DATA ANALYTICS AND APPLIED MATHEMATICS
E-ISSN: 2773-4854
VOLUME 4, ISSUE 1, 2023, 1 – 7
DOI: https://doi.org/10.15282/daam.v4i1.9171

RESEARCH ARTICLE

Customer analytics for online retailers using weighted k-means and RFM analysis

A. Serwah[1], K.W. Khaw[1]*, S.P.Y. Cheang[1] and A. Alnoor[2]

[1]School of Management, Universiti Sains Malaysia, 11800 USM Penang, Malaysia
[2]Management Technical College, Southern Technical University, Basrah, Iraq

ABSTRACT - In recent years, there has been a significant trend toward data-driven enterprises in the business world. This trend is exemplified by the frustration reported by 74% of customers when they encounter ads that are not relevant to them, as reported by Infosys. This emphasizes the importance of personalization in marketing efforts. In order to effectively personalize marketing efforts, businesses often track and analyze the actions of consumers when they interact with websites or click on ads. However, creating completely personalized content for every individual is not practical due to the vast number of people and limited resources and time. In this study, a new approach has been used to segment customers based on the combination of RFM analysis and weighted k-means clustering to help an online retailer better target its customers. The results with weighted k-means are significantly higher with a silhouette score of 0.40 compared to 0.30 of the traditional k-means.

ARTICLE HISTORY
Received : 10th Mar. 2023
Revised : 03rd April 2023
Accepted : 19th April 2023
Published : 30th April 2023

KEYWORDS
RFM analysis,
Weighted k-means,
k-means,
Customer segmentation,
E-commerce,

1.0 INTRODUCTION

Marketing segmentation is based on the idea that people are different and therefore require personalized marketing strategies. "Buying personal" is a crucial element of marketing strategy, and with the rise of e-commerce platforms and online data, there has been a new era of digital existence that allows for a new way of interactions between the customers and the retailers, however, it is not practical to define every single person of interest to a potential client individually, so a more effective approach is to group people based on common characteristics such as demographics or consumption attributes [1-3]. Customer segmentation can help enterprises provide personalized products and services to different consumer groups [4-6], which can ultimately lead to increased customer satisfaction and profitability [7-9].

In recent years, a lot of studies have implemented one of the Machine Learning (ML) algorithms, i.e. k-means clustering to segment customers associated with RFM (recency, frequency, and monetary value) [10-15] which is usually used for creating what is called a “loyalty program” to segment customers into groups based on the score of their purchasing behavior [16]. with the help of ML to automate the process of assigning the customer into these groups, k-means was implemented however the k-means is very sensitive to outliers [10], [17] which may lead to bias in the results. a new variant of the k-means clustering algorithm is called ‘weighted k-means” which is used to assign different weights for each data point [18-19]. This can be useful in situations where certain data points are more important or relevant than others, or where the distance between data points should be weighted differently [18].

Customer segmentation is a commonly used technique in the retail industry, as it allows companies to tailor their marketing efforts to specific groups of customers and better address their needs and preferences [20]. By better understanding their customers, retailers can improve customer loyalty, retention, and overall profitability [21]. utilizing weighted k-means clustering to identify customer segments and assess the performance of this approach in comparison to traditional k-means clustering. The results of this study have significant implications for the online retail industry and provide valuable insights for the effective implementation of customer segmentation strategies. By utilizing these techniques, retailers can gain a deeper understanding of their customers, create more personalized marketing campaigns, and improve customer engagement, retention, and sales [13], [21].

In this paper, a study has been conducted on the customer segmentation efforts of an online retailer. The purpose of this study is to better understand the characteristics, behaviors, and needs of the retailer's customers to identify meaningful segments and develop targeted marketing and engagement strategies Given the intense competition and rapidly evolving market faced by online retailers in London, it is crucial for them to adapt to changing customer preferences and demands continuously. Effective customer segmentation, which involves dividing customers into smaller groups based on shared characteristics, is one way to achieve this. Therefore, this study aims to showcase the effectiveness of using RFM analysis and weighted K-means in customer segmentation for online retailers in the United Kingdom to better aligned with customer needs and make a personalized experience for them.

The dataset utilized in this study was obtained from the UCI Machine Learning Repository Centre for Machine Learning and Intelligent Systems and pertains to online transactions from an online retail company based in London, United Kingdom. Specifically, the dataset consists of records of online transactions occurring between December 1st,

***CORRESPONDING AUTHOR | K.W. Khaw** | ✉ khaiwah@usm.my

2010, and December 9th, 2011. The dataset is comprised of 8 distinct features and contains a total of 541,909 records. It is worth noting that this dataset has been publicly available since 2015.

Table 1. Data with description

Name of the features	Description
Invoice	Invoice number. Nominal. A 6-digit integral number is uniquely assigned to each transaction. If this code starts with the letter 'c,' it indicates a cancellation.
StockCode	A 5-digit integral number is uniquely assigned to each distinct product.
Description	Description of every project.
Quantity	The quantities of each product (item) per transaction.
InvoiceDate	Invoice date and time. The day and time when a transaction was generated.
UnitPrice	Product price per unit in sterling (Â£).
CustomerID	Customer number. A 5-digit integral number is uniquely assigned to each customer.
Country	Customer number. A 5-digit integral number is uniquely assigned to each customer.

2.0 METHODOLOGY

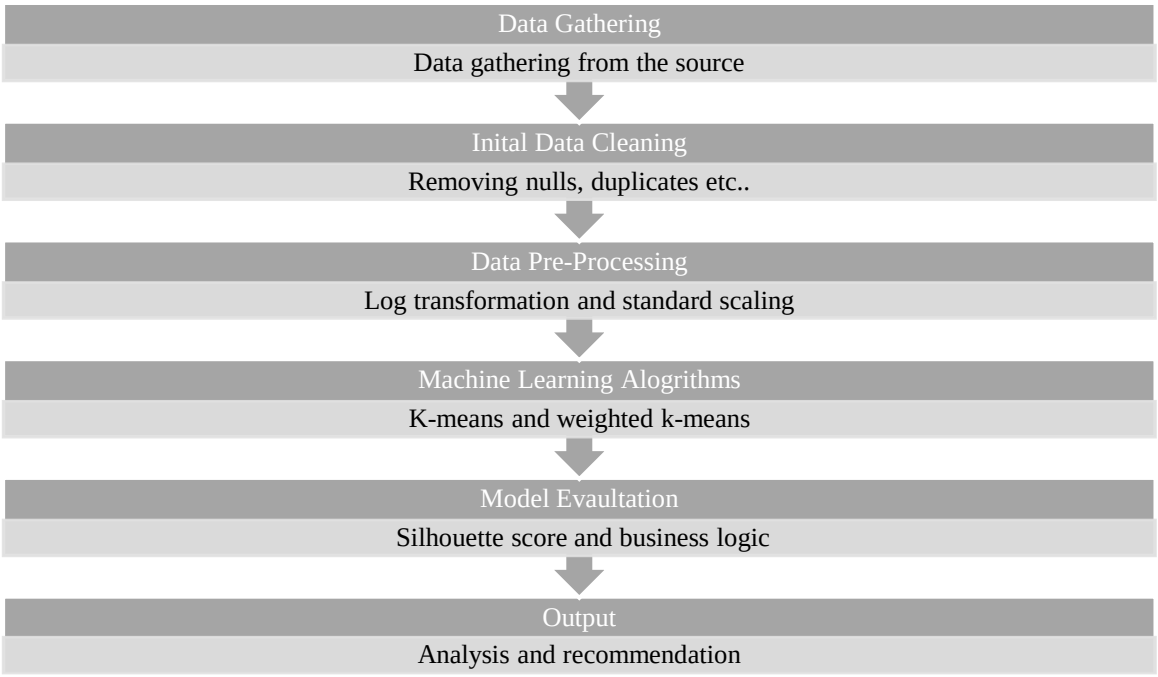

Figure 1. Model workflow

The sequence diagram represents the steps involved in analyzing and interpreting data. The process begins with the input of raw data, which is cleaned and preprocessed using techniques such as log transformation and standard scaling. The preprocessed data is then fed into a machine-learning model, consisting of K-means clustering and weighted K-means, which groups the data into clusters based on similarity. The model is evaluated using the silhouette score and business logic, and the output is generated based on the results of this evaluation. This output could be a set of recommendations, a prediction, or some other form of analysis.

2.1 Data cleaning

The 'invoice' column has more than 16% canceled orders which can bias the final results can be due to several reasons one of them being the orders can be canceled due to the weather conditions or the shipping process hence the choice of dropping the null values has been implemented

2.2 Data pre-processing

After cleaning the data, the next major workflow is RFM analysis, to conduct the RFM analysis further data manipulation is needed to be performed.

RFM analysis is the process of dividing customers by calculating three metrics

- Recency: How recently did the customer make a purchase? Customers who have made a purchase recently are more valuable than those who made a purchase a long time ago.
- Frequency: How often does the customer make a purchase? Customers who make more frequent purchases are more valuable than those who make infrequent purchases.
- Monetary value: How much does the customer spend per purchase? Customers who spend more per purchase are more valuable than those who spend less [15], [22].

CustomerID	Recency	Freq	Montearyvalue	R	F	M	score	segment	tier
12346	326	1	77183.60	1	1	1	3	111	bronze
12347	2	182	4310.00	4	4	4	12	444	gold
12348	75	31	1797.24	2	2	2	6	222	silver
12349	19	73	1757.55	3	3	3	9	333	gold
12350	310	17	334.40	1	1	1	3	111	bronze

Figure 2. RFM with tier segment

2.3 Data Pre-processing for RFM clustered normally-based model

After the initial data cleaning, weighted k-means have a few assumptions that must be met to significant results the most important assumption is the data to in a normal distribution form "mean of 1 and a standard deviation is 0" "Monetary value" is influenced by the outliers so this provokes the assumption.

Data scaling using the Sklearn standard scaler can help to ensure that features are on the same scale and that there is no bias introduced by one feature having a much larger scale than the others [8].

$$np \log() \tag{1}$$

The natural logarithm is the base-e logarithm, where e is the mathematical constant approximately equal to 2.71828.

2.4 RFM clustered-based model

The objective of this study is to utilize the K-means clustering algorithm on a pre-processed dataset to determine if the data can be meaningfully divided into distinct segments based on recency, frequency, and monetary values. The K-means algorithm operates by calculating the Euclidean distance between points in the dataset.

$$d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]} \tag{2}$$

Table 2. K-means clustering steps [5], [11], [22]

K-means clustering steps	
1.	Choose the number of clusters, k, that you want to generate.
2.	Select k random points as the initial centroids for each cluster.
3.	Assign each data point to the closest centroid, forming k clusters.
4.	Calculate the mean of the points in each cluster and use it as the new centroid for that cluster.
5.	Reassign each data point to the closest new centroid.
6.	Repeat Steps 4 and 5 until the centroids do not change or a predetermined number of iterations is reached.
7.	The final clusters are the clusters that are formed after the centroids stop changing.

weighted Euclidean distance measure used in the weighted k-means algorithm.

$$D(x, c) = \sqrt{[w_1(x_1 - c_1)^2 + w_2(x_2 - c_2)^2 + \cdots + w_n(x_n - c_n)^2]} \tag{3}$$

Table 3. Weighted k-means clustering steps [18-19]

Weighted k-means clustering steps	
1.	Choose the number of clusters K that is optimal for the business
2.	Select k random points as the initial centroids for each cluster.
3.	Assign each data point to the closest centroid, forming k clusters.
4.	Calculate the mean of the points in each cluster and use it as the new centroid for that cluster.
5.	Calculate the weights for each data point based on its distance from the centroid. The weight for each data point is inversely proportional to the distance from the centroid.
6.	Reassign each data point to the closest new centroid using the weights calculated in Step 5.
7.	Repeat Steps 4-6 until the centroids do not change or a predetermined number of iterations is reached.
8.	The final clusters are the clusters that are formed after the centroids stop changing.

3.0 RESULTS AND DISCUSSION

The results of the comparison between weighted k-means and traditional k-means indicate that weighted k-means perform better in terms of silhouette score. This measure of cluster quality reflects the degree of coherence and compactness within each cluster, and a higher silhouette score suggests a higher level of similarity among the points within each cluster. In this study, the frequency of purchases was identified as the most important factor in determining the clusters, with customers who bought more frequently being assigned higher weights. This suggests that the level of loyalty to the online retailer can be effectively measured by the frequency of purchases, as customers who buy more often may be considered more loyal. These findings have significant implications for customer segmentation and loyalty analysis in the online retail industry, as they suggest that targeting customers based on their frequency of purchases may be a more effective strategy for improving customer retention and loyalty.

Based on the RFM analysis and applying the weighted k-means clustering the customers were divided into three clusters cluster 0 which represents the customer with medium frequency and monetary value to the and recency on average, cluster 1 which represents the loyal customers class 2 represents the customers that at risk for churning.

segment	Recency	Freq	Montearyvalue	score	N OF OBSERVATION	% of observation
0	102.436470	38.797980	626.360358	6.856991	1881	0.433610
1	39.271523	206.315894	4958.299226	10.504636	1510	0.348087
2	157.803590	8.633580	234.417350	3.939810	947	0.218303

Figure 3. RFM cluster-based method

With 40% cluster 0 which represents the customer's medium value, the goal for the retailer is to convert them to be in cluster 1 which represents the "loyal customers" that can be done with personalized offers by analyzing their purchasing behavior.

The comparison between traditional k-means and weighted k-means in this study showed that weighted k-means had a higher silhouette score of 0.40 compared to 0.3 for traditional k-means. This suggests that the clusters formed by weighted k-means are more coherent and compact, indicating a higher level of similarity within each cluster. It is worth noting that the performance of traditional k-means may be affected by the presence of outliers, as these points can significantly influence the position of the centroids and the formation of the clusters. In contrast, the use of weights in weighted k-means allows for the influence of these points to be mitigated, resulting in more robust and stable clusters. The decision to assign higher weights to the recency of buying in this study was based on the business logic that customers who buy recently are more valuable to the retailer as the products and the services remain in their minds. The improved performance of weighted k-means in this study supports this assumption and suggests that the use of weighted k-means may be a more effective approach for customer segmentation and loyalty analysis in online retailers and the e-commerce industry.

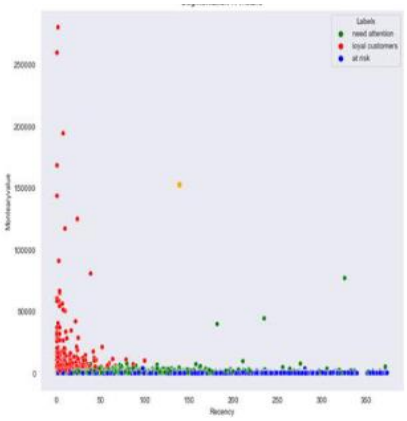

Figure 4. Traditional k-means

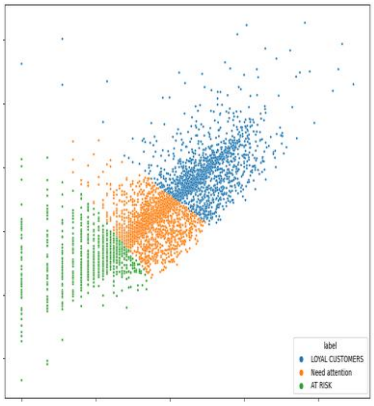

Figure 5. Weighted k-means

4.0 CONCLUSION

Customer segmentation can be a very challenging task, as it involves identifying and understanding the different characteristics, behaviors, and needs of customers to group them into meaningful segments. This process requires careful analysis of customer data and the use of appropriate techniques and tools to accurately identify and differentiate between different segments.

Customers can be divided into three categories based on the results of RFM analysis: "Need attention," "Loyal customers," and "At risk." The "Need attention" group is more likely to leave the retailer and search for alternatives. The "Loyal customers" group is highly loyal to the retailer, characterized by more frequent purchases and more recent transactions. "Fewer days" between transactions indicate a higher value for the retailer. The goal for the retailer is to convert the "Need attention" group into "Loyal customers" and prevent them from becoming "At risk." The objective is to retain the loyalty of these customers through targeted marketing efforts.

Two ML algorithms are used in this study k-means and weighted k-means. K-means clustering is a popular unsupervised machine-learning technique used to group similar data points into clusters. In the context of customer segmentation, k-means clustering can be used to group customers with similar purchase behaviors into different segments. The technique works by iteratively assigning data points to clusters based on their similarity in a given feature space and adjusting the cluster centroids until convergence is reached.

Weighted k-means clustering is a variation of the k-means clustering technique that assigns different weights to each variable in the clustering analysis. In the context of customer segmentation, weighted k-means clustering can be used to assign different weights to the three RFM factors based on the business's priorities and goals. By assigning different weights, businesses can adjust the clustering analysis to focus on the factors that are most relevant to their business goalsthe most significant factor considered in this study and utilized in the weighted k-means model is the frequency of customer purchases.

Based on the comparison between weighted RFM and traditional k-means, it can be concluded that weighted RFM performs better with a silhouette score of 0.40 compared to the silhouette score of 0.3 for k-means. This suggests that the clusters formed by weighted RFM are more coherent and compact, leading to a higher silhouette score. Therefore, weighted RFM can be considered a more effective clustering method in this particular case due to that the frequency of purchases is considered the most significant factor in the weighted k-means model as it provides a reliable indicator of customer loyalty. Customers who make frequent purchases are more likely to be loyal to the retailer, and as such, are more valuable to the business. Therefore, by clustering customers based on their purchase frequency, the retailer can identify and target those who are at risk of leaving and focus on retaining their loyalty through targeted marketing campaigns.

The study faces some limitations, primarily stemming from the presence of outliers in the data collected. Specifically, certain transactions in the dataset exhibit anomalous values for the monetary values, which may have negatively impacted the performance of the RFM model and clustering algorithm. Additionally, the dataset used in this study is limited to sales data from the United Kingdom, which may not be representative of other countries due to differences in lifestyle and consumer preferences. Furthermore, the absence of certain key demographic variables, such as customer age, may have influenced the clustering approach. it is important to note that the most significant factor considered in this study and utilized in the weighted k-means model is the frequency of customer purchases, although other businesses may choose to prioritize different factors, it is worth noting that the primary objective of this study is to explore and identify an optimal model for forecasting, and as such, it is classified as an exploratory study, rather than a confirmatory study.

6.0 ACKNOWLEDGMENT

Special thanks were extended to Eng. Aliaa Zaki for the invaluable support and guidance.

7.0 REFERENCES

[1] Y. Li, X. Chu, D. Tian, J. Feng and W. Mu, "Customer segmentation using k-means clustering and the adaptive particle swarm optimization algorithm," Applied Soft Computing, vol. 113, no. Part B, p. 107924, 2021. https://doi.org/10.1016/j.asoc.2021.107924

[2] B. Turkmen, "Customer segmentation with machine learning for online retail industry," The European Journal of Social & Behavioural Sciences, vol. 31, no. 2, pp. 111-136, 2022. https://doi.org/10.15405/ejsbs.316

[3] R. Shirole, L. Salokhe and S. Jadhav, "Customer segmentation using RFM model and k-means clustering," International Journal of Scientific Research in Science and Technology, vol. 8, no. 3, pp. 591-597, 2021. https://doi.org/10.32628/IJSRST2183118

[4] F. Yoseph and M. Heikkila, "Segmenting retail customers with an enhanced RFM and a hybrid regression/clustering method," in Proceedings of 2018 International Conference on Machine Learning and Data Engineering, Australia, 2019, pp. 108-116. https://doi.org/10.1109/iCMLDE.2018.00029

[5] Dedi, M.I. Dzulhaq, K.W. Sari, S. Ramdhan, R. Tullah and Sutarman, "Customer segmentation based on RFM value using k-means algorithm," in Proceedings of 2019 Fourth International Conference on Informatics and Computing, Indonesia, 2020, pp. 1-7. https://doi.org/10.1109/ICIC47613.2019.8985726

[6] B. Rizki, N.G. Ginasta, M.A. Tamrin and A. Rahman, "Customer loyality segmentation on point of sale system using recency-frequency-monetary (RFM) and k-means," Jurnal Online Informatika, vol. 5, no. 2, pp. 130-136, 2020. https://doi.org/10.15575/join.v5i2.511

[7] D. Kamthania, A. Pahwa and S.S. Madhavan, "Market segmentation analysis and visualization using k-mode clustering algorithm for e-commerce business," Journal of Computing and Information Technology, vol. 26, no. 1, pp. 57-68, 2018. https://doi.org/10.20532/cit.2018.1003863

[8] M. Dhamecha, *Advances in Intelligent Systems and Computing*. Singapore: Springer, 2021, pp. 61-69. https://doi.org/10.1007/978-981-15-9516-5_5

[9] A.J. Christy, A. Umamakeswari, L. Priyatharsini and A. Neyaa, "RFM ranking–an effective approach to customer segmentation," Journal of King Saud University-Computer and Information Sciences, vol. 33, no. 10, pp. 1251-1257, 2021. https://doi.org/10.1016/j.jksuci.2018.09.004

[10] K. Tabianan, S. Velu, and V. Ravi, "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data," Sustainability, vol. 14, no. 12, p. 7243, 2022. https://doi.org/10.3390/su14127243

[11] R. Gustriansyah, N. Suhandi and F. Antony, "Clustering optimization in RFM analysis based on k-means," Indonesian Journal of Electrical Engineering and Computer Science, vol. 18, no. 1, pp. 470-477, 2020. https://doi.org/10.11591/ijeecs.v18.i1.pp470-477

[12] M. Aryuni, E.D. Madyatmadja and E. Miranda, E, "Customer segmentation in XYZ bank using k-means and k-medoids clustering, in Proceedings of 2018 International Conference on Information Management and Technology, Indonesia, 2018, pp. 412-416. https://doi.org/10.1109/ICIMTech.2018.8528086

[13] I. Maryani, D. Riana, R.D. Astuti, A. Ishaq, Sutrisno, E.A. Pratama, "Customer segmentation based on RFM model and clustering techniques with k-means algorithm, in Proceedings of 2018 Third International Conference on Informatics and Computing, Indonesia, 2019, pp. 1-6. https://doi.org/10.1109/IAC.2018.8780570

[14] R.W.S. Brahmana, F.A. Mohammed and K. Chairuang, "Customer segmentation based on RFM model using k-means, k-medoids, and DBSCAN methods," Lontar Komputer: Journal Ilmiah Teknologi Informasi, vol. 11, no. 1, pp. 32-43, 2020. https://doi.org/10.24843/LKJITI.2020.v11.i01.p04

[15] Y.-L. Chen, M.-H. Kuo, S.-Y. Wu and K. Tang, "Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data," Electronic Commerce Research and Applications, vol. 8, no. 5, pp. 241-251, 2009. https://doi.org/10.1016/j.elerap.2009.03.002

[16] J. Wu, L. Shi, W.-P. Lin, S.-B. Tsai, Y. Li, L. Yang and G. Xu, "An empirical study on customer segmentation by purchase behaviors using an RFM model and k-means algorithm," Mathematical Problems in Engineering, vol. 2020, p. 8884227, 2020. https://doi.org/10.1155/2020/8884227

[17] S. Gupta, R. Kumar, K. Lu, B. Moseley, S. Vassilvitskii, "Local search methods for k-means with outliers," Proceedings of the VLDB Endowment, vol. 10, no. 7, pp. 757-768, 2017. https://doi.org/10.14778/3067421.3067425

[18] K. Kerdprasop, N. Kerdprasop, and P. Sattayatham, *Data Warehousing and Knowledge Discovery*. Berlin, German: Springer, 2005, pp. 488-497. https://doi.org/10.1007/11546849_48

[19] J.Z. Huang, M.K. Ng, H. Rong and Z. Li, "Automated variable weighting in k-means type clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 5, pp. 657-668, 2005. https://doi.org/10.1109/TPAMI.2005.95

[20] D. Chen, S.L. Sain and K. Guo, "Data mining for the online retail industry: a case study of RFM model-based customer segmentation using data mining," Journal of Database Marketing & Customer Strategy Management, vol. 19, pp. 197-208, 2012. https://doi.org/10.1057/dbm.2012.17

[21] A. Sheshasaayee and L. Logeshwari, "Implementation of clustering technique based RFM analysis for customer behaviour in online transactions," in Proceedings of 2018 2nd International Conference on Trends in Electronics and Informatics, India, 2018, pp. 1166-1170. https://doi.org/10.1109/ICOEI.2018.8553873

[22] Y.-M. Cheung, "K*-means: a new generalized k-means clustering algorithm," Pattern Recognition Letters, vol. 24, no. 15, pp. 2883-2893, 2003. https://doi.org/10.1016/S0167-8655(03)00146-6