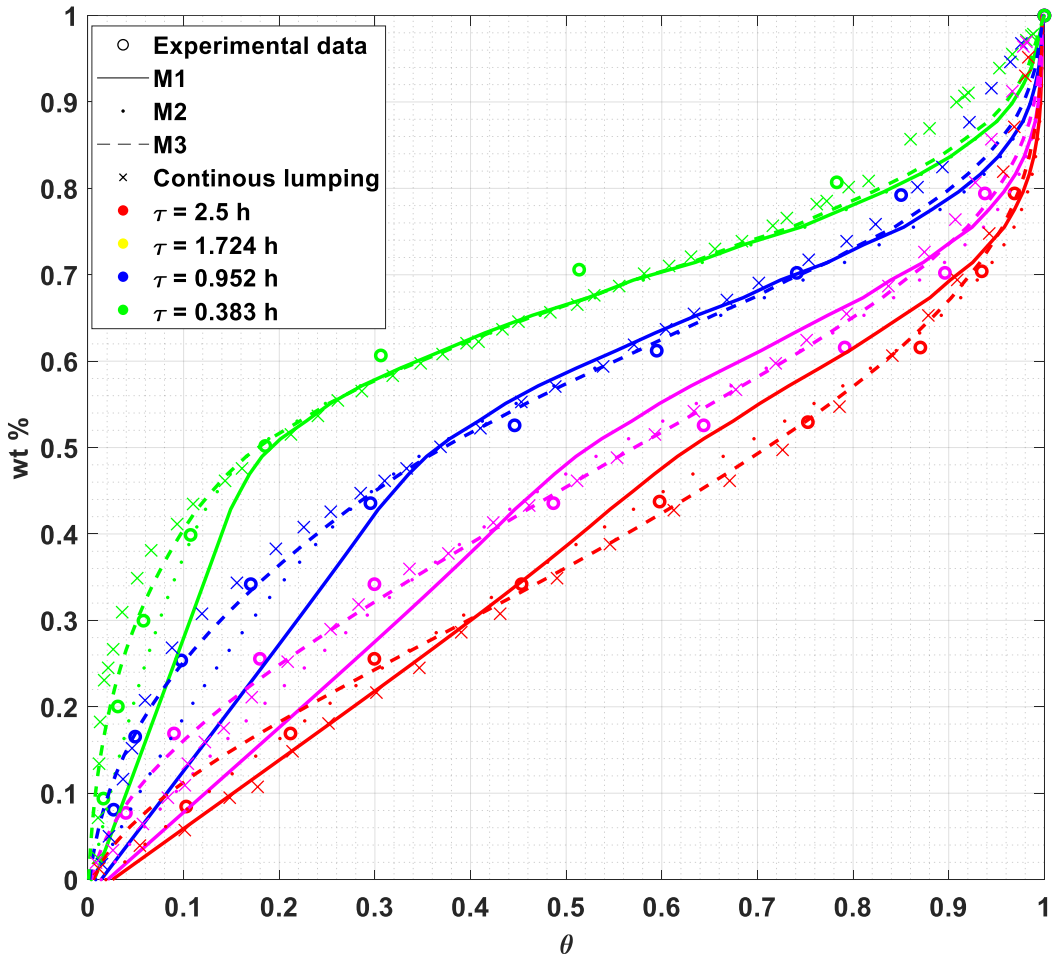

Figure 4.4-2 Comparison with experimental data from Bennett and Bourne [94] and modelled Continuous Lumping data from [97]

In Figure 4.4-3 the concentration changes of the heaviest (θ=1) and the lightest (θ=0) components are shown in time and space (calculated with M3). As it was expected, the mass fraction of the lightest component is equal to 0 at t = 0 everywhere in the reactor, and it is 0 during the whole experiment in the feed. Similar statement can be reported in case of the heaviest pseudo component: the mass fraction is not changing in the feed and the initial value is 0 everywhere in the reactor at the beginning of the simulation. By increasing the residence time at t = $t_{end,}$ the conversion of the heaviest component and the formation of the lightest component increases, since the bulk of the reaction mixture have spent more time

in the reactor. As it was mentioned, the simulation time is not equal to the residence time since the simulation was terminated only when the reactor reached the steady state condition. The equity of the residence and simulation time can only be expected when the hydrodynamic model is identical to the plug flow, which is not the case in this investigation. Moreover, the hydrodynamic model is slightly different in the case of every liquid velocity, since the convection part of the model is more pronounced if the fluid velocity is higher (residence time is lower) and at the same time the hydrodynamic model is more similar to the plug flow. This can be one possible root cause that model performance is getting worst with the increased residence time, since if the Reynold number is high, in a real reactor the CRTD (cumulative residence time distribution) is comparable to plug flow [97]. Aware of this limitation, one can optimize the residence time according to the target selectivity by applying finer discretization on the reactor volume to achieve the same hydrodynamic behavior in the simulation in case of lower space velocities.

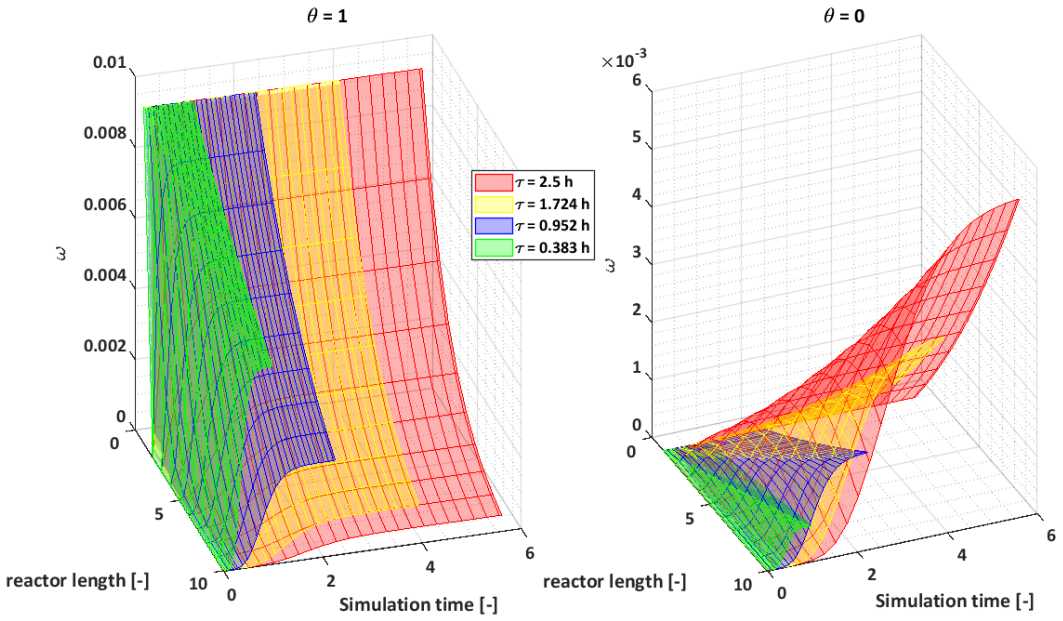

Figure 4.4-3 Transient behavior of the lightest (θ=0) and the heaviest (θ=1) pseudo components during the simulation in the case of different residence times.

4.5 Conclusion

In this work, three different modelling approaches were investigated to predict the hydrocracking trends respect to four different residence times. The model performance were measured by comparing simulation results to measurements which were previously published in [94] and used to develop the original continues lumping approach applied on a hydrocracking reactor [34].

The three approaches only differ in the applied selectivity function which is used to estimate the yield of each component in the different cracking reactions. The three different selectivity functions are based on the following assumptions:

- First approach (M1): the rate of cracking for every component is the same.
- Second approach (M2): the strengths of the C-C bond is the highest at the centre C-C and decreases monotonically toward the end bonds.
- Third approach (M3): the amount of the formed species are represented by the same skewed Gaussian-type distribution function.

With M2 and M3, the normal hydrocracking trends were predicted well, but the assumption of M1 is not in line with the beta-scission, so with this model a prediction cannot be made with a tolerable precision. The predictions are the most accurate in the case of higher residence times, which can be caused by the change in the hydrodynamic model due to the different space velocities. As conclusion, the developed approaches (M2 and M3) are suitable to optimize the residence time according to the target selectivity, and a more precise model can be developed if one takes into account the effect of liquid velocity on the cumulative residence time distribution.

5 Exploration of application domains for thermodynamic models through mixture of experts learning

The proper selection of thermodynamic models (TM) is a starting point for an accurate process simulation. It can occur in a process simulation that the proper thermodynamic model changes with the operating conditions. Therefore, the application domains for all appropriate models have to be determined and the models have to be used respectively to the domains. In this work, six TMs are investigated in case of hydrogen solubility in several n-paraffin, olefins, and aromatic compounds. In petroleum industry, the most commonly used TMs are the Soave-Redlich-Kwong EOS, and the Peng-Robinson EOS. The Zudkevitch Joffee, Chao Seader and Grayson Streed models are recommended to use in case of high hydrogen content. As the first step to determine the application domains, several measurements were collected from literature, and the solubility of hydrogen was estimated with the abovementioned TMs. Based on the prediction error, a mixture of experts-based expectation maximization (EM) algorithm was used to explore the optimal combination of a set of TMs through the corresponding application domains. As multivariate Gaussian distributions with zero covariance represent the mixture-of experts, the resulted axis-parallel clusters can be easily visualized as a set of univariate normal distributions determining the suggested application regions. The results illustrate that the developed Gaussian mixture model not only significantly improves the prediction performance of the TMs, but the extracted information also supports the systematic development of the models.

5.1 Introduction

To calculate the hydrogen solubility in different mixtures even in pure components is a challenging mission. However, in many processes (e.g. hydrocracking) the concentration of hydrogen in the liquid phase plays a very important role. The solubility of hydrogen is mainly depend on pressure and temperature, therefore with calculating the optimal amount of dissolved hydrogen and the softening operating conditions can help to reduce the operating cost of the process.

One way to calculate the amount of dissolved hydrogen is the application of the appropriate thermodynamic model (TM), but the proper selection of TMs is a recurrent problem in process simulations. According to the *Aspen HYSYS Property Package Selection Assistant*, there are four possible packages to calculate the properties of hydrocarbons in case of non-vacuum conditions, these are: SRK, Peng-Robinson (PR), Chao-Seader (CS), and Grayson-Streed (GS).

The conditions of applicability of the recommended and an extra (Zudkevitch-Joffee (ZJ)) property packages are summarized in Table 5.1-1. Three of these models were developed to calculate the properties of systems with high hydrogen content, these are: ZJ [98] CS [99] and GS [100].

Table 5.1-1 Conditions of applicability of the TMs recommended by Aspen HYSYS Property Package Selection Assistant

		CS	GS	PR	SRK	ZJ
T [K]		255 to 533	255 to 698	> 2	> 130	-
p [bar]		<100	<200	<1,000	<350	>10
For all hydrocarbons (except CH_4)		$0.5<T_{ri}<1.3$ $P_{mixture}<0.8$	$0.5<T_{ri}<1.3$ $P_{mixture}<0.8$	-	-	-
If CH_4 or H_2 is present:	*-molal average T_r*	< 0.93	< 0.93	-	-	-
	CH_4 mole fraction	<0.3	<0.3	-	-	-
	-mole fraction dissolved gases	<0.2	<0.2	-	-	-
Predicting K values for:	*-Paraffinic or olefinic mixtures, liquid phase aromatic mole fraction:*	< 0.5	< 0.5	-	-	-
	-Aromatic mixtures, liquid phase aromatic mole fraction:	> 0.5	> 0.5	-	-	

According to Table 5.1-1, different models are recommended to use based on the operating conditions. However, there are several domains where multiple models are suitable for modeling, and there is no pure measure to compare their goodness based on the "selection guideline". For example, every model is suitable in temperature range between 255 and 533 K and pressure range between 10 and 100 bar.

In addition, not all models can be compared to each other, rather a pairwise analysis is possible (CS and GS, PR and SRK have similar characterization). Based on these considerations, one can conclude that the provided framework for the selection of proper TM is not straightforward (at least in case of predicting the hydrogen solubility), despite to the fact, that selecting the appropriate model is the first task for describing successfully the physical properties [101]. Moreover, the available decision trees for the selection of TMs lead to several cases, when multiple models can be used [102].

The application of the proper model is not the only way to improve the prediction: if there are multiple candidates, one can take the advantage of all models and combine them into a mixture model. There are several techniques to combine

multiple models e.g.: simple weighted average of the predictions of individual models or the mixture of experts approach, where the weights are the function of the feature space [103].

As we would like to explore the application domains of the TMs, we designed a Gaussian mixture of experts model that combines the TMs, in which the products of univariate Gaussian functions representing the optimal regions of the models. The proposed generative model can be easily visualized with the distribution of the planned applications, so the developed tool highlights how the available sets of TMs should be combined for a given set of applications. According to our knowledge, this is the first work discussing how the trained Gaussian mixture of expert approach can be utilized for the determination of application domains of TMs.

In this chapter, we compared the aforementioned TMs calculation obtained by Aspen HYSYS to measurement data (634) from literature [104], [105]. The collected measurements consist of data on the solubility of H2 in paraffins, olefins and aromatic compounds. The results illustrate that the developed Gaussian mixture model gives excellent prediction performance, and the proposed visualization provides interpretable information about how TMs should be applied for specific applications.

5.2 Gaussian mixture of thermodynamic models

The key idea of the proposed approach is that the distribution of the validation data used to represent the planned prediction tasks is approximated by a Gaussian mixture of the available TMs. In the studied specific problem the N pairs of validation data is represented as $\{(\boldsymbol{x}_k, y_k)\}_{k=1}^{N}$, where y_k is the measured hydrogen solubility and $\boldsymbol{x_k}$ denotes the explanatory variables including molecular descriptors like the carbon number (C_n) and the number of the hydrogen atoms (H_n) and the operating conditions of the experiments (pressure and temperature). The set of the predictions of the available TMs are also represented as paired samples $\left\{\left(\boldsymbol{x}_k, \hat{y}_{k,j}\right)_{j=1}^{n}\right\}_{k=1}^{N}$ where $\hat{y}_{k,j}$ is the k^{th} predicted hydrogen solubility by the j^{th} thermodynamic model, hence the k^{th} prediction error of the j^{th} TM can be

calculated as $e_{k,j} = (y_k - \hat{y}_{k,j})^2$. The identification of the model is formulated as a clustering problem in which the prediction error is used to measure the distance of the models and the validation data. The clustering is based on the Expectation Maximization algorithm that minimizes the sum of the weighted squared distances: $J = \sum_{j=1}^{n}\sum_{k=1}^{N}(u_{j,k})^m D^2(\boldsymbol{x}_k, y_k, \eta_j)$, where η_j represents the parameters of the j^{th} cluster including the j^{th} TM. The proposed clustering algorithm can be interpreted in a probabilistic framework, the distance is inversely proportional to the $p(\boldsymbol{x}_k, y_k|\eta_j)$ probability that the $\boldsymbol{x}_k$ data point belongs to the i^{th} cluster:

$$\frac{1}{D^2(\boldsymbol{x}_k, y_k, \eta_j)} = \alpha_j \frac{1}{\sqrt{2\pi\sigma_{j,l}^2}} e^{\left(-\frac{e_{k,j}^2}{2\sigma_{j,l}^2}\right)} \frac{1}{\sqrt{2\pi|\Sigma_j|}} e^{\left(-\frac{1}{2}\cdot(\boldsymbol{x}_i-\boldsymbol{v}_j)^T\cdot\Sigma_j^{-1}(\boldsymbol{x}_i-\boldsymbol{v}_j)\right)} \qquad 5.2\text{-}1$$

The first α_j term represents the a priori probability of the cluster, while second is the distance between the k^{th} data point and the j^{th} model. The third term defines the distance between the cluster prototype and the data in the feature space of the variables describing the applicability domain of the models represented by the mean and the covariance matrices calculated as:

$$\boldsymbol{v}_j = \frac{\sum_i (u_{j,k})^m \boldsymbol{x}_i}{\sum_i (u_{j,k})^m}, \Sigma_j = \frac{\sum_i (u_{j,k})^m (\boldsymbol{x}_i - \boldsymbol{v}_j)^T(\boldsymbol{x}_i - \boldsymbol{v}_j)}{\sum_i (u_{j,k})^m}, \alpha_j = \frac{\sum_i (u_{j,k})^m}{N} \qquad 5.2\text{-}2$$

The $u_{j,k}$ weights are updated in every iteration:

$$u_{j,k} = \frac{1}{\sum_{l=1}^{n}\left(D^2(\boldsymbol{x}_k, y_k, \eta_j)/D^2(\boldsymbol{x}_k, y_k, \eta_l)\right)^{2/(m-1)}}, \forall j, k \qquad 5.2\text{-}3$$

The Alternating Optimization (AO) of these clusters is identical to the Expectation Maximization (EM) (maximum likelihood estimation) identification of the mixture of these Gaussian models when the fuzzy weighting exponent *m = 2.*

5.3 Application domains of thermodynamic models

To investigate the goodness of the TMs, 634 solubility data from literature was collected. The data set contains of H_2 solubility in paraffinic (458), olefinic (49),

and aromatic (127) compounds with carbon range between 1-16, 2-8 ,6-13, temperature range between 90.5-623 K, 123.15-436.15 K, 208.15-621.75 K and pressure range between 7.08-784.5 bar, 20.3-304 bar, and 20.3-507 bar respectively. In Figure 5.3-1 every data point is marked according to the smallest prediction error.

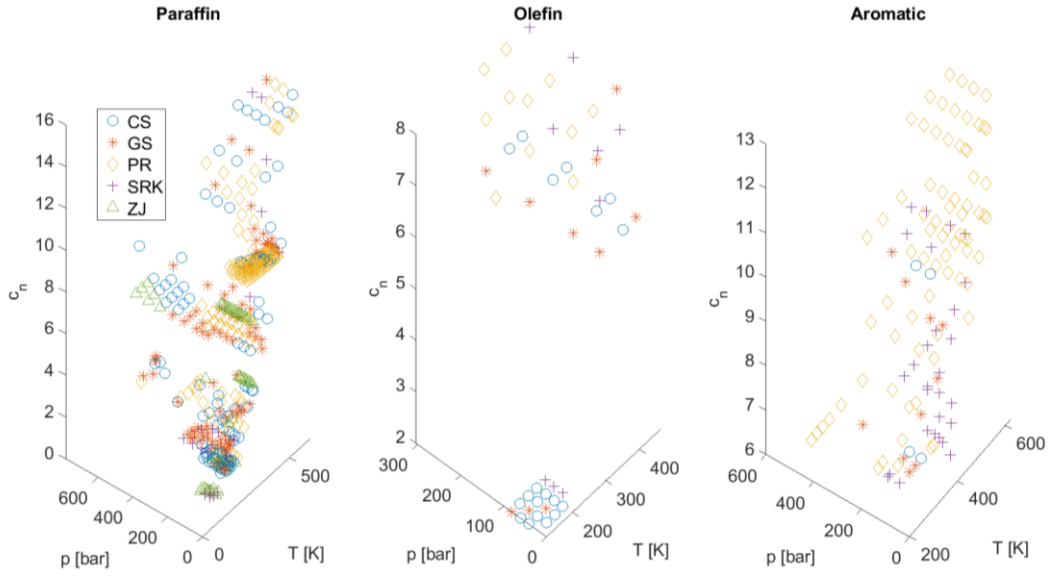

Figure 5.3-1 Different markers represents the TMs which obtained with the smallest prediction error at each investigation condition.

In paraffinic compounds all models are required to predict the H_2 solubility, but it can be noted, that different models dominate different, but well specified domains. However, in case of olefinic and aromatic compounds the usage of ZJ model is not recommended at all, and the application domains distributed lightly. There are a few but coupled measurements for olefinic compounds, hence it is hard to make proper conclusions. The only thing is clear that the CS GS and PR are the dominant models and the application domains mainly distributed along the carbon number and the temperature. The PR is the best approach to reproduce the measurements in aromatic compounds, but in case of lower pressure and carbon number mainly the SRK performs better.

The selection of the proper model is challenging based on Figure 5.3-1. To determine well defined application domains, Gaussian mixture model was used. As the TMs are treated like black box models, the application domains are investigated

along arbitrary chosen variables. The selection of pressure and temperature as variables is natural, but for identifying different molecule types and size, carbon number (C_n) and the number of the hydrogen atoms (H_n) in the molecules was applied.

The mean square errors for 10-fold cross-validation based training are presented in Figure 5.3-2. Despite that the linear mixed model is not suitable to determine the application domains, it is a good benchmark to compare the prediction errors. The Gaussian mixed model has better performance than the individual models, but slightly worse than the linear combination. Based on the mean square errors, the developed mixed model is a reasonable choice to investigate the application domains of the TMs.

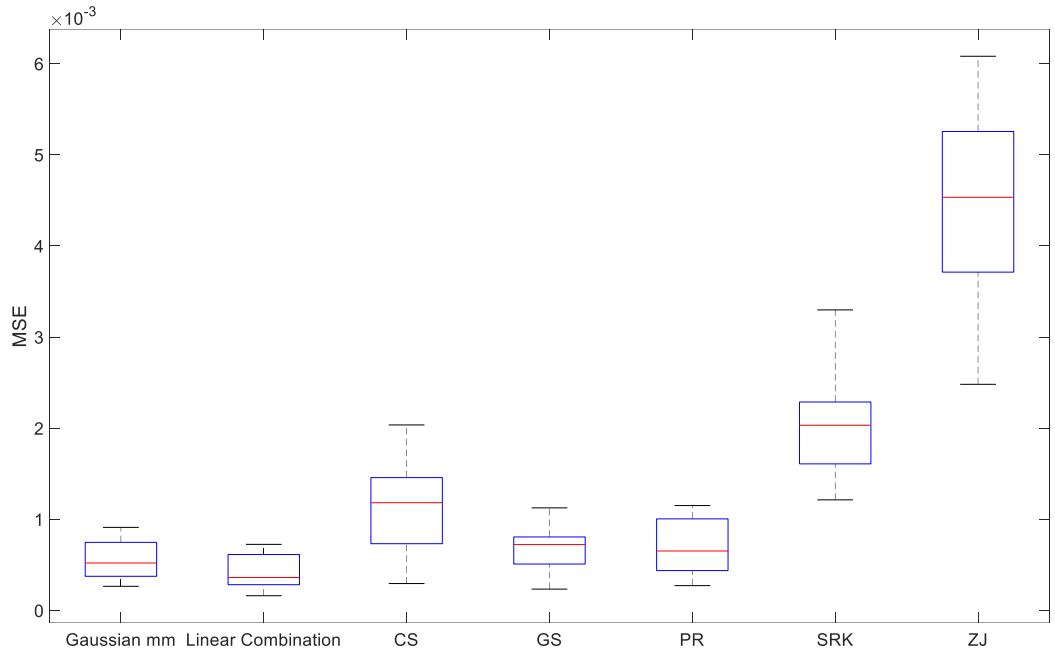

Figure 5.3-2 The mean square error of individual TMs and mixture models.

In Figure 5.3-3 the distribution of the data along the variables is presented as histograms, and the calculated application domains are presented as Gaussian distributions. The multivariate distribution used by the algorithm can be obtained with multiplying the individual distributions. The height of curves represents the "probability of being appropriate" for the TMs at a specified modeling condition, moreover this height is proportion to the weights of the TMs in the mixed model.

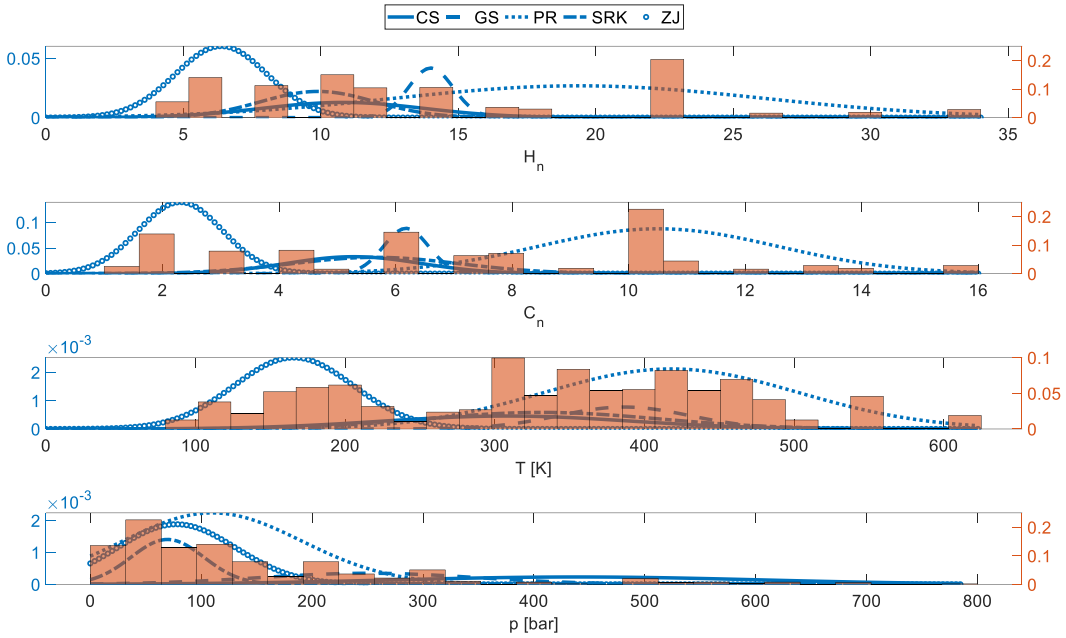

Figure 5.3-3 The proposed application domains of TMs and the distribution of the data set along the variables.

The conclusions which can be made based on Figure 5.3-3 can be divided into two main groups:

1) The ones which are confirm or at least not disprove the fluid package selection guideline:
    a. PR has the widest application domain and dominates most of the domains of investigated variables.
    b. ZJ is recommended only in case of light hydrocarbons and extremely low temperature, but as the solubility exists in liquid phase and the light hydrocarbons have low dew point, these statements can be coupled.
    c. The performance of CS and GS are nearly the same, but in higher temperatures GS predicts better.
2) The others which are disprove some of the statements of Table 5.1-1:
    a. From those models which were developed for calculate properties in high H_2 content, only the ZJ dominates a wide range of domain.

b. The CS performs best at extremely high pressure outside the recommended application range, but it can be the result of the lack of data in extremely high pressures.
c. The performance of SRK is far worst then the performance of PR.

5.4 Conclusion

A tool was developed to explore the optimal combination of a set of thermodynamic models for a set of planned applications or validation data, but it should be mentioned, that the well distributed the data is in the planned application domain, the better the exploration is. The visualization of the proposed Gaussian mixture of expert model with the distribution of the data along the variables determines the suggested application domains of the models. The application domains of five TMs were investigated based on the hydrogen solubility in paraffinic, olefinic, and aromatic compounds. The results illustrate that the developed method not only explores how the models should be combined, but it significantly improves the prediction performance of the TMs.

6 Retention time alignment of gas chromatographic data

Gas chromatography (GC) is an effective tool for the analysis of complex mixtures with a huge number of components. To keep tracking the chemical changes during the processes like plastic waste pyrolysis usually different sample states are profiled, but retention time drifts between the chromatograms make the comparability difficult. The aim of this study is to develop a fast and simple method to eliminate the time drifts between the chromatograms using easily accessible priori information. The proposed method is tested on GC chromatograms obtained by analysis of pyrolysis product (Mg/Y catalyst) of shredded real waste HDPE/PP/LDPE mixture. A modified k-means algorithm was developed to account the retention time drifts between samples (different sample states). The outcome of the retention time alignment is an averaged retention time for each peak from all the chromatograms which makes the comparison and further analysis (such as “fingerprinting”) easier or possible.

6.1 Introduction

Pyrolysis is one of the most investigated routes used to minimize plastic waste and convert it into a valuable product. A huge number of components (about 300-400 peaks on chromatogram) can be found in the pyrolysis product which can be characterized by using GC. When multiple samples are profiled, retention time shift occurs between the chromatograms due to some instrument-related phenomena (e.g. injection-timing problem, varying flow rate, temperature disturbances/gradient) or due to the chemical interaction between the samples and the instrument (selectivity changes over time). Despite that the instrument-induced retention time shifts have been lessened through the advanced electronic control systems; an appreciable amount of time drift remains in the chromatographic data [46].

One of the reasons to keep tracking the chemical changes during processes with profiling different sample states is to assist the development of a reliable kinetic model. In this case, the determination of the target chromatogram is not possible, and the uncertainty can be increased with user chosen parameters of chromatogram analysis. Thus, the abovementioned methods are not suitable for retention time alignment (in this special case) and the development of a method is required in which these disadvantages are eliminated. The fact that k-means algorithm was originally designed for minimizing variance and not the arbitrary distances, makes the method unpopular to use for time series. However, this paper shows that with some modification and with the appropriate preprocess of data, it is also a powerful tool for handling time shifts in chromatograms. The experiments analysed in this chapter were performed at different temperature levels using different zeolite based catalysts, additional details can be found in [106]. Based on these experiments a lumped kinetic model was developed and published in [28], and the uncertainty of the model was diminished by reducing the size of the reaction network in [107].

The starting point for a traditional lumping model is in macroscopic level (e.g. boiling point), so the amount of information that can obtained is quite limited [108]. One possible way to allow more obtainable information from the model is to define the pseudo components more properly, e.g based on molecular rather than physical

properties. The molecular properties of the experimental products can be obtained directly from chromatographic data. Our aim is to develop an algorithm to perform the alignment of peaks from different chromatograms (so eliminate the time drifts) which characterized the product of a complex reaction system in time, which makes easier to define the proper pseudo-components.

6.2 Preprocessing the data

Suppose that $X = \{x_{1,1}, x_{1,2}, ... x_{1,m}, x_{2,1}, x_{2,2}, \ldots x_{2,m}, \ldots x_{n,m}\}$ is a given data set of n retention times of chromatographic peaks from m measurements. The object of a clustering algorithm without any constraints is to grouping a set of objects (peaks) into k clusters ($c = \{c_1, c_2, \ldots c_k\}$), in such way that objects in the same group are more similar to each other than to those in other groups. In this section we present a method that allows the proper alignment of peaks from different chromatograms obtained by analyzing different sample states.

First of all we would like to highlight the most important properties of the investigated dataset:

- obtained by the GC based product analysis of waste plastic pyrolysis carried out in a two-stage laboratory scale reactor system. The 50 g solid plastic waste was measured into the reactor at the start of all experiments and 15 $dm^3\,h^{-1}$ nitrogen flow was maintained that drove volatiles through the second. The experiments in which the investigated chromatograms were performed at 425 °C using Mg/Y catalyst, additional details can be found in [28], [105].
- data contains 7 chromatograms in different sample states (sampled as the experiment progressed, at: 10, 20, 30, 40, 50, 60 and 70 min);
- paraffinic peaks were identified in advance.

As it was stated in a previous study of this system, only a small changes can be noticed in the chromatograms of pyrolysis product samples taken at different time steps [109]. Hence, the collected data can be applied to test the proposed clustering

algorithm, since the primary aim of this algorithm to find peaks in every chromatogram which can be the same molecule.

The identification of the paraffinic peaks is an easy but essential task, as these peaks serve as points of reference during the peak alignment process. The chromatograms are divided into segments by these reference points. Moreover, the alignment of the reference points is unequivocal, hence through the segments the task of retention time alignment can be divided into subtasks. The dataset is plotted in Figure 6.2-1, where the dashed lines are reference points (i.e. paraffinic peaks) and the sections between them are the same segments in all chromatograms (the highlighted segments are the C_{10} fractions). These segments are coherent so they can be grouped, and the retention time alignment within these segment groups are the subtasks.

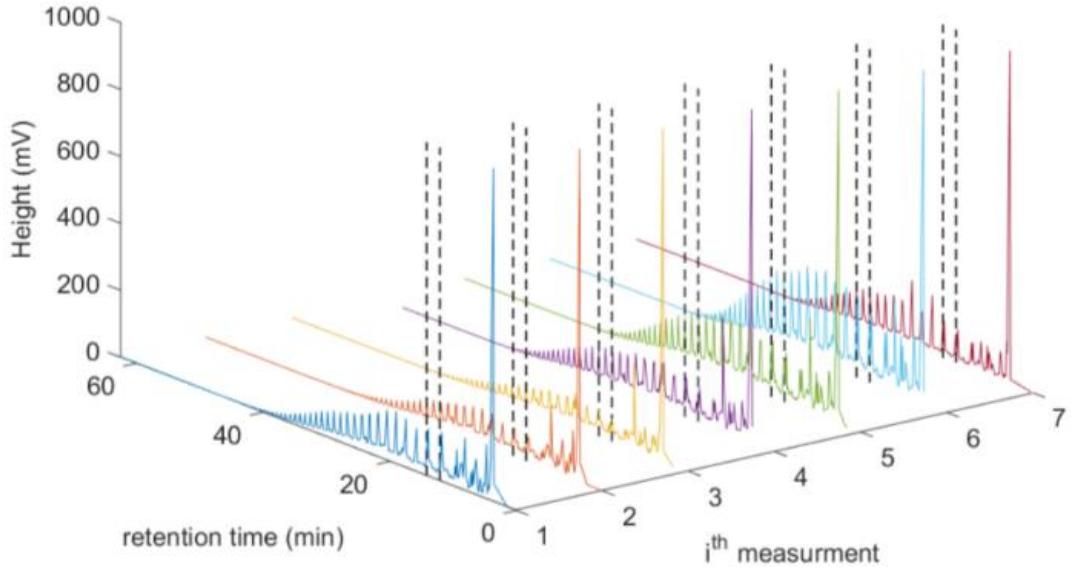

Figure 6.2-1 The chromatographic data. The segments between the dashed lines denote the C_{10} fractions.

In Figure 6.2-2 (a), the retention times of data from C_{10} fractions from all chromatograms is illustrated. The size of the circles denotes the origins of the data points, for example the smallest circles are from 1st measurement, and the largest ones are from the 7th sample. Figure 6.2-2 (b) shows the data from Figure 6.2-2 (a)

when it is normalised to 0 - 1 range for each segment in the segment group separately according to Eq. 6.2-1 (the retention time of paraffinic peak heading is 0 and the retention time of paraffinic peak trailing is 1, but the latter is not shown).

$$\hat{x}_{n,m} = \frac{x_{n,m} - x_{pa,h}}{x_{pa,t} - x_{pa,h}} \qquad 6.2\text{-}1$$

Where $x_{pa,h}$ is the retention time of paraffinic peak heading and $x_{pa,t}$ is the retention time of paraffinic peak trailing $x_{n,m}$.

The normalisation balanced the retention time drifts to such extent that some of the coherent data points can be grouped manually without any further ado. The transformed data set is only one dimensional, there is no clear pattern in time shifts, and coherent data points seem to be similar to clusters where the variance need to be minimized. All the above-mentioned facts led us to use k-means for the retention time alignment.

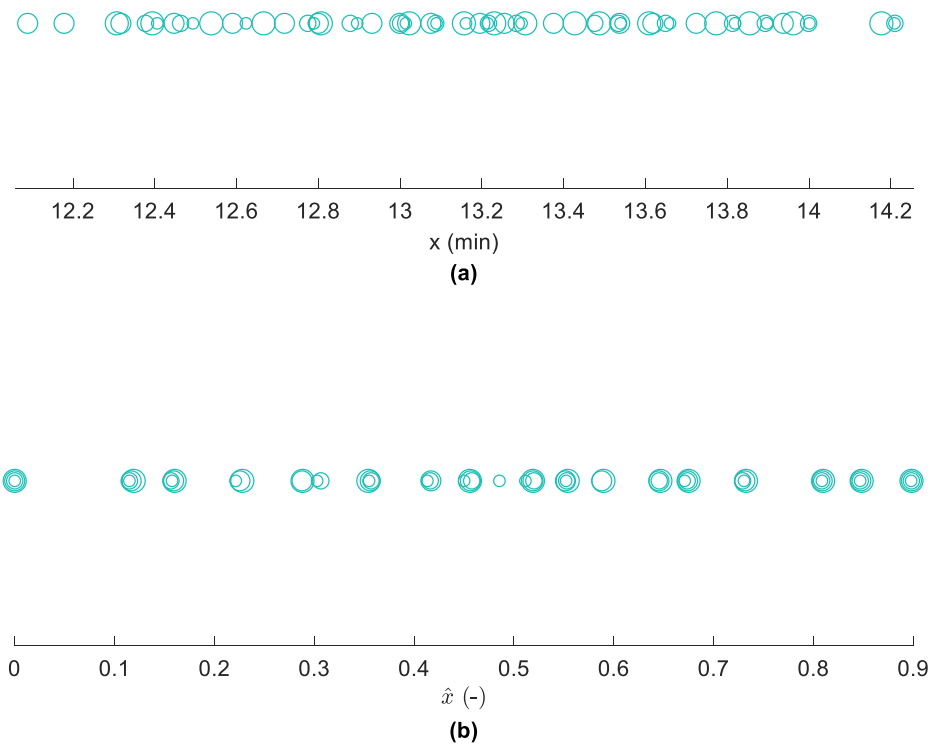

Figure 6.2-2 The retention times of C10 fractions from all chromatograms before (a) and after (b) the normalisation.

6.3 Modified K-means algorithm for retention time alignment

K-means is a well-known clustering algorithm which partitions data into clusters based on the distance from each data point to different centroids. The algorithm requires the number of maximum iterations, the initial centroids, and the number of clusters. The **standard algorithm** can be described in three steps [110]:

1. Initialization: initialization of the centroids (μ_j) (usually random data points from the data set) according to Eq. 6.3-1

$$\mu_j^1 = \left\{x_p : x_p \in \boldsymbol{X}, \mu_i^1 \neq x_p, 1 \leq i \leq k, i \neq j\right\} \quad 6.3\text{-}1$$

2. Assignment: each data point is assigned to the nearest cluster according to squared Euclidean distances (*t* denotes the iteration step):

$$c_j^t = \left\{x_p : \left\|x_p - \mu_j^t\right\|^2 \leq \left\|x_p - \mu_i^t\right\|^2, 1 \leq i \leq k\right\} \quad 6.3\text{-}2$$

3. Update: calculating the centroids for the next iteration based on the data assigned to each cluster:

$$\mu_j^{t+1} = \frac{1}{\left|c_j^t\right|} \sum_{\forall x_i \in c_j^t} x_i \quad 6.3\text{-}3$$

The algorithm terminates when the number of maximum iterations is reached (or the cluster centres do not change significantly), otherwise it iterates back to step 2.

In real world applications the maximum size of the clusters, or must-link/cannot-link constraints (data points that should or should not be grouped together) are available as background knowledge. A modified k-means algorithm which can handle the maximum cluster size problem is published in [111]. However, if data points were to be eliminated from clusters to satisfy the constraint, an iteration will be used constructed in which the algorithm rather finds the nearest centre to the points, than assign the nearest points to the centre. This way a point could be

assigned to a wrong cluster and the size of the cluster could reach the maximum, so another point which is closer to the cluster centre will be forced to be assigned to another cluster. A modified k-means algorithm with must-link/cannot-link constraint is published in [111], however in this study we provide a detailed approach from an engineering point of view.

In the proposed algorithm the assignment step is complemented (Figure 6.3-1), so it can handle both constraints in an inner iteration. If there is a maximum cluster size constraint and $|c_j|$ denotes the size of the j^{th} cluster and ζj denotes the maximum size of the j^{th} cluster, than an extra constraint is has to be satisfied: $|cj| \leq \zeta j$. The maximum cluster size is guaranteed as follows:

1. each data points are assigned to the nearest cluster according to squared Euclidean distances;
2. sort the assigned points for each cluster in ascending order according to the distances;
3. from 1 to maximum cluster size the assigned points remain in the clusters (or less if there are not as many assigned points), the others are saved for the next iteration;
4. the clusters that reached their capacity do not take part in the next iteration;
5. back to step 1 until all the data points are assigned to a cluster.

The fulfilment of cannot-link constraint is divided into two parts. The first one: in every (inner) iteration step the currently assigned points (for each cluster) do not violate the constraint. If there is a constraint violation, only the nearest data point to the cluster centre remains in the cluster from those that should not be linked, the others are saved for the next iteration. Hence, it is needed to be executed after sorting the points according to distances. In practice, the constraint violations are detected through an additional property. This means that a number is assigned to each data point (based on their original chromatogram) as a property, and two points cannot be linked if the same number is assigned to them. The second part of the cannot-link constraint fulfilment is the inspection of clusters created in the previous iterations. Those clusters need to be identified to which the current individual data

points should not be assigned, and to ensure that such data points will stay out of the clusters. The constraint violations are detected in the same way as previously based on the additional property. To ensure to avoid the violation, if a data point should not be assigned to a cluster, the number which represents its distance from the cluster centre will be replaced by an infinite number. Hence, it is needed to be executed from the second iteration step before sorting the points according to distances. A simplified flow chart of the algorithm is shown in Figure 6.3-1.

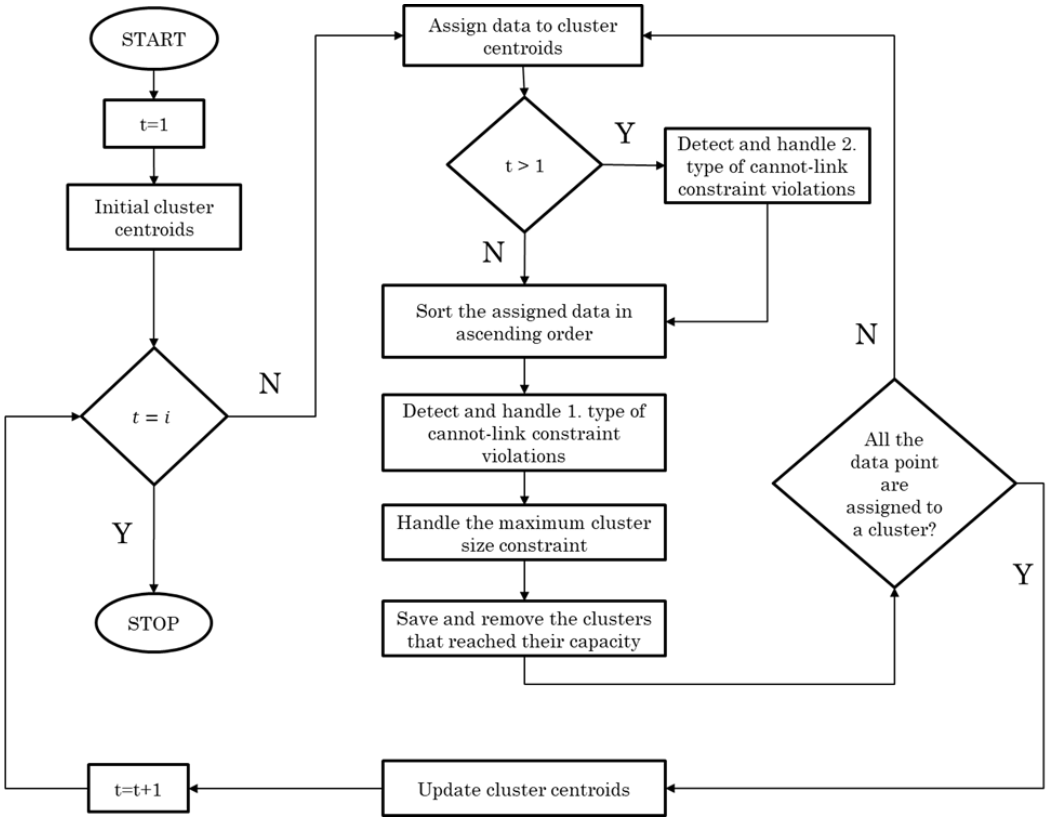

Figure 6.3-1 Simplified flow chart of modified k-means algorithm

6.4 Determining the optimal number of clusters and initial cluster centroids

The determination of the number of the clusters is essential but the appropriate method varies from task to task. In this section a proper method is provided when the algorithm is applied to processing GC data obtained by analysis of hydrocarbon

products. The number of the clusters is determined by the investigation of segments from the current segment group (subtask), and it is equal to the maximum number of peaks in one segment (this segment is denoted as S_0). This is the minimum number of the clusters, but later it can be increased based on the cluster variances to avoid that different chemical substances are grouped together. The initial centroids are the normalized retention times from S_0. The reason why the number of clusters should be increased is that any segment from the current segment group could contain a data point, which is not equivalent to any data points from S_0 (this data point is a chemical substance which is not present in S_0). After the clustering, the outlier clusters are determined according to their variances (Grubbs's test was utilized [112]). If there is at least one outlier cluster, the clustering has to be performed again with an additional cluster. In this case the initial centroids are the centroids which were determined in the previous clustering iteration and an additional random data point from the outlier cluster or clusters. The clustering is repeated until no outlier cluster is detected.

6.5 Results

In this section the method is tested on chromatograms obtained by the analysis of pyrolysis products of real waste plastics in different sample states. In our case, the maximum size of the clusters is 7 as the data set contains 7 different chromatograms. Additionally, we defined a cannot-link constraint because the data points (chromatographic peaks) from the same chromatogram cannot be in one cluster. Figure 6.5-1 is similar to Figure 6.2-2 (b), but normalization was performed for all subtasks (segment groups). Figure 6.5-1 confirms the statement that the normalization balanced the retention time-drifts such an extent that the modified k-means algorithm can be applied.

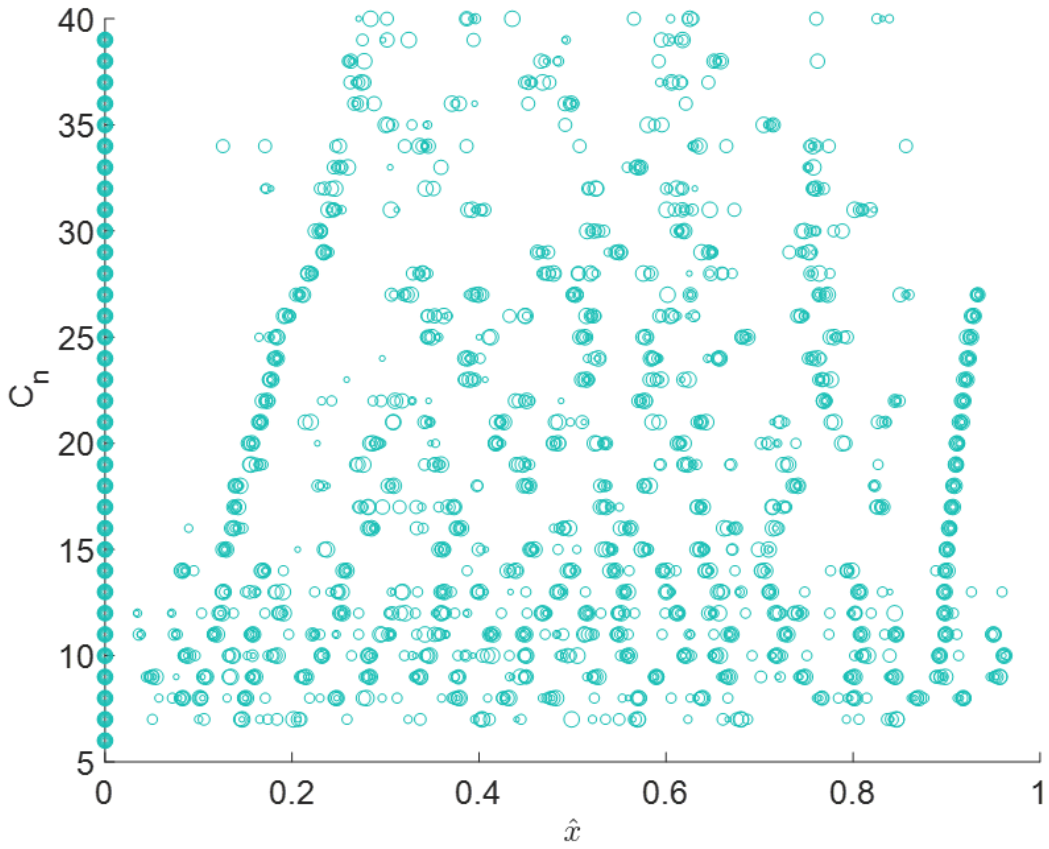

Figure 6.5-1 The normalised retention times for all chromatographic data, y coordinate denotes the fractions.

As the chromatograms are divided by the reference points (as described in Section 4), clustering was performed for each segment group separately along the normalized retention time. Hence, data points with the same y coordinate from Figure 4 (except data points with x = 0 coordinate) can take part in the clustering at the same time.

The results are shown in Figure 6.5-2, the clusters are circled and marked with colours as well, and the width of the cluster is proportional to the cluster variance. Higher variance clusters were formed in fractions with fewer peaks i.e.: in C_7 - C_8 and C_{35+} fractions. Figure 6.5-2 shows that the developed algorithm partitioned the data points effectively and can handle the overlapping.

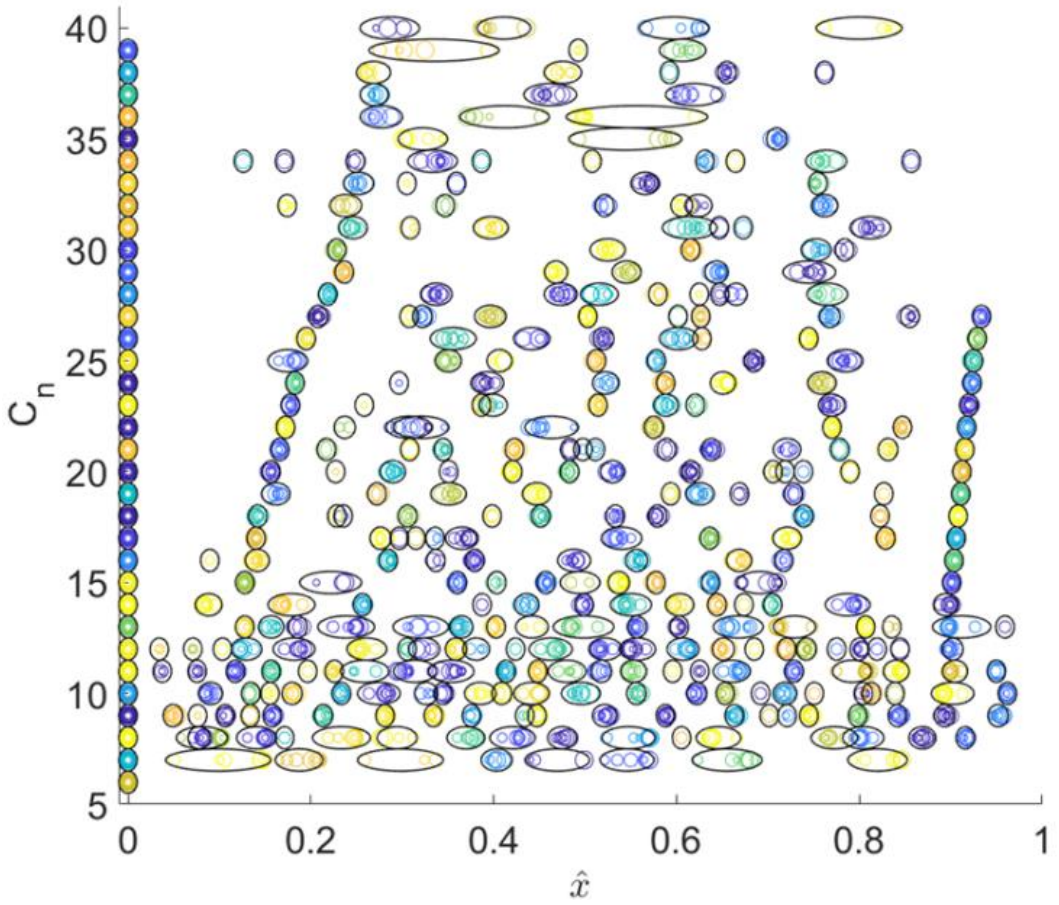

Figure 6.5-2 The resulted clusters, i.e. the components in the pyrolysis product. The individual clusters are circled and marked with different colors as well

In Figure 6.5-3 the alignment of the C_{10} fractions is shown. Hence the clustering was performed in one dimension (normalized retention time), the height of the peaks is not important so their value in the figure is one. In this subtask, 107 chromatographic peaks were grouped into 22 clusters, meaning there are 22 different chemical substances within the C_{10} fraction were formed during the experiment. In total, 382 clusters were determined, i.e. 382 individual components are detected. 49% of the clusters contain seven peaks, which means that the presence of almost half of the components continuously presented in the product mixture during the experiment.

As it is shown in Figure 6.5-4 (a), 11% have one, 8% have two and 8% of the clusters have three elements. Hence, the presence of 27% of the components is

temporary in terms of the sample states, the presence of the rest of the components (24%) is permanent. The pie charts in Figure 6.5-4 (b) shows the distribution of cluster sizes along the measurements. Since the height of the peaks were not constrained, every cluster took part in the investigation. Through this analysis the noisiest chromatograms can be detected and marked as outliers.

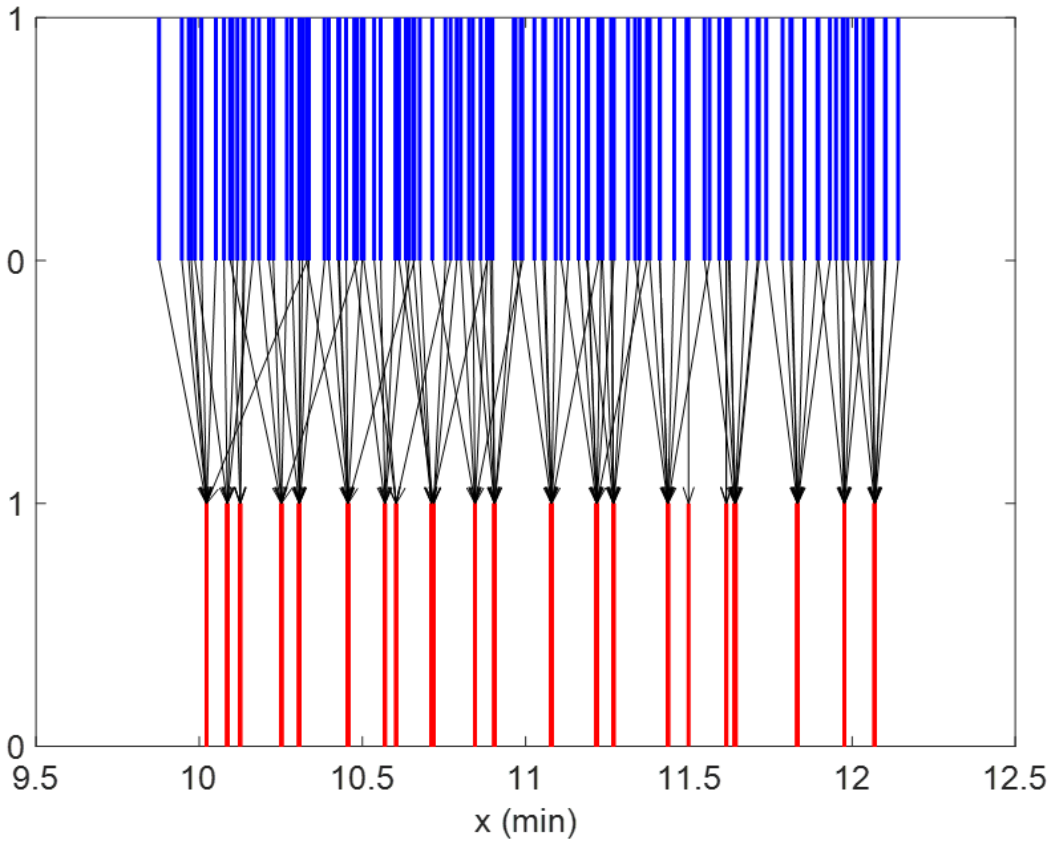

Figure 6.5-3 The alignment of peaks in C10 fraction from seven different chromatograms (different sample states)

The outliers are the first, fourth and fifth chromatograms as the proportion of small sized clusters is the highest in these chromatograms. The proportion of clusters with one or two elements is 52 % in the fourth chromatogram, and this proportion is significant in case of the first (37 %) and fifth (30 %) chromatogram. Based on the above-mentioned facts, the proposed method is suitable for analysing

the chromatograms and determines the outliers, hence the experiments can be repeated considering the results to avoid the outlier samples.

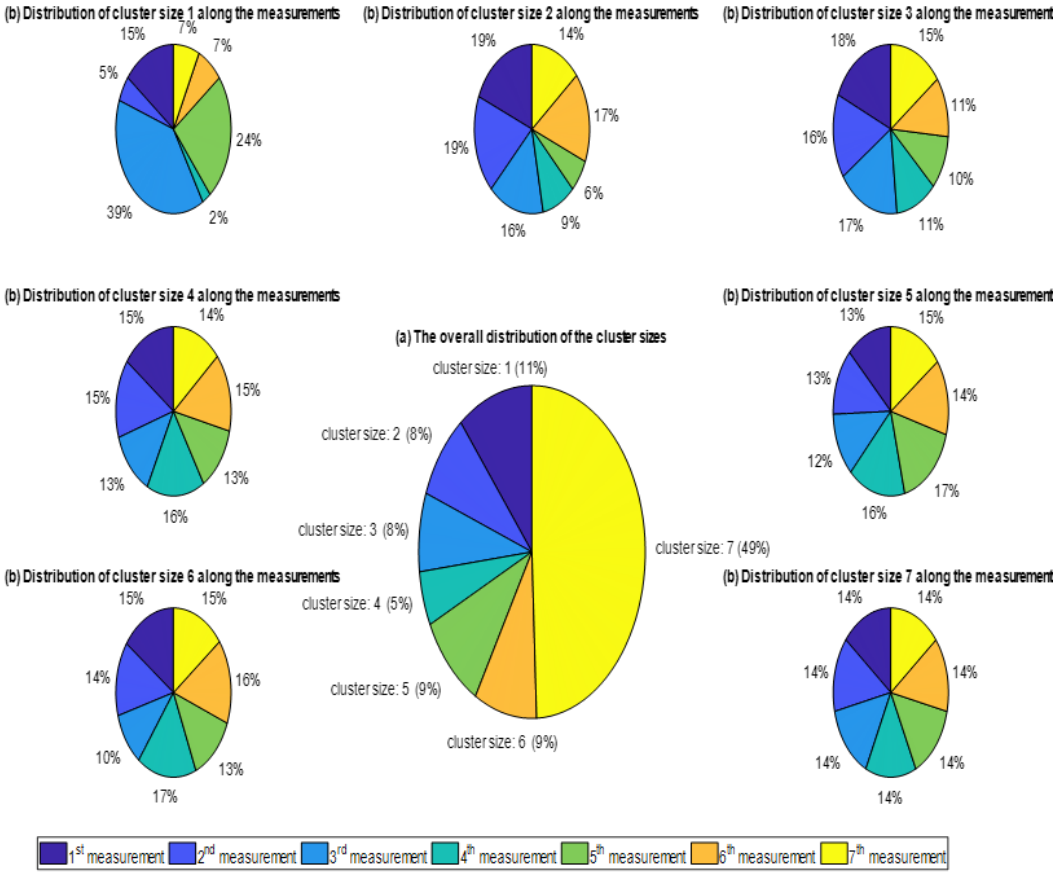

Figure 6.5-4 (a) The distribution of cluster sizes along the overall data (b) The distribution of the individual cluster sizes along the measurements.

The corrected retention times belonging to the elements of the individual clusters are equal to the cluster centroids. In this case the connection between the peaks in the chromatograms is a clear bijective function. Therefore, the retention time drifts have been eliminated and the chromatograms have become comparable as it is shown in Figure 6.5-5. The retention time is a characteristic parameter in qualitative analysis. Ideally, peaks with the same retention time denote the same molecule. However, the peak area under the curve is proportional to the concentration. Figure 6.5-5 is an example for the visualisation of chemical changes during the pyrolysis

process. Points with the same coordinates denote the same molecules and their colours are applied to mark their concentration in the sample.

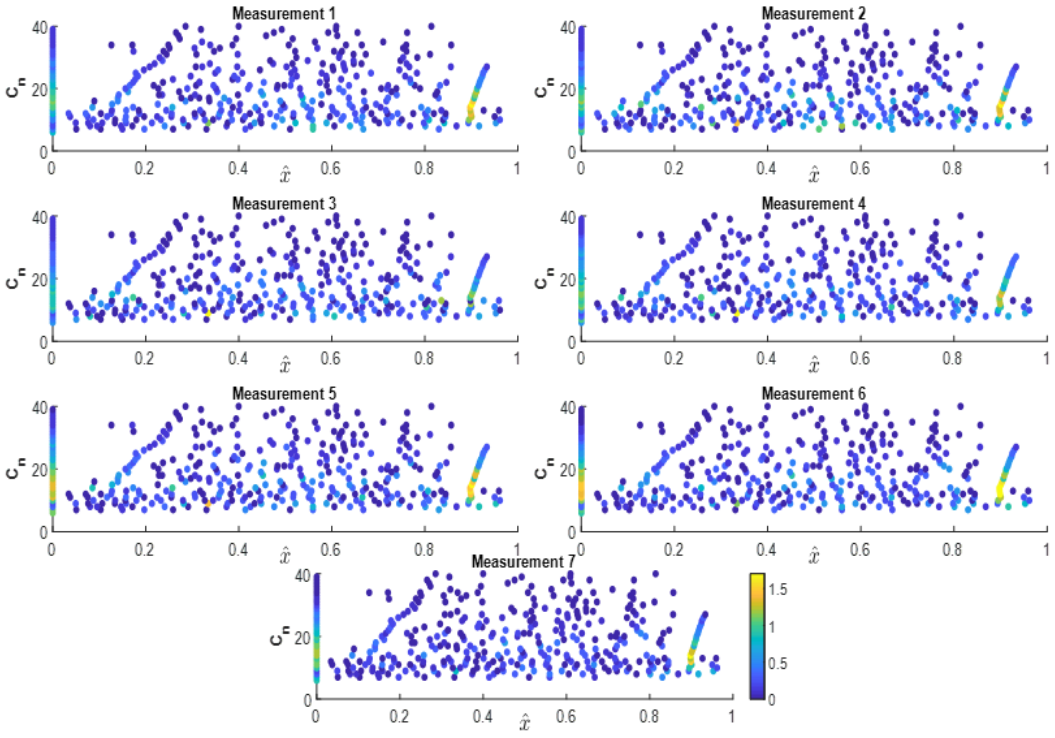

Figure 6.5-5 Visualisation of the chemical changes during the pyrolysis process. Points with the same coordinates denote the same molecules and their colour is proportional to the concentration.

6.6 Conclusion

In special cases such as chromatograms, the developed algorithm is appropriate for the alignment of time series. The main criterion for the application is that the time series have reference points. Based on the properties of segments between these reference points, the number of clusters can be determined and, in an iteration, can be corrected based on the cluster variances. The main advantages of the developed algorithm compared to other methods are that no target chromatogram is needed, and the result is not influenced by any user chosen parameters. The method was tested in the analysis of the chromatographic data coming from thermo-catalytic pyrolysis of waste plastics. The results showed that with proper pre-

processing of the data the developed algorithm is appropriate for handling the retention time drifts and can assign to each other to become traceable how the component concentrations changing in time.

7 Improving Molecular Composition Estimates using Kovats Retention Index and Molecular Similarities

Catalytic pyrolysis presents a promising avenue for mitigating plastic waste accumulation by converting it into valuable products. In this study, we investigate the application of computational methods integrating molecular similarities and Kovats retention index to enhance the accuracy of qualitative analysis in catalytic pyrolysis processes. Utilizing gas chromatography data and high-level measurement results, we evaluate the determined molecular compositions and assess their consistency across various measurement conditions. Despite encountering challenges such as algorithm failures due to high computational costs, our analysis reveals significant insights into the molecular composition of pyrolysis products. Through the application of molecular similarity methods, we demonstrate the potential to refine the estimation of molecular compositions, particularly in scenarios where retention index database accuracy is uncertain. Our findings underscore the importance of further refining computational methods and formulating additional constraints based on high-level measurements to enhance the accuracy of molecular composition estimates.

7.1 Introduction

Catalytic pyrolysis represents an efficient approach to reduce plastic waste and transform it into valuable products. When polyethylene undergoes pyrolysis, it yields a range of hydrocarbon compounds including alkanes, alkenes, aromatics, and hydrogen gas. The composition of these compounds in the pyrolysis product varies depending on factors like pyrolysis temperature, residence time, pressure, and catalyst presence. Furthermore, secondary and side reactions during the process can result in the formation of additional compounds. [52] In organic chemistry, homologous series refer to groups of molecules sharing the same fundamental structure, with their similarity quantified by various molecular fingerprint-based similarity indexes.

In this chapter, we aim to develop a computational method based on molecular similarities and the Kovats retention index to improve the accuracy of qualitative analysis relying on GC data. The idea is that, during the determination of the composition of the pyrolysis product, the uncertainty arising from inaccuracies in retention indices found in databases can be reduced by utilizing similarity indices. The goal is to ensure that the identified molecules are as similar to each other as possible. The method assumes that, during the pyrolysis of plastic waste, which is mainly consists of polyethylene and polypropylene, the dominant reactions lead to the formation of homologous series of components and no significant amount of heteroatoms are presented in the liquid pyrolysis product.

The resulted task is a large-scale, constrained optimization problem. Population-based algorithms, such as genetic algorithms, which can solve constrained nonlinear optimization problems, have restricted capability in finding feasible solutions. Therefore, the task is formulated as an integer linear programming (ILP) problem, and the feasibility of the solution is assured by the established constraints.

7.2 Data

Data collection is an essential part of the developed method to identify the molecular composition of a given mixture. As the calculation can be divided into two parts, the necessary data can be classified into two categories as well:

- Raw GC chromatograms.
- Kovats retention indexes from online database and molecular similarities of multiple molecule pairs.

At this point I would like to highlight that the used plastic waste experimental data were measured and previously published by Sója et. al. [113].

7.2.1 Measurement data

The experimental setup of the plastic waste pyrolysis process is described in detail by Miskolczi et al. [106], only a brief summarization of the conditions is presented in Table 7.2-1. The experiments were performed in a two-stage laboratory scale reactor system using 50 g of raw material, with the presence of 2.5 g catalyst in the first and 20 g catalyst in second reactor. The reactors had separate temperature control system: while the temperature of the second reactor was set to 380 °C in all experiments. The measured temperature of the first reactor corresponds to the melted polymer due to that thermocouple in the vessel was in contact with the melted polymer. The product was cooled down to 50 °C with a tube-intube water-cooled heat exchanger and the condensed outflow were sampled in specific reaction times and analysed by gas chromatography. In total 150 samples from 27 experiments were analysed. The timing of each sample depended on the temperature, and the last sample was taken once the reactor reached the steady state condition. Sampling intervals occurred at 8, 4, and 2-minute intervals corresponding to temperatures of 425, 455, and 485°C, respectively.

Table 7.2-1 Experimental conditions

Raw material	Composition of the raw material	Temperature of the first reactor	Applied catalyst
Shredded and then crashed plastic waste with average particle size of 3 mm.	41% HDPE, 42% PP and 17% LDPE	425/455/485 °C	Mg/Y; Sn/Y; Ce/Y; Zn/Y; Fe(II)/Y; H/Y; Ni/Y; Cu/Y, Fe(II)/Y

7.2.2 Kovats retention indexes and similarities

A large number of Kovats retention indexes have been determined experimentally and published in different databases. In the current analysis, the data which contains 83462 Kovats indexes of 26102 hydrocarbons measured with Standard non-polar column was collected from *https://pubchem.ncbi.nlm.nih.gov/*. We assumed that during pyrolysis most of the heteroatoms are converted into gaseous product: compounds containing heteroatoms were filtered out, hence the final data set contains 23940 Kovats indexes and SMILES (simplified molecular-input line-entry system) of 3495 hydrocarbons. We considered that all that unknown molecules in the analysed chromatograms can be found in this database. The similarity of the collected hydrocarbons was calculated using the implemented fingerprint-based approaches in RDKit toolbox [114]: Tanimoto, Dice, Cosine, Sokal, Russel, Kulczynski, McConnaughey, and Tversky similarity indices. The RDKit-specific fingerprinting algorithm was inspired by the Daylight fingerprint [115].

7.2.3 Preprocessing the data

In Chapter VI. an algorithm was introduced with which the time drifts between chromatograms can be eliminated using easily accessible priori information

(paraffinic peaks are determined in advance) [116]. The essence of the algorithm is that under the same conditions, multiple GC analysis is performed on the liquid product (in different times), allowing the traceability of the temporal impact of the reactions: the number of chromatographic peaks did not change significantly in the individual samples; however, there was a slight drift in their retention times. Therefore, this drift needed to be corrected to allow the traceability of the concentration change of a same molecule in the mixture over time. This algorithm was used to calculate an average Kovats retention index for every gas chromatographic peak considering every associated sample from each experiment. The calculation resulted in a list of retention indexes for every (27) experiment.

The alignment of molecules and retention indexes in the collected database is not a bijective function, meaning that one molecule can have several retention indexes, and several molecules can have the same retention index. Therefore, the second step of data preprocessing - in case of every experiment - is to select 5 candidate molecules for every peak based on the absolute deviation in the retention index. An example of the result of the first two steps of data preparation in case one experiment is shown in Table 7.2-2.

Table 7.2-2 Resulted table of the first two steps of data preprocessing

Nr. Of peak	SMILE	Deviation in KRI
1.	CC(C)CC(C)C' 'CCCC(C)(C)C' 'CCC(C)(C)C=C' 'CC=CC(C)(C)C' 'CC(C)C(C)(C)C'	0.1022 0.2321 0.2678 0.2678 0.4678
2.	CCCC(C)(C)C' 'CC(C)CC(C)C' 'CC(C)C(C)(C)C' 'CC(=C)C(C)(C)C' 'CCC(C)(C)C=C'	0.1878 0.2878 0.3121 0.6878 0.6878
3.	‘CC(C)C(C)(C)C' 'CC(=C)C(C)(C)C' 'CC(C)CC(=C)C' 'CC(CC=C)C=C' 'CC=CC(C)(C)C'	0.3958 0.3958 0.3958 0.3958 0.5958
⋮	⋮	⋮

The third step is to calculate the similarity of every candidate molecule in the list to each other as it is shown in Table 7.2-3 (implemented in Python), where the value of similarity is equal to 1 if two molecules are identical and equal to 0 if the molecules are completely different. Here, we would like to highlight that one molecule can appear in several rows and columns in the table since one molecule can be a candidate for multiple peaks. In the next steps of the method, we address this issue by formulating constraints on both the occurrence frequency of molecules and the number of molecules associated with individual peaks.

Table 7.2-3 Example of calculated molecular similarities in case of one experiment

	CC(C)CC(C)C	**CCCC(C)(C)C**	**CCC(C)(C)C=C**	**…**
CC(C)CC(C)C	1.000	0.636	0.278	**…**
CCCC(C)(C)C	0.636	1.000	0.389	**…**
CCC(C)(C)C=C	0.278	0.389	1.000	**…**
⋮	⋮	⋮	⋮	…

7.3 Proposed methodology

This chapter describes a possible solution to estimate the molecular composition of a given mixture of hydrocarbons, produced during a pyrolysis process.

For better understanding, a brief summarization of the essential inputs and the calculation chain is shown in Figure 7.3-1, where: GC CG – gas chromatographic chromatograms; KRI -Kovats retention index, SMILES - simplified molecular-input line-entry system.

- **Data collection**: The raw data consists of SMILES and corresponding Kovats retention indices from an external database and GC chromatograms (multiple analysis of the product at different times) in which the paraffinic peaks were determined in advance.
- **Data preprocessing:**
    1. Align the peaks from individual chromatograms and compute the average Kovats retention index for all corresponding peaks.
    2. Using this calculated average Kovats index along with the collected database, choose five potential molecules for each chromatographic peak.
    3. Calculate the similarities between the candidate molecules.

- **Solution:** Formulate the task as an integer linear programming (ILP) problem and assure the feasibility of the solution by the established constraints.
- The dark grey colour indicates the steps which were published elsewhere.

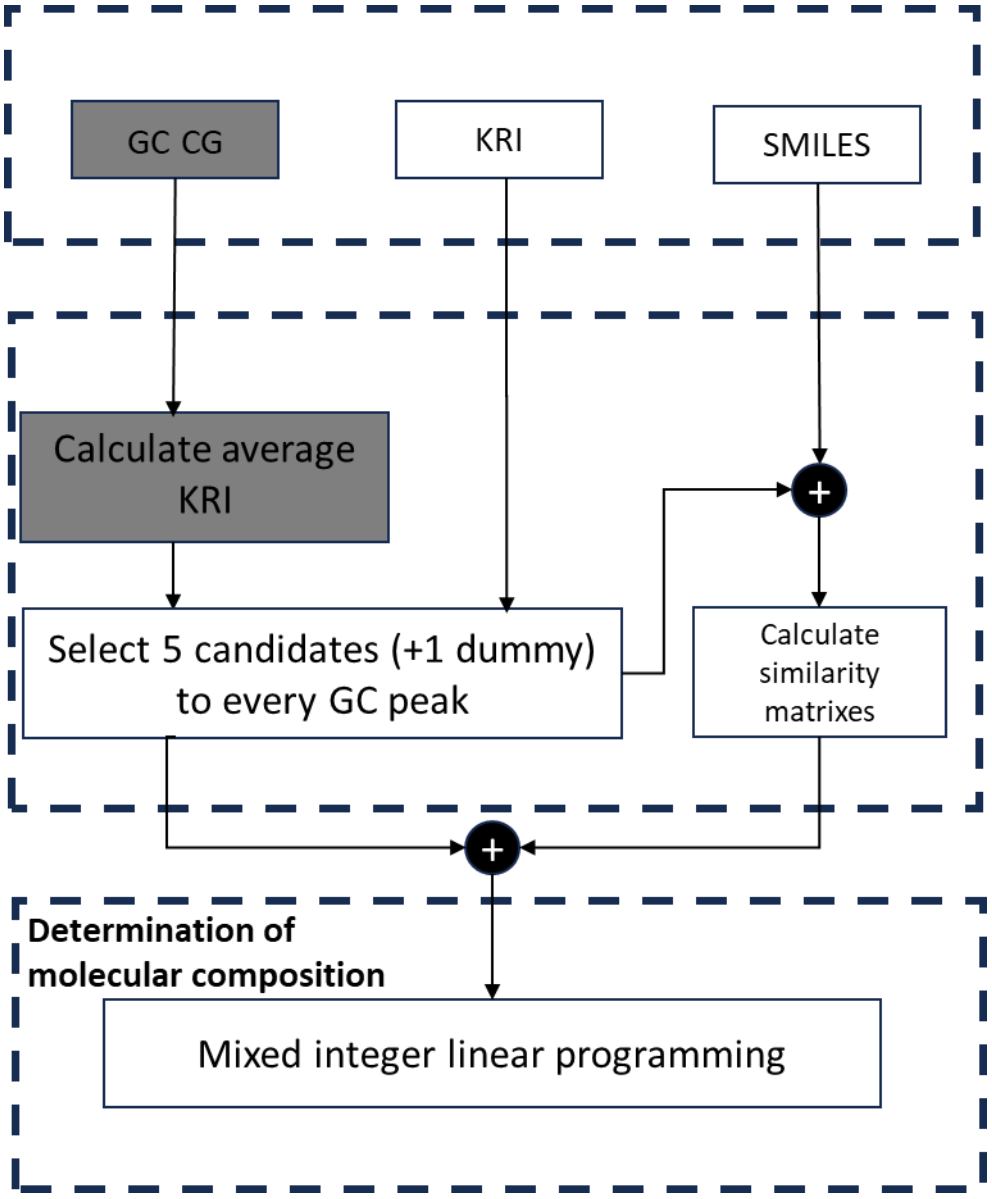

Figure 7.3-1 Single flow chart of the proposed methodology

To determine which candidates are presented in the mixture, an ILP problem was formulated that generally formulated as:

$$\min_x f^T x \; subject\; to \begin{cases} A \cdot x \leq b \\ Aeq \cdot x \leq beq \\ LB \leq x \leq UB \end{cases} \quad 7.3\text{-}1$$

In this problem, the values of x can only be 0 and 1 indicating the absence or presence of the candidates respectively. Ideally, the value of the coefficients contained in f would consist of the absolute deviation of Kovats retention indexes (to be minimized, shown in Table 7.2-2) and the molecular similarities (to be maximized).

Since the problem can be best explained by example, let's assume a much simpler problem: there are only three peaks with three candidate molecules in case of all peaks, as it is shown in Table 7.3-1. In Table 7.3-1 - Table 7.3-4, *S* abbreviation indicates similarity, *MOL* indicates molecule and *D* indicates deviation in Kovats index. Subscripted letters denote the candidates, numbers denote the corresponding peaks. As it was highlighted previously, one molecule can be a candidate for multiple peaks. For example, MOL_C is candidate for $Peak_1$ and $Peak_2$ as well as MOL_D and MOL_E are candidates for $Peak_2$ and $Peak_3$. The task here is to select one candidate for each peak in a way that one candidate can only be selected once.

Table 7.3-1 Simplified example of the problem

Problem		
	Candidates	**Deviations**
$Peak_1$	MOL_A MOL_B MOL_C	$D_{A,1}$ $D_{B,1}$ $D_{C,1}$
$Peak_2$	MOL_C MOL_D MOL_E	$D_{C,2}$ $D_{D,2}$ $D_{E,2}$
$Peak_3$	MOL_D MOL_E MOL_F	$D_{D,3}$ $D_{E,3}$ $D_{F,3}$

Since the methodology combines different types of information, besides the deviation in KRI, the calculated molecular similarities are also needed. In Table 7.3-3. and Table 7.3-4. the structure of the deviation in KRI and the structure of the calculated molecular similarities are shown respectively. Table 7.3-2 shows the corresponding existence variables, which values can only be 0 or 1. The presence of a molecule in the mixture is determined by the main diagonal. Therefore, the value of the elements of the smaller matrixes containing elements of the main diagonal cannot be 1 (e.g. e_2, e_3, e_{10}, e_{12}, e_{19}, e_{21}), and the smaller matrixes without diagonal elements are determines the similarity between the peaks (e.g. e_4, e_5, e_6, e_{13}, e_{14}, e_{15}, e_{22}, e_{23}, e_{24}).

Table 7.3-2 Corresponding existence variables. The value of ei can only be 0 or 1.

		Peak₁			**Peak₂**			**Peak₃**		
		MOL_A	MOL_B	MOL_C	MOL_C	MOL_D	MOL_E	MOL_D	MOL_E	MOL_F
Peak₁	MOL_A	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9
	MOL_B	e_{10}	e_{11}	e_{12}	e_{13}	e_{14}	e_{15}	e_{16}	e_{17}	e_{18}
	MOL_C	e_{19}	e_{20}	e_{21}	e_{22}	e_{23}	e_{24}	e_{25}	e_{26}	e_{27}
Peak₂	MOL_C	e_{28}	e_{29}	e_{30}	e_{31}	e_{32}	e_{33}	e_{34}	e_{35}	e_{36}
	MOL_D	e_{37}	e_{38}	e_{39}	e_{40}	e_{41}	e_{42}	e_{43}	e_{44}	e_{45}
	MOL_E	e_{46}	e_{47}	e_{48}	e_{49}	e_{50}	e_{51}	e_{52}	e_{53}	e_{54}
Peak₃	MOL_D	e_{55}	e_{56}	e_{57}	e_{58}	e_{59}	e_{60}	e_{61}	e_{62}	e_{63}
	MOL_E	e_{64}	e_{65}	e_{66}	e_{67}	e_{68}	e_{69}	e_{70}	e_{71}	e_{72}
	MOL_F	e_{73}	e_{74}	e_{75}	e_{76}	e_{77}	e_{78}	e_{79}	e_{80}	e_{81}

In Table 7.3-3. the values are changing row wise, meaning that the deviation in KRI is independent from the molecular composition of the mixture. The value of the molecular similarity is between 0 and 1, but the value of the deviation can be much larger. In this case, the effect of the deviation in KRI would be the dominant on the objective function, so these values must be normalized between 0 and 1 as well.

Table 7.3-3 Corresponding deviations in KRI

	MOL_A	$D_{A,1}$	$D_{A,1}$	$D_{A,1}$	$D_{A,1}$	$D_{A,1}$	$D_{A,1}$	$D_{A,1}$	$D_{A,1}$	$D_{A,1}$
Peak$_1$	MOL_B	$D_{B,1}$	$D_{B,1}$	$D_{B,1}$	$D_{B,1}$	$D_{B,1}$	$D_{B,1}$	$D_{B,1}$	$D_{B,1}$	$D_{B,1}$
	MOL_C	$D_{C,1}$	$D_{C,1}$	$D_{C,1}$	$D_{C,1}$	$D_{C,1}$	$D_{C,1}$	$D_{C,1}$	$D_{C,1}$	$D_{C,1}$
	MOL_C	$D_{C,2}$	$D_{C,2}$	$D_{C,2}$	$D_{C,2}$	$D_{C,2}$	$D_{C,2}$	$D_{C,2}$	$D_{C,2}$	$D_{C,2}$
Peak$_2$	MOL_D	$D_{D,2}$	$D_{D,2}$	$D_{D,2}$	$D_{D,2}$	$D_{D,2}$	$D_{D,2}$	$D_{D,2}$	$D_{D,2}$	$D_{D,2}$
	MOL_E	$D_{E,2}$	$D_{E,2}$	$D_{E,2}$	$D_{E,2}$	$D_{E,2}$	$D_{E,2}$	$D_{E,2}$	$D_{E,2}$	$D_{E,2}$
	MOL_D	$D_{D,3}$	$D_{D,3}$	$D_{D,3}$	$D_{D,3}$	$D_{D,3}$	$D_{D,3}$	$D_{D,3}$	$D_{D,3}$	$D_{D,3}$
Peak$_3$	MOL_E	$D_{E,3}$	$D_{E,3}$	$D_{E,3}$	$D_{E,3}$	$D_{E,3}$	$D_{E,3}$	$D_{E,3}$	$D_{E,3}$	$D_{E,3}$
	MOL_F	$D_{F,3}$	$D_{F,3}$	$D_{F,3}$	$D_{F,3}$	$D_{F,3}$	$D_{F,3}$	$D_{F,3}$	$D_{F,3}$	$D_{F,3}$

In Table 7.3-4. the values of the elements of the main diagonal are equal to 1 since the row and the column determines the same molecule and the other similarities are symmetric to the main diagonal.

Table 7.3-4 Corresponding molecular similarities

		Peak$_1$			**Peak$_2$**			**Peak$_3$**		
		MOL_A	MOL_B	MOL_C	MOL_C	MOL_D	MOL_E	MOL_D	MOL_E	MOL_F
	MOL_A	1	S_{AB}	S_{AC}	S_{AC}	S_{AD}	S_{AE}	S_{AD}	S_{AE}	S_{AF}
Peak$_1$	MOL_B	S_{AB}	1	S_{BC}	S_{BC}	S_{BD}	S_{BE}	S_{BD}	S_{BE}	S_{BF}
	MOL_C	S_{AC}	S_{BC}	1	1	S_{CD}	S_{CE}	S_{CD}	S_{CE}	S_{CF}
	MOL_C	S_{AC}	S_{BC}	1	1	S_{CD}	S_{CE}	S_{CD}	S_{CE}	S_{CF}
Peak$_2$	MOL_D	S_{AD}	S_{BD}	S_{CD}	S_{CD}	1	S_{ED}	1	S_{DE}	S_{DF}
	MOL_E	S_{AE}	S_{BE}	S_{CE}	S_{CE}	S_{DE}	1	S_{DE}	1	S_{EF}
	MOL_D	S_{AD}	S_{BD}	S_{CD}	S_{CD}	1	S_{DE}	1	S_{DE}	S_{DF}
Peak$_3$	MOL_E	S_{AE}	S_{BE}	S_{CE}	S_{CE}	S_{DE}	1	S_{DE}	1	S_{EF}
	MOL_F	S_{AF}	S_{BF}	S_{CF}	S_{CF}	S_{DF}	S_{EF}	S_{DF}	S_{EF}	1

During the solution, the e_i (Table 7.3-2.) variables must be determined in such a way, that their value can be 0 (absence) or 1 (presence). The value of the coefficient (in the objective function) associated with the existence variable is the sum of the elements which in Table 7.3-3 and in Table 7.3-4 are located in the identical position.

If the problem is solved with a non-linear objective function, only the values of the variables in the main diagonal must be determined. In this case, the position of the true values (1) determines the value of the coefficients. However, the population-based algorithms which are suitable to solve a constrained nonlinear optimization problem (e.g. genetic algorithms) have limited capability to find feasible solutions. At the same time, the linear problem requires much more constraints to be specified. The constraints of the ILP problem determined by the task are consists of both linear inequality and linear equality constraints. The equality constraints ensure that:

- A candidate must be selected for each peak: the sum of the existence variables correspond to one peak must be equal to one. For example:

$$e_1 + e_2 + e_3 + e_{10} + e_{11} + e_{12} + e_{19} + e_{20} + e_{21} = 1 \qquad 7.3\text{-}2$$

- The main diagonal determines the composition: the sum of the elements of the main diagonal is equal to the number of peaks.

$$e_1 + e_{11} + e_{21} + e_{31} + e_{41} + e_{51} + e_{61} + e_{71} + e_{81} = 3 \qquad 7.3\text{-}3$$

- The corresponding similarities are selected - the row sums equal to the corresponding column sums e.g.:

$$\begin{gathered} e_1 + e_{10} + e_{19} + e_{28} + e_{37} + e_{46} + e_{55} + e_{64} + e_{73} \\ -e_1 - e_2 - e_3 - e_4 - e_5 - e_6 - e_7 - e_8 - e_9 = 0 \end{gathered} \qquad 7.3\text{-}4$$

The inequality constraints ensure that a candidate is selected only once, so the sum of variables representing the same molecules must be smaller or equal to one e.g.:

$$e_{21} + e_{31} \leq 1 \qquad 7.3\text{-}5$$

The final problem can be formulated as:

$$\min_e (S + D)^T e \; subjet \; to \begin{cases} A \cdot e \leq b \\ Aeq \cdot e \leq beq \\ 0 \leq e \leq 1 \end{cases} \qquad 7.3\text{-}6$$

Where the elements of e are integers.

The original problem was formulated in the same way for each 27 experiments using all eight variations of molecular similarities. To avoid formulating a problem without feasible solutions, the candidates of each peak were analysed. If one set of candidates were identical to another, a “dummy candidate” was added to both sets. In the resulted optimisation problems, the number of the variables were around 3 million (350 peaks, 5 candidates for each peak. That will result an existence variable matrix with 1750 rows and 1750 columns.). The integer linear programming problem were implemented and solved in MATLAB 2023b, using the implemented *intlinprog* function. Due to the large number of the variables, the average RAM consumption was between 60-80 GB.

7.4 Results

The determined molecular compositions are evaluated using two types of evaluation criteria:

- Comparison of computed results with high-level measurement results, such as total aromatic and olefin content.
- Comparison of computed results with each other to assess the consistency of the developed method: during the experiments, multiple catalysts were tested at different temperatures. The optimization problem was formulated and solved for all measurements, and the plausibility of the determined molecular composition was checked through the consistency of the results, meaning that how similar the determined composition is in case of the same applied catalyst at different temperatures.

Before proceeding with the evaluation, it's important to note that in three out of 27 experimental data sets (Zn - 425°C, H – 455 °C, Ni – 455 °C), the algorithm failed due to excessively high computational costs, specifically, high RAM consumption.

Figure 7.4-1 presents a pairwise comparison of the results obtained using different molecular similarity methods across all measurements. The colorbar indicates the degree of similarity in molecular composition determined by the various methods. The similarity value is a continuous variable ranging from 0 to 1, where 0 indicates completely different compositions and 1 signifies identical compositions. This value is calculated as follows: the number of identical molecules in two calculated compositions divided by the total number of molecules.

Upon initial observation, the prominent emergence of blue lines is noticeable. These blue lines indicate that the Russel similarity method differs the most from all other methods, although the lowest match between two compositions is above 0.90, the average best match is 97.45%, observed between the Dice and Kulczynski similarities. By employing various similarity indices, the most similar composition, with a match of 96.52%, was determined in the 16^{th} measurement (H – 485°C),

while the poorest match, achieving 94.53%, was observed in the 15th measurement (Ce(IV) – 455°C).

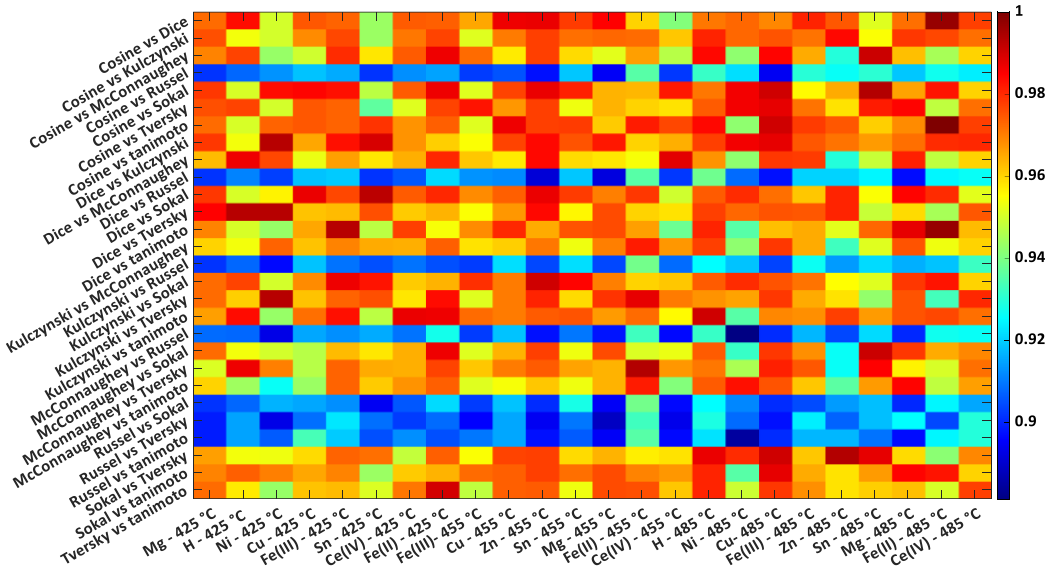

Figure 7.4-1 Similarity between compositions determined using different similarity indices.

The molecules under investigation were categorized according to four fundamental structural properties: containing a double bond, forming a ring, being aromatic, or existing as an isomer. Here, 'isomer' refers to branched-chain hydrocarbons. Figure 7.4-2 presents a Venn diagram-based comparison of the molecular compositions determined using different similarity indices. The diagrams depict the categorized representation of the identified molecules across all measurements. The intersection of all sets in every case—except for Russel similarity—consists of 25 molecules. This intersection contains aromatic hydrocarbons with branched hains and double bonds. The most populous set is the alkyls, where one-third of the molecules are solely ismers, and another third of the molecules not only contain branched chains but also have double bonds. In total, approximately 91% of the identified molecules are isomers. Additionally, around 5% of the molecules are straight-chained olefins, and some cycloalkanes and cycloalkenes were also found. The most significant conclusion drawn from the

figure is that the compositions determined by the individual methods exhibit very similar molecular structures.

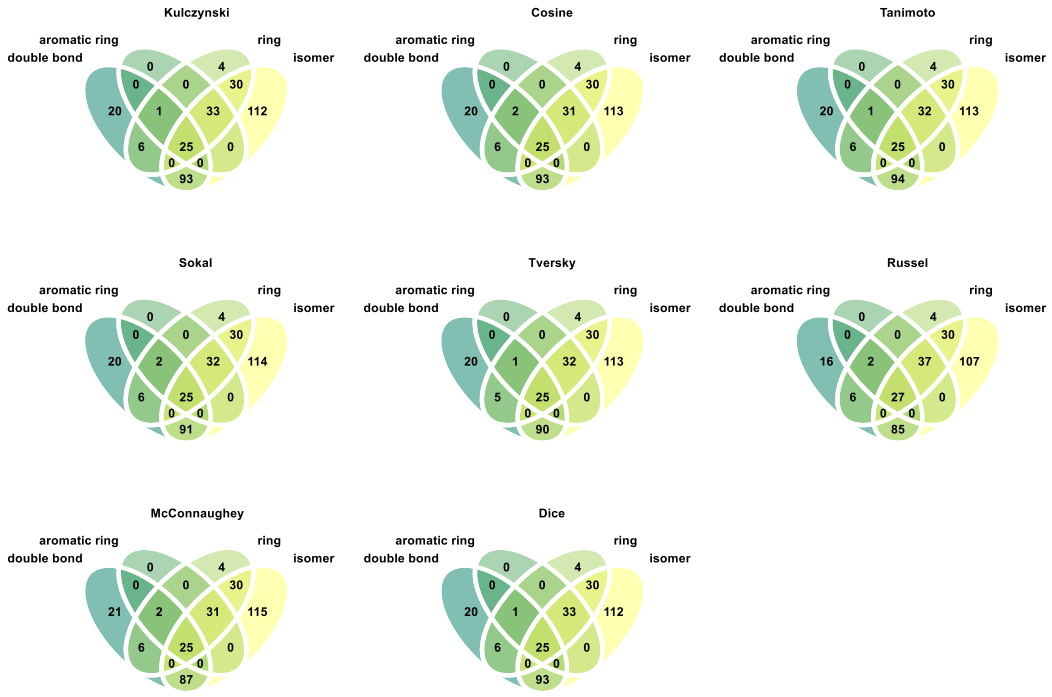

Figure 7.4-2 Venn diagram-based comparison of the molecular compositions

Figure 7.4-3 and Figure 7.4-4 illustrates the comparison of computed compositions with high-level measurement results. The three empty subplots represent measurements where the algorithm failed. From the figures, it is evident that the estimated olefin concentration is much more accurate than the estimation of the aromatic content, and the accuracy in both cases decreases with increasing temperature. The mean errors compared to the measured olefin concentration are: Cosine – 4.7 m/m%; Dice – 4.8 m/m%; Kulczynski – 5 m/m%; McConnaughey – 4.6 m/m%; Russel – 4.7 m/m% ; Sokal – 4.7 m/m%; Tversky – 4.8 m/m %; Tanimoto – 4.9 m/m %. The maximum errors are between 10.1 – 11.33 m/m % while the minimum errors are between 0.17 – 0.5 m/m%.

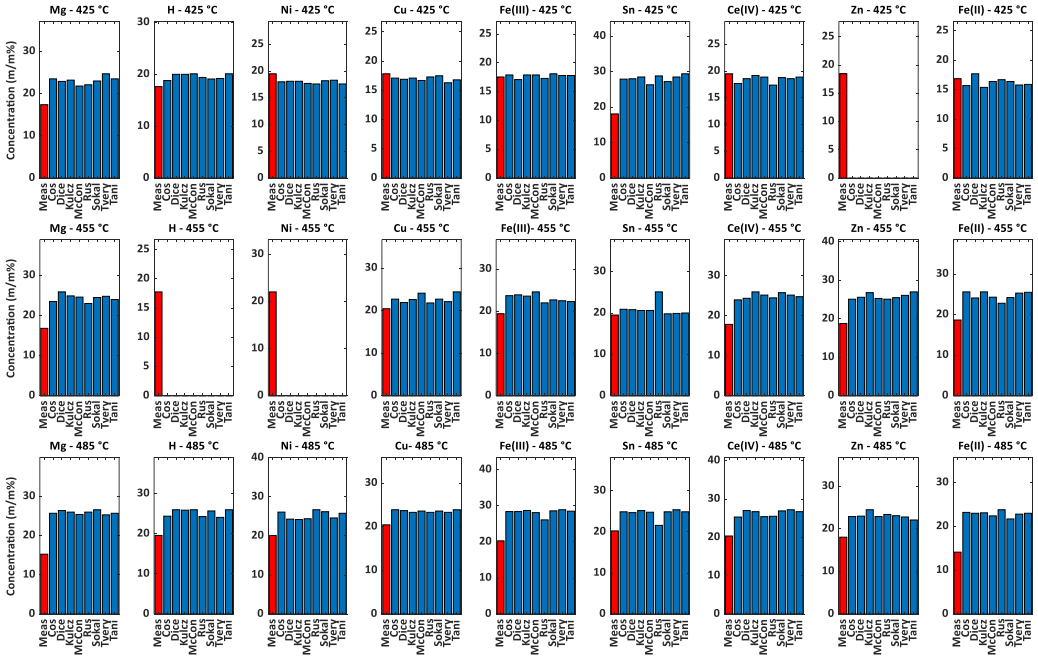

Figure 7.4-3 Comparison of the measured and calculated total olefin content. Where, Meas – Measurement; Cos-Cosine; Dice-Dice; Kulcz – Kulczynski; McCon – McConnaughey; Rus – Russel; Sokal – Sokal; Tver -Tversky; Tani – Tanimoto similarity

The inaccuracy in prediction of aromatic content may be attributed to multiple reasons, but two are particularly probable. Firstly, the assumption that all molecules present in the pyrolysis product are included in the collected database may not hold true, leading to the selection of false molecules. Secondly, the similarity between aromatic molecules may be higher than that between other types of molecules. Supporting this hypothesis: 56% of the similarities between candidate aromatic molecules are above 0.5, whereas only 38% of the similarities between olefinic candidates are above 0.5.

The results obtained with different molecular similarity measures show good accordance regarding both molecular structures and predicted concentrations. The best average match (75%) in the determined composition, using the same catalyst at different temperatures, was achieved with the Cu catalyst, while the overall average was 71%.

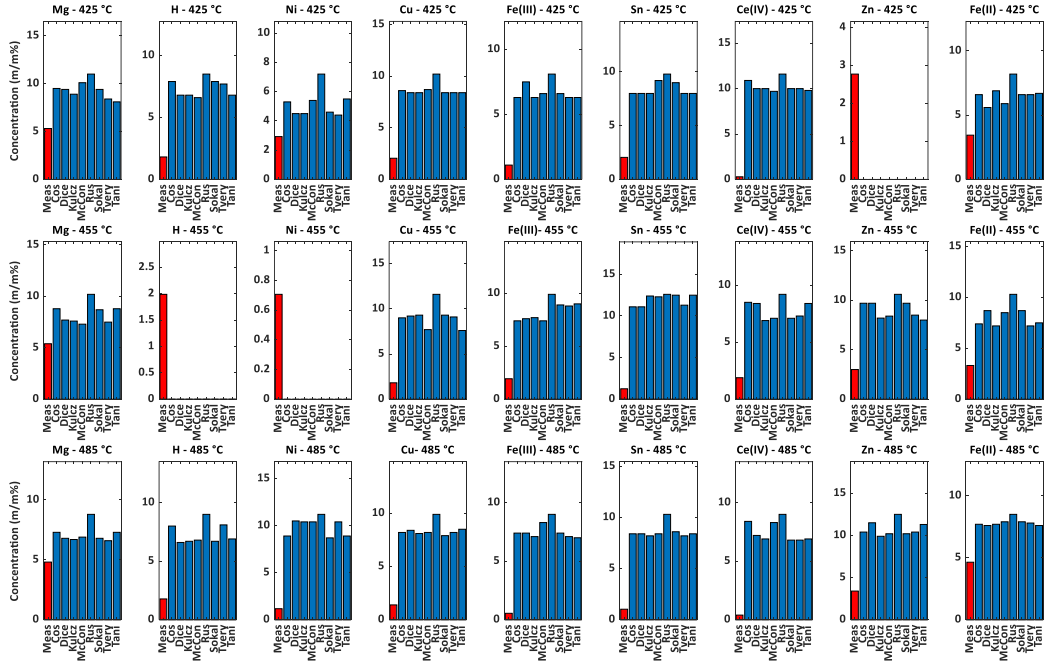

Figure 7.4-4 Comparison of the measured and calculated total aromatic content. Where, Meas – Measurement; Cos-Cosine; Dice-Dice; Kulcz – Kulczynski; McCon – McConnaughey; Rus – Russel; Sokal – Sokal; Tver -Tversky; Tani – Tanimoto similarity.

7.5 Conclusion

In this chapter, we have explored the application of computational methods based on molecular similarities and Kovats retention index to enhance the accuracy of qualitative analysis relying on gas chromatography data, particularly in catalytic pyrolysis of plastic waste. Through comparison with high-level measurement results and amongst each other, we have evaluated the determined molecular compositions. Notably, despite encountering challenges such as algorithm failures due to high computational costs in certain measurements, our analysis revealed significant insights.

Our findings underscore the importance of molecular similarity methods in refining the estimation of molecular compositions, especially in scenarios where

the accuracy of retention indices in databases is uncertain. By leveraging various similarity indices, we have demonstrated that the identified molecules exhibit highly similar molecular structures, with some variations observed based on different measurement conditions. Additionally, we have identified potential reasons contributing to inaccuracies, such as limitations in the database coverage and variations in similarity between aromatic and other types of molecules.

There is a clear opportunity for further development by formulating additional constraints based on high-level measured concentrations to refine the accuracy of molecular composition estimates. Such advancements will not only enhance our understanding of catalytic pyrolysis processes but also contribute to the optimization of waste-to-value conversion strategies.

Overall, our study highlights the potential of computational methods in facilitating more precise qualitative analysis of complex chemical mixtures, paving the way for sustainable waste management practices and the valorization of plastic waste into valuable products.

8 Summary and future work

The thesis includes chemical process analysis and modeling, with a primary focus on sustainable energy production and waste management. It encompasses several key areas of investigation across multiple chapters.

In the third chapter, catalyst deactivation phenomena during hydrocracking of a sunflower oil and kerosene mixture are analyzed. The aim is to develop lumped models capable of describing complex chemical systems, particularly in the context of aviation fuel production. Experimental data is utilized to identify and integrate catalyst deactivation models into kinetic modeling, providing insights into process optimization and catalyst performance.

In the fourth chapter, the research delves into continuous lumping approaches to model intricate hydrocracking chemistry. The emphasis is on developing models that can capture changes over time and space, optimizing residence time and selectivity for improved process efficiency.

The fifth chapter investigates the selection of thermodynamic models for accurate process simulations, particularly in hydrogen solubility studies. A Gaussian mixture model is developed to optimize model selection based on varying operating conditions, enhancing the reliability and accuracy of thermodynamic predictions.

In the sixth chapter, a novel method is introduced to eliminate retention time drifts in gas chromatography analysis of pyrolysis products. By applying a modified k-means algorithm, chromatographic data alignment is achieved, facilitating more straightforward qualitative analysis and process understanding.

Finally, the seventh chapter explores computational methods integrating molecular similarities and retention indices to enhance qualitative analysis in catalytic pyrolysis. This approach aims to refine the estimation of molecular compositions under diverse process conditions, contributing to more precise and efficient waste valorization strategies.

Overall, the thesis underscores the importance of advanced computational techniques, and innovative modeling approaches in advancing sustainable energy production and waste management practices within the chemical engineering domain. Each chapter contributes unique insights and methodologies towards optimizing chemical processes for environmental and economic sustainability.

Further steps may include the introduction of a novel model, based on the extended continuous lumping approach and the developed approach for hydrogen solubility estimation. The combination of those two approaches would be resulted in an accurate model which can estimates the pressure dependency as well. Moreover, the developed molecular composition estimation can be the cornerstone for a new, single-event type modelling approach, in which the estimated composition would be the basis of reaction pathway identification.

9 Theses

Thesis #1. I extended the discrete-lumped modeling approach by incorporating catalyst deactivation models to provide a more complex description of catalyst fouling.

I have investigated catalyst deactivation phenomena in case of hydrocracking of sunflower oil and kerosene mixture. The investigation includes development of lumped models to describe the chemical system, moreover catalyst deactivation models were integrated into the kinetic model to improve estimation performance.

Related publications: 1.

Thesis #2. I revised the continuous-lumped modeling approach to better describe the spatial and temporal changes in a hydrocracker reactor.

I have developed a novel continuous lumping approach to model hydrocracking chemistry, enabling to analyze changes over time and space to optimize residence time and selectivity for enhanced process efficiency. Furthermore, I have developed a Gaussian mixture model to optimize the model selection for solubility estimation of hydrogen under varying operating conditions.

Related publications: 3,5.

Thesis #3 I developed new computational methods for qualitative analysis in catalytic pyrolysis by integrating molecular similarity measures and retention indices to refine the estimation of molecular compositions.

I have developed computational methods integrating molecular similarities and retention indices to refine the estimation of molecular compositions in catalytic pyrolysis. The first part of this development involved a fast and simple clustering algorithm to eliminate time drifts between chromatograms using easily accessible prior information. The second part consisted of formulating and solving a linear programming problem that applies molecular similarity measures to further refine the estimation of molecular compositions.

Related publications: 2, 4.

10 Publications related to this thesis

Articles in international journals

1. Hamadi, Omar & Varga, Tamás & Till, Zoltán & Eller, Zoltán & Hancsók, Jenő. (2019). Model based investigation of catalyst fouling in case of special hydrocracking of sunflower oil and kerosene mixture. Energy & Fuels. 33. 10.1021/acs.energyfuels.8b04085.

2. Hamadi, Omar & Varga, Tamás. (2022). Semi-supervised Clustering Algorithm for Retention Time Alignment of Gas Chromatographic Data. Periodica Polytechnica Chemical Engineering. 66. 10.3311/PPch.18834.

3. Hamadi, Omar & Varga, Tamás. (2023). Novel distributed parameter model-based continuous lumping approach: An application to a pilot-plant hydrocracking reactor. Chemical Engineering Science. 271. 118572. 10.1016/j.ces.2023.118572.

4. Hamadi, Omar & Varga, Tamás. (2024). Computational Insights into Catalytic Pyrolysis: Refining Molecular Composition Estimates using Kovats Retention Index and Molecular Similarities. Industrial & Engineering Chemistry Research. 10.1021/acs.iecr.4c03040

Articles in conference publications

5. Hamadi, Omar & Varga, Tamás & Abonyi, János. (2020). Application Domain Discovery of Thermodynamic Models by Mixture of Experts Learning. 10.1016/B978-0-12-823377-1.50066-5.

11 References

[1] H.-B. Park, K.-D. Kim, and Y.-K. Lee, ‘Promoting asphaltene conversion by tetralin for hydrocracking of petroleum pitch’, *Fuel*, vol. 222, pp. 105–113, Jun. 2018, doi: 10.1016/j.fuel.2018.02.154.

[2] M. S. Rana, V. Sámano, J. Ancheyta, and J. A. I. Diaz, ‘A review of recent advances on process technologies for upgrading of heavy oils and residua’, *Fuel*, vol. 86, no. 9, pp. 1216–1231, Jun. 2007, doi: 10.1016/j.fuel.2006.08.004.

[3] M. S. Rigutto, R. Van Veen, and L. Huve, ‘Zeolites in Hydrocarbon Processing’, in *Studies in Surface Science and Catalysis*, vol. 168, Elsevier, 2007, pp. 855–XXVI. doi: 10.1016/S0167-2991(07)80812-3.

[4] R. Saab, K. Polychronopoulou, L. Zheng, S. Kumar, and A. Schiffer, ‘Synthesis and performance evaluation of hydrocracking catalysts: A review’, *Journal of Industrial and Engineering Chemistry*, vol. 89, pp. 83–103, Sep. 2020, doi: 10.1016/j.jiec.2020.06.022.

[5] Y. Choi, J. Lee, J. Shin, S. Lee, D. Kim, and J. K. Lee, ‘Selective hydroconversion of naphthalenes into light alkyl-aromatic hydrocarbons’, *Applied Catalysis A: General*, vol. 492, pp. 140–150, Feb. 2015, doi: 10.1016/j.apcata.2014.12.001.

[6] J.-I. Park, J.-K. Lee, J. Miyawaki, Y.-K. Kim, S.-H. Yoon, and I. Mochida, ‘Hydro-conversion of 1-methyl naphthalene into (alkyl)benzenes over alumina-coated USY zeolite-supported NiMoS catalysts’, *Fuel*, vol. 90, no. 1, pp. 182–189, Jan. 2011, doi: 10.1016/j.fuel.2010.09.002.

[7] M. Romero *et al.*, ‘Preliminary experimental study on biofuel production by deoxygenation of Jatropha oil’, *Fuel Processing Technology*, vol. 137, pp. 31–37, Sep. 2015, doi: 10.1016/j.fuproc.2015.04.002.

[8] M. M. Gui, K. T. Lee, and S. Bhatia, ‘Feasibility of edible oil vs. non-edible oil vs. waste edible oil as biodiesel feedstock’, *Energy*, vol. 33, no. 11, pp. 1646–1653, Nov. 2008, doi: 10.1016/j.energy.2008.06.002.

[9] G. W. Huber, P. O’Connor, and A. Corma, ‘Processing biomass in conventional oil refineries: Production of high quality diesel by hydrotreating vegetable oils in heavy vacuum oil mixtures’, *Applied Catalysis A: General*, vol. 329, pp. 120–129, Oct. 2007, doi: 10.1016/j.apcata.2007.07.002.

[10] A. S. Nizami *et al.*, ‘Developing waste biorefinery in Makkah: A way forward to convert urban waste into renewable energy’, *Applied Energy*, vol. 186, pp. 189–196, Jan. 2017, doi: 10.1016/j.apenergy.2016.04.116.

[11] S. M. Sadrameli, ‘Thermal/catalytic cracking of hydrocarbons for the production of olefins: A state-of-the-art review I: Thermal cracking review’, *Fuel*, vol. 140, pp. 102–115, Jan. 2015, doi: 10.1016/j.fuel.2014.09.034.

[12] M. Syamsiro *et al.*, 'Fuel Oil Production from Municipal Plastic Wastes in Sequential Pyrolysis and Catalytic Reforming Reactors', *Energy Procedia*, vol. 47, pp. 180–188, 2014, doi: 10.1016/j.egypro.2014.01.212.

[13] R. Miandad, M. A. Barakat, A. S. Aburiazaiza, M. Rehan, and A. S. Nizami, 'Catalytic pyrolysis of plastic waste: A review', *Process Safety and Environmental Protection*, vol. 102, pp. 822–838, Jul. 2016, doi: 10.1016/j.psep.2016.06.022.

[14] A. Nawaz and S. A. Razzak, 'Co-pyrolysis of biomass and different plastic waste to reduce hazardous waste and subsequent production of energy products: A review on advancement, synergies, and future prospects', *Renewable Energy*, vol. 224, p. 120103, Apr. 2024, doi: 10.1016/j.renene.2024.120103.

[15] L. P. de Oliveira, D. Hudebine, D. Guillaume, and J. J. Verstraete, 'A Review of Kinetic Modeling Methodologies for Complex Processes', *Oil Gas Sci. Technol. – Rev. IFP Energies nouvelles*, vol. 71, no. 3, p. 45, May 2016, doi: 10.2516/ogst/2016011.

[16] H. M. S. Lababidi and F. S. AlHumaidan, 'Modeling the Hydrocracking Kinetics of Atmospheric Residue in Hydrotreating Processes by the Continuous Lumping Approach', *Energy Fuels*, vol. 25, no. 5, pp. 1939–1949, May 2011, doi: 10.1021/ef200153p.

[17] J. C. W. Kuo and J. Wei, 'Lumping Analysis in Monomolecular Reaction Systems. Analysis of Approximately Lumpable System', *Ind. Eng. Chem. Fund.*, vol. 8, no. 1, pp. 124–133, Feb. 1969, doi: 10.1021/i160029a020.

[18] J. Wei and J. C. W. Kuo, 'Lumping Analysis in Monomolecular Reaction Systems. Analysis of the Exactly Lumpable System', *Ind. Eng. Chem. Fund.*, vol. 8, no. 1, pp. 114–123, Feb. 1969, doi: 10.1021/i160029a019.

[19] S. A. Qader and G. R. Hill, 'Hydrocracking of Gas Oil', *Ind. Eng. Chem. Proc. Des. Dev.*, vol. 8, no. 1, pp. 98–105, Jan. 1969, doi: 10.1021/i260029a017.

[20] D. I. Orochko, I. Ya. Perezhigina, S. P. Rogov, M. V. Rysakov, and G. N. Chernakova, 'Applied over-all kinetics of hydrocracking of heavy petroleum distillates', *Chem Technol Fuels Oils*, vol. 6, no. 8, pp. 561–565, Aug. 1970, doi: 10.1007/BF00714107.

[21] J. F. Mosby, R. D. Buttke, J. A. Cox, and C. Nikolaides, 'Process characterization of expanded-bed reactors in series', *Chemical Engineering Science*, vol. 41, no. 4, pp. 989–995, 1986, doi: 10.1016/0009-2509(86)87184-6.

[22] A. R. Ayasse, H. Nagaishi, E. W. Chan, and M. R. Gray, 'Lumped kinetics of hydrocracking of bitumen', *Fuel*, vol. 76, no. 11, pp. 1025–1033, Sep. 1997, doi: 10.1016/S0016-2361(97)00104-X.

[23] B. E. Stangeland, 'A Kinetic Model for the Prediction of Hydrocracker Yields', *Ind. Eng. Chem. Proc. Des. Dev.*, vol. 13, no. 1, pp. 71–76, Jan. 1974, doi: 10.1021/i260049a013.

[24] V. W. Weekman and D. M. Nace, ‘Kinetics of catalytic cracking selectivity in fixed, moving, and fluid bed reactors’, *AIChE J.*, vol. 16, no. 3, pp. 397–404, May 1970, doi: 10.1002/aic.690160316.

[25] L.-S. Lee, Y.-W. Chen, T.-N. Huang, and W.-Y. Pan, ‘Four-lump kinetic model for fluid catalytic cracking process’, *Can. J. Chem. Eng.*, vol. 67, no. 4, pp. 615–619, Aug. 1989, doi: 10.1002/cjce.5450670414.

[26] M. Morales-Blancas, F. S. Mederos-Nieto, I. Elizalde, J. Felipe Sánchez-Minero, and F. Trejo-Zárraga, ‘Discrete lumping kinetic models for hydrodesulfuration and hydrocracking of a mixture of FCC feedstock and light gasoil’, *Chem. Pap.*, Apr. 2022, doi: 10.1007/s11696-022-02219-8.

[27] A. A. Forghani, M. Jafarian, P. Pendleton, and D. M. Lewis, ‘Mathematical modelling of a hydrocracking reactor for triglyceride conversion to biofuel: model establishment and validation: Analysis of a hydrocracking reactor for triglyceride conversion’, *International Journal of Energy Research*, vol. 38, no. 12, pp. 1624–1634, Oct. 2014, doi: 10.1002/er.3244.

[28] Z. Till, T. Varga, J. Sója, N. Miskolczi, and T. Chován, ‘Kinetic identification of plastic waste pyrolysis on zeolite-based catalysts’, *Energy Conversion and Management*, vol. 173, pp. 320–330, Oct. 2018, doi: 10.1016/j.enconman.2018.07.088.

[29] A. E. Lechleitner, T. Schubert, W. Hofer, and M. Lehner, ‘Lumped Kinetic Modeling of Polypropylene and Polyethylene Co-Pyrolysis in Tubular Reactors’, *Processes*, vol. 9, no. 1, p. 34, Dec. 2020, doi: 10.3390/pr9010034.

[30] G. Astarita and S. I. Sandler, Eds., *Kinetic and thermodynamic lumping of multicomponent mixtures: proceedings of an ACS Symposium on Kinetic and Thermodynamic Lumping of Multicomponent Mixtures, Atlanta, GA, April 15, 1991*. Amsterdam ; New York: Elsevier, 1991.

[31] ‘On the theory of reactions in continuous mixtures’, *Phil. Trans. R. Soc. Lond. A*, vol. 260, no. 1112, pp. 351–393, Sep. 1966, doi: 10.1098/rsta.1966.0054.

[32] T. de Donder, *L’Affinite*, 2nd ed. Paris: Gautier Villars, 1931.

[33] M. Y. Chou and T. C. Ho, ‘Continuum theory for lumping nonlinear reactions’, *AIChE J.*, vol. 34, no. 9, pp. 1519–1527, Sep. 1988, doi: 10.1002/aic.690340914.

[34] C. S. Laxminarasimhan, R. P. Verma, and P. A. Ramachandran, ‘Continuous lumping model for simulation of hydrocracking’, *AIChE J.*, vol. 42, no. 9, pp. 2645–2653, Sep. 1996, doi: 10.1002/aic.690420925.

[35] I. Elizalde and J. Ancheyta, ‘On the detailed solution and application of the continuous kinetic lumping modeling to hydrocracking of heavy oils’, *Fuel*, vol. 90, no. 12, pp. 3542–3550, Dec. 2011, doi: 10.1016/j.fuel.2011.03.044.

[36] I. Elizalde, M. A. Rodríguez, and J. Ancheyta, ‘Application of continuous kinetic lumping modeling to moderate hydrocracking of heavy oil’, *Applied*

Catalysis A: General, vol. 365, no. 2, pp. 237–242, Aug. 2009, doi: 10.1016/j.apcata.2009.06.018.

[37] I. Elizalde, M. A. Rodríguez, and J. Ancheyta, 'Modeling the effect of pressure and temperature on the hydrocracking of heavy crude oil by the continuous kinetic lumping approach', *Applied Catalysis A: General*, vol. 382, no. 2, pp. 205–212, Jul. 2010, doi: 10.1016/j.apcata.2010.04.050.

[38] I. Elizalde and J. Ancheyta, 'Modeling the Simultaneous Hydrodesulfurization and Hydrocracking of Heavy Residue Oil by using the Continuous Kinetic Lumping Approach', *Energy Fuels*, vol. 26, no. 4, pp. 1999–2004, Apr. 2012, doi: 10.1021/ef201916s.

[39] I. Elizalde and J. Ancheyta, 'Modeling catalyst deactivation during hydrocracking of atmospheric residue by using the continuous kinetic lumping model', *Fuel Processing Technology*, vol. 123, pp. 114–121, Jul. 2014, doi: 10.1016/j.fuproc.2014.02.006.

[40] E. Kováts, 'Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone', *Helvetica Chimica Acta*, vol. 41, no. 7, pp. 1915–1932, Jan. 1958, doi: 10.1002/hlca.19580410703.

[41] R. Ghavami and S. Faham, 'QSRR Models for Kováts' Retention Indices of a Variety of Volatile Organic Compounds on Polar and Apolar GC Stationary Phases Using Molecular Connectivity Indexes', *Chroma*, vol. 72, no. 9–10, pp. 893–903, Nov. 2010, doi: 10.1365/s10337-010-1741-4.

[42] T. R. NOVIANDY, 'THE PREDICTION OF KOVATS RETENTION INDICES OF ESSENTIAL OILS AT GAS CHROMATOGRAPHY USING GENETIC ALGORITHM-MULTIPLE LINEAR REGRESSION AND SUPPORT VECTOR REGRESSION', *Journal of Engineering Science and Technology*, vol. 0306-0326., 2022.

[43] B. d'Acampora Zellner, C. Bicchi, P. Dugo, P. Rubiolo, G. Dugo, and L. Mondello, 'Linear retention indices in gas chromatographic analysis: a review', *Flavour & Fragrance J*, vol. 23, no. 5, pp. 297–314, Sep. 2008, doi: 10.1002/ffj.1887.

[44] F. Gong, Y.-Z. Liang, Y.-S. Fung, and F.-T. Chau, 'Correction of retention time shifts for chromatographic fingerprints of herbal medicines', *Journal of Chromatography A*, vol. 1029, no. 1–2, pp. 173–183, Mar. 2004, doi: 10.1016/j.chroma.2003.12.049.

[45] H. Parastar, M. Jalali-Heravi, and R. Tauler, 'Comprehensive two-dimensional gas chromatography (GC×GC) retention time shift correction and modeling using bilinear peak alignment, correlation optimized shifting and multivariate curve resolution', *Chemometrics and Intelligent Laboratory Systems*, vol. 117, pp. 80–91, Aug. 2012, doi: 10.1016/j.chemolab.2012.02.003.

[46] K. J. Johnson, B. W. Wright, K. H. Jarman, and R. E. Synovec, 'High-speed peak matching algorithm for retention time alignment of gas

chromatographic data for chemometric analysis', *Journal of Chromatography A*, vol. 996, no. 1–2, pp. 141–155, May 2003, doi: 10.1016/S0021-9673(03)00616-2.

[47] P. Zhu, W. Ding, W. Tong, A. Ghosal, K. Alton, and S. Chowdhury, 'A retention-time-shift-tolerant background subtraction and noise reduction algorithm (BgS-NoRA) for extraction of drug metabolites in liquid chromatography/mass spectrometry data from biological matrices', *Rapid Communications in Mass Spectrometry*, vol. 23, no. 11, pp. 1563–1572, Jun. 2009, doi: 10.1002/rcm.4041.

[48] Y. Koh, K. K. Pasikanti, C. W. Yap, and E. C. Y. Chan, 'Comparative evaluation of software for retention time alignment of gas chromatography/time-of-flight mass spectrometry-based metabonomic data', *Journal of Chromatography A*, vol. 1217, no. 52, pp. 8308–8316, Dec. 2010, doi: 10.1016/j.chroma.2010.10.101.

[49] T. G. Bloemberg, J. Gerretzen, A. Lunshof, R. Wehrens, and L. M. C. Buydens, 'Warping methods for spectroscopic and chromatographic signal alignment: A tutorial', *Analytica Chimica Acta*, vol. 781, pp. 14–32, Jun. 2013, doi: 10.1016/j.aca.2013.03.048.

[50] R. Bro, C. A. Andersson, and H. A. L. Kiers, 'PARAFAC2—Part II. Modeling chromatographic data with retention time shifts', *Journal of Chemometrics*, vol. 13, no. 3–4, pp. 295–309, May 1999, doi: 10.1002/(SICI)1099-128X(199905/08)13:3/4<295::AID-CEM547>3.0.CO;2-Y.

[51] M. D. Robinson *et al.*, 'A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments', *BMC Bioinformatics*, vol. 8, no. 1, p. 419, 2007, doi: 10.1186/1471-2105-8-419.

[52] P. T. Williams and E. A. Williams, 'Fluidised bed pyrolysis of low density polyethylene to produce petrochemical feedstock', *Journal of Analytical and Applied Pyrolysis*, vol. 51, no. 1–2, pp. 107–126, Jul. 1999, doi: 10.1016/S0165-2370(99)00011-X.

[53] S. H. Wijaya, F. M. Afendi, I. Batubara, L. K. Darusman, M. Altaf-Ul-Amin, and S. Kanaya, 'Finding an appropriate equation to measure similarity between binary vectors: case studies on Indonesian and Japanese herbal medicines', *BMC Bioinformatics*, vol. 17, no. 1, p. 520, Dec. 2016, doi: 10.1186/s12859-016-1392-z.

[54] M. J. Warrens, *Similarity coefficients for binary data: properties of coefficients, coefficient matrices, multi-way metrics and multivariate coefficients*. [S.l.: s.n.], 2008.

[55] S.-H. Cha, C. Tappert, and S. Yoon, 'Enhancing Binary Feature Vector Similarity Measures', *JPRR*, vol. 1, no. 1, pp. 63–77, 2006, doi: 10.13176/11.20.

[56] B. Zhang and S. N. Srihari, ‘Binary Vector Dissimilarity Measures for Handwriting Identification’, presented at the Electronic Imaging 2003, T. Kanungo, E. H. Barney Smith, J. Hu, and P. B. Kantor, Eds., Santa Clara, CA, Jan. 2003, pp. 28–38. doi: 10.1117/12.473347.

[57] Holliday, ‘Grouping of Coefficients for the Calculation of Inter-Molecular Similarity and Dissimilarity using 2D Fragment Bit-Strings’, *cchts*, vol. 5, no. 2, 2002, doi: 10.2174/1386207024607338.

[58] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, and P. Willett, ‘Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets’, *J. Chem. Inf. Model.*, vol. 52, no. 11, pp. 2884–2901, Nov. 2012, doi: 10.1021/ci300261r.

[59] A. D. S. Meyer, A. A. F. Garcia, A. P. D. Souza, and C. L. D. Souza Jr., ‘Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (Zea mays L)’, *Genet. Mol. Biol.*, vol. 27, no. 1, pp. 83–91, 2004, doi: 10.1590/S1415-47572004000100014.

[60] C. Seung-Seok, C. Sung-Hyuk, and Charles C Tappert, ‘A survey of binary similarity and distance measures’, *Journal of systemics, cybernetics and informatics*, vol. 8, no. 1, pp. 43–48, Jan. 2010.

[61] Consonni, Viviana and Todeschini, Roberto, ‘New similarity coefficients for binary data’, vol. 68, no. 581–592, Jan. 2012.

[62] E. Kosman and K. J. Leonard, ‘Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species’, *Molecular Ecology*, vol. 14, no. 2, pp. 415–424, Feb. 2005, doi: 10.1111/j.1365-294X.2005.02416.x.

[63] A. Tversky, ‘Features of similarity.’, *Psychological Review*, vol. 84, no. 4, pp. 327–352, Jul. 1977, doi: 10.1037/0033-295X.84.4.327.

[64] D. Rogers and M. Hahn, ‘Extended-Connectivity Fingerprints’, *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, May 2010, doi: 10.1021/ci100050t.

[65] ‘Daylight’. [Online]. Available: https://www.daylight.com/dayhtml/doc/theory/theory.finger.html

[66] Y.-S. Ooi, R. Zakaria, A. R. Mohamed, and S. Bhatia, ‘Catalytic Cracking of Used Palm Oil and Palm Oil Fatty Acids Mixture for the Production of Liquid Fuel: Kinetic Modeling’, *Energy & Fuels*, vol. 18, no. 5, pp. 1555–1561, Sep. 2004, doi: 10.1021/ef049948v.

[67] Y. K. Ong and S. Bhatia, ‘The current status and perspectives of biofuel production via catalytic cracking of edible and non-edible oils’, *Energy*, vol. 35, no. 1, pp. 111–119, Jan. 2010, doi: 10.1016/j.energy.2009.09.001.

[68] N. Taufiqurrahmi and S. Bhatia, ‘Catalytic cracking of edible and non-edible oils for the production of biofuels’, *Energy & Environmental Science*, vol. 4, no. 4, p. 1087, 2011, doi: 10.1039/c0ee00460j.

[69] Z. D. Yigezu and K. Muthukumar, ‘Biofuel production by catalytic cracking of sunflower oil using vanadium pentoxide’, *Journal of Analytical and Applied Pyrolysis*, vol. 112, pp. 341–347, Mar. 2015, doi: 10.1016/j.jaap.2015.01.002.

[70] S. Sunarno, R. Rochmadi, P. Mulyono, M. Aziz, and A. Budiman, ‘Kinetic Study of Catalytic Cracking of Bio-oil over Silica-alumina Catalyst’, *BioResources*, vol. 13, no. 1, Jan. 2018, doi: 10.15376/biores.13.1.1917-1929.

[71] H. F. Meier, V. R. Wiggers, G. R. Zonta, D. R. Scharf, E. L. Simionatto, and L. Ender, ‘A kinetic model for thermal cracking of waste cooking oil based on chemical lumps’, *Fuel*, vol. 144, pp. 50–59, Mar. 2015, doi: 10.1016/j.fuel.2014.12.020.

[72] B. Periyasamy, ‘Reaction pathway analysis in thermal cracking of waste cooking oil to hydrocarbons based on monomolecular lumped kinetics’, *Fuel*, vol. 158, pp. 479–487, Oct. 2015, doi: 10.1016/j.fuel.2015.05.066.

[73] M. Anand and A. K. Sinha, ‘Temperature-dependent reaction pathways for the anomalous hydrocracking of triglycerides in the presence of sulfided Co–Mo-catalyst’, *Bioresource Technology*, vol. 126, pp. 148–155, Dec. 2012, doi: 10.1016/j.biortech.2012.08.105.

[74] A. K. Sinha *et al.*, ‘Development of Hydroprocessing Route to Transportation Fuels from Non-Edible Plant-Oils’, *Catalysis Surveys from Asia*, vol. 17, no. 1, pp. 1–13, Mar. 2013, doi: 10.1007/s10563-012-9148-x.

[75] A. A. Forghani, M. Jafarian, P. Pendleton, and D. M. Lewis, ‘Mathematical modelling of a hydrocracking reactor for triglyceride conversion to biofuel: model establishment and validation: Analysis of a hydrocracking reactor for triglyceride conversion’, *International Journal of Energy Research*, vol. 38, no. 12, pp. 1624–1634, Oct. 2014, doi: 10.1002/er.3244.

[76] O. Péter Hamadi, T. Varga, Z. Till, Z. Eller, and J. Hancsók, *Model based investigation of catalyst fouling in case of special hydrocracking of sunflower oil and kerosene mixture*, vol. 33. 2019. doi: 10.1021/acs.energyfuels.8b04085.

[77] G. Félix, J. Ancheyta, and F. Trejo, ‘Sensitivity analysis of kinetic parameters for heavy oil hydrocracking’, *Fuel*, vol. 241, pp. 836–844, Apr. 2019, doi: 10.1016/j.fuel.2018.12.058.

[78] C. E. Galarraga, C. Scott, H. Loria, and P. Pereira-Almao, ‘Kinetic Models for Upgrading Athabasca Bitumen Using Unsupported NiWMo Catalysts at Low Severity Conditions’, *Industrial & Engineering Chemistry Research*, vol. 51, no. 1, pp. 140–146, Jan. 2012, doi: 10.1021/ie201202b.

[79] P. Forzatti, ‘Catalyst deactivation’, *Catalysis Today*, vol. 52, no. 2–3, pp. 165–181, Sep. 1999, doi: 10.1016/S0920-5861(99)00074-7.

[80] C. H. Bartholomew and R. J. Farrauto, *Fundamentals of industrial catalytic processes*, 2nd ed. Hoboken, N.J: Wiley, 2006.

[81] B. W. Wojciechowski, ‘THE KINETIC FOUNDATIONS AND THE PRACTICAL APPLICATION OF THE TIME ON STREAM THEORY OF CATALYST DECAY’, *Catalysis Reviews*, vol. 9, no. 1, pp. 79–113, Jan. 1974, doi: 10.1080/01614947408075370.

[82] C. G. Rudershausen and C. C. Watson, ‘Variables affecting activity of molybdena-alumina hydroforming catalyst in aromatization of cyclohexane’, *Chemical Engineering Science*, vol. 3, no. 3, pp. 110–121, Jun. 1954, doi: 10.1016/0009-2509(54)80016-9.

[83] E. Wolf, ‘Kinetics of deactivation of a reforming catalyst during methylcyclohexane dehydrogenation in a diffusion reactor’, *Journal of Catalysis*, vol. 46, no. 2, pp. 190–203, Feb. 1977, doi: 10.1016/0021-9517(77)90199-3.

[84] S. Szepe and O. Levenspiel, ‘Proceedings of the 4th European Symposium on Chemical Reaction Engineering’, in *Chemical reaction engineering: proceedings*, 1st ed., European Federation of Chemical Engineering, Ed., Oxford, New York: Pergamon Press, 1971.

[85] L. Jossens, ‘Fouling of a platinum-rhenium reforming catalyst using model reforming reactions’, *Journal of Catalysis*, vol. 76, no. 2, pp. 265–273, Aug. 1982, doi: 10.1016/0021-9517(82)90257-3.

[86] M. Pacheco, ‘On the development of a catalyst fouling model’, *Journal of Catalysis*, vol. 88, no. 2, pp. 400–408, Aug. 1984, doi: 10.1016/0021-9517(84)90017-4.

[87] M. Pacheco, ‘On a general correlation for catalyst fouling’, *Journal of Catalysis*, vol. 86, no. 1, pp. 75–83, Mar. 1984, doi: 10.1016/0021-9517(84)90349-X.

[88] I. Pitault, D. Nevicato, M. Forissier, and J.-R. Bernard, ‘Kinetic model based on a molecular description for catalytic cracking of vacuum gas oil’, *Chemical Engineering Science*, vol. 49, no. 24, pp. 4249–4262, 1994, doi: 10.1016/S0009-2509(05)80018-1.

[89] C. Chu, ‘Effect of Adsorption on Fouling of Catalyst Pellets’, *Industrial & Engineering Chemistry Fundamentals*, vol. 7, no. 3, pp. 509–514, Aug. 1968, doi: 10.1021/i160027a024.

[90] E. K. T. Kam, P. A. Ramachandran, and R. Hughes, ‘The effect of film resistances on the fouling of catalyst pellets—I Pseudo-steady state analysis’, *Chemical Engineering Science*, vol. 32, no. 11, pp. 1307–1315, 1977, doi: 10.1016/0009-2509(77)85025-2.

[91] A. Monzón, E. Romeo, and A. Borgna, ‘Relationship between the kinetic parameters of different catalyst deactivation models’, *Chemical Engineering Journal*, vol. 94, no. 1, pp. 19–28, Jul. 2003, doi: 10.1016/S1385-8947(03)00002-0.

[92] R. Prins, ‘Eley–Rideal, the Other Mechanism’, *Top Catal*, vol. 61, no. 9–11, pp. 714–721, Jun. 2018, doi: 10.1007/s11244-018-0948-8.

[93] S. Le Digabel, ‘Algorithm 909: NOMAD: Nonlinear Optimization with the MADS Algorithm’, *ACM Transactions on Mathematical Software*, vol. 37, no. 4, pp. 1–15, Feb. 2011, doi: 10.1145/1916461.1916468.

[94] R. BENNETT and K. BOURNE, ‘HYDROCRACKING FOR MIDDLE DISTILLATE. A STUDY OF PROCESS REACTIONS AND CORRESPONDING PRODUCT YIELDS AND QUALITIES’, *AMER. CHEM. SOC., DIV. PETROLEUM CHEM., PREPR*, vol. VOL 17, no. NO 4, pp. G45–G62, 1972.

[95] S. Hamdi, W. Schiesser, and G. Griffiths, ‘Method of lines’, *Scholarpedia*, vol. 2, no. 7, p. 2859, 2007, doi: 10.4249/scholarpedia.2859.

[96] S. Le Digabel, ‘Algorithm 909: NOMAD: Nonlinear Optimization with the MADS Algorithm’, *ACM Transactions on Mathematical Software*, vol. 37, no. 4, pp. 1–15, Feb. 2011, doi: 10.1145/1916461.1916468.

[97] J. Ancheyta, S. Sánchez, and M. A. Rodríguez, ‘Kinetic modeling of hydrocracking of heavy oil fractions: A review’, *Catalysis Today*, vol. 109, no. 1–4, pp. 76–92, Nov. 2005, doi: 10.1016/j.cattod.2005.08.015.

[98] J. R. Hart, I. Guymer, F. Sonnenwald, and V. R. Stovin, ‘Residence Time Distributions for Turbulent, Critical, and Laminar Pipe Flow’, *J. Hydraul. Eng.*, vol. 142, no. 9, p. 04016024, Sep. 2016, doi: 10.1061/(ASCE)HY.1943-7900.0001146.

[99] D. Zudkevitch and J. Joffe, ‘Correlation and prediction of vapor-liquid equilibria with the redlich-kwong equation of state’, *AIChE Journal*, vol. 16, no. 1, pp. 112–119, Jan. 1970, doi: 10.1002/aic.690160122.

[100] K. C. Chao and J. D. Seader, ‘A general correlation of vapor-liquid equilibria in hydrocarbon mixtures’, *AIChE Journal*, vol. 7, no. 4, pp. 598–605, Dec. 1961, doi: 10.1002/aic.690070414.

[101] R. Torres, J.-C. de Hemptinne, and I. Machin, ‘Improving the Modeling of Hydrogen Solubility in Heavy Oil Cuts Using an Augmented Grayson Streed (AGS) Approach’, *Oil & Gas Science and Technology – Revue d’IFP Energies nouvelles*, vol. 68, no. 2, pp. 217–233, Mar. 2013, doi: 10.2516/ogst/2012061.

[102] E. Carlson, ‘Don’t Gamble with Physical Properties for Simulations’, *Chemical Engineering Progress*, vol. 92, pp. 35–46, 0 1996.

[103] A. Al-Matar, ‘Selecting Fluid Packages (Thermodynamic Model) for HYSYS/ Aspen Plus/ ChemCAD Process Simulators’, 2015, *Unpublished*. doi: 10.13140/RG.2.1.3461.4487.

[104] S. E. Yuksel, J. N. Wilson, and P. D. Gader, ‘Twenty Years of Mixture of Experts’, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 8, pp. 1177–1193, Aug. 2012, doi: 10.1109/TNNLS.2012.2200299.

[105] T. Tsuji *et al.*, ‘Hydrogen solubility in triolein, and propane solubility in oleic acid for second generation BDF synthesis by use of hydrodeoxygenation

reaction', *Fluid Phase Equilibria*, vol. 362, pp. 383–388, Jan. 2014, doi: 10.1016/j.fluid.2013.11.006.

[106] C. L. Young, Ed., *Hydrogen and deuterium*, 1st ed. in Solubility data series, no. v. 5-6. Oxford ; New York: Pergamon Press, 1981.

[107] N. Miskolczi, J. Sója, and E. Tulok, 'Thermo-catalytic two-step pyrolysis of real waste plastics from end of life vehicle', *Journal of Analytical and Applied Pyrolysis*, vol. 128, pp. 1–12, Nov. 2017, doi: 10.1016/j.jaap.2017.11.008.

[108] Z. Till, T. Chován, and T. Varga, 'Uncertainties of Lumped Reaction Networks in Reactor Design', *Ind. Eng. Chem. Res.*, vol. 59, no. 22, pp. 10531–10541, Jun. 2020, doi: 10.1021/acs.iecr.0c00549.

[109] P. J. Becker, N. Serrand, B. Celse, D. Guillaume, and H. Dulot, 'Comparing hydrocracking models: Continuous lumping vs. single events', *Fuel*, vol. 165, pp. 306–315, Feb. 2016, doi: 10.1016/j.fuel.2015.09.091.

[110] Z. Till, T. Varga, J. Sója, N. Miskolczi, and T. Chován, 'Structural assessment of lumped reaction networks with correlating parameters', *Energy Conversion and Management*, vol. 209, p. 112632, Apr. 2020, doi: 10.1016/j.enconman.2020.112632.

[111] N. Ganganath, C.-T. Cheng, and C. K. Tse, 'Data Clustering with Cluster Size Constraints Using a Modified K-Means Algorithm', in *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, Shanghai, China: IEEE, Oct. 2014, pp. 158–161. doi: 10.1109/CyberC.2014.36.

[112] Wagstaf, Kiri and Cardie, Claire and Rogers, Seth and Schrödl, and Stefan, 'Constrained K-means Clustering with Background Knowledge', *Proceedings of 18th International Conference on Machine Learning*, no. 577–584, Jan. 2001.

[113] F. E. Grubbs, 'Sample Criteria for Testing Outlying Observations', *Ann. Math. Statist.*, vol. 21, no. 1, pp. 27–58, Mar. 1950, doi: 10.1214/aoms/1177729885.

[114] J. Sója, 'Szénhidrogénfrakciók előállítása autóipari műanyaghulladékok hőbontásával', 2020, doi: 10.18136/PE.2020.764.

[115] Greg Landrum *et al.*, *rdkit/rdkit: Release_2023.09.5*. (Feb. 08, 2024). Zenodo. doi: 10.5281/ZENODO.10633624.

[116] 'RDKitbook'. [Online]. Available: https://www.rdkit.org/docs/RDKit_Book.html

[117] O. P. Hamadi and T. Varga, 'Semi-supervised Clustering Algorithm for Retention Time Alignment of Gas Chromatographic Data', *Period. Polytech. Chem. Eng.*, vol. 66, no. 3, pp. 414–421, May 2022, doi: 10.3311/PPch.18834.

The Nomenclature is grouped based on to the topics:

Discrete lumping		
TG	concentration of Triglycerides	[g/g]
C_D	concentration of Diesel	[g/g]
C_K	concentration of Kerosene	[g/g]
C_{GO}	concentration of Gasoline	[g/g]
C_G	concentration of Gas	[g/g]
k_i	reaction rate constant	[1/min]
a	activity	[-]
Ψ_d	deactivation function	[-]
Ψ^*_d	deactivation function*	[-]
d	deactivation order	[-]
a_s	residual activity	[-]
t	time	[min]
C_{cat}	catalyst concentration	[g/g]
$C_{TG\&cat.}$	catalyst+ Triglycerides (adsorbed TG)	[g/g]
$k_{TG+cat.}$	reaction rate constant for adsorbing	[1/min]
$k_{TG-cat.}$	reaction rate constant for desorbing	[1/min]
V_i	volume of pseudo-components	[cm^3]
ρ_i	density of pseudo-components	[g/cm^3]
M_i	Average molar mass of pseudo-components	[g/mol]
N_A	Avogadro number	[1/mol]
r_i	radius of pseudo-components	cm
A_i	surface of pseudo-components	cm^2/g

$A_{cat.}$	surface of the catalyst	cm^2/g
K_i	Extent of the adhesion of pseudo-components	[-]
c_i	surface concentration of pseudo-components	[g/g]
y_{exp}	measurement concentrations	[g/g]
y_{model}	calculated concentrations	[g/g]
y^{max}_{exp}	maximum of the measurement concentrations	[g/g]
A_0	Pre-exponential factor	[1/min]
Ea	Activation Energy	[kJ/kg]
T	Temperature	[K]
β	Correction of pre-exponential factor	[-]
Continuous lumping		
w_i	Weight fraction	[g/100g]
t	Time	[s]
v_x	Flow velocity	[m/s]
x	Axial position	[m]
TBP (θ)	True boiling point	[°C]
TBP(l)	The lowest true boiling point	[°C]
TBP(h)	The highest true boiling point	[°C]
k_{max}	Reaction rate constant for the heaviest component	[1/min]
k	Reaction rate constant	[1/min]
α, α_2, α_3, u, s	Parameters of the selectivity-distribution.	[-]
P(*k*,*K*)	The yield of a species with reactivity *k* resulting from the hydrocracking of components with reactivity *K*.	[-]
D(k)	Species-type distribution fcn.	[-]

σ	Model parameter of P(k,K)	[-]
l	Number of the pseudo components	[-]
n	Nominal molecular weight	[-]
y_{exp} , y_{model}	the experimental and the predicted weight fractions.	[g/100g]
Exploration of application domains for thermodynamic models		
N	Number of pairs of validation data	[-]
x	Explanatory variables (carbon number, number of H atoms)	[-]
C_n	Carbon number	[-]
H_n	Number of hydrogen atoms	[-]
$\hat{y}_{k,j}$	k^{th} Predicted hydrogen solubility by the j^{th} thermodynamic model	[Mole fraction (-)]
y_k	k^{th} predicted hydrogen solubility	[-]
$e_{k,j}$	prediction error of the j^{th} TM	[-]
η_j	parameters of the j^{th} cluster	[-]
$D^2(\mathbf{x}_k, y_k, \eta_j)$	Squared distance between the validation data point and the j^{th} cluster.	[-]
α_j	Weighting factor for the j^{th} cluster. It accounts for the relative importance or contribution of each TM to the overall distribution.	[-]
$\sigma^2_{j,l}$	Variance of the prediction error of the j^{th} TM.	[-]
Σ_j	Covariance matrix of the j^{th} cluster, defines the spread of the data in the space of explanatory variables.	[-]
MSE	Mean square error	[-]
Retention time alignment of gas chromatographic data		
x	Data point	[-]
X	Data set	[-]

k	Number of clusters	[-]
c	Cluster	[-]
n	Number of peaks	[-]
m	Number of measurements	[-]
$x_{n,m}$	n^{th} peak from m^{th} measurement	[min]
$x_{pa,t}$	Retention time of paraffinic peak trailing $x_{n,m}$	[min]
$x_{pa,h}$	Retention time of paraffinic peak heading $x_{n,m}$	[min]
$\hat{x}$	Normalised retention time	[-]
μ_j	Centroid of cluster j.	[-]
C_n	Carbon number	[-]
Kovats Retention Index and Molecular Similarities		
KRI	Kovats retention index	[-]
LB	Lower boundaries	[-]
UB	Upper boundaries	[-]
A	Coefficients for non-equity constarints	[-]
Aeq	Coefficients for equity constraints	[-]
x	variables	[-]
MOL	Molecule	[-]
e	Existence variable	[-]
D	Deviation	[-]
S	Molecular similarity	[-]

Acknowledgement

Dedicated to everyone who believed in me!

Thank you to my beloved mother, father, grandmother, and sisters!

I am deeply grateful to my mentor and supervisor, Varga Tamás, who supported me throughout this long journey. From the very beginning, you were the mentor I needed—guiding me, showing me the way, and helping me grow. I will deeply miss our conversations, brainstorming sessions, and all the “aha moments.” Thank you for everything; I will truly miss you!

I would also like to express my heartfelt thanks to Alex Kummer for the helping hand that enabled me to complete this dissertation. Thank you, Alex!

Finally, I extend my gratitude to my colleagues in the Department of Process Engineering for their support, as well as to my co-authors for their valuable time and contributions.

"It isn’t the mountains ahead to climb that wear you out*; it’s the pebble in your shoe."*

Muhammad Ali