https://doi.org/10.3311/PPch.18834
Creative Commons Attribution ⓘ

Periodica Polytechnica Chemical Engineering, 66(3), pp. 414–421, 2022

Semi-supervised Clustering Algorithm for Retention Time Alignment of Gas Chromatographic Data

Omar Péter Hamadi[1]*, Tamás Varga[1]

[1] Research Centre for Biochemical, Environmental and Chemical Engineering, Faculty of Engineering, University of Pannonia, Egyetem u. 10, H-8200 Veszprém, Hungary
* Corresponding author, e-mail: hamadio@fmt.uni-pannon.hu

Received: 25 June 2021, Accepted: 14 October 2021, Published online: 10 February 2022

Abstract

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.

Keywords

constrained k-means, cannot-link, maximum-cluster size, pyrolysis

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 [1].

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 [2], GC × GC data [3], diesel fuel GC profiles [1], drug metabolites LC/MS data [4], and metabonomic GC/MS data [5]. The most commonly used methods to eliminate the time drifts are the wrapping algorithms and principal component analysis (PCA). A clear summary of wrapping methods for chromatographic signal alignment is available in [6]. PARAFAC2 is a generalization of PCA, which is a powerful and popular tool for handling retention time shifts [7]. 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 [8].

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

Cite this article as: Hamadi, O. P., Varga, T. "Semi-supervised Clustering Algorithm for Retention Time Alignment of Gas Chromatographic Data", Periodica Polytechnica Chemical Engineering, 66(3), pp. 414–421, 2022. https://doi.org/10.3311/PPch.18834

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 were performed at different temperature levels using different zeolite based catalysts, additional details can be found in [9].

Based on these experiments a lumped kinetic model was developed and published in [10], and the uncertainty of the model was diminished by reducing the size of the reaction network in [11]. 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 [12]. 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 aforementioned 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.

2 Proposed methodology

Suppose that $X = \{x_{1,1}, x_{1,2}, \ldots, x_{1,m}, x_{2,1}, x_{2,2}, \ldots, x_{2,m}, \ldots, x_{n,m}\}$ 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.

2.1 Preprocessing the data

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 [1, 13];
- 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 we stated in our previous modelling study of this system, only a small changes can be noticed in the chromatograms of pyrolysis product samples taken at different time steps [2]. 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 Fig. 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 alignments within these segment groups are the subtasks.

In Fig. 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. Fig. 2 (b) shows the data from Fig. 2 (a) when it is normalized to 0–1 range for each segment in the segment group separately according to Eq. (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.)

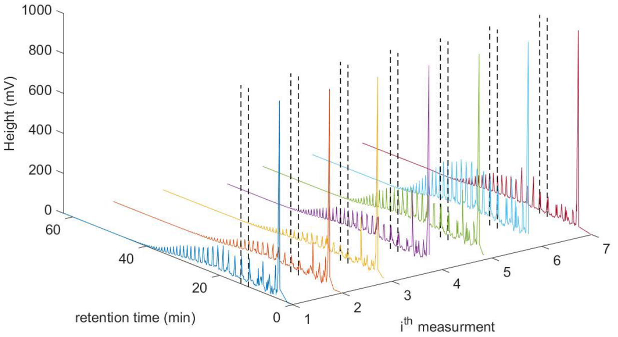

Fig. 1 The chromatographic data. The segments between the dashed lines denote the C_{10} fractions.

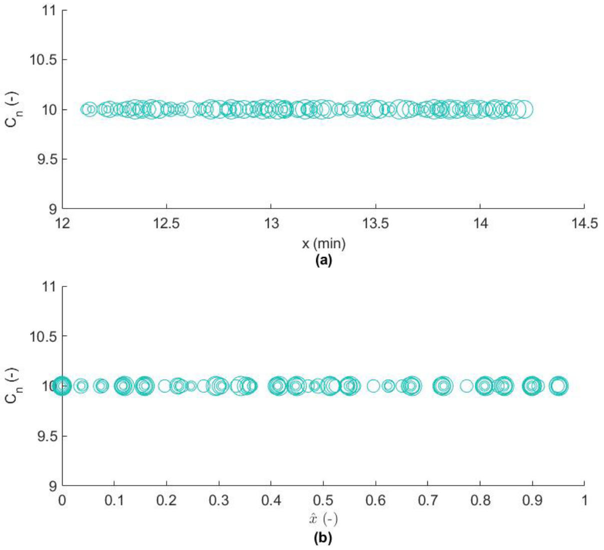

Fig. 2 The retention times of C_{10} fractions from all chromatograms before (a) and after (b) the normalization

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

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 normalization 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 needs to be minimized. All the above-mentioned facts led us to use k-means for the retention time alignment.

2.2 Modified K-means algorithm

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 [3]:

1. Initialization: initialization of the centroids (μ_j) (usually random data points from the data set) according to Eq. (2).

$$\mu_j^1 = \left\{x_p : x_p \in \boldsymbol{X}, \mu_i^1 \neq x_p, 1 \le i \le k, i \neq j\right\} \tag{2}$$

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

$$c_j^t = \left\{x_p : x_p - \mu_j^{t2} \le x_p - \mu_i^{t2}, 1 \le i \le k\right\} \tag{3}$$

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

$$\mu_j^{t+1} = \frac{1}{\left|c_j^t\right|} \sum_{\forall x_i \in c_j^t} x_i \tag{4}$$

The proposed algorithm (Fig. 3) terminates when the number of maximum iterations is reached (or the cluster centers 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 [4]. However, if data points were to be eliminated from clusters in order to satisfy the constraint, an iteration will be used constructed in which the algorithm rather finds the nearest center to the points, than assign the nearest points to the center. 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 center will forced to be assigned to another cluster. A modified k-means algorithm with must-link/cannot-link constraint is published in [4], however in this study we provide a detailed approach from an engineering point of view.

In the proposed algorithm the assignment step is complemented (Fig. 3), 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 jth cluster and ζj denotes the maximum size of the jth cluster, than an extra constraint is has to be satisfied: $|c_j| \le \zeta j$. 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.

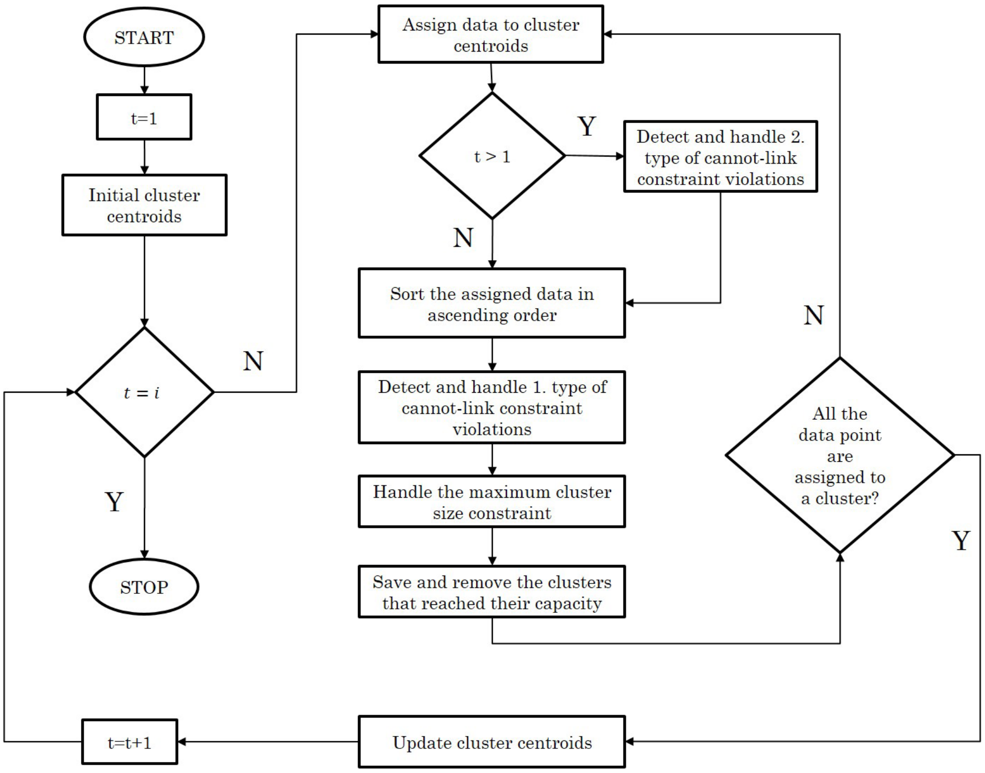
START
t=1
Initial cluster centroids
t = i
N
Y
STOP
Assign data to cluster centroids
t > 1
Y
Detect and handle 2. type of cannot-link constraint violations
N
Sort the assigned data in ascending order
Detect and handle 1. type of cannot-link constraint violations
Handle the maximum cluster size constraint
Save and remove the clusters that reached their capacity
N
All the data point are assigned to a cluster?
Y
t=t+1
Update cluster centroids

Fig. 3 Simplified flow chart of modified k-means algorithm

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 center 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 center 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 Fig. 3.

2.3 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). 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.

3 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. Fig. 4 is similar to Fig. 2 (b), but normalization was performed for all subtasks (segment groups). Fig. 4 confirms the statement that the normalization balanced the retention time-drifts such an extent that the modified k-means algorithm can be applied.

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 Fig. 4 (except data points with $x = 0$ coordinate) can take part in the clustering at the same time.

The results are shown in Fig. 5, the clusters are circled and marked with colors 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. Fig. 5 shows that the developed algorithm partitioned the data points effectively and can handle the overlapping.

In Fig. 6 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 Fig. 7 (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 Fig. 7 (b) shows the distribution of cluster sizes along the measurements. Since the heights 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.

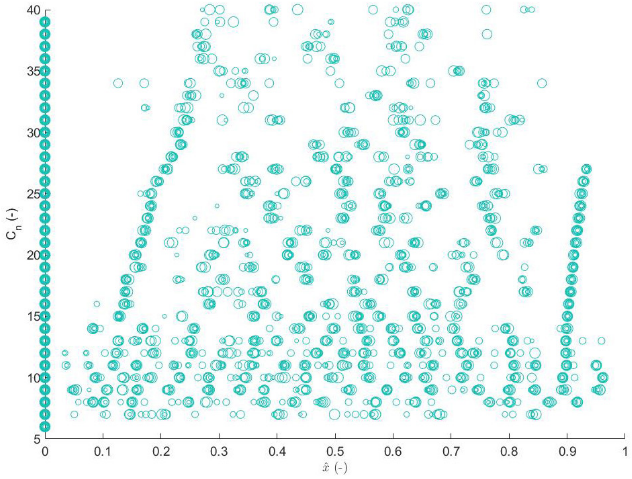

Fig. 4 The normalized retention times for all chromatographic data, y coordinate denotes the fractions

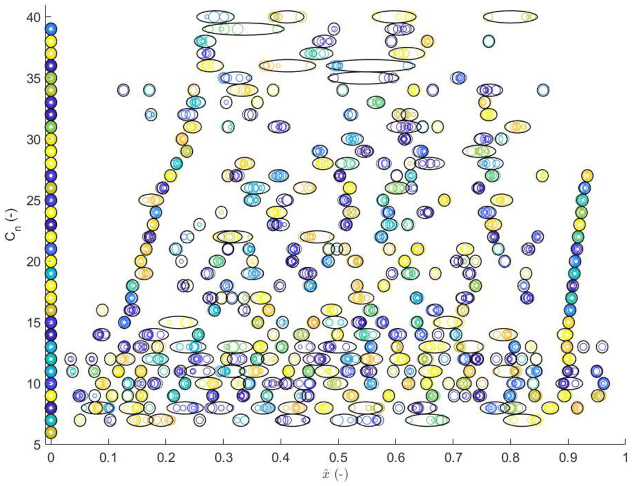

Fig. 5 The resulted clusters, i.e. the components in the pyrolysis product. The individual clusters are circled and marked with different colors as well

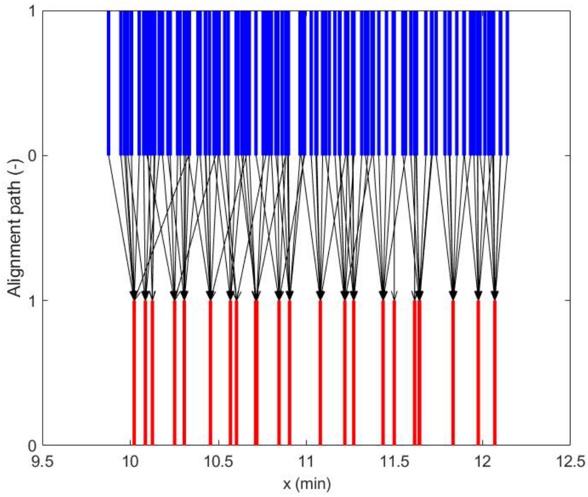

Fig. 6 The alignment of peaks in C_{10} fraction from seven different chromatograms (different sample states)

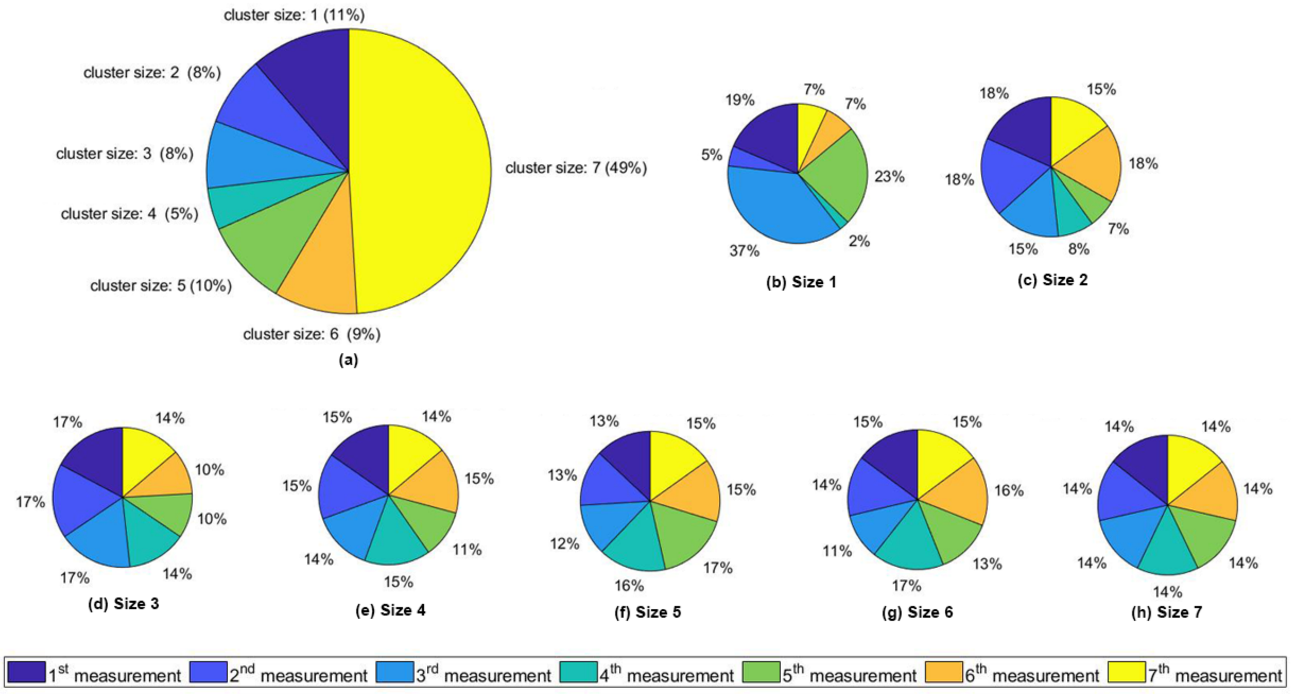

Fig. 7 (a) The distribution of cluster sizes along the overall data (b)–(h) The distribution of the individual cluster sizes 1–7 along the measurements

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

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 Fig. 8. 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. Fig. 8 is an example for the visualization of chemical changes during the pyrolysis process. Points with the same coordinates denote the same molecules and their colors are applied to mark their concentration in the sample.

4 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.

Acknowledgements

We acknowledge the financial support of Széchenyi 2020 under the GINOP-2.3.2-15-2016-00053. Tamás Varga's contribution to this paper was supported by the Janos Bolyai Research Scholarship of the Hungarian Academy of Sciences.

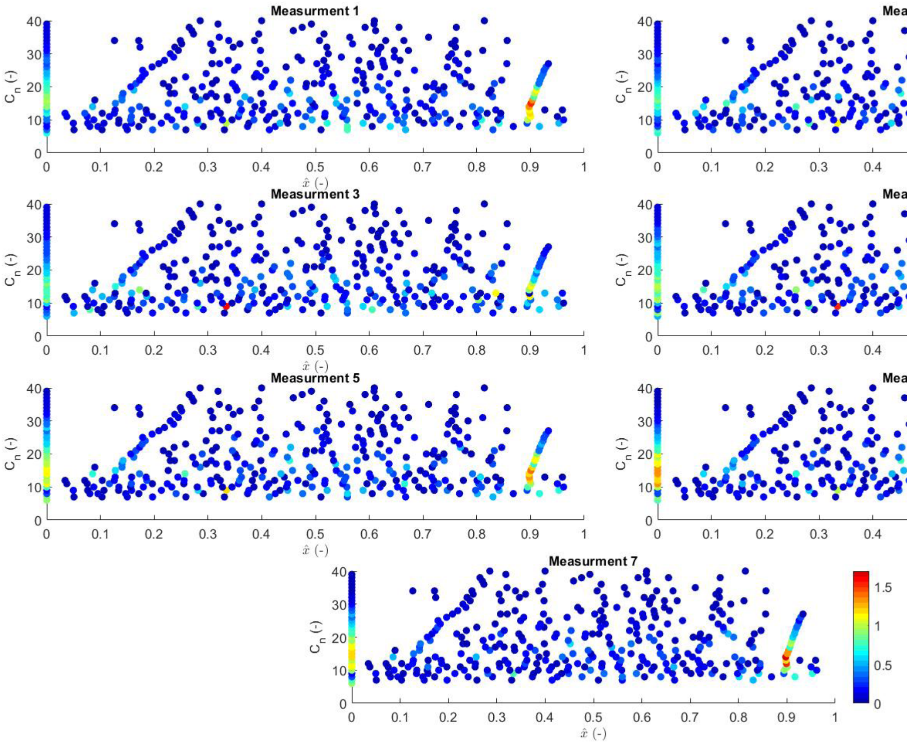

Fig. 8 Visualization of the chemical changes during the pyrolysis process. Points with the same coordinates denote the same molecules and their color is proportional to the concentration

References

[1] Johnson, K. J., Wright, B. W., Jarman, K. H., Synovec, R. E. "High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis", Journal of Chromatography A, 996(1–2), pp. 141–155, 2003. https://doi.org/10.1016/S0021-9673(03)00616-2

[2] Gong, F., Liang, Y.-Z., Fung, Y.-S., Chau, F. T. "Correction of retention time shifts for chromatographic fingerprints of herbal medicines", Journal of Chromatography A, 1029(1–2), pp. 173–183, 2004. https://doi.org/10.1016/j.chroma.2003.12.049

[3] Parastar, H., Jalali-Heravi, M., Tauler, R. "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, 117, pp. 80–91, 2012. https://doi.org/10.1016/j.chemolab.2012.02.003

[4] Zhu, P., Ding, W., Tong, W., Ghosal, A., Alton, K., Chowdhury, S. "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, 23(11), pp. 1563–1572, 2009. https://doi.org/10.1002/rcm.4041

[5] Koh, Y., Pasikanti, K. K., Yap, C. W., Chan, E. C. Y. "Comparative evaluation of software for retention time alignment of gas chromatography/time-of-flight mass spectrometry-based metabonomic data", Journal of Chromatography A, 1217(52), pp. 8308–8316, 2010. https://doi.org/10.1016/j.chroma.2010.10.101

[6] Bloemberg, T. G., Gerretzen, T. G., Lunshof, A., Wehrens, R., Buydens, L. M. C. "Warping methods for spectroscopic and chromatographic signal alignment: A tutorial", Analytica Chimica Acta, 781, pp. 14–32, 2013. https://doi.org/10.1016/j.aca.2013.03.048

[7] Bro, R., Andersson, C. A., Kiers, H. A. L. "PARAFAC2—Part II. Modeling chromatographic data with retention time shifts", Journal of Chemometrics, 13(Special Issue3–4), pp. 295–309, 1999. https://doi.org/10.1002/(SICI)1099-128X(199905/08)13:3/4%3C295::AID-CEM547%3E3.0.CO;2-Y

[8] Robinson, M. D., De Souza, D. P., Keen, W. W., Saunders, E. C., McConville, M. J., Speed, T. P., Likić, V. A. "A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments", BMC Bioinformatics, 8(1), Article number: 419, 2007. https://doi.org/10.1186/1471-2105-8-419

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

[10] Till Z., Varga T., Sója J., Miskolczi N., Chován T. "Structural assessment of lumped reaction networks with correlating parameters", Energy Conversion and Management, 209, Article number: 112632, 2020. https://doi.org/10.1016/j.enconman.2020.112632

[11] Till, Z., Varga, T., Sója, J., Miskolczi, N., Chován, T. "Reduction of lumped reaction networks based on global sensitivity analysis", Chemical Engineering Journal, 375, Article number: 121920, 2019. https://doi.org/10.1016/j.cej.2019.121920

[12] Becker, P. J., Serrand, N., Celse, B., Guillaume, D., Dulot, H. "Comparing hydrocracking models: Continuous lumping vs. single events", Fuel, 165, pp. 306–315, 2016. https://doi.org/10.1016/j.fuel.2015.09.091

[13] Till, Z., Varga, T., Sója, J., Miskolczi, N., Chován, T. "Kinetic identification of plastic waste pyrolysis on zeolite-based catalysts", Energy Conversion and Management, 173, pp. 320–330, 2018. https://doi.org/10.1016/j.enconman.2018.07.088

[14] Ganganath, N., Cheng, C. T., Tse, C. K. "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, 2014, pp. 158–161. https://doi.org/10.1109/CyberC.2014.36

[15] Wagstaff, K., Cardie, C., Seth, R., Schrödl, S. "Constrained K-means Clustering with Background Knowledge", In: Brodley, C. E., Danyluk, A. P. (eds.) ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann Publishers, San Francisco, CA, USA, 2001, pp. 577–584.