The implementation of K-Means clustering in kovats retention index on gas chromatography

T R Noviandy¹, A Maulana¹, N R Sasmita², R Suhendra¹, Muslem³, G M Idroes⁵, M Paristiwati⁶, Z Helwani⁷, E Yandri⁸, S Rahimah⁹, Muhammad¹⁰, Irvanizam¹, R Idroes³,⁴,*

¹Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Kopelma Darussalam, Banda Aceh 23111, Indonesia
²Computational and Applied Statistics Research Group, Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Kopelma Darussalam, Banda Aceh 23111, Indonesia
³Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Kopelma Darussalam, Banda Aceh 23111, Indonesia
⁴Department of Pharmacy, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Kopelma Darussalam, Banda Aceh 23111, Indonesia
⁵Department of Chemical Engineering, Faculty of Engineering, Universitas Syiah Kuala, Kopelma Darussalam, Banda Aceh 23111, Indonesia
⁶Department of Chemistry, Faculty of Mathematics and Sciences, Universitas Negeri Jakarta, Jl. Rawamangun Muka, Jakarta 13220, Indonesia
⁷Department of Chemical Engineering, Universitas Riau, Pekanbaru 28293, Indonesia
⁸Graduate School of Renewable Energy, Darma Persada University, Jl. Radin Inten 2, Pondok Kelapa, East Jakarta 13450, Indonesia
⁹Department of Food Industrial Technology, Faculty of Agroindustrial Technology Universitas Padjadjaran, Jatinangor 54363, Indonesia
¹⁰Faculty of Teacher Training and Education, Abulyatama University, Lampoh Keudee, Aceh Besar 23372, Indonesia

*E-mail: rinaldi.idroes@unsyiah.ac.id

Abstract. In this study, the retention index data of 146 compounds that are found in coal and petroleum-derived liquid fuels were grouped using the K-means clustering method, and the similarities between each cluster were analyzed. The psycho-chemical properties of each compound in the cluster were identified and compared with other clusters. Each compound's retention index is grouped based on the similarity between the column polarity and heating rate of one compound to another. Based on the results of tests carried out on nine different k values, it is known that the grouping with the value of k = 3 is the best determined from the obtained silhouette score = 0.568, where this score is higher than the score obtained on the other k values. The results of clustering with k = 3 obtained three clusters, namely cluster C1, cluster C2, and cluster C3. Cluster C1 and cluster C2 consist of chemical compounds that have a relatively low carbon number and molecular mass, but in cluster C2 the molecular mass of the compound is lower than in cluster C1. In contrast, the C3 cluster consists of chemical compounds that have a relatively high carbon number and molecular mass.
1. Introduction

The Kovats Retention Index was introduced by Ervin Kovats in 1958 as a second parameter that is fixed and resistant to the change of experimental conditions [1]. The Kovats retention index has been extensively developed for various compounds and is applied to gas chromatography [2] and liquid chromatography [3]. The retention index's essence is to obtain the psycho-chemical [4] and thermodynamic information [5] from a chromatographic analysis that the retention time cannot provide. The nature of the retention time that is dependent on changing the analysis conditions causes the retention time to be limited as a marker of separation. The identification of separate compounds must be made using spectroscopic methods as a tandem such as FTIR [6], NMR [7], or MS [8–11].

In general, the retention index is calculated using data of retention time and dead time with an algorithm's support. The dead time value can be determined by measuring the inert gas or homologous series method [12]. In comparison, the retention time is measured from the compound to be determined by its retention index. However, this method significantly depends on the accuracy of the dead time determination because it will affect the accuracy of the retention index calculation [13,14]. The retention index value can also be predicted based on its chemical structure. Idroes et al. have succeeded in predicting the Kovats retention index of flavors and fragrances [15,16] and essential oils [17] compounds using the genetic algorithm (GA) [18] combined with multiple linear regression (MLR).

As a representation of chemical information, retention index values can have a relationship and pattern to these chemical properties. These relationships can be analyzed using statistical methods. Some statistical methods commonly used to analyze the relationship between chemical parameter values and their properties are PCA [19,20] and clustering. Clustering is a method of partitioning the dataset into groups called clusters. Clustering is done by separating data in such a way that points in the same cluster are similar, and points in other clusters are different [21]. The purpose of clustering is to form a compact representation of the data (a subset selection), assessment of the diversity, and detect redundant variables [22]. Husna et al. have grouped the retention index of compounds in liquid fuels derived from coal and petroleum using the hierarchical cluster analysis (HCA) method [23], and Idroes et al. have grouped flavor and fragrance retention index compounds using the biplot method [24].

K-means clustering is a clustering method that looks for cluster centers that represent certain regions of the data. K-means clustering has numerous advantages, such as not requiring a distance matrix as required by hierarchical clustering and having fast computation time[25]. K-means clustering has been applied to the field of chemometrics, such as cluster analysis on Chinese medicine samples and Chinese tea [26] and pattern recognition analysis of chromatographic fingerprints [27].

In this study, the retention index data of 146 compounds that are found in coal and petroleum-derived liquid fuels were grouped using the K-means clustering method, and the similarities between each cluster were analyzed. Coal and petroleum-derived liquid fuels compounds are used because they are difficult to identify using only mass spectra, so this method is combined with retention indices. This combination can be used to identify individual compounds in complex mixtures with high confidence [28]. The purpose of clustering in this study is to classify similar compounds based on their psycho-chemical properties. The psycho-chemical properties of each compound in a cluster were identified and compared with other clusters.

2. Data and methods

2.1. Data collection

The dataset used in this study was the retention index data of 146 compounds that are found in coal and petroleum-derived liquid fuels obtained from Lai & Song literature [28]. The columns used are intermediately polar capillary column coated with 50% phenyl-50% methyl polysiloxane (Rtx-50), and a slightly polar column coated with 5% phenyl-95% methyl polysiloxane (DB-5). The column temperature is programmed linearly at heating rates of 2, 4, and 6 °C min⁻¹.
2.2. Data Clustering

K-means clustering is performed using Orange Data Mining software [29]. To determine the best $k$ value, tests were carried out on nine different values of $k$, ranging from 2 to 10. Grouping for each value of $k$ was done by ten times repetitions and the number of iterations of 300 iterations for each repetition. The best $k$ value will be determined based on the silhouette score obtained. The equation used to calculate the silhouette value for each entity $y_i$ can be seen in equation 1.

$$s(y_i) = \frac{b(y_i) - a(y_i)}{\max\{a(y_i) - b(y_i)\}}$$

(1)

Where $b(y_i)$ is the minimum dissimilarity to all clusters and $a(y_i)$ is the average dissimilarity of $y_i$ to all points in the same cluster [30].

3. Results and discussions

Retention index data of 146 compounds found in coal and petroleum-derived liquid fuels were grouped using K-means clustering based on the similarity between column polarity and each compound’s heating rate. The results of the test conducted on nine different $k$ values are available in Table 1. It can be seen that the $k = 3$ is the best determined from the obtained silhouette score = 0.568, where this score is higher than the scores obtained on the other $k$ values. This means that at $k = 3$, the level of similarity between a compound to another compound in the same cluster and the level of dissimilarity between a compound against other compounds in a different cluster is higher than the other $k$ values.

Table 1. Silhouette score for each $k$.

| $k$ | Silhouette Score |
|-----|------------------|
| 2   | 0.547            |
| 3   | **0.568**        |
| 4   | 0.508            |
| 5   | 0.506            |
| 6   | 0.506            |
| 7   | 0.506            |
| 8   | 0.462            |
| 9   | 0.447            |
| 10  | 0.445            |

The results of clustering with $k = 3$ obtained three clusters, namely cluster C1, cluster C2, and cluster C3. The results of grouping for each compound from 146 compounds found in coal and petroleum-derived liquid fuels can be seen in Table 2. Cluster C1 consists of 64 compounds, cluster C2 consists of 66 compounds, and cluster C3 consists of 16 compounds. The visualization of the clustering results is seen in Figure 1.

Table 2. Results of clustering.

| No | Compound                  | Cluster |
|----|---------------------------|---------|
| 1  | n-Tridecane               | C1      |
| 2  | 1-Tridecene               | C1      |
| 3  | 1,3,5_Triethylbenzene     | C1      |
| 4  | n-Hexylbenzene            | C1      |
| 5  | Bicyclohexyl              | C1      |
| 6  | Naphthalene               | C1      |
| No | Compound                                      | Cluster |
|----|----------------------------------------------|---------|
| 7  | 2,4,6-Trimethylphenol                        | C1      |
| 8  | n-Tetradecane                                | C1      |
| 9  | 1-Tetradecene                                | C1      |
| 10 | Benzothiophene                               | C1      |
| 11 | 2,3-Dihydroindole                            | C1      |
| 12 | Cyclohexylbenzene                            | C1      |
| 13 | 1-Indanol                                    | C1      |
| 14 | n-Octycyclohexane                            | C1      |
| 15 | Quinoline                                    | C1      |
| 16 | n-Pentadecane                                | C1      |
| 17 | 2-Methylnapthalene                           | C1      |
| 18 | 1-Pentadecene                                | C1      |
| 19 | 1-Methylnapthalene                           | C1      |
| 20 | 8-Methylquinoline                            | C1      |
| 21 | 1,2-Dicyclohexylethane                      | C1      |
| 22 | 5,6,7,8-Tetrahydro-3-methylquinoline         | C1      |
| 23 | n-Octylbenzene                               | C1      |
| 24 | n-Hexadecane                                 | C1      |
| 25 | 1,2,3,4-Tetrahydroquinoline                 | C1      |
| 26 | 2,6-Di-tert-butylphenol                     | C1      |
| 27 | 2-Ethynaphthalene                            | C1      |
| 28 | 1-Hexadecane                                 | C1      |
| 29 | 2,7_Dimethynaphthalene                      | C1      |
| 30 | 2,6_Dimethynaphthalene                      | C1      |
| 31 | 1,1'-Biphenyl (diphenyl)                     | C1      |
| 32 | 2-Methylbiphenyl                             | C1      |
| 33 | 1-Ethynaphthalene                            | C1      |
| 34 | 1,3_Dimethynaphthalene                      | C1      |
| 35 | 1,6_Dimethynaphthalene                      | C1      |
| 36 | 2-Isopropynaphthalene                        | C1      |
| 37 | 1,4_Dimethynaphthalene                      | C1      |
| 38 | 1,5_Dimethynaphthalene                      | C1      |
| 39 | 1-Isopropynaphthalene                        | C1      |
| 40 | n-Decylcyclohexane                           | C1      |
| 41 | 1,2_Dimethynaphthalene                      | C1      |
| 42 | n-Heptadecane                                | C1      |
| 43 | 5,6,7,8-Tetrahydro-1-naphthol                | C1      |
| 44 | 3-Methylbiphenyl                             | C1      |
| 45 | 1,8_Dimethynaphthalene                      | C1      |
| 46 | 4-Methylbiphenyl                             | C1      |
| 47 | Acenaphthene                                 | C1      |
| 48 | 1,2_Dihylenylethane (bibenzyl)               | C1      |
| 49 | Dibenzofuran                                 | C1      |
| 50 | n-Decylbenzene                               | C1      |
| 51 | n-Octadecane                                 | C1      |
| 52 | 1-Octadecane                                 | C1      |
| 53 | 3,3'-Dimethylbiphenyl                        | C1      |
| 54 | 2-Naphthol                                   | C1      |
| 55 | 4,4' Dimethylbiphenyl                        | C1      |
| 56 | Fluorene                                     | C1      |
| No  | Compound                                      | Cluster |
|-----|----------------------------------------------|---------|
| 57  | n-Nonadecane                                 | C1      |
| 58  | 2,6-Diisopropynaphthalene                    | C1      |
| 59  | Dibenzyl ether                               | C1      |
| 60  | 9,10-Dihydroanthracene                       | C1      |
| 61  | 9,10-Dihydrophenanthrene                     | C1      |
| 62  | 1,2,3,4,5,6,7,8-Octahydroacridine            | C1      |
| 63  | Phenanthrene                                 | C1      |
| 64  | Anthracene                                   | C1      |
| 65  | n-Pentane                                    | C2      |
| 66  | n-Hexane                                     | C2      |
| 67  | 1-Hexene                                     | C2      |
| 68  | n-Heptane                                    | C2      |
| 69  | Cyclohexane                                  | C2      |
| 70  | 1-Heptene                                    | C2      |
| 71  | Cyclohexene                                  | C2      |
| 72  | Methylcyclohexane                            | C2      |
| 73  | Benzene                                      | C2      |
| 74  | n-Octane                                     | C2      |
| 75  | trans-1,4-Dimethylcyclohexane                | C2      |
| 76  | c&1,3-Dimethylcyclohexane                    | C2      |
| 77  | 1-Octene                                     | C2      |
| 78  | trns-1,2-Dimethylcyclohexane                 | C2      |
| 79  | 1-Methylcyclohexene                          | C2      |
| 80  | cis-1,4-Dimethylcyclohexane                  | C2      |
| 81  | runs-1,3-Dimethylcyclohexane                 | C2      |
| 82  | cis-1,2-Dimethylcyclohexane                  | C2      |
| 83  | Ethylcyclohexane                             | C2      |
| 84  | Toluene                                      | C2      |
| 85  | n-Nonane                                     | C2      |
| 86  | Pyridine                                     | C2      |
| 87  | 1-Nonene                                     | C2      |
| 88  | n-Propylcyclohexane                          | C2      |
| 89  | Ethylbenzene                                 | C2      |
| 90  | 1,4-Dimethylbenzene (p-xylene)               | C2      |
| 91  | 1,3-Dimethylbenzene (m-xylene)               | C2      |
| 92  | n-Decane                                     | C2      |
| 93  | 1,2-Dimethylbenzene (o-xylene)               | C2      |
| 94  | 1-Decene                                     | C2      |
| 95  | trans-Octahydro-lH-indene                   | C2      |
| 96  | tert-Butylcyclohexane                        | C2      |
| 97  | Isopropylbenzene (cumene)                    | C2      |
| 98  | cis-Octahydro-l H-indene                    | C2      |
| 99  | n-Propylbenzene                              | C2      |
| 100 | n-Butylcyclohexane                           | C2      |
| 101 | 1-Ethyl-3-methylbenzene                     | C2      |
| 102 | 1-Ethyl-4-methylbenzene                     | C2      |
| 103 | 1,3,5_Trimethylbenzene (mesitylene)          | C2      |
| 104 | tert-Butylbenzene                            | C2      |
| 105 | 4,5,6,7_Tetrahydroindan                     | C2      |
| 106 | 1-Ethyl-2-methylbenzene                     | C2      |
| No  | Compound                        | Cluster |
|-----|---------------------------------|---------|
| 107 | n-Undecane                      | C2      |
| 108 | 1,2,4-Trimethylbenzene          | C2      |
| 109 | 1-Undecene                      | C2      |
| 110 | set-Butylbenzene                | C2      |
| 111 | trans-Decalin                   | C2      |
| 112 | 1,1′-Bicyclopentyl              | C2      |
| 113 | 1,2,3-Trimethylbenzene          | C2      |
| 114 | Phenyl ethyl ether              | C2      |
| 115 | n-Butylbenzene                  | C2      |
| 116 | n-Pentylcyclohexane             | C2      |
| 117 | cis-Decalin                     | C2      |
| 118 | Indan                           | C2      |
| 119 | n-Dodecane                      | C2      |
| 120 | 1-Dodecene                      | C2      |
| 121 | 2-Methylindan                   | C2      |
| 122 | 1-Methylindan                   | C2      |
| 123 | 1,2,4,5-Tetramethylbenzene      | C2      |
| 124 | n-Hexylcyclohexane              | C2      |
| 125 | 1,4-Diisopropylbenzene          | C2      |
| 126 | 5-Methylindan                   | C2      |
| 127 | 1,2,3,4-Tetramethylbenzene      | C2      |
| 128 | 2,6-Dimethylphenol              | C2      |
| 129 | 4-Methylindan                   | C2      |
| 130 | Tetralin                        | C2      |
| 131 | n-Eicosane                      | C3      |
| 132 | n-Henicosen                     | C3      |
| 133 | 1,2,3,4-Tetrahydrocarbazole     | C3      |
| 134 | n-Docosane                      | C3      |
| 135 | 1-Phenyl-naphthale              | C3      |
| 136 | n-Tricosane                     | C3      |
| 137 | n-Tetracosane                   | C3      |
| 138 | Fluoranthen                     | C3      |
| 139 | n-Pentacosane                   | C3      |
| 140 | Pyrene                          | C3      |
| 141 | 9,10-Dimethylanthracene         | C3      |
| 142 | n-Hexacosane                    | C3      |
| 143 | p-Terphenyl                     | C3      |
| 144 | n-Triacontane                   | C3      |
| 145 | Chrysene                        | C3      |
| 146 | n-Hentriacontane                | C3      |
The results indicate that each cluster has its characteristics that are formed based on their psychophysical properties. There are two chemical parameters that show the different characteristics between the clusters. The C3 cluster consists of chemical compounds that have a relatively high carbon number and molecular mass. Cluster C1 and cluster C2 consist of chemical compounds that have a relatively low carbon number and molecular mass. However, the molecular mass of compounds in cluster C2 is lower than compounds in cluster C1.

4. Conclusion
Based on the results of the clustering of 146 compounds that are found in coal and petroleum-derived liquid fuels with several $k$ ranging from 2 to 10, the best cluster results were obtained at $k = 3$ with a silhouette score of 0.568. From the three divided clusters, it can be seen that cluster C1 and cluster C2 consist of chemical compounds that have a relatively low carbon number and molecular mass, but in cluster C2 the molecular mass of the compound is lower than in cluster C1. In contrast, the C3 cluster consists of chemical compounds that have a relatively high carbon number and molecular mass.

Reference
[1] Kovats von E. 1958 Helv. Chim. Acta., 41 1915–32.
[2] Pratiwi S U T, Lagendijk E L, Weert S de, Idroes R, Hertiani T, Hondel C Van den 2015 Int. J. Appl. Res. Nat. Prod., 8 1–13.
[3] Idroes R. 2009 Indones. J. Pharm., 133–140.
[4] Nowotnik D P, Narra R K. 1993 Liq. Chromatogr., 16 3919–32.
[5] Didaoui L, Touabet A, Badjih Hadj Ahmed A Y, Meklati B Y, Engewald W. 1999 J. High Resolut. Chromatogr., 22 559–564.
[6] Berdeaux O, Fontagné S, Sémond E, Velasco J, Sébédio J L, Dobarganes C. 2012 Chem. Phys. Lipids, 165 338–347.
[7] Albert K. 2014. Handbook of Spectroscopy Second Ed. 1679–1716.
[8] Earlia N, Rahmad R, Amin M, Prakoeswa C, Khairan K, Idroes R. 2019 Sains Malaysiana. 48 1019–24.
[9] Helwani Z, Ramli M, Saputra E, Bahruddin B, Yolanda D, Fatra W, Idroes G M, Muslem M, Mahlia T M I, Idroes R. 2020 Catalysts., 10 (164) 1-13.
[10] Estevam E C, Griffin S, Nasim M J, Zieliński D, Aszyk J, Osowicka M, Dawidska N, Idroes R, Bartoszek A, Jacob C. 2015 Nat. Prod. Commun., 10 1934578X1501001025.
[11] Earlia N, Suhendra R, Amin M, Prakoeswa C R S, Idroes R. 2019 Sci World J., (2019) 1-7.

Figure 1. K-means clustering results visualization.
[12] Idroes R, Muslem, Saiful, Mahmudi, Idroes G M, Suhendra R, Irvanizam, Zamzami M P. 2019 IOP Conf. Series: Mater. Sci. Eng., 364 1-7
[13] Idroes R, Husna I, Muslem, Mahmudi, Rusyana A, Helwani Z, Idroes G M, Suhendra R, Yandri E, Rahimah S, Sasmita N R. 2019 IOP Conf.Series: Earth Environ. Sci., 364 1-7.
[14] Idroes R, Muslem, Mahmudi, Saiful, Idroes G M, Suhendra R, Irvanizam. 2020. Heliyon. 6 1-8.
[15] Idroes R, Noviandy T R, Maulana A, Suhendra R, Sasmita N R, Muslem M, Idroes G M, Irvanizam I. 2019 Int. Rev. Model. Simulations, 12 373-382.
[16] Maulana A, Noviandy T R, Idroes R, Sasmita N R, Suhendra R, Yandri E, Rahimah S, Sasmita N R. 2019 IOP Conf. Series: Earth Environ., 364 1-7.
[17] Noviandy T R, Maulana A, Sasmita N R, Suhendra R, Irvanizam I, Muslem M, Idroes G M, Yusuf M, Sofyan H, Abidin T F, Idroes R. 2020 J. Eng. Sci. Technol., xx-xx
[18] Idroes R, Maulana A, Noviandy T R, Suhendra R, Sasmita N R, Lala A, Irvanizam. 2020. IOP Conf.Series: Earth Environ., 796 012033.
[19] Nadia Y, Ramli M, Muslem, Japnur A F, Rusyana A, Idroes G M, Suhendra R, Muhammad, Sasmita N R, Tallei T E, Idroes R. 2019 IOP Conf.Series: Earth Environ., 364 012040.
[20] Wahidah S, Khaire, Lelifajri, Idroes R, Rahmadi, Lala A, Mahmudi, Muslem, Japnur A F. 2019 IOP Conf.Series: Earth Environ., 364 012016.
[21] Müller A C, Guido S. 2016 Introduction to machine learning with Python: a guide for data scientists. California: O’Reilly Media, Inc.
[22] Drab K, Daszykowski M. 2014 J. AOAC Int 97 29–38.
[23] Husna I, Rusyana A, Muslem, Idroes G M, Suhendraand R, Idroes R. 2020 IOP Conf. Series: Mater. Sci. Eng., 523 012007.
[24] Kundu M, Kundu P K, Damarla S K. 2017 Chemometric Monitoring: Product Quality Assessment, Process Fault Detection, and Applications. Florida: CRC Press.
[25] Aliakbarzadeh G, Sereshti H, Parastar H. 2016. Anal. Bioanal. Chem., 408 3295–3307.
[26] Lai W. 1995. Fuel, 74 1436–51.
[27] Demšar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Možina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Žagar L, Žbontar J, Žitnik M, Zupan B. 2013. J. Mach. Learn. Res., 14 2349–2353.
[28] de Amorim R C, Hennig C. 2015. Inf. Sci., 324 126–145.