International Journal of Advances in Applied Sciences (IJAAS)
Vol. 14, No. 2, June 2025, pp. 513~522
ISSN: 2252-8814, DOI: 10.11591/ijaas.v14.i2.pp513-522

Birth data clustering to segmentation delays in birth certificate registration

Erfan Hasmin[1], Aedah Abd Rahman[2]
[1]Department of Informatics, Faculty of Computer Science, Dipa Makassar University, Makassar, Indonesia
[2]School of Science and Technology, Asia e University, Kuala Lumpur, Malaysia

Article Info

Article history:

Received Oct 23, 2024
Revised Apr 25, 2025
Accepted May 10, 2025

Keywords:

Birth registration
Data-based policy
Elbow
K-means
Silhouette

ABSTRACT

Timely and accurate birth registration is essential for ensuring access to vital public services. This study focuses on clustering birth data to identify patterns in registration delays, using data mining techniques such as the K-means algorithm. By clustering birth data from Makassar City, Indonesia, based on various demographic and birth-related criteria, the study segments the data into groups that reflect both timely and delayed registrations. The optimal number of clusters is determined using the elbow and silhouette methods. Results show that a three-cluster configuration effectively captures patterns in birth registration delays, offering critical insights for policymakers. These findings provide a foundation for improving birth registration processes, ensuring more timely registration, and guiding data-driven public policy decisions.

Corresponding Author:

Erfan Hasmin
Department of Informatics, Faculty of Computer Science, Dipa Makassar University
Perintis Kemerdekaan IX, Tamalanrea, Makassar 90241, Indonesia
Email: erfan.hasmin@undipa.ac.id

1. INTRODUCTION

Policymakers rely on accurate and timely data to shape their decisions and oversee the progress of policies and programs. Essential statistics pertaining to the quantity and geographic spread of births, along with the reasons behind these deaths, are crucial for guiding strategic planning in various sectors, such as healthcare, education, workforce, urban development, finance, economic growth, trade, social safety nets, environmental management, and demographic analysis [1]. Policies that are not based on data can cause various problems. Policies may not be effective in solving problems without accurate and relevant data [2]. Without policies based on data, it can result in a lack of access to civil registration, such as births, which results in the inability of individuals to access fundamental rights such as education, health services, and inheritance rights [3].

Efforts and initiatives are underway to enhance the registration of births across diverse scenarios, geographical areas, and social contexts within the community. In fact, governmental policies are formulated based on data, with a preference for data analysis as a crucial factor in decision-making. Current data mining algorithms can better analyze specific data than other data analysis methods [4]. The need for data clustering in civil registration data involves several reasons that are essential in the context of data management, public policy, and population understanding which include population grouping which can help in better understanding the needs and preferences of each group, which is very valuable in policy planning public and community services, personalization of services so that civil registration services can adapt services to better meet the needs of individuals or groups, understanding demographic trends to understand complex demographic trends, such as changes in population structure, migration, or changes in family composition

and of course taking data-based decisions so that civil registration service policies become more based on solid data to reduce the risk of ineffective policies [5]. It is anticipated that this approach will mitigate issues arising from policies aimed at increasing birth registration, which have historically lacked a foundation in data patterns and segmentation. Such deficiencies often lead to inefficiencies in resource allocation and budget utilization.

For this reason, the data mining approach employed to categorize birth registration data involves clustering the data using multiple criteria. This multi-criteria clustering enables the government to develop public policies that align with the actual situation [6]. Clustering birth data according to appropriate criteria, including birth events, child biodata, parents, and regional demographics. The data clusters formed can be a basis for determining policies and activities to improve birth registration. Birth data analysis using the neural network method emphasizes the need for data mining analysis on civil registration data to understand e-governance data better [7]. As well as the use of data mining techniques on government data is proven to make better planning and decision-making [8]. Utilization of birth data for the development and determination of policies and activities needs to be done to increase the number of registered births. Over three years, birth data can be examined using data mining techniques to extract new insights that could inform the formulation of policies to enhance birth registration systems.

The segmentation of birth registration data is of paramount importance for research, especially considering the challenges inherent in conducting population data analysis. This complexity stems from the need to incorporate data on critical life events, such as births. By segmenting this data, it becomes feasible to identify and delineate key patterns, particularly those related to delays in birth registration. This approach provides a foundation for enhancing the system's efficiency and timeliness. Research related to data mining, especially clustering is done at civil registration offices to improve community services. Clustering techniques are applied to find better ways of managing complaints. Clustering, particularly the K-means technique, has been employed to aggregate and visualize extensive, unstructured data complaints in a cloud format according to the discovered root causes [9]. The main conclusions obtained from the grouping of data complaints in the civil registry office confirm that data mining is efficient in obtaining the root causes of complaints [10]. Research related to the use of birth data in records for use The use of population data at the civil registration office is processed by applying data mining clustering using the K-means algorithm to cluster and describe the clustering of children's data based on the number of children's birth certificate ownership in each sub-district where the clustering results can be used as material for planning and evaluating targets in birth certificate ownership services and child identity card to accelerate the achievement of birth certificate and child identity card ownership targets set by the government [11]. The results of the formed K-means clustering produced groups of sub-districts in each cluster, categorized based on the scope of non-ownership of birth certificates [12].

Research related to data mining analysis on population data by mining population data belonging to the city government of Al Khums in Libya [13]. This study utilizes population data in Al Khums municipality in Libya by mining classification data using the k-nearest neighbors (KNN) algorithm and grouping data techniques with the K-means method for poverty level measurement (through income), population rate increase measurement (through marriage) and population rate decline measurement (through death). The study's conclusion claimed to be the first of its kind, is based on helping decision-makers in municipalities make informed decisions. The results of this cluster can be used as a reference for the population and civil registration office in mapping birth certificate data in Indonesia. And relation to the grouping of birth data with the research title application of the K-means method for clustering child data based on birth certificate ownership and maternal and child healthcare (MCH) in the study found the best number of clusters for the same birth data cluster in the first study, namely 4 clusters [14], While the research on clustering tested four methods-elbow, gap statistic, silhouette coefficient, and canopy-to determine the K value in the K-means method [15]. None of these studies focus on delays in birth registration, despite available data that could support such analysis. These delays may stem from data access or security issues at the civil registry office. However, this research offers segmentation insights that help policymakers understand the distribution of registration delays, enabling targeted interventions and data-driven decisions to improve the civil registration system and promote timely registrations.

2. METHOD

2.1. K-means clustering algorithm

The K-means algorithm is a non-hierarchical clustering method that partitions data into distinct groups based on shared features, ensuring high similarity within each cluster [16]. Data with dissimilar characteristics, exhibiting low inter-class similarity, are assigned to separate clusters [17]. Subsequently, the

algorithm computes the distance between each data point and each cluster center centers [18] the Euclidean distance formula is named, (d_{ik}) in (1).

$$d_{ik=}\sqrt{\sum_{j=1}^{m}\left\{x_{ij-}c_{kj}\right\}^2} \quad (1)$$

The calculation of the initial data distance using (1) is performed against the centroid values of each cluster based on the seven criteria utilized.

Distance between the first data and the centroid point of cluster_0 (C0),

$$C0=\sqrt{((2-2)^2)+((1-1)^2)+((2-2)^2)+\left((3-3)^2+((1-1)^2)+((1-1)^2)+((5-5)^2)\right)}$$
$$C0=0$$

Distance between the first data and the centroid point of cluster_1 (C1),

$$\mathrm{C1}=\sqrt{((2-1)^2)+((1-1)^2)+((2-2)^2)+\left((3-2)^2+((1-2)^2)+((1-1)^2)+((5-3)^2)\right)}$$
$$\mathrm{C1}=7$$

Distance between the first data and the centroid point of cluster_2 (C2),

$$C2=\sqrt{((2-1)^2)+((1-5)^2)+((2-1)^2)+\left((3-1)^2+((1-2)^2)+((1-1)^2)+((5-4)^2)\right)}$$
$$C2=24$$

A data point will be assigned to the k-th cluster if its distance to the center of the k-th cluster is the minimum among all distances to the centers of other clusters. The calculation can be performed using automation or by determining the minimum value of the objective equation. Subsequently, the collected data is organized into several clusters, each comprising its respective members. Minimum value in (2).

$$\min\sum_{k=1}^{k} d_{ik} \quad (2)$$

Implement (2) by determining the smallest value from the results of the centroid calculations.

Cluster=0 → because C0 is smaller than C1 and C2.

The calculation of the new cluster center value involves determining the mean value of the data points that belong to the cluster, as specified by (3).

$$c_{kj}=\frac{1}{p}\sum_{i=1}^{p} x_{ij} \quad (3)$$

Which $x_{ij} \in C_{kj}$or x_{ij} is in the k-th cluster, and *p* is the number of members of the k-th cluster.

The new cluster center point is determined using the cluster center point formula on (3) based on data from each cluster member.

New point cluster_0=((2+1))/2=1.5 ((1+1))/2=1 ((2+2))/2=2 ((3+4))/2=3.5 ((1+2))/2=1.5 ((1+1))/2=2 ((5+6))/2=5.5

=1.5, 1, 2, 3.5, 1.5, 2, 5.5

The fundamental algorithm utilized in K-means is i) specify the number of clusters (k) and set an arbitrary cluster center, ii) calculate the distance of each record to the cluster center using, ii) group data into clusters with the shortest distance, iv) calculate the new cluster center, v) steps 2 through 4 will be repeated so no more data moves to another cluster [19].

2.2. Elbow algorithm

The elbow method is employed to ascertain the optimal number of k clusters by analyzing the outcomes of the comparison where the number of clusters exhibits an inflection point, resembling an elbow [20]. This method performs unlimited test clusters to solve the same case, making the optimal number of clusters easier to determine [21]. The comparison value between the number of clusters is calculated by calculating the Sum of Square Error (SSE) in each cluster [22].

$$SSE=\sum_{k=1}^{K}\sum_{x_i}||\,x_i-C_k||^2 \quad (4)$$

Which K is number of clusters, Xi is i data, and Ck is cluster centroid.

Plotting the performance differences for k values ranging from 2 to 7, it becomes evident that the most significant improvement occurs at k=3. This suggests that three clusters best represent the underlying structure of the data, balancing model complexity and explanatory power. Selecting this optimal cluster count helps ensure meaningful and interpretable groupings, which can enhance subsequent analysis and decision-making processes. Table 1 shows that the largest performance difference occurs at k=3, with a value of 0.671, indicating the optimal cluster count.

Table 1. Birth data cluster distance performance value

k	Average centroid distance	Difference
2	3.94	0
3	3.269	0.671
4	2.8	0.469
5	2.462	0.338
6	2.089	0.373
7	1.863	0.226

2.3. Silhouette algorithm

The silhouette coefficient is a widely used metric for evaluating the quality and strength of clusters in a dataset [23]. It combines two key aspects: cohesiveness, which measures how closely related objects are within the same cluster, and separation, which assesses how distinct or well-separated a cluster is from other clusters. By balancing these two factors, the silhouette coefficient provides a single value that reflects how appropriately the data has been clustered [24]. A higher coefficient indicates that clusters are both compact and well-separated, which is desirable for meaningful clustering results. This method helps in validating the effectiveness of clustering algorithms and in selecting the optimal number of clusters.

Calculate the average distance of document i with all documents in other clusters.

$$d(i,C) = \frac{1}{|A|}\sum_{j \in C} d(i,j) \quad (5)$$

Calculate the silhouette coefficient value.

$$S(i) = \frac{b(i)-a(i)}{max\ (a(i),b(i))} \quad (6)$$

The results of the birth data cluster are then optimized using the silhouette method to determine the optimal number of clusters. The following are the results of the silhouette coefficient calculation based on the calculation of the values k=2 to k=7. results of the silhouette coefficient calculation process on the birth cluster data, the maximum silhouette coefficient result is when k=3 with a silhouette value=0.573, as seen in Table 2.

Table 2. Birth data silhouette coefficient

k	Silhouette coefficient	Rank
2	0.470	3
3	0.573	1
4	0.474	2
5	0.462	4
6	0.461	5
7	0.452	6

3. RESULTS AND DISCUSSION

Forming birth data clusters offers valuable insights, particularly for policy planning. This process involves grouping birth records based on similarities in specific features, serving as the initial step in subsequent analysis [25]. In the clustering process, it is necessary to establish the number of clusters. The function of forming the number of clusters is to determine how many clusters will be used in the clustering process. This can affect the final result of the analysis and the resulting interpretation [24]. The formation of birth data clusters is divided based on seven relevant criteria, including i) place of birth, ii) type of birth, iii) birth, iv) attendant, v) birth order, vi) gender, vii) age of the child when the birth was registered.

3.1. Framework

The first stage of this research design involves pre-processing birth data from the Makassar City population and civil registration database (2019-2023). This step is critical to ensure the data's accuracy, completeness, and consistency before any analysis is conducted. The research framework is divided into two main parts, as illustrated in Figure 1 the first part focuses on data preparation and cleaning, while the second part involves applying analytical methods to uncover patterns and insights.

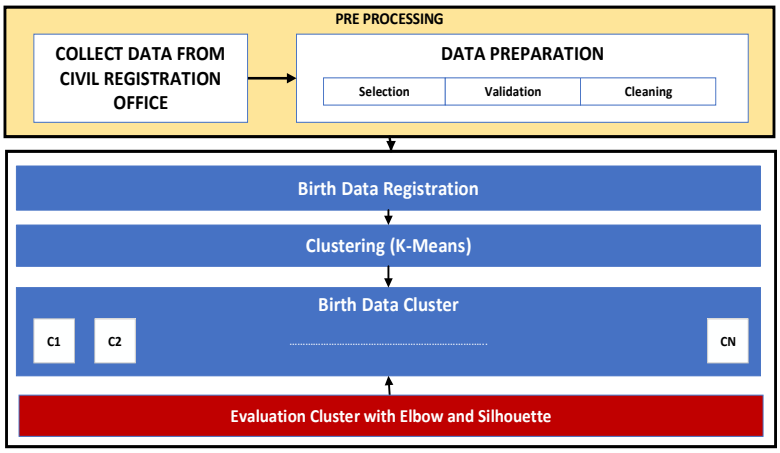

Figure 1. Design framework

The data validation process is an essential stage in managing the dataset to ensure the accuracy and consistency of the information contained in the dataset. This includes checks to detect duplicates, missing values, or other inconsistencies in the dataset [26]. This process involves a series of steps to identify, overcome, and remove problems in the dataset such as identification of missing values, noisy data cleaning, remove outliers (removing observation results that are significantly different from the majority of the data in the dataset), and resolve inconsistencies [27]. Next, the K-means clustering model is shown in Figure 2.

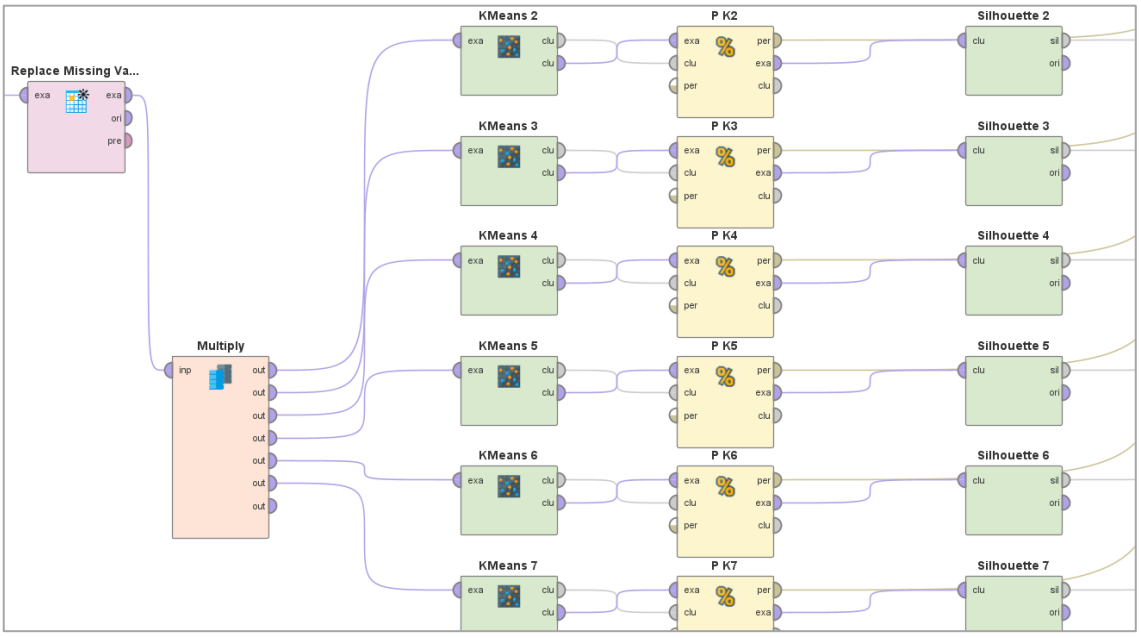

Figure 2. K-means clustering model

The K-means application starts by selecting the number of clusters to group 6,003 birth records, tested from 2 to 7 clusters using seven criteria. A simulation with 10 data points illustrates this clustering process. The design explores different cluster counts to find the optimal grouping, determined using the elbow and silhouette methods.

3.2. Dataset

The Makassar City population and civil registration office birth data contains 12,334 records with three statuses: accepted, rejected, and incomplete. Only accepted data is processed. Data cleaning involves checking for empty fields and missing values to ensure error-free data. Records with missing values are removed. Detailed status information is shown in Table 3.

Table 3. Birth dataset

No	Count	Description
1	12,334	Original birth data from civil registration
2	7,193	Original birth data with accepted status
3	6,003	Birth data after data cleaning

The results of the data validation ensure that the data to be processed is free from errors and corresponds to the vulnerable value based on this validation process. Of the seven criteria used in birth data, the criterion for the age of the child (label/target criteria) when the birth was registered provides information on the data, including the type of birth registration that is on time or late. The distribution of data on the criteria of the child's age at birth is detailed in Table 4.

Table 4. Age of the child when the birth was registered dataset

No	Value	Count	Status
1	0 month	1,991	On-time
2	1-3 month	1,882	Delay
3	4-6 month	1,039	Delay
4	7-12 month	1,073	Delay
5	13-24 month	1,000	Delay
6	More than 24 months	208	Delay

The clustering results above show the distribution of data into several clusters based on the number of k (number of clusters), which varies from 2 to 7. The higher the value of k, the more the data is split into smaller and more specific clusters. The distribution results for cluster 0 to cluster 6 can be seen in Table 5.

Table 5. K-means clustering results on birth data

k	2	3	4	5	6	7
cluster 0	3,973	3,668	2,959	2,659	1,005	1,065
cluster 1	2,030	2,059	276	269	1,825	1,005
cluster 2		276	858	1,586	1,586	272
cluster 3			1,910	1,213	276	1,798
cluster 4				276	1,039	521
cluster 5					272	1,066
cluster 6						276

3.3. Optimal number of clusters

Determining the optimal cluster is an essential step in clustering analysis to ensure the data is divided into the right groups. In this analysis using the elbow, cluster 3 emerged as the optimal choice, as it showed the largest decrease in variance when compared to clusters with fewer or more groupings, indicating a point of diminishing returns. Furthermore, the silhouette method, which evaluates the cohesion and separation of clusters by calculating the silhouette coefficient, corroborated the results of the elbow method by also selecting cluster 3 as the optimal configuration. The results of the comparison of the elbow and silhouette methods can be seen in Table 6.

Table 6. Comparison of birth data optimum cluster with elbow and silhouette

k	Elbow		Silhouette	
	average centroid	difference	silhouette coefficient	rank
2	3.94	0	0.436	3
3	3.269	0.671	0.573	1
4	2.8	0.469	0.474	2
5	2.462	0.338	0.462	4
6	2.089	0.373	0.461	5
7	1.863	0.226	0.452	6

3.4. K-means clusters and distribution of birth registration delays

The cohesiveness approach assesses the proximity of relationships among objects within a cluster, while the separation method evaluates the distinction between clusters. This combined approach confirms that the chosen clustering structure balances internal consistency and external differentiation, leading to meaningful and interpretable groupings in the dataset. Based on both the elbow and silhouette analyses, a three-cluster solution is deemed most effective for segmenting the birth data as shown in Figure 3.

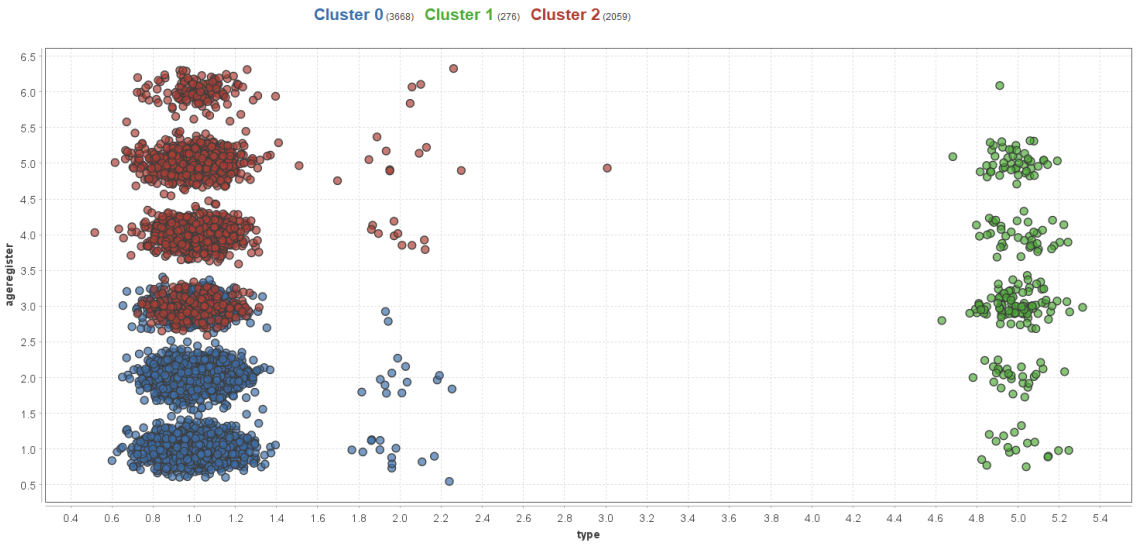

Figure 3. Distribution of birth data with 3 clusters

This three-cluster distribution provides an insightful view of the data, capturing significant groupings and likely underlying patterns. Figure 3 shows the distribution of birth data with a view of 3 clusters, where cluster 0 is the first cluster with 3,668 data items, cluster 1 is the second cluster, which is displayed with green items with 2,059 data items, and cluster 2 is the third cluster which is shown in red with 276 data items. The distribution of data for each cluster on the label criteria can be seen in Table 7.

Table 7. Distribution of label criteria for each cluster

No	Cluster	Criteria label (Month)					
		0 on-time	1-3	4-6	6-12 late	12-24	>24
1	Birth cluster 0	1,705	1,464	499	0	0	0
2	Birth cluster 1	0	0	253	839	784	183
3	Birth cluster 2	21	32	121	43	58	1

We found that the pattern of data distribution for the registration criteria of birth events with a type of 0 months (on-time) is concentrated in cluster 0. Additionally, the late registration criteria for the 1-3 month and 4-6-month ranges are also concentrated in cluster 0. In contrast, cluster 1 shows a different distribution of data regarding the registration criteria for birth data. This three-cluster configuration offers deep segmentation of the birth data, allowing for identifying significant patterns among the leading groups in the

data. Furthermore, the distribution table of registration criteria shows that these clusters play an important role in understanding the distribution of delays in birth registration. Cluster 0 concentrates on data involving timely registration (0 months) and late registration in the 1-3 months and 4-6 months. Cluster 1 shows the distribution of data on birth registration with different characteristics. These results underline the effectiveness of the three-cluster configuration in optimally clustering birth data. With segmentation results that successfully identify the structure in the data [28] and enable clustering for further processing [29], cluster 0 provides information that can be explored to understand birth events registered on time and those registered late, with delays ranging from 1 to 6 months. Similarly, cluster 1 can be examined further to provide insights into birth events registered late, with delays ranging from 6 months to more than 24 months.

4. CONCLUSION

This study presents findings involving the inclusion of target labels or criteria, specifically whether a birth is registered in the clustering process, marking a departure from traditional multi-criteria clustering methods that typically exclude such labels. This innovative approach resulted in data distributions within each cluster that were distinctly concentrated around the specified label or target criteria, enhancing the clarity and relevance of the clustering outcomes. The analysis of birth registration data revealed significant patterns, with cluster 0 representing cases of on-time or minimally delayed registrations (1-6 months) and cluster 1 highlighting instances of substantial delays (6-24 months). These segmentation results offer policymakers valuable insights into the distribution and nature of birth registration delays, facilitating more targeted interventions and data-driven strategies to improve civil registration systems and encourage timely registrations. Moreover, the study underscores the potential benefits of incorporating label or target criteria as integral factors in clustering processes to deepen the understanding of birth registration trends. Looking ahead, future research could focus on applying more advanced clustering techniques, such as dynamic clustering models, to better capture temporal variations in delayed registrations. Additionally, integrating machine learning approaches like neural networks and classification algorithms with clustering may enhance the accuracy of predicting unregistered births, ultimately contributing to more effective policy formulation and resource allocation. This research thus lays a foundation for further exploration into hybrid analytical methods that combine unsupervised and supervised learning to address complex challenges in population data management.

ACKNOWLEDGEMENTS

Author thanks to Makassar City Civil Registry Office for trusting me to process birth registration data for this research and to the Dipanegara Foundation for their trust, contribution, and financial support, which have been very meaningful for me in completing this research.

FUNDING INFORMATION

This research was supported by funding from the Dipanegara Foundation. Additional resources were provided by the Makassar City Civil Registration Office, which granted access to the population database (2019-2023). The funding bodies had no role in the design of the clustering methodology, data analysis, or interpretation of results related to birth registration patterns.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Erfan Hasmin	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Aedah Abd Rahman	✓	✓			✓			✓		✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest related to this research. The Makassar City Civil Registration Office provided access to the data but had no role in the study design, data analysis, interpretation, or manuscript preparation. All authors have disclosed any potential competing interests and confirm that none exist that could have influenced the outcomes of this research.

INFORMED CONSENT

This research was conducted with formal approval from the Makassar City Civil Registration Office. The study was authorized under the permission number 000.9_733/Disdukcapil/IX/2023, ensuring that all data collection and analysis procedures complied with ethical standards and regulations.

ETHICAL APPROVAL

This research received ethical approval from the Makassar City Civil Registration Office under the official permission number 000.9_733/Disdukcapil/IX/2023. All procedures involving data collection and analysis were conducted in accordance with ethical standards and guidelines to protect participant confidentiality and ensure responsible use of personal information. The study adhered to relevant regulations and maintained transparency and integrity throughout the research process.

DATA AVAILABILITY

The data supporting the findings of this study are available from the Population and Civil Registration Office of Makassar City. Restrictions apply to the availability of this data, which is used under license for this research. The data is not publicly available except with permission from the Population and Civil Registration Office of Makassar City because it contains the personal data of Indonesian citizens. Sample data is available (https://www.kaggle.com/datasets/erfanhasmin/birth-and-death-data-sample-in-civil-registration) with permission to help understand the context and data criteria used in this study.

REFERENCES

[1] A. S. Venkataramani, R. O'Brien, G. L. Whitehorn, and A. C. Tsai, "Economic influences on population health in the United States: toward policymaking driven by data and evidence," *PLoS Medicine*, vol. 17, no. 9, 2020, doi: 10.1371/journal.pmed.1003319.

[2] M. Favaretto, E. De Clercq, and B. S. Elger, "Big data and discrimination: perils, promises and solutions. A systematic review," *Journal of Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0177-4.

[3] T. Yu *et al*., "MOPO: model-based offline policy optimization," *Advances in Neural Information Processing Systems*, 2020.

[4] M. H. Tekieh and B. Raahemi, "Importance of data mining in healthcare: a survey," *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*, pp. 1057–1062, 2015, doi: 10.1145/2808797.2809367.

[5] R. Silva, "Population perspectives and demographic methods to strengthen CRVS systems: introduction," *Genus*, vol. 78, no. 1, 2022, doi: 10.1186/s41118-022-00156-8.

[6] A. Viloria and O. B. P. Lezama, "Improvements for determining the number of clusters in K-means for innovation databases in SMEs," *Procedia Computer Science*, vol. 151, pp. 1201–1206, 2019, doi: 10.1016/j.procs.2019.04.172.

[7] P. Y. Desai, "Implementation of artificial neural network algorithm on vehicle registration data," *International Journal of Advanced Research in Engineering and Technology*, vol. 10, no. 1, pp. 83–87, 2019, doi: 10.34218/IJARET.10.1.2019.008.

[8] K. Konstas, P. T. Chountalas, E. A. Didaskalou, and D. A. Georgakellos, "A pragmatic framework for data-driven decision-making process in the energy sector: insights from a wind farm case study," *Energies*, vol. 16, no. 17, 2023, doi: 10.3390/en16176272.

[9] S. Bhosale, S. Patankar, K. Kadam, R. Dhere, and P. M. Desai, "Survey on civil complaints management system by using machine learning techniques," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 606–610, 2021, doi: 10.48175/ijarsct-1449.

[10] M. Anityasari and I. D. A. Indriasari, "Text mining implementation in complaint management: a case study at Surabaya City office for population administration and civil registration (COPACR)," *AIP Conference Proceedings*, vol. 2693, no. 1, 2023, doi: 10.1063/5.0120572.

[11] A. F. Gebremedhin, A. Dawson, and A. Hayen, "Evaluations of effective coverage of maternal and child health services: a systematic review," *Health Policy and Planning*, vol. 37, no. 7, pp. 895–914, 2022, doi: 10.1093/heapol/czac034.

[12] R. G. Aboagye, J. Okyere, A. A. Seidu, B. O. Ahinkorah, E. Budu, and S. Yaya, "Determinants of birth registration in sub-Saharan Africa: evidence from demographic and health surveys," *Frontiers in Public Health*, vol. 11, 2023, doi: 10.3389/fpubh.2023.1193816.

[13] A. D. Ghanim and Z. S. Zubi, "Data mining methods in municipality social data system (DMSDS)," *International Journal of Computers*, vol. 7, 2022.

[14] O. O. Balogun *et al*., "Effectiveness of the maternal and child health handbook for improving continuum of care and other maternal and child health indicators: A cluster randomized controlled trial in Angola," *Journal of Global Health*, vol. 13, 2023, doi: 10.7189/jogh.13.04022.

[15] C. Yuan and H. Yang, "Research on K-value selection method of K-means clustering algorithm," *J*, vol. 2, no. 2, pp. 226–235, 2019, doi: 10.3390/j2020016.

[16] M. Vichi, C. Cavicchia, and P. J. F. Groenen, "Hierarchical means clustering," *Journal of Classification*, vol. 39, no. 3, pp. 553–577, 2022, doi: 10.1007/s00357-022-09419-7.
[17] M. Annas and S. N. Wahab, "Data mining methods: K-means clustering algorithms," *International Journal of Cyber and IT Service Management*, vol. 3, no. 1, pp. 40–47, Mar. 2023, doi: 10.34306/ijcitsm.v3i1.122.
[18] J. Hämäläinen, T. Kärkkäinen, and T. Rossi, "Improving scalable K-means++," *Algorithms*, vol. 14, no. 1, pp. 1–20, 2021, doi: 10.3390/a14010006.
[19] T. M. Ghazal *et al.*, "Performances of K-means clustering algorithm with different distance metrics," *Intelligent Automation and Soft Computing*, vol. 30, no. 2, pp. 735–742, 2021, doi: 10.32604/iasc.2021.019067.
[20] A. M. Jabbar, K. R. Ku-Mahamud, and R. Sagban, "An improved ACS algorithm for data clustering," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 17, no. 3, pp. 1506–1515, 2020, doi: 10.11591/ijeecs.v17.i3.pp1506-1515.
[21] D. Jollyta, S. Efendi, M. Zarlis, and H. Mawengkang, "Analysis of an optimal cluster approach: a review paper," *Journal of Physics: Conference Series*, vol. 2421, no. 1, 2023, doi: 10.1088/1742-6596/2421/1/012015.
[22] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-means clustering method and elbow method for identification of the best customer profile cluster," *IOP Conference Series: Materials Science and Engineering*, vol. 336, no. 1, 2018, doi: 10.1088/1757-899X/336/1/012017.
[23] F. Wang, J. Chen, and F. Liu, "Keyframe generation method via improved clustering and silhouette coefficient for video summarization," *Journal of Web Engineering*, vol. 20, no. 1, pp. 147–170, 2021, doi: 10.13052/jwe1540-9589.2018.
[24] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," *Proceedings-2020 IEEE 7th International Conference on Data Science and Advanced Analytics, DSAA 2020*, pp. 747–748, 2020, doi: 10.1109/DSAA49011.2020.00096.
[25] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *Journal of Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00327-4.
[26] S. M. Bhagavathi *et al.*, "Weather forecasting and prediction using hybrid C5.0 machine learning algorithm," *International Journal of Communication Systems*, vol. 34, no. 10, 2021, doi: 10.1002/dac.4805.
[27] K. Rahul and R. K. Banyal, "Detection and correction of abnormal data with optimized dirty data: a new data cleaning model," *International Journal of Information Technology and Decision Making*, vol. 20, no. 2, pp. 809–841, 2021, doi: 10.1142/S0219622021500188.
[28] M. Chaudhry, I. Shafi, M. Mahnoor, D. L. R. Vargas, E. B. Thompson, and I. Ashraf, "A systematic literature review on identifying patterns using unsupervised clustering algorithms: a data mining perspective," *Symmetry*, vol. 15, no. 9, 2023, doi: 10.3390/sym15091679.
[29] M. Z. Rodriguez *et al.*, "Clustering algorithms: a comparative approach," *PLoS ONE*, vol. 14, no. 1, 2019, doi: 10.1371/journal.pone.0210236.
[30] J. Pereira, P. Contreras, D. C. Morais, and P. Arroyo-López, "Multi-criteria ordered clustering of countries in the global health security index," *Socio-Economic Planning Sciences*, vol. 84, 2022, doi: 10.1016/j.seps.2022.101331.

BIOGRAPHIES OF AUTHORS

Erfan Hasmin is a lecturer in the Department of Informatics, at Dipa Makassar University, Makassar, Indonesia. Erfan does research in decision support systems and data mining. Have expertise in software development and application of mathematical methods to programming. He can be contacted via email: erfan.hasmin@undipa.ac.id.

Aedah Abd Rahman is a Profesor in School of Science and Technology, Asia e University, Kuala Lumpur, Malaysia. Have expertise in data mining and software engineering. She can be contacted at email: aedah.abdrahman@aeu.edu.my.