Mathematics Letters, 2 (2): 75-88, 2023
eISSN: 2948-3735

META ANALYSIS ON THE IMPROVISATION OF FUZZY C-LEAST MEDIAN USING CANBERRA DISTANCE IN MULTIPLE LINEAR REGRESSION TOWARDS MALAYSIAN HOUSEHOLD INCOME

Anis Nelissa Abdul Hamid[1], **Shahirulliza Shamsul Ambia**[2*] **and Sumarni Abu Bakar**[3]

College of Computing, Informatics and Mathematics, Universiti Teknologi MARA (UiTM), Malaysia

[1]anisnelissa854@gmail.com, [2*]sliza@ tmsk.uitm.edu.my, [3]sumarni164@uitm.edu.my

(indicates the corresponding author*

ABSTRACT

Predicting the degree of ambiguity in the data is the main challenge in data clustering. Therefore, the objective of this literature review and meta-analysis is to summarise the existing research to highlight the importance of distance metrics in clustering tools. A lot of improvements have been done to both hard clustering and soft clustering. Recently, these clustering have been integrated with the Multiple Linear regression (MLR) model to achieve better performance. Data retrieval for this systematic review was carried out on May 7, 2023, using Scopus, Web of Science, Google Scholar, and ScienceDirect in compliance with PRISMA criteria. The inclusion and exclusion criteria for this analysis resulted in the inclusion of a total of 17 papers. The majority of the included research found that distance metrics do play an important role in clustering and Canberra metrics surpass Euclidean metrics capabilities while Fuzzy C-Least Median (FCLM) appears to be a more robust and accurate model than Fuzzy C-Means (FCM) clustering. This literature review and meta-analysis also found that demographic factors had a significant effect on the household income distribution in Malaysia.

Keywords*: Canberra, Clustering, Fuzzy C-Least Median, Household Income, Meta-Analysis, Multiple Linear Regression*

Received for review: 09-09-2023; Accepted: 22-10-2023; Published: 06-11-2023

1. Introduction

Reviewing, analysing, and classifying the currently available literature on distance metrics, clustering and regression models applied to Malaysia's household income are the ultimate objectives of this paper. This literature review and meta-analysis examined the research trends, including the purpose of study, research methodologies, significant data used and the theoretical and conceptual frameworks for the research.

Table 1: Themes of this study

No	Themes of this study
1	Purpose of the study
2	Research methodology
3	Significant data

Clustering aids in finding patterns in data and is helpful for picture segmentation, anomaly detection, pattern identification, and exploratory data analysis. The Fuzzy C-Means (FCM) clustering algorithm is one of the most widely used clustering techniques which uses Euclidean distance metrics as a similarity measurement (Arora et al., 2018). Furthermore, distance function plays a crucial part in clustering techniques. The clusters are formed according to the distance between data points and the cluster centres are formed for each cluster. The distance measure between the cluster centre and the data to be clustered can be used using several methods, such as Manhattan, Euclidean, Cosine, Mahalanabis, and Hamming etc. However, further research by Setyawan and Ilham (2019) showed that FCM which uses the Euclidean distance, usually causes clustering errors, especially when processing multidimensional and noisy data. In addition, the complexity of regression studies can increase to the point where no present algorithms can predict values with any degree of reliability. In accordance with Pai (2021), Multiple Linear Regression (MLR) is insufficient for managing complex relationships. Therefore, clustering and regression can work together in some cases to create a highly accurate model. Clustering the data before running the MLR model, can successfully reduce the variability in household income due to income gap. In addition, income inequality in Malaysia has increased in recent years, especially in the aftermath of the pandemic (Zainal, 2023). High income inequality can be harmful to a country's economic progress, making it an especially pressing concern for developing nations like Malaysia (De Dominicis et al. 2008; Qin et al. 2009).

2. Literature Review

A research by Kaushik (2023) clarifies that clustering aids in finding patterns in data and is helpful for picture segmentation, anomaly detection, pattern identification, and exploratory data analysis. Rashid et al. (2018) continue further to indicate that the generated object clusters should have high levels of internal homogeneity (within-cluster) and outward heterogeneity (between-cluster). As a result, it is an effective tool for comprehending data and can help to disclose insights that may not be obvious through other analytical techniques, in accordance with Kaushik (2023). K-Means (KM) clustering was discovered to be an unsupervised learning technique by Banoula (2023). KM technique, which clusters the data before running the MLR model, appears to successfully reduce the variability in household income due to income gap.

However, a study by Yee et al. (2023) recommends that future researchers try implementing FCM clustering into MLR to get the best model with the lowest error rates. A study by Kumar (2022) further explained FCM clustering is a popular soft clustering algorithm that outperforms KM clustering on overlapping data sets. In contrast to the KM algorithm, in which data points can only belong to one cluster, data points in the FCM algorithm can have a higher likelihood of belonging to more than one cluster. According to Wahab et al. (2018), the hybrid MLR model and FCM method outperform the MLR model toward the paddy yield data in Malaysia. Therefore, it showed that clustering can indeed be helpful for supervised machine-learning tasks. However, Mallik et al. (2021) recently introduced an extension of FCM called Fuzzy C-Least Median (FCLM) of square algorithms by enhancing the crisp clustering technique and considered the least median instead of mean value on selecting the cluster centre. This study also discovered that by removing the impact of outliers and lowering the error, FCLM successfully enhanced the FCM technique.

Arora et al. (2018) further explained that similarity is an important component of clustering algorithms computed based on the distance measure between the cluster centre and the data to be clustered. From this literature review, this study found that when comparing Euclidean distance with the other distance without Manhattan metrics, Euclidean is still included as one of the best results (Arora et al., 2018). However, Panda et al. (2012) and Rusdiana et al. (2021) found that when Manhattan distance is included in the comparisons with the other distance metrics, Manhattan is found to be more effective than Euclidean distance in

clustering. Not just that, when Canberra distance is included with the other distance metrics alongside Manhattan and Euclidean, it is found that Canberra surpasses both Euclidean and Manhattan distance (Vélez-Falconí et al., 2020; Dik et al., 2014). Therefore, Canberra distance would be selected to be implemented in the FCLM model. The information through statistical analysis is important to top management in decision making to optimise the economic situation. Furthermore, it is crucial that the data be modelled using an appropriate technique to minimise the error rates arising from the income gap so that the most significant factors reflecting the household income can be identified.

There are numerous existing recommendations for literature reviews, claims Snyder (2019). All sorts can be beneficial and appropriate to achieve a certain aim, depending on the approach required to meet the purpose of the review. According to a study by Gough, Oliver, and Thomas (2017), a literary review is a sort of review that compiles several research studies and summarises them to provide a thorough analysis of a research subject. The researchers also list the primary tasks of a literary review as finding pertinent research, systematically evaluating research reports, synthesising results, and comprehending research findings. In accordance with Moher (2009), bias can be reduced by utilising explicit and systematic techniques when analysing articles and all relevant data, resulting in accurate findings from which judgements can be formed.

3. Methodology/Implementation

Figure 1 illustrates how the studies were chosen in accordance with Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA). PRISMA, or publication standard, aids readers and reviewers in the logical steps of searching for pertinent research articles and directs writers on how to summarise the review process of chosen publications (Vu-Ngoc et al., 2018). The systematic literature evaluation based on publication standards was initiated by coming up with a good research question. The pertinent journal databases that will be utilised are then finalised. Finally, this study describes the systematic strategy used to pick pertinent research papers. This strategy included three main steps: identification, screening, and eligibility.

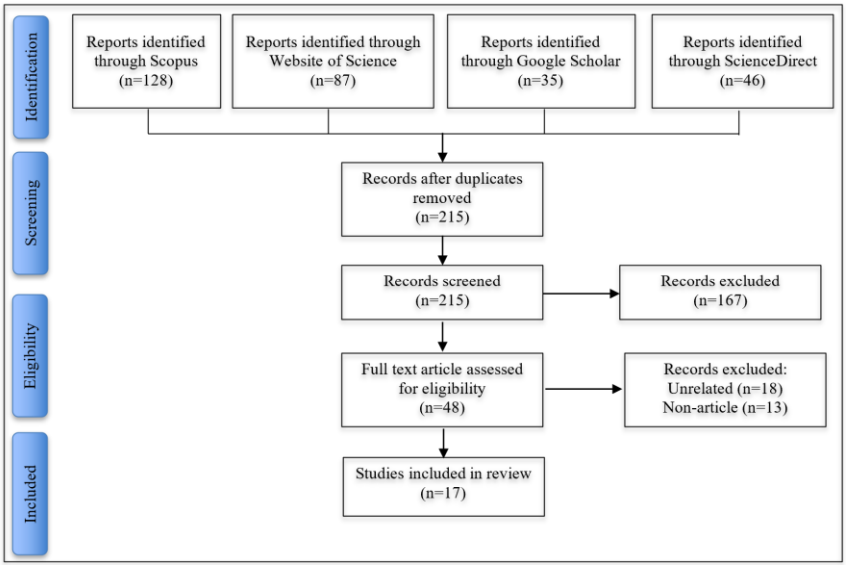

Figure 1: PRISMA Flow Diagram for Selection Process

From the literature review, Table 2 shows the data collected such as the title of the articles, year published, the author's name and the journal name for the articles included in this study.

Table 2: Identity of the articles to be analysed further.

No	Article Title	Year	Author's Name	Journal Name
1	Statistical analysis of household income data in Perak, Malaysia	2023	Revathi Ganesan, Nurulkamal Masseran	AIP Conference Proceedings
2	K-means clustering analysis and multiple linear regression model on household income in Malaysia	2023	Gan Pei Yee, Mohd Saifullah Rusiman, Shuhaida Ismail, Suparman, Firdaus Mohamad Hamzah, Muhammad Ammar Shafi	IAES International Journal of Artificial Intelligence (IJ-AI)
3	A clustering approach to identify multidimensional poverty indicators for the bottom 40 percent group	2021	Mariah Abdul Rahman, Nor Samsiah SaniI, Rusnita Hamdan, Zulaiha Ali Othman, Azuraliza Abu Bakar	PLOS (Public Library of Science) ONE
4	Comparison of Distance Metrics on Fuzzy C-Means Algorithm Through Customer Segmentation	2021	Uus Rusdiana, Iin Ernawati, Noor Falih, Artika Arista	International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS
5	An Efficient Fuzzy C-Least Median Clustering Algorithm	2021	Moksud Alam Mallik, Nurul Fariza Zulkurnain, Mohammed Khaja Nizamuddin, Aboosalih KC	IOP Conference Series: Materials Science and Engineering
6	The comparisons of Clustering Algorithms K-Means and Fuzzy C-Means for segmentation retina blood vessel	2020	Wiharto Wiharto, Esti Suryani	Acta Informatica Medica
7	Relationship between B40 Household Income and Demographic Factors in Malaysia	2020	Humaida Banu Samsudin, Amirul Aqil Nadzrulizam	International Journal of Innovative Technology and Exploring Engineering (IJITEE)
8	Comparative Study of Distance Measures for the Fuzzy C-means and K-means Non-Supervised Methods Applied to Image Segmentation	2020	Martín Vélez-Falconía, Josué Marína, Selena Jiméneza, Lorena Guachi-Guachi	ICAIW 2020: Third International Conference on Applied Informatics 2020
9	Prediction of Student's Academic Performance using k-Means Clustering and Multiple Linear Regressions	2019	Oladele Tinuke Omolewa, Aro Taye Oladele, Adegun Adekanmi Adeyinka, Ogundokun Roseline Oluwaseun	Journal of Engineering and Applied Sciences
10	Effect of Clustering Data in Improving Machine Learning Model Accuracy	2019	Samih M.Mostafa, Hirofumi Amano	Journal of Theoretical and Applied Information Technology

11	K-Means vs. Fuzzy C-Means: A Comparative Analysis of Two Popular Clustering Techniques on the Featured Mobile Applications Benchmark	2018	Tuğrul Cabir Hakyemez, Aysun Bozanta, Mustafa Coşkun	IMISC 2018 Conference Proceedings
12	A Technique of Fuzzy C-Mean in Multiple Linear Regression Model toward Paddy Yield	2018	Nur Syazwan Wahab, Mohd Saifullah Rusiman, Mahathir Mohamad, Nur Amira Azmi, Norziha Che Him, M.Ghazali Kamardan, Maselan Ali	IOP Conference Series: Journal of Physics
13	Fuzzy c-Means Clustering Strategies: A Review of Distance Measures	2018	Jyoti Arora, Kiran Khatter, Meena Tushir	Conference Paper Advances in Intelligent Systems and Computing book series
14	Fractional Metrics for Fuzzy c-Means	2014	Amina Dik, Khalid Jebari, Abdelaziz Bouroumi, Aziz Ettouhami	International Journal of Computer and Information Technology
15	Dealing with Distances and Transformations for Fuzzy C-Means Clustering of Compositional Data	2012	Javier Palarea-Albaladejo, Josep Antoni Mart´ın-Fern´andez, Jes´us A. Soto	Journal of Classification
16	Classification of Households with Respect to Poverty by Using Cluster Analysis	2011	Zahoor Ahmad, Zainab Ejaz	Proc. ICCS-11, Lahore, Pakistan
17	Determinants of Income Class in Philippine Households: Evidence from the Family Income and Expenditure Survey 2009	2009	Stephen Jun Villejo, Mark Tristan Enriquez, Michael Joseph Melendres, Dexter Eric Tan, Peter Julian Cayton	The Philippine Statistician

The data collected for meta-analysis on clustering and regression models applied to Malaysia's household income as shown in table 2, analyses 18 articles that have been published from the year 2009 until the year 2023 in reputable international journals and selected based on predetermined criteria. Most of the references come from the journal article, conference paper and proceeding papers. There is also a few references from chapters in books such as the "Classification of Households with Respect to Poverty by Using Cluster Analysis" by Ahmad and Ejaz (2011). The selected articles will be more thoroughly explained in the following section.

4. Data Collection and Analysis

The main aim of this study is to carry out a literature review and meta-analysis of the empirical studies on the extension of the MLR using clustering approach with better distance metrics on the household income distribution in Malaysia. This research used a variety of electronic databases to access data sources. Scopus, Web of Science, Google Scholar and ScienceDirect are the primary online databases that have been used for literature search and data retrieval. One of the most important databases for social science is considered to be Web of Science, followed by Scopus. According to a study by Mongeon & Paul-Hus (2016), Scopus has a wider range of journals covered than Web of Science. In the meantime, as demonstrated by Chadegani et al. (2013), Web of Science generates a greater volume of articles with high impact factors.

The high accessibility rate of articles also encourages the use of Google Scholar and ScienceDirect in addition to these two databases for finding pertinent articles. The selected articles were all released by reputable scientific journals.

There are several keywords used in data search, namely Canberra, clustering, fuzzy c-least median, household income, multiple linear regression, and Malaysia. In addition, this study also used several criteria in analysing data. The systematic searching strategies was conducted according to Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) by the author Vu-Ngoc et al. (2018) where identification, screening and eligibility are the three main processes. The research question for this systematic review was created using PICo as a guideline. PICo is a technique that helps the authors create the best research question.

The three main parts of PICo are population, phenomenon of interest, and context, according to research by Stern et al. (2014). This study have adhered to the review's guidelines based on these three factors namely Malaysia (Population), the distance metrics used in clustering approach (Interest) and the significant demographic factor affecting household income (Context) which then direct this study to develop the main research question – Does distance metrics play an important role in clustering the demographic factors that affect the household income in Malaysia? Table 3 shows the data collected from the articles included in this study.

Table 3: Analysis of the articles based on themes

Author & Year	Purpose of the study	Methodology	Data
Ganesan and Masseran (2023)	To analyse the 2016 household income data in Perak, Malaysia, on the basis of demographic factors and economic activities of income classification groups referred to as B40, M40, and T20.	1) MLR 2) Gini coefficient 3) Ordinal regression	1) Marital status 2) Gender 3) Ethnicity 4) Education level 5) Household size
Samsudin and Nadzrulizam (2020)	This study makes the comparison on the range and average of income amongst the B40 household income of each state in Malaysia.	1) Descriptive analysis 2) MLR	1) Residential area 2) Educations 3) Working status
Ahmad and Ejaz (2011)	To explore the factors which are playing a significant role for the cluster of non-poor household and poor household and also investigate the direction and significance of importance variables	1) Two step cluster analysis	1) Education 2) Household income 3) Dependency ratio 4) family size
Villejo et al. (2014)	To develop a model that can classify households based on their demographic profile and household assets in the Philippines.	1) Clustering Method 2) Multinomial logistics model	1) Education 2) Employment 3) Rural development
Rahman et al. (2021)	This study proposes a B40 clustering-based K-Means with cosine similarity architecture to identify the right indicators and dimensions that will provide data driven MPI measurement	1) B40 clustering-based K-Means with cosine similarity architecture	1) Education 2) Living standard 3) Employment
Omolewa et al. (2019)	To identify factors that lead to a student's success or failure and forecasting student's performance using k-means clustering and MLR	1) K-means clustering 2) MLR	Student's test scores, quiz and assignment

Mostafa et al. (2019)	To perform comparative study of the K-Means Clustering algorithm and popular prediction methods namely Linear Regression, Ridge, Lasso, and Elastic.	K-means clustering: -MLR, Ridge, Lasso, and Elastic Result: All model with clustering approach is better for almost all dataset	1) Iris 2) Diamond 3) Diabetes 4) Admission 5) Boston 6) California
Yee et al. (2023)	To propose the hybrid model where K-means and multiple linear regression (MLR) for clustering and predicting household income in Malaysia.	1) K-means clustering (KM) 2) MLR 3) Mean square error Result: KM in conjunction with the MLR model outperformed the MLR model without clustering	1) Age 2) Gender 3) Marital status 4) Education 5) Certificate 6) Activity 7) Occupation 8) Industry 9) Household Size
Wahab et al. (2018)	To investigate the factors that affect the paddy yield.	1) Fuzzy c-means clustering 2) Multiple linear regression	Variants of the top soil
Wiharto and Suryani (2020)	This study aims to analyse the performance of the algorithms of k-means and FCM for retinal blood vessel segmentation.	1) Fuzzy c-means clustering 2) K-means clustering	Retinal blood vessel segmentation
Hakyemez and Bozanta (2018)	To investigate and compare the efficiency and effectiveness of two separate clustering algorithms, namely k-means and fuzzy c- means in grouping the mobile applications.	1) Fuzzy c-means clustering 2) K-means clustering	Mobile Apps dataset
Arora et al. (2018)	The purpose of the experimental part was to test the operation of the FCM algorithm by applying different distance metrics on synthetic and real dataset.	1) Fuzzy c-means clustering 2) Distance Metrics: Euclidean, Standard Euclidean, Mahalanobis, Standard Mahalanobis, Minkowski, Chebyshev, Best result for all dataset: Euclidean, Minkowski, Chebyshev Distance	1) Synthetic Data 2) Different Volume Rectangle 3) High Dimensional Data: Iris, Wine and Wisconsin dataset.
Palarea-Albaladejo et al. (2012)	To identify which distance measures satisfied main geometric principles of the simplex sample space: scale invariance and subcompositional dominance.	1) Fuzzy c-means clustering 2) Distance Metrics: Euclidean, Angular, Bhattacharyya, C-KL, City Block,Hellinger, J - Divergence,	Nutritional Dataset

		Mahalanobis, Aitchison Distance Result for all dataset: Aitchison and C-KL Distance	
Rusdiana et al. (2021)	To examine the optimal distance metrics method in the clustering algorithm	1) Fuzzy c-means clustering 2) Distance Metrics: Euclidean, Manhattan, Chebyshev,Minkowski Optimal Result: Manhattan, Minkowski	Customer Segmentation
Vélez-Falconí et al. (2020)	To compare the effectiveness of the distance	1) Fuzzy c-means clustering 2) Distance Metrics: Euclidean, Manhattan,Spearman, Canberra (most effective)	Image Segmentation
Dik et al. (2014)	To identify whether fractional metric can improve the performance of FCM	1) Fuzzy c-means clustering 2) Distance Metrics: Euclidean, Manhattan, Canberra, Chebyshev, BrayCurtis Result: Canberra distance (Iris, Wine and BreastTissue)	1) Iris 2) BCW 3) Wine 4) Heart 5) BreastTissue 6) Indian Dataset
Mallik et al. (2021)	To propose Fuzzy C-least median of squares algorithm which is an improvement to Fuzzy C-means (FCM) algorithm	1) Fuzzy c-means (FCM) 2) Fuzzy C-Least Median (FCLM) Result: FCLM better than FCM	World Wide Web

Next, the analysis of the article based on themes is elaborated in the next section, which is section 5 with relation to the result and discussion.

5. Results and Discussion

In connection with this literary review and meta-analysis, the data analysis used in this study is content analysis. According to Elango and Kumaravel (2022), content analysis as a research technique was used to communicate the evidence-based content on objective, systematic and quantitative description. Further explained that the content analysis method helps the researcher to systematically identify its properties and include large amounts of textual information and make valid conclusions from texts. There are three steps taken in analysing data with this technique, namely: (a) purpose analysis, (b) methodology analysis and (c) significant data analysis.

a. Purpose of the studies

From table 3, we can conclude that the purpose of the selected articles is quite similar, which is to investigate the effectiveness of clustering onto regression modelling and analyse the optimal distance metrics used in clustering algorithm. The selected article also focuses on exploring the significant factors that affect the household income and which modelling can give the best fit for Malaysian household income data. However, it can be concluded that clustering is indeed helpful in improving regression analysis and distance metrics can be used to improve the clustering approach. This outcome encourages more research on clustering algorithms towards household income distribution in Malaysia.

b. Methodology

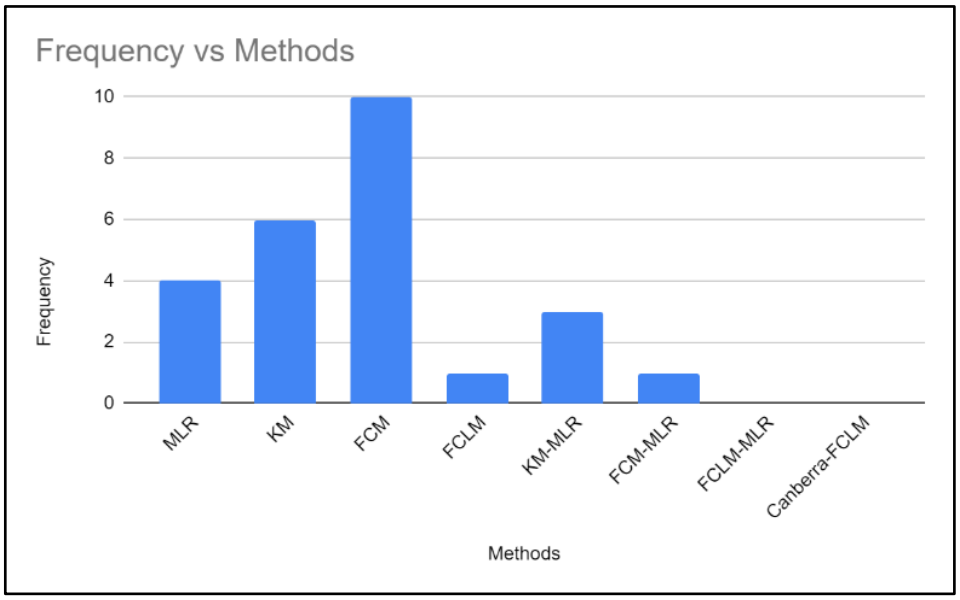

Figure 2: Methodology Used in the Studies

Figure 2 provides an overview of the research methodology used in the articles reviewed. Out of 17 articles, in order to determine the significant factor that affects the household income, 4 research utilise MLR while 15 research use clustering methods such as KM, FCM and FCLM with frequency of 6, 10 and 1 respectively. Furthermore, a total of 3 research publications used a hybrid model by extending the MLR using the KM clustering approach, while 1 research paper extended the MLR using the FCM approach. These studies from Yee et al. (2023); Omolewa et al. (2019); M. Mostafa et al. (2019); Wahab et al. (2018) found out that the hybrid model possesses better robustness and better accuracy compared to the MLR without KM and FCM clustering approach as the clustering analysis can reduce the variation between household income. Thus, results in a more accurate estimate of household income. Furthermore, there's also a few studies that compare the effectiveness of FCM with KM clustering. Based on the study included, FCM provides better effectiveness than KM clustering. Meanwhile, no studies had yet been conducted to extend the MLR using FCLM clustering especially onto Malaysian household income. There's also no study yet to implement Canberra distance onto FCLM approach. Therefore, research that combines linear regression with FCLM clustering approach using Canberra similarity should be conducted to improve the performance of MLR. Therefore, I believe that it's important to investigate whether or not this hybrid model is effective with the household income data in Malaysia.

c. Significant data

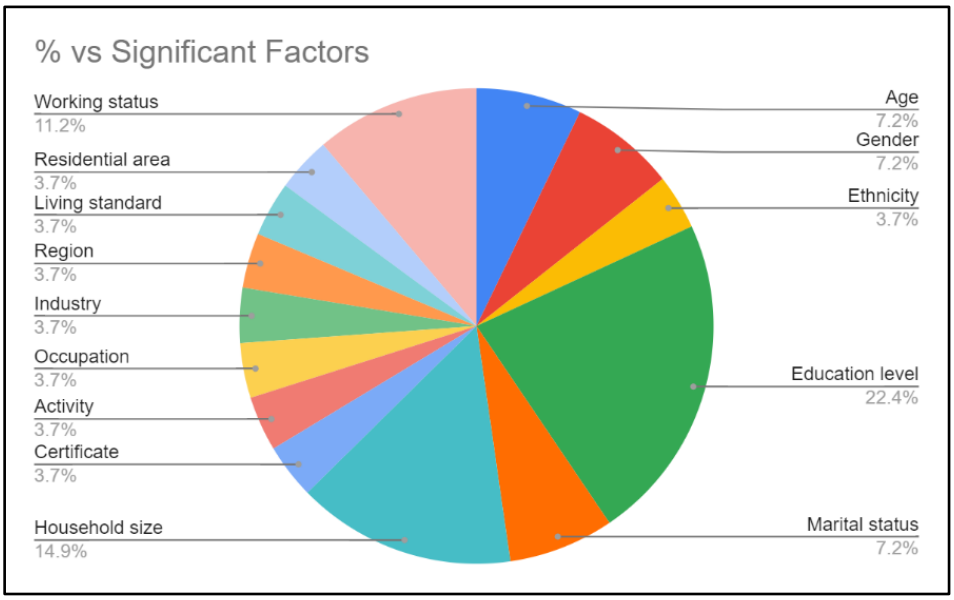

Figure 3: Significant Factors That Affect the Household Income from The Study

Income is one of the most important factors that determine the level of wealth and poverty of individuals. According to the study of past research as shown in figure 3, it is found that all of the article shared the similarity that is all of them were focused on the demographic factors such as age, gender, ethnicity, education level, marital status, household size, certificate, activity, occupation, industry, region, living standard, residential area and working status. However, according to the frequency and the percentage of more than 10% generated from the literature review, there are 3 significant factors that most affect the household income in Malaysia that have been identified which are household size, education level and working status of the head of household. As a result of the findings of all the included studies, education is thought to be the most important component with a percentage of 22.4%. Therefore, future researchers should take these factors into account.

6. Conclusion and Recommendations

This study is the literature review and meta-analysis on clustering and regression models applied to Malaysia's household income, which analyses 17 articles that have been published from the year 2009 until the year 2023 in reputable international journals and selected based on predetermined criteria. This article provides a clear picture of the literature review and meta-analysis on household income distribution in Malaysia and current trends by identifying, screening and checking the eligibility of the research results based on research themes. There are three main themes found based on data analysis: the purpose of the study, methodology and the significant factors that affect the household income from the study. The limitation of this literary review is the range of study which included articles from 2009-2023. Another limitation is the relatively small number of articles reviewed (17 articles). Faber and Fonseca (2014) contend that samples shouldn't be either too large or too small because both have drawbacks that could undermine the conclusions drawn from the studies. However, systematic literature reviews in the social sciences disciplines aim to find as much pertinent research as possible on a specific research question by using explicit methods to identify what can be reliably said on the basis of these studies. Methods should be clear and organised in order to produce a range of trustworthy results. This reduces the bias that can exist in other methods of reviewing the research evidence. This review may assist future researchers in analysing the improvisation of clustering approach, hence encourage them to modify the model in order to reduce the variability in household income due to income gap more effectively. In addition, further action made by the government might give a positive impact towards the health and well-being of individuals and families which are significantly influenced by income (Carter and Howard, 2016).

7. Acknowledgement

The authors would like to thank the College of Computing, Informatics and Mathematics, Universiti Teknologi MARA for supporting this project.

8. Funding

The author(s) received no specific funding for this work.

9. Conflict of Interest

The authors have no conflicts of interest to declare.

10. References

Ahmad, Z. & Ejaz, Zainab. (2011). Classification of households with respect to poverty by using cluster analysis. 10.13140/2.1.4604.6728.

Arora, J., Khatter, K., & Tushir, M. (2018). Fuzzy c-Means Clustering Strategies: A Review of Distance Measures. In *Advances in intelligent systems and computing* (pp. 153–162). Springer Nature. https://doi.org/10.1007/978-981-10-8848-3_15

Banoula, M. (2023). K-means Clustering Algorithm: Applications, Types, & How Does It Work? *Simplilearn.com*. https://www.simplilearn.com/tutorials/machine-learning-

tutorial/k-means-clustering-algorithm#what_is_meant_by_the_kmeans_clustering_algorithm

Carter, V. J., & Howard, M. W. (2016, July 14). *Income inequality | Definition, Kinds, & Facts*. Encyclopedia Britannica.https://www.britannica.com/topic/income-inequality

Chadegani, A. A., Salehi, H., Yunus, M. M., Farhadi, H., Fooladi, M., Farhadi, M., & Ebrahim, N. A. (2013). A Comparison between Two Main Academic Literature Collections: Web of Science and Scopus Databases. *Asian Social Science*, *9*(5). https://doi.org/10.5539/ass.v9n5p18

De Dominicis L., Florax R.J. & De Groot H.L. 2008. A meta-analysis on the relationship between income inequality and economic growth. Scottish Journal of Political Economy 55(5): 654-682.

Dik, A., Jebari, K., Bouroumi, A., & Ettouhami, A. (2014). Fractional metrics for fuzzy c-means. *Int. J. Comput. Infor. Tech*, *3*, 1490-149

Elango, M., & Kumaravel, K. (2022). Content Analysis of OER: A Literature Review. *Shanlax International Journal of Education*, *10*(3), 61–70.https://doi.org/10.34293/education.v10i3.4872

Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. *Dental Press Journal of Orthodontics*, *19*(4), 27–29. https://doi.org/10.1590/2176-9451.19.4.027-029.ebo

Ganesan, R., & Masseran, N. (2023). Statistical analysis of household income data in Perak, Malaysia. In AIP Conference Proceedings. American Institute of Physics. https://doi.org/10.1063/5.0109917

Gough, D., Oliver, S. and Thomas, J. (2017) 'Introducing Systematic Reviews' in *An Introduction to Systematic Reviews.* 3rd edn. ed. by Gough, D., Oliver, S. and Thomas, J. Los Angeles: Sage Publications LtdM.

Hakyemez, T. C., & Bozanta, A. (2018). K-Means vs. Fuzzy C-Means: A Comparative Analysis of Two Popular Clustering Techniques on the Featured. *ResearchGate*.

Kaushik, S. (2023). Clustering | Introduction, Different Methods, and Applications (Updated 2023). *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of clustering/#:~:text=Clustering%20is%20the%20task%20of,and%20assign%20them%20into%20clusters.

Kumar, S. (2022). C-Means Clustering Explained. *Built In*. https://builtin.com/data-science/c-means

Mallik, M. A., Zulkurnain, N. F., Nizamuddin, M., & Aboosalih, K. C. (2021). An Efficient Fuzzy C-Least Median Clustering Algorithm. *IOP Conference Series: Materials Science and Engineering*, *1070*, 012050. https://doi.org/10.1088/1757-899x/1070/1/012050

Moher, D. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Annals of Internal Medicine*, *151*(4), 264. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228.https://doi.org/10.1007/s11192-015-1765-5

Mostafa, Samih & Amano, Hirofumi. (2019). Effect of clustering data in improving machine learning model accuracy. Journal of Theoretical and Applied Information Technology. 97. 2973-2981.

Omolewa, O. R., Oladele, A. T., Adeyinka, A. A., & Ogundokun, R. O. (2019). Prediction of Student's Academic Performance using k-Means Clustering and Multiple Linear Regressions. Journal of Engineering and Applied Sciences, 14(22), 8254–8260. https://doi.org/10.36478/jeasci.2019.8254.8260

Pai, N. K. M. (2021). Building sharp regression models with K-Means Clustering + SVR. *Paperspace Blog*. https://blog.paperspace.com/svr-kmeans-clustering-for-regression/

Palarea-Albaladejo, J., Martín-Fernández, J. A., & Soto, J. (2012). Dealing with Distances and Transformations for Fuzzy C-Means Clustering of Compositional Data. *Journal of Classification*, *29*(2), 144–169. https://doi.org/10.1007/s00357-012-9105-4

Panda, S., Sahu, S. K., Jena, P. K., & Chattopadhyay, S. (2012). Comparing Fuzzy-C Means and K-Means Clustering Techniques: A Comprehensive Study. In *Advances in intelligent and soft computing* (pp. 451–460). Springer Science+Business Media. https://doi.org/10.1007/978-3-642-30157-5_45

Qin D., Cagas M.A., Ducanes G., He X., Liu R. & Liu S. 2009. Effects of income inequality on China's economic growth. Journal of Policy Modelling 31(1): 69-86.

Rahman, M. A., Sani, N. F. M., Hamdan, R., Othman, Z. A., & Bakar, A. A. (2021). A clustering approach to identify multidimensional poverty indicators for the bottom 40 percent group. *PLOS ONE*, *16*(8), e0255312. https://doi.org/10.1371/journal.pone.0255312

Rashid, N.K.A. & Sulaiman, N.F.C. & Rahizal, N.A.. (2018). Survivability through basic needs consumption among muslim households B40, M40 and T20 income groups. Pertanika Journal of Social Sciences and Humanities. 26. 985-998.

Rusdiana, U., Ernawati, I., Falih, N., & Arista, A. (2021). *Comparison of Distance Metrics on Fuzzy C-Means Algorithm Through Customer Segmentation*.

Samsudin, H. B., & Nadzrulizam, A. A. (2020). Relationship between B40 Household Income and Demographic Factors in Malaysia. International Journal of Innovative Technology and Exploring Engineering, 10(2), 113–117.https://doi.org/10.35940/ijitee.b8286.1210220

Setyawan, A. A., & Ilham, A. (2019). A novel framework of the fuzzy c-means distances problem based weighted distance. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1907.13513

Snyder, H. R. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339.https://doi.org/10.1016/j.jbusres.2019.07.039

Stern, C., Jordan, Z., & McArthur, A. (2014). Developing the Review Question and Inclusion Criteria. *American Journal of Nursing*, *114*(4), 53–56. https://doi.org/10.1097/01.naj.0000445689.67800.86

Villejo, Stephen Jun & Enriquez, Mark & Melendrez, Michael & Tan, Dexter & Cayton, Peter Julian. (2014). Determinants of Income Class in Philippine Households: Evidence from the Family Income and Expenditure Survey 2009. The Philippine Statistician. 63. 87-104.

Vu-Ngoc, H., Elawady, S. S., Mehyar, G. M., Abdelhamid, A. M., Mattar, O. M., Halhouli, O., Vuong, N. L., Ali, C. D. M., Hassan, U. H., Kien, N. T., Hirayama, K., & Huy, N. T. (2018). Quality of flow diagram in systematic review and/or meta-analysis. *PLOS ONE*, *13*(6), e0195955. https://doi.org/10.1371/journal.pone.0195955

Wahab, N. H. A., Rusiman, M. S., Mohamad, M., Azmi, N. F., Him, N. C., Kamardan, M. G., & Ali, M. (2018). A Technique of Fuzzy C-Mean in Multiple Linear Regression Model toward Paddy Yield. *Journal of Physics*, *995*, 012010. https://doi.org/10.1088/1742-6596/995/1/012010

Wiharto, W., & Suryani, E. (2020). The comparison of clustering algorithms K-Means and fuzzy C-Means for segmentation retinal blood vessels. *Acta Informatica Medica : AIM : Journal of the Society for Medical Informatics of Bosnia & Herzegovina : ČAsopis DrušTva Za Medicinsku Informatiku BiH*, *28*(1), 42. https://doi.org/10.5455/aim.2020.28.42-47

Yee, G. P., Rusiman, M. S., Ismail, S., Suparman, S., Hamzah, F. M., & Shafi, M. (2023). K-means clustering analysis and multiple linear regression model on household income in Malaysia. IAES International Journal of Artificial Intelligence (IJ-AI), 12(2), 731. https://doi.org/10.11591/ijai.v12.i2.pp731-738

Zainal, F. (2023, February 26). Baby steps to address wealth gap. *The Star*. https://www.thestar.com.my/news/nation/2023/02/27/baby-steps-to-address-wealth-gap