Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Gender, Education, and Financial Insecurity Among the Older Adults in Oman: Insights from a Comparative Clustering Analysis

Hamza Zidoum

zidoum@squ.edu.om

Sultan Qaboos University

Thuraya Al Nabhani

Sultan Qaboos University

Research Article

Keywords: Older adults population segmentation, Oman, Unsupervised machine learning, Clustering algorithms, Gender disparity, Socioeconomic inequality

Posted Date: June 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9731781/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Gender, Education, and Financial Insecurity Among the Older Adults in Oman: Insights from a Comparative Clustering Analysis

Abstract

Oman is undergoing a significant demographic transition, with its older adults population (aged 60 and above) projected to more than double—from 166,464 individuals (4.79%) in 2023 to 376,716 (11.80%) by 2040. Despite this rapid growth, older adults social support services remain largely uniform, exemplified by the Social Protection Fund's fixed monthly benefit of 115 OMR regardless of individual socioeconomic circumstances. This study applies unsupervised machine learning to segment Oman's older adults population into meaningful subgroups, providing data-driven insights to support more targeted, equitable welfare policy. Using a nationally representative dataset of 1,700 individuals aged 60 and above—drawn from a 1% stratified sample provided by the National Centre for Statistics and Information (NCSI)—five clustering algorithms were systematically compared: K-Means, K-Prototypes, Hierarchical Clustering (with Gower distance), Self-Organizing Maps (SOM), and DBSCAN. Performance was evaluated using the Silhouette Score and Calinski-Harabasz Index. K-Means clustering with k=2 achieved the highest Calinski-Harabasz Index (550), while Hierarchical Clustering with Gower distance achieved the highest Silhouette Score (0.62). Three algorithms—K-Means, K-Prototypes, and Hierarchical Clustering—independently converged on near-identical two-cluster solutions, substantially strengthening confidence in the identified subgroups. The two clusters reveal a pronounced gendered socioeconomic bifurcation: a financially stable group of predominantly married males with retirement income and higher educational attainment, and a vulnerable group of predominantly widowed or divorced females characterised by illiteracy, absence of retirement income, and low household ownership rates. These findings expose critical disparities within Oman's older adults population and provide an evidence base for developing more targeted, gender-sensitive, and need-responsive social protection programmes.

Keywords: *Older adults population segmentation · Oman · Unsupervised machine learning · Clustering algorithms · Gender disparity · Socioeconomic inequality*

1 Introduction

Global demographic trends reveal an accelerating shift towards population ageing. Declining fertility rates and rising life expectancy—particularly in low- and middle-income nations—are reshaping age structures worldwide. The World Health Organization projects that by 2050, the proportion of people aged 60 or older will nearly double from 12% (2015) to 22%, placing unprecedented demands on social protection systems, healthcare infrastructure, and community services (WHO 2022). This trajectory mirrors accelerating global trends: as the United Nations World Social Report 2023 documents, population ageing is now a defining feature of development across all income groups, with older women consistently forming a majority of the ageing population (United Nations, 2023).

Oman mirrors this global trajectory with particular intensity. Life expectancy increased dramatically between 1950 and 2024, while the total fertility rate fell sharply from 7.25 children per woman in 1950 to 2.47 in 2024 (World Bank 2024). As a consequence, the older adults population is projected to grow from 166,464 individuals (4.79% of the total population) in 2023 to 376,716 (11.80%) by 2040 (NCSI 2024). This more than twofold increase demands proactive planning across older adults care, financial security, and social welfare.

The demographic imperative is compounded by a structural policy gap. Oman's older adults population is heterogeneous in its socioeconomic characteristics—differentiated by gender, marital status, educational attainment, income, geographic location, and household ownership. Yet existing support services remain largely undifferentiated. The Social Protection Fund (SPF), for instance, provides a fixed monthly benefit of 115 OMR to all eligible older adults individuals, irrespective of their diverse socioeconomic circumstances (SPF 2024). As SPF specialists interviewed for this study emphasised, this one-size-fits-all approach fails to capture the complex realities of Oman's ageing population and limits the potential for targeted intervention.

The academic literature on older adults population segmentation has grown substantially over the past decade, with studies from the United States (Ahn et al. 2020), China (Wei and Li 2022; Zheng et al. 2023), and clinical settings internationally (Khalid and Prieto-Alhambra 2019; Elbattah and Molloy 2016) demonstrating the value of data-driven clustering for uncovering meaningful subgroups among older adults. However, no study has systematically compared multiple clustering algorithms applied specifically to Oman's older

adults population using demographic and socioeconomic data—a gap that is particularly significant given Oman's distinct cultural, institutional, and developmental context within the Gulf Cooperation Council (GCC).

This paper addresses that gap. We apply and systematically evaluate five clustering algorithms—K-Means, K-Prototypes, Hierarchical Clustering (with Gower distance), Self-Organizing Maps (SOM), and DBSCAN—to a nationally representative dataset of 1,700 older adults Omani citizens. The study is guided by three objectives: (1) to segment the older adults population based on demographic and socioeconomic characteristics using multiple unsupervised learning approaches; (2) to evaluate and compare clustering performance using internal validation metrics; and (3) to interpret the resulting clusters to identify meaningful subgroups and generate policy-relevant insights for older adults service planning in Oman.

The remainder of the paper is structured as follows. Section 2 reviews the relevant literature on older adults population clustering, situating this study within the broader gerontological and methodological literature. Section 3 describes the data, preprocessing pipeline, clustering algorithms, and evaluation metrics. Section 4 presents the results for each algorithm. Section 5 discusses the findings, their theoretical implications, and their policy relevance. Section 6 concludes and identifies directions for future research.

2 Background and Theoretical Context

2.1 Ageing Policy and Socioeconomic Heterogeneity

The growing literature on population ageing in low- and middle-income countries recognises that older adults are not a homogeneous group. Gender, education, employment history, and geographic location interact across the life course to produce divergent welfare outcomes in old age (Bongaarts 2009). A 2024 global analysis across 143 countries using the Global Aging Society Index found that only 5 out of 143 countries scored higher for women than men on ageing adaptation outcomes, underscoring the urgent need for gender-disaggregated, context-sensitive evidence of the kind our study provides (Chen et al., 2025). Recent cross-national evidence underscores how such undifferentiated benefit structures fail to address the intersecting social inequities that shape whether older adults can access adequate support (Bolster-Foucault et al., 2024). In GCC contexts, these dynamics take on distinctive features: rapid socioeconomic development in recent decades has generated large cohort effects in educational attainment, labour market participation, and pension eligibility,

producing older adults populations characterised by substantial within-group variation that conventional descriptive analysis may obscure (Liu et al. 2023).

In Oman specifically, female labour force participation was historically limited, and women's access to formal pension schemes consequently lags behind men's. The SPF's Older adults Benefits Programme was established in recognition of the vulnerability of older citizens, but its uniform benefit structure has not been refined to account for distributional heterogeneity. SPF practitioners interviewed during this study consistently identified a shortage of evidence-based segmentation as a barrier to more responsive service design—a gap that applied clustering analysis is well-positioned to address.

2.2 Clustering Approaches in Older adults Population Research

Unsupervised machine learning, and clustering analysis in particular, offers a principled method for uncovering latent subgroup structures within heterogeneous populations. Unlike traditional descriptive statistics or bivariate analysis, clustering can simultaneously capture non-linear interactions among multiple variables, revealing subgroup profiles that are not apparent from univariate or bivariate approaches (Liu et al. 2023).

Ahn et al. (2020) applied K-Means (k=3) to segment 328 older adults in the United States based on reasons for ageing-in-place, identifying groups distinguished by financial concern, housing comfort, and balanced social wellbeing. Wei and Li (2022) similarly used K-Means to segment a Chinese pension market sample by socioeconomic and social engagement variables, revealing four clusters with different income, health, and loneliness profiles. Elbattah and Molloy (2016) demonstrated the clinical utility of K-Means in segmenting older hip fracture patients by care outcome profiles. Khalid and Prieto-Alhambra (2019) extended this to feature selection and clustering using electronic health records, while Zheng et al. (2023) used hierarchical clustering to segment older adults individuals by thermophysiological demand.

A systematic review by Liu et al. (2023) of 79 clinical segmentation studies found that fewer than 30% included multi-algorithm comparisons—the majority applying a single clustering method. This limits the ability to assess the generalisability and robustness of identified subgroups. The present study directly addresses this gap by applying and comparing five algorithms, including approaches designed for mixed-type data (K-Prototypes, Gower distance), on a single nationally representative dataset.

2.3 The Case for Mixed-Type Clustering Methods

Most standard clustering algorithms, including K-Means and DBSCAN, assume numerical input variables. Demographic and socioeconomic datasets, however, are typically mixed-type—containing both numerical features (such as age) and categorical features (such as gender, marital status, and education level). Applying Euclidean-distance-based methods directly to one-hot encoded categorical data can inflate dimensionality, introduce spurious ordinal relationships, and distort cluster assignments.

Huang (1998) addressed this limitation by proposing K-Prototypes, which combines Euclidean distance for numerical features with a simple matching dissimilarity for categorical features, enabling native handling of mixed data without encoding. Similarly, Gower distance (Gower 1971) computes a normalised per-feature similarity that accommodates both numerical (range-normalised) and categorical (binary match/mismatch) features. The inclusion of both approaches in the present study represents a methodological contribution beyond prior work in this area.

3 Materials and Methods

Figure 1 provides a visual summary of the main steps followed throughout this study, from data collection to cluster interpretation and insights extraction. It outlines the methodology adopted to segment the older adults population in Oman using unsupervised machine learning techniques. The process involves data preparation, application of various clustering algorithms, evaluation of clustering performance, and analysis of the resulting clusters.

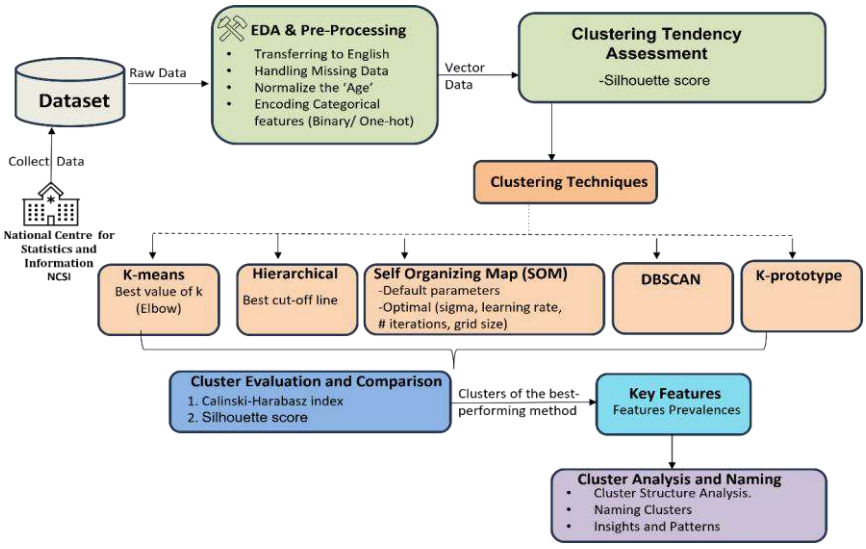


Fig.1 Methodology Overview

3.1 Data Source and Description

The dataset was provided by the National Centre for Statistics and Information (NCSI) of Oman. It comprises 1,700 individual records drawn from a 1% stratified sample of Omani citizens aged 60 years and above, collected cross-sectionally in March 2024. Stratification ensured gender representation across the sample. The dataset contains eight variables (Table 1) capturing both numerical and categorical dimensions of older adults demographic and socioeconomic status.

Table 1 Dataset feature summary

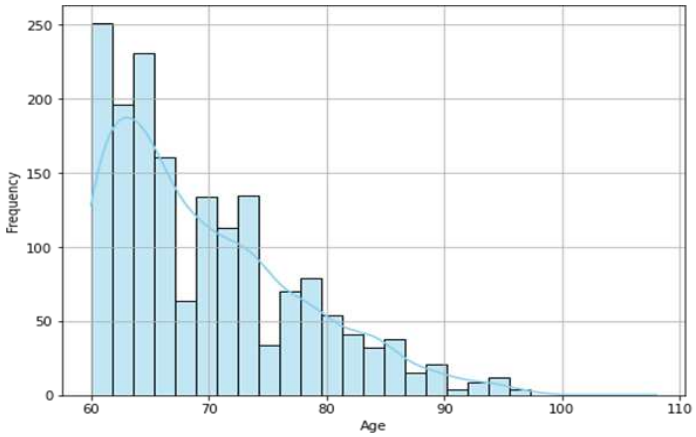
No.	Feature	Type	Description / Categories
1	Age	Numerical	60 years and above
2	Marital Status	Categorical	Married, Single, Divorced, Widowed
3	Gender	Categorical	Male / Female
4	Governorate	Categorical	One of 11 governorates in Oman
5	Wilayat	Categorical	One of 61 wilayats in Oman
6	Educational Level	Categorical	Illiterate; Reads/Writes; Primary; Preparatory; Secondary; Diploma; Adv. Diploma; Bachelor's; Master's; Doctoral
7	Retirement Salary	Categorical	Yes / No
8	Household Unit Owner	Categorical	Yes / No

Feature selection was guided by the existing literature on older adults population segmentation, which consistently identifies age, gender, marital status, educational attainment, income, and residential characteristics as key differentiators of older adults subgroups (Ahn et al. 2020; Wei and Li 2022). These variables collectively capture individual-level socioeconomic resources and broader social context, making them analytically appropriate for clustering-based segmentation.

3.2 Exploratory Data Analysis

Prior to clustering, comprehensive exploratory data analysis (EDA) was conducted. The age distribution was right-skewed (mean = 69.7 years; SD = 8.19; range: 60–108), with the majority of participants in the 60–74 age bracket (Figure1). Gender was approximately balanced, with females comprising 55% of the sample. Married individuals represented the largest marital status category, followed by widowed respondents. Geographically, Al

Batinah North, Al Dakhiliyah, and Muscat were the most represented governorates. As seen in Figure 2, the age distribution of the study's older adults population exhibits a right-skewed pattern, with the majority of people falling between the ages of 60 and 74. The dataset's minimum age is 60, while its highest age is 108. With a standard deviation of 8.19 and an average age of roughly 69.7 years, the sample's age distribution is moderately variable. The majority of participants appear to be in their late 60s and early 70s, as indicated by the 25th percentile of 63 years, the median (50th percentile) of 68 years, and the 75th percentile of 74 years. This distribution validates that the dataset is suitable for segmentation analysis aimed at the older adults and fits the expected demographic profile of an ageing population. Feature selection was guided by the existing literature on older adults population segmentation, which consistently identifies age, gender, marital status, educational attainment, income, and residential characteristics as key differentiators of older adults subgroups (Ahn et al., 2020; Wei & Li, 2022). These variables collectively capture both individual-level socio-economic resources and broader social context, making them analytically appropriate for clustering-based segmentation.



	Age
count	1700.000000
mean	69.723367
std	8.193940
min	60.000000
25%	63.000000
50%	68.000000
75%	74.000000
max	108.000000

Fig.2 Age Distribution (N = 1700)

Cramér's V association analysis revealed moderate-to-strong associations among several categorical variables (Figure 3). Gender showed particularly strong associations with Household Unit Ownership (V = 0.71) and moderate associations with Retirement Salary (V = 0.45) and Marital Status (V = 0.46), suggesting that gender is a central axis of socioeconomic differentiation within the older adults population. Retirement Salary was also moderately associated with both Education Level and Marital Status (V = 0.31 each). Pearson correlation analysis (on ordinally encoded variables) showed that Education Level and Retirement Salary both exhibited moderate negative correlations with Age (r = −0.35 and

−0.21 respectively), indicating that younger older adults individuals tend to have higher educational attainment and are more likely to receive retirement income. A pre-clustering Silhouette Score of 0.423 using Gower distance confirmed sufficient latent structure to justify clustering analysis.

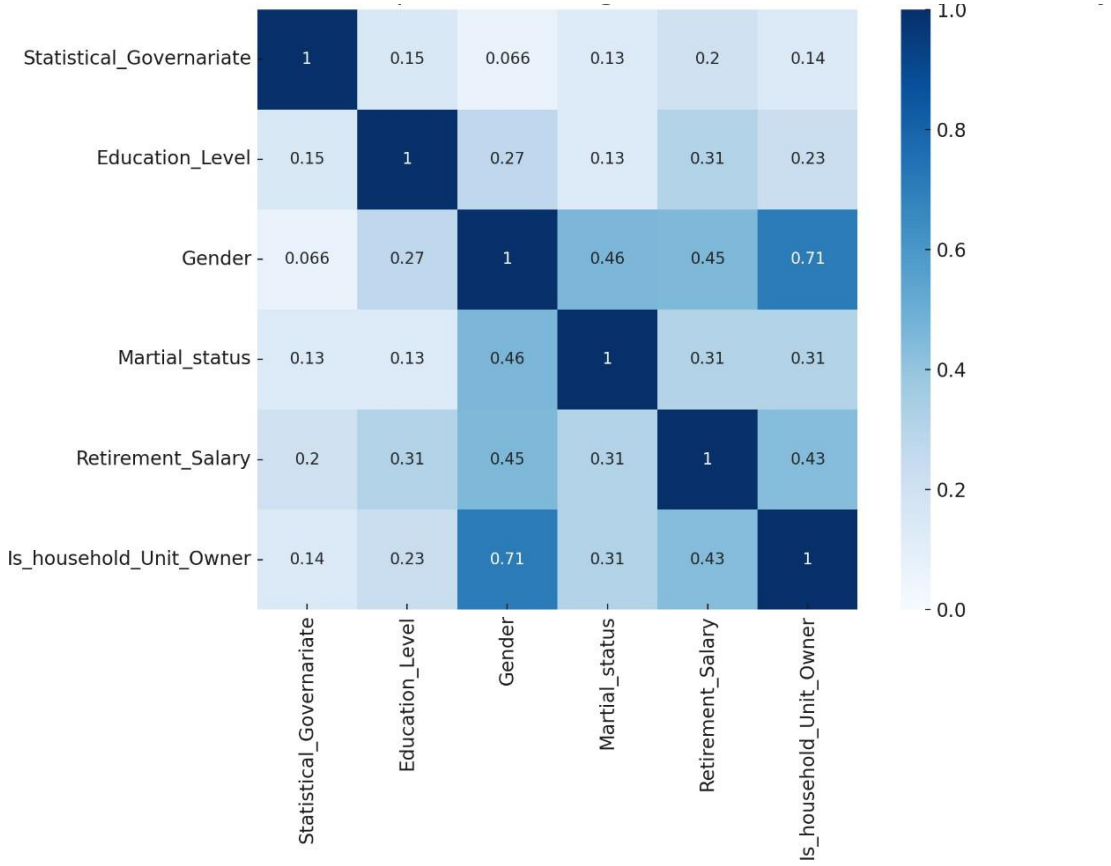


Fig.3 Cramér's V Association Heatmap Between Categorical Variables

3.3 Data Pre-Processing

The preprocessing pipeline involved four sequential stages. First, all variable labels and values were translated from Arabic to English to ensure analytical compatibility. Second, missing values were addressed using contextually appropriate imputation: the single missing Age value was replaced with the sample mean; ten missing Education Level values were imputed using the modal education category within the corresponding age group (reflecting the documented negative correlation, r = −0.35); and the 363 missing Retirement Salary values (21.4% of the sample)—concentrated among older, less-educated individuals, reflecting genuine patterns of informal-sector employment and non-pensionable work histories—were imputed using the modal category within each education level stratum, as recommended by SPF data specialists.

Third, categorical variables were encoded based on cardinality. Binary encoding was applied to features with two categories (Gender, Retirement Salary, and Household Unit Ownership) to avoid dimensionality inflation. One-hot encoding was applied to variables with more than two categories (Marital Status, Education Level, and Governorate) to prevent the inference of spurious ordinal relationships. Fourth, the numerical Age variable was normalised using Min-Max scaling to ensure proportional contribution in distance-based algorithms.

3.4 Clustering Algorithms

Five clustering algorithms were applied, representing distinct methodological families: partition-based (K-Means, K-Prototypes), agglomerative hierarchical (Hierarchical Clustering with Gower distance), model-based (SOM), and density-based (DBSCAN).

K-Means is a partition-based algorithm that iteratively assigns data points to k clusters by minimising within-cluster sum of squared Euclidean distances. The optimal number of clusters was determined through a multi-criteria approach: the Elbow Method, the Calinski-Harabasz Index, and the Silhouette Score, evaluated for k = 2 through 10.

K-Prototypes extends K-Means by combining Euclidean distance for numerical features with simple matching dissimilarity for categorical features, enabling native handling of mixed-type data without requiring encoding. The weighting parameter γ balances the contribution of categorical dissimilarity relative to numerical distance.

Hierarchical Clustering with Gower distance uses agglomerative single linkage with Gower distance as the dissimilarity measure—a metric designed for mixed-type data that computes a normalised per-feature similarity averaged across all variables. The optimal cluster count was determined by evaluating Silhouette Score and Calinski-Harabasz Index across k = 2 to 10.

Self-Organizing Maps (SOM) project high-dimensional input onto a two-dimensional grid while preserving topological relationships. Eleven systematic experiments were conducted varying neighbourhood radius (sigma: 0.5–2.0), grid size (8×8 to 15×15), and learning rate (0.1–0.9), evaluated via Silhouette Score and Calinski-Harabasz Index.

DBSCAN identifies clusters as regions of high density separated by low-density regions. The parameters epsilon (ε: 2.0–4.5) and min_samples (5–15) were systematically varied, with performance evaluated using both internal metrics. (Figure 4)

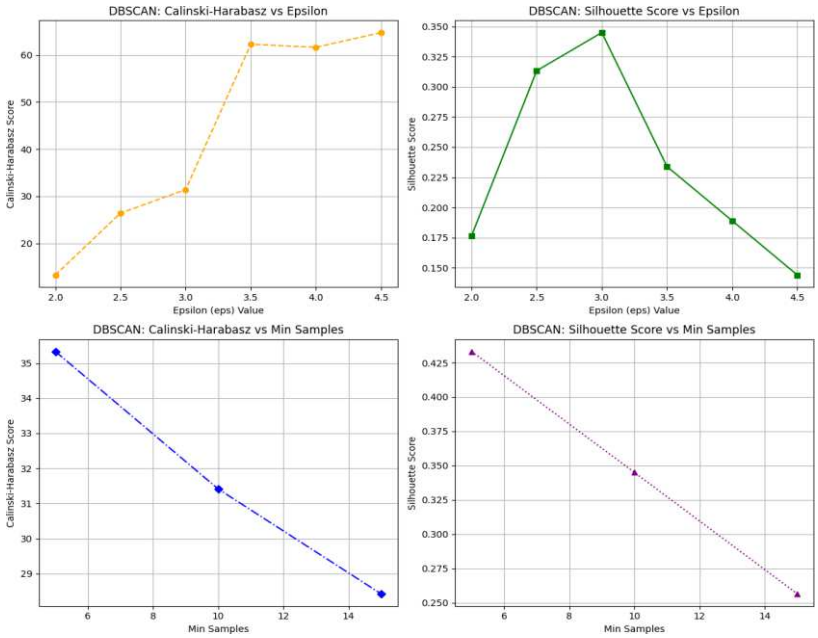


Fig.4 DBSCAN Performance: Effect of Epsilon & Min Samples

3.5 Evaluation Metrics

Clustering performance was assessed using two complementary internal validation metrics. The Calinski-Harabasz (CH) Index measures the ratio of between-cluster variance to within-cluster variance; higher values indicate more compact and well-separated clusters. The Silhouette Score measures how similar each observation is to its own cluster relative to the nearest alternative cluster, with values ranging from −1 (misclassified) to +1 (well-separated). Together, these metrics provide a comprehensive assessment of cluster quality without requiring external labels.

4 Results

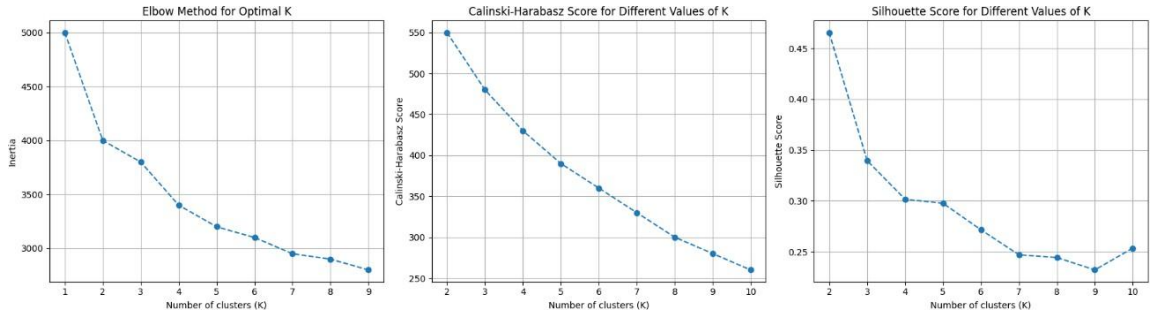


Fig. 5 K-Means Evaluation (Elbow, CH, Silhouette)

4.1 K-Means Clustering

The Elbow Method identified k = 2 as the inflection point, with inertia declining steeply before plateauing (figure 5). This was corroborated by both the CH Index (maximum of 550 at k=2) and the Silhouette Score (peak of approximately 0.47 at k=2). These three independent criteria consistently identified a two-cluster solution as optimal.

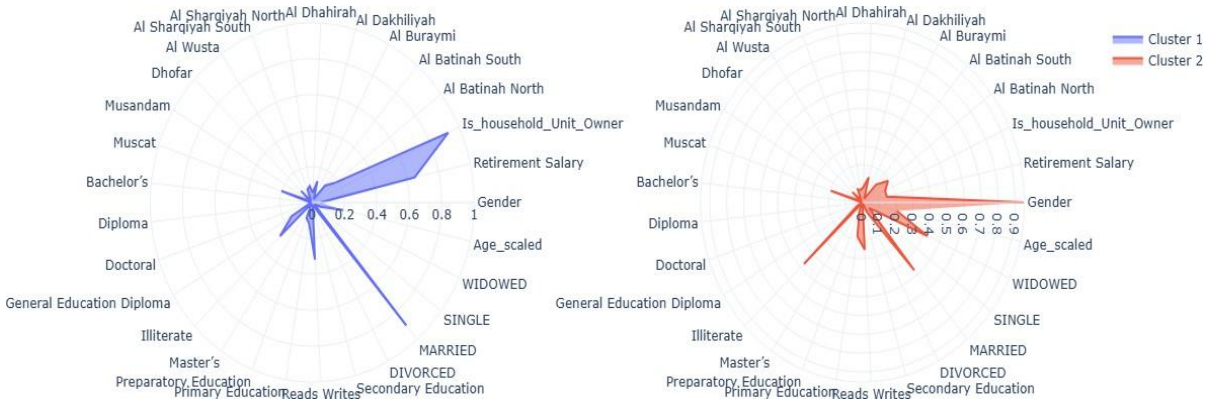


Fig.6 Radar Plot of Clusters Characteristics for K-Means, k=2

Cluster 1 (n = 775; 45.6% of sample) was characterised by predominantly married males (Gender_Male prevalence = 0.95), high household ownership (0.92), substantial retirement salary receipt (0.64), and relatively higher educational attainment (illiteracy prevalence of 0.27). Cluster 2 (n = 925; 54.4%) comprised predominantly females, with markedly lower rates of household ownership (0.14) and retirement salary receipt (0.14), higher rates of illiteracy (0.47) and widowhood (0.42), and lower educational attainment overall (Figure 6). Experiments with k=3 and k=4 yielded progressively more granular segmentation—including younger, more educated males concentrated in Muscat (k=3) and widowed, illiterate females primarily in Al Batinah North (k=4)—but these configurations were not supported by the quantitative evaluation metrics. t-SNE visualisation confirmed

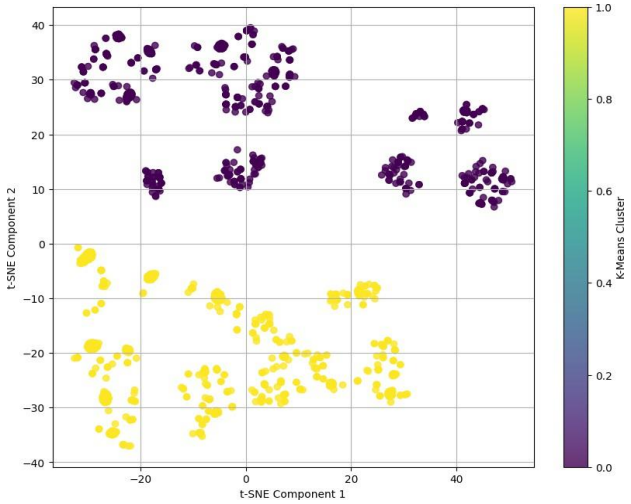


Fig.7 t-SNE Plot with K-Means Clusters

clear spatial separation between the two k=2 clusters, indicating robust internal cohesion (Figure 7).

4.2 K-Prototypes Clustering

Applied with k=2 on the original mixed-type dataset, K-Prototypes identified cluster profiles highly consistent with K-Means results. Cluster 0 (n = 919) was characterised by older individuals (normalised mean age = 0.296) residing predominantly in Al Batinah North, with illiterate educational backgrounds, female gender, married status, and absence of both retirement salary and household ownership. Cluster 1 (n = 781) comprised relatively younger individuals (normalised mean age = 0.192), predominantly from Muscat, with basic literacy (Reads/Writes), male gender, married status, and presence of both retirement income and property ownership. The Gower-based Silhouette Score was 0.34 and the CH Index was 59.59.

4.3 Hierarchical Clustering with Gower Distance

Hierarchical clustering achieved the highest Silhouette Score among all algorithms tested (0.62 at k=2), indicating strong internal cohesion and cluster separation. The CH Index also peaked at k=2 (score = 60). The two identified clusters closely mirrored those produced by K-Means and K-Prototypes: Cluster 0 was predominantly older females from Al Batinah North who were illiterate, married, without retirement salary, and not household owners; Cluster 1 comprised relatively younger males from Muscat with basic literacy, married status, retirement income, and household ownership.

4.4 Self-Organizing Maps

Systematic hyperparameter tuning revealed that a grid size of 10×10, sigma = 1.0, and learning rate = 0.7 produced the best performance (Silhouette Score = 0.36; CH Index = 112.89). However, the resulting U-Matrix showed scattered neuron activations with no clearly defined dense regions, and the clusters lacked interpretable structure. Across all parameter combinations tested, SOM performance remained considerably below that of K-Means, reflecting the algorithm's difficulty in capturing distinct socioeconomic groupings within this moderately low-dimensional, predominantly categorical dataset.

4.5 DBSCAN

DBSCAN performed poorly regardless of parameter configuration. At the best-performing settings (ε = 3.0–3.5; min_samples = 5), the algorithm generated between 32 and 74 clusters—far exceeding any interpretable segmentation—and classified between 128 and 174 observations as noise (7.5–10.2% of the sample). DBSCAN's reliance on density-based separation and its sensitivity to the high effective dimensionality of the one-hot encoded feature space limited its applicability to this dataset.

4.6 Comparative Performance Summary

Table 2 summarises the comparative performance of all five clustering algorithms. Three algorithms—K-Means, K-Prototypes, and Hierarchical Clustering with Gower distance—produced interpretable, policy-relevant clusters; the remaining two exhibited significant limitations in this context.

Table 2 Comparative performance of clustering algorithms

Algorithm	CH Index	Silhouette	Interpretable Clusters
K-Means (k=2)	550	0.47	Yes — Two distinct, meaningful groups
K-Prototypes (k=2)	59.59	0.34	Yes — Consistent with K-Means
Hierarchical (Gower)	60	0.62	Yes — Highest Silhouette Score
SOM (tuned)	90–125*	0.26–0.39*	Weak — Low interpretability
DBSCAN (tuned)	15–65*	0.15–0.43*	No — Over-fragmented, high noise

** Performance varies significantly across parameter settings.*

5 Discussion

5.1 Robustness of the Two-Cluster Solution

A central finding is the remarkable consistency of cluster profiles across three methodologically distinct algorithms: K-Means, K-Prototypes, and Hierarchical Clustering with Gower distance. Despite differences in data preprocessing requirements, distance metrics, and algorithmic mechanisms, all three methods independently converged on near-identical two-cluster solutions. This convergence substantially strengthens confidence in the validity of the identified subgroups and suggests that the two dominant older adults profiles reflect genuine, robust structure in the data rather than artefacts of any particular algorithmic assumption—a standard of multi-method validation rarely achieved in prior clustering studies of older adults populations (cf. Liu et al. 2023).

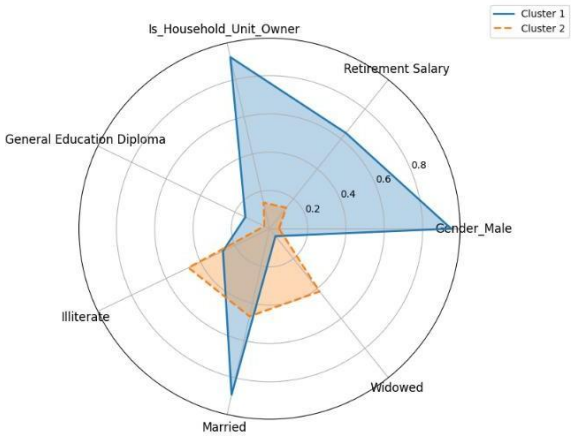


Fig. 8 Feature Prevalence Prevalence (K-Means, k=2)

5.2 Gendered Socioeconomic Disparities and Their Structural Origins

The identified clusters reveal pronounced gendered disparities in socioeconomic security among Oman's older adults. The 'Stable and Financially Secure Males' cluster is marked by high rates of retirement income receipt (64%), homeownership (92%), and relative educational attainment. In contrast, the 'Vulnerable and Disadvantaged Females' cluster exhibits much lower rates of retirement income (14%) and homeownership (14%), alongside high rates of illiteracy (47%) and widowhood (42%).

These disparities reflect structural inequalities embedded in Oman's recent socioeconomic history: lower female labour force participation rates in previous decades,

concentration of female employment in informal or non-pensionable sectors, and prevailing patterns of property ownership that favour men. The over-representation of widowed women in the vulnerable cluster is particularly significant, as widowhood frequently corresponds to the loss of a spouse's income and a potential reduction in household resources—a dynamic well-documented in ageing literature from other national contexts (Wei and Li 2022; Khalid and Prieto-Alhambra 2019).

These findings extend comparable international evidence to a GCC context. Khalid and Prieto-Alhambra (2019) identified older women with high fracture risk as a distinct high-need subgroup in healthcare clustering; Wei and Li (2022) noted that older adults individuals with low income and limited social connections constituted a high-vulnerability cluster requiring specialised support. The present study provides a nationally specific characterisation of gendered vulnerability that is directly actionable within Oman's policy landscape. Older women are substantially less likely than older men to receive a contributory pension — a disparity directly attributable to lower historical labour force participation and interrupted working lives, patterns mirrored in our Oman data (United Nations DESA, 2025).

5.3 Geographic and Educational Dimensions of Vulnerability

Beyond gender, the clusters exhibit consistent geographic and educational differentiation. The vulnerable cluster is disproportionately concentrated in Al Batinah North—one of Oman's largest governorates by rural population—while the stable cluster is more represented in Muscat. This urban-rural gradient suggests that geographic remoteness compounds socioeconomic vulnerability among older adults women, likely reflecting differential access to formal employment, social services, and educational opportunities in earlier life stages.

The high prevalence of illiteracy (47%) in the vulnerable cluster (figure 8), contrasted with the relatively higher educational attainment in the stable cluster, underscores the long-term consequences of historical educational inequalities. Education level's moderate association with retirement income receipt (Cramér's V = 0.31) demonstrates how early-life disadvantages compound across the life course to produce concentrated vulnerability in old age—a well-established life-course dynamic in social gerontology (Bongaarts 2009).

5.4 Algorithm Performance and Methodological Implications

K-Means achieved the highest CH Index (550), reflecting excellent between-group separation, while Hierarchical Clustering with Gower distance achieved the highest

Silhouette Score (0.62), indicating superior within-cluster cohesion. These complementary strengths suggest that both algorithms capture different facets of the clustering structure. K-Means' strong between-cluster separation recommends it for policy segmentation applications where clear group boundaries are needed; Hierarchical Clustering's cohesion-focused advantage may be preferable for applications requiring tight internal group consistency.

The relatively poorer performance of SOM and DBSCAN is instructive for methodological practice. SOM's difficulty likely reflects the relatively low effective dimensionality of the encoded dataset and the predominantly categorical nature of variables—conditions for which SOM's topological mapping offers limited advantage. DBSCAN's failure to produce meaningful clusters reflects the dataset's lack of clear density-based separation in high-dimensional one-hot encoded space—consistent with the known limitations of DBSCAN in high-dimensional settings (Liu et al. 2023). These findings suggest that density-based and neural topological approaches are better suited to datasets with larger sample sizes, richer variable sets, or predominantly continuous features.

5.5 Policy Implications

The segmentation framework produced by this study has direct implications for older adults welfare policy in Oman. The current uniform benefit structure of the Social Protection Fund does not reflect the differential needs and vulnerabilities revealed by this analysis. The following cluster-specific priorities emerge from the findings.

For the 'Vulnerable Females' cluster, immediate policy priorities include: enhanced financial assistance programmes calibrated to individual income and asset status; accessible and affordable housing programmes, including elder-friendly rental options; free literacy and digital skills programmes to address educational disadvantage; and dedicated support for widowed women facing acute income loss. For the 'Stable Males' cluster, targeted interventions should include financial planning services to support long-term welfare maintenance and preventive health and social inclusion programmes.

The Ministry of Housing and the Social Development Ministry can both draw on these cluster profiles to design more equitable and responsive service delivery models. More broadly, the study illustrates how nationally representative survey data, combined with unsupervised machine learning, can generate actionable segmentation frameworks for evidence-based policymaking at relatively low analytical cost.

5.6 Limitations

Several limitations should be acknowledged. The dataset comprises only eight variables and excludes health status indicators, disability levels, and psychological wellbeing measures—variables that prior research suggests are important differentiators of older adults subgroup needs (Khalid and Prieto-Alhambra 2019). The cross-sectional design precludes analysis of how individuals transition between subgroups over time. The substantial proportion of imputed Retirement Salary values (21.4%) may introduce bias, and the choice of single linkage for hierarchical clustering—while justified by its simplicity—is known to be susceptible to chaining effects and may not produce the most compact clusters. Cluster validation relied exclusively on internal metrics; external validation against expert-defined categories or service utilisation data was not available. Finally, the study covers Omani citizens only and does not address the large expatriate older adults population.

6 Conclusion

This study demonstrates the value of unsupervised machine learning as an analytical tool for understanding heterogeneity within ageing populations and informing evidence-based social policy. Using a nationally representative dataset of 1,700 Omani older adults individuals, five clustering algorithms were applied and systematically evaluated. K-Means, K-Prototypes, and Hierarchical Clustering with Gower distance consistently identified two robust, meaningful subgroups—a level of cross-algorithm convergence that substantially strengthens the credibility of the findings.

The two identified clusters reveal significant gendered, educational, and geographic disparities: a financially stable, predominantly male group with formal retirement income and homeownership, and a vulnerable, predominantly female group characterised by widowhood, illiteracy, and financial insecurity. These findings provide a data-driven foundation for developing more targeted, equitable, and need-sensitive support programmes for Oman's growing older adults population, and directly challenge the adequacy of uniform benefit structures.

The study contributes methodologically by demonstrating that K-Means and Hierarchical Clustering with Gower distance outperform SOM and DBSCAN in mixed-type, moderately low-dimensional sociodemographic datasets, and by illustrating the value of multi-algorithm comparison as a validation strategy. Future research should extend this framework by incorporating health-related indicators, longitudinal tracking of cluster

membership transitions, and hybrid supervised-unsupervised modelling approaches to further refine older adults subgroup identification and facilitate direct policy translation.

Acknowledgements

The authors extend sincere gratitude to the National Centre for Statistics and Information (NCSI) of Oman for providing the dataset that underpins this research. Appreciation is also due to the Social Protection Fund (SPF)—particularly Omaima Al Lamki (Social Benefits Specialist) and Badr Al Burdai (Social Benefits Executive)—for their expert insights into older adults support services in Oman.

Declarations

Conflict of Interest The authors declare no conflict of interest.

Funding This research was supported by Sultan Qaboos University. The funding body had no role in the design, analysis, interpretation, or reporting of this study.

Data Availability The dataset used in this study was provided by the National Centre for Statistics and Information (NCSI) of Oman. Requests for access should be directed to NCSI.

Ethics Approval The study used anonymised secondary data provided under a formal data-sharing agreement with NCSI. No individual participants were contacted. Cluster interpretation was conducted with sensitivity to privacy and with care to avoid discriminatory labelling.

References

Ahn, M., Kwon, H. J., & Kang, J. (2020). Supporting aging-in-place well: Findings from a cluster analysis of the reasons for aging-in-place and perceptions of well-being. Journal of Applied Gerontology, 39(1), 3–15. https://doi.org/10.1177/0733464817748779

Bolster-Foucault, C., Vedel, I., Busa, G., Hacker, G., Sourial, N., & Quesnel-Vallée, A. (2024). Social inequity in ageing in place among older adults in Organisation for Economic Cooperation and Development countries: a mixed studies systematic review. Age and Ageing, 53(8), afae166. https://doi.org/10.1093/ageing/afae166

Bongaarts, J. (2009). Human population growth and the demographic transition. Philosophical Transactions of the Royal Society B, 364(1532), 2985–2990.

Elbattah, M., & Molloy, O. (2016). Data-driven patient segmentation using K-means clustering. In Proceedings of the Australasian Computer Science Week Multiconference (ACSW) (pp. 1–8). https://doi.org/10.1145/3014812.3014874

Chen, C., Lim, J., Koh, J. *et al.* (2025). A global analysis of adaptation to societal aging across low-, middle- and high-income countries using the Global Aging Society Index. *Nat Aging* **5**, 113–121. https://doi.org/10.1038/s43587-024-00772-3

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. Biometrics, 27(4), 857–874.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 2(3), 283–304.

Khalid, S., & Prieto-Alhambra, D. (2019). Machine learning for feature selection and cluster analysis in drug utilisation research. Current Epidemiology Reports, 6(3), 364–372. https://doi.org/10.1007/s40471-019-00211-7

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43(1), 59–69.

Liu, P., Wang, Z., Liu, N., & Peres, M. A. (2023). A scoping review of the clinical application of machine learning in data-driven population segmentation analysis. Journal of the American Medical Informatics Association, 30(9), 1573–1582. https://doi.org/10.1093/jamia/ocad111

National Centre for Statistics and Information (NCSI). (2024). The expected number of societal categories in Oman at 2040. https://www.ncsi.gov.om

Social Protection Fund (SPF). (2025). Social Protection Fund — Old Age Benefit. https://www.spf.gov.om/en/protection_programs/old-age-benefit/

Wei, X., & Li, X. (2022). Research on the segmentation of pension market based on K-means clustering algorithm. In Proceedings of the 3rd International Conference on Computer Science and Management Technology (ICCSMT) (pp. 296–299). https://doi.org/10.1109/iccsmt58129.2022.00069

United Nations. (2023). World Social Report 2023: Leaving No One Behind in an Ageing World. United Nations Department of Economic and Social Affairs. **DOI:** https://doi.org/10.18356/9789210019682

United Nations DESA. (2025). Gender matters in an ageing world: The case for gender-responsive policies. UN DESA Policy Brief No. 186. https://desapublications.un.org/policy-briefs/un-desa-policy-brief-no-186-gender-matters-ageing-world-case-gender-responsive

World Bank. (2025). Life expectancy at birth, Oman. https://data.worldbank.org/indicator/SP.DYN.LE00.IN?locations=OM

World Health Organization. (2022). Ageing and health. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

Zheng, G., Wei, C., Yue, X., & Li, K. (2023). Application of hierarchical cluster analysis in age segmentation for thermal comfort differentiation of older adults people in summer. Building and Environment, 234, 109981. https://doi.org/10.1016/j.buildenv.2023.109981