MULTI-PARAMETER CLUSTERING OF CITIZENS IN AN ELECTRONIC DEMOGRAPHY PLATFORM: A MIXED-TYPE DISTANCE APPROACH TO THE DIGITAL DEMOGRAPHIC PORTRAIT

Farhad Firudin YUSIFOV
Elvin Shahin GURBANOV
Institute of Information Technology
Ministry of Science and Education of the Republic of Azerbaijan, Baku, Azerbaijan
E-mail: elvingurbanov@list.ru

Received: *17 April 2026* ***Revised:*** *27 Mai 2026* ***Accepted:*** *25 Juune 2026*

UOT: 004.8:519.237 **DOİ:**https://doi.org/10.32010/CHPO7648

Abstract: Electronic demography platforms describe each citizen by many administrative variables of different kinds, and grouping citizens into a small number of interpretable segments is a natural basis for digital-inclusion policy. The objective of this study is to develop and validate a clustering method for the mixed-type digital demographic portrait, a ten-feature description (five continuous, one ordinal, one nominal and three binary variables) produced by an earlier feature-selection stage of this research. Because the portrait mixes measurement types, the standard k-means algorithm is inappropriate; we therefore base the analysis on the Gower distance and compare partitioning around medoids, the k-prototypes algorithm, average-linkage hierarchical clustering, and a one-hot k-means baseline. The number of segments is chosen with the silhouette width and the Calinski-Harabasz and Davies-Bouldin indices, and stability is assessed by bootstrap resampling using the adjusted Rand index. On 8,620 adults from a synthetic dataset, the validity indices and the stability analysis support a three-segment solution. Partitioning around medoids on the Gower distance gives the most balanced, interpretable and stable segmentation, whereas the k-prototypes and one-hot k-means baselines separate the segments less clearly. The three segments, an urban digitally-engaged group, a rural moderate-digital group and an older low-digital group, show a clear gradient in e-service engagement (67, 56 and 45 per cent). Because the data are synthetic, the results illustrate the method rather than describe a real population. The pipeline provides a reproducible, mixed-type segmentation method for e-demography.

Keywords: e-demography; mixed-type clustering; Gower distance; partitioning around medoids; citizen segmentation; digital divide.

Introduction

Governments increasingly link civil-registration, education, employment, social-protection, health and digital-service records into integrated electronic demography platforms, which give a continuously updated, person-level picture of the population [1,2]. The growth of these data sources has been described as a data revolution for demography, with both opportunities and methodological risks [3,4], and machine-learning models are now applied to national demographic data for prediction and profiling [5]. A recurring need in this setting is segmentation: grouping citizens into a small number of interpretable archetypes that can guide service design and digital-inclusion policy.

Effective segmentation requires two ingredients. The first is a compact, validated description of each citizen. In a preceding stage of the present research programme, 83 administrative variables were reduced to a ten-feature digital demographic portrait using a redundancy-aware feature-selection procedure: the features most informative about e-service engagement were retained, redundant variables and a deliberate negative control were screened out, and the

ranking drew on the minimum-redundancy maximum-relevance principle [6]. The second ingredient is a clustering method suited to that description. Existing population-segmentation studies in health and public administration often apply clustering to a fixed variable set [7,8], and digital-divide research consistently finds that age, education and connectivity structure who uses digital services [9-12]; yet the methodological question of how to cluster a portrait that mixes measurement types is rarely addressed directly.

This is the central difficulty. The portrait contains continuous scores (internet access, digital literacy), an ordinal variable (device richness), a multi-category nominal variable (occupation) and several binary indicators (urban residence, pension status, social-media use). The familiar k-means algorithm assumes Euclidean geometry and equal-scale numeric inputs, so applying it to one-hot-encoded categories distorts distances and over-weights high-cardinality fields. Clustering mixed-type data is a recognised challenge with a dedicated literature [13], and it calls for distance measures and algorithms designed for heterogeneous variables.

The aim of this study is to develop and validate a clustering method for the mixed-type digital demographic portrait, and so to provide the core algorithmic component of a wider research programme on multi-parameter clustering of citizens in e-demography. Four objectives follow from this aim. First, to define a principled distance and a set of candidate clustering algorithms appropriate for mixed-type data. Second, to determine the number of segments supported by internal validity criteria. Third, to compare the algorithms and characterise the resulting segments. Fourth, to assess how stable the segmentation is under resampling. Because the analysis uses synthetic data, every result is a property of the generating process and illustrates the method rather than describing any real population.

Methods

Data and the digital demographic portrait

The problem addressed first is to fix the input on which clustering operates. The analysis uses a probabilistically generated synthetic dataset of 10,000 individuals described by 83 administrative variables; synthetic generation removes any disclosure risk while preserving realistic distributions and dependencies. Because minors are not independent users of public services, the clustering is performed on adults aged 18 and over (n = 8,620). The ten portrait features and their measurement types are listed in Table 1.

Table 1. The ten-feature digital demographic portrait used for clustering, by measurement type.

Feature	Domain	Measurement type	Levels or range
Internet-access score	Digital access	Continuous	3-100
Digital-literacy score	Digital access	Continuous	3-100
Age	Demographic	Continuous	18-84
Daily internet use (hours)	Digital access	Continuous	0-9
Work experience (years)	Employment	Continuous	0-66
Device richness	Digital access	Ordinal	4 ordered levels
Occupation	Employment	Nominal	5 categories
Urban residence	Household	Binary	urban / rural
Pension status	Social protection	Binary	pensioner / not
Social-media use	Digital access	Binary	yes / no

These features were derived in the preceding stage of the research programme rather than chosen by hand: all 83 variables were ranked for their relevance to e-service engagement using several complementary importance measures, the rankings were combined in a redundancy-aware way so that mutually correlated variables were not all retained, and a deliberately inserted

irrelevant variable (blood type) was screened out as a negative control. The smallest feature set that preserved the predictive information about engagement, while still spanning the demographic, household, employment, social-protection and digital-access domains, was kept as the portrait. E-service engagement, a binary indicator of using at least one public e-service, is held out of the clustering and used only afterwards to interpret the segments and to perform a post hoc consistency check. The conclusion of this step is a clean, mixed-type, ten-dimensional description of each adult.

Distance and clustering algorithms

The aim here is to cluster heterogeneous variables without distorting their meaning. Dissimilarity between citizens is measured with the Gower coefficient, which handles mixed types by combining a range-normalised absolute difference for numeric and ordinal variables with simple matching for categorical variables, giving every feature a comparable contribution [14]. On this distance we apply partitioning around medoids (PAM), which represents each segment by an actual citizen and is robust to outliers [15]. We compare three further methods: the k-prototypes algorithm, which is designed for mixed numeric and categorical data and optimises a combined numeric-and-matching cost [16]; average-linkage hierarchical clustering on the same Gower distance; and, as a deliberately naive baseline, k-means applied to standardised numeric features with one-hot-encoded categories. The numeric features are standardised and the ordinal device variable is coded by its order. For the distance-based methods, a fixed random subsample of 3,000 adults was used to keep the full distance matrix computationally tractable, and the scalable methods were additionally run on the complete adult dataset as a sensitivity analysis. The conclusion is a panel of four algorithms spanning distance-based, prototype-based, hierarchical and naive approaches.

Choosing the number of segments and internal validation

The problem is to select the number of segments k without reference to any external label. For k from 2 to 8 and for every algorithm we compute the average silhouette width on the Gower distance, which rewards partitions whose members are closer to their own segment than to the nearest other segment [17]. Two complementary indices computed on the standardised feature space, the Calinski-Harabasz ratio (higher is better) and the Davies-Bouldin index (lower is better), are used to corroborate the choice [18,19]. The conclusion is the value of k that is jointly best across these criteria.

Stability and method agreement

The final methodological problem is whether the segmentation is reproducible. Following the resampling approach to cluster validation [20], we draw repeated bootstrap subsamples, re-cluster each with PAM on the Gower distance, and measure how closely the resampled partition matches the reference partition with the adjusted Rand index, which corrects chance agreement [21]. Mean adjusted Rand values are reported for each k, and the same index is used to quantify agreement between the four algorithms. All analyses were implemented in Python with NumPy, pandas, scikit-learn, the gower and kmodes libraries and Matplotlib, with fixed random seeds, so that the pipeline is designed to be reproducible.

Results

A three-segment solution is supported by the validity criteria

On the Gower distance, the silhouette width for partitioning around medoids peaks at three segments and declines thereafter, and on the standardised feature space the Calinski-Harabasz ratio is maximised and the Davies-Bouldin index minimised at the same value (Table 2, Figure 1). The silhouette curves for the alternative algorithms instead peak at two segments; however, the two-segment solutions separate only a small residual group and are far less informative than the three-segment structure. The silhouette values are modest throughout (about 0.32 at the optimum for partitioning around medoids), which indicates segments that are distinguishable but overlapping rather than sharply separated, an honest reflection of a population that varies continuously in digital capability. Weighing the validity indices, the stability analysis reported below, interpretability and cluster balance together, three segments best describe the portrait.

Table 2. Internal cluster-validity criteria by number of clusters k (adults, random subsample). Silhouette is computed on the Gower distance; the row for the selected k = 3 is shown in bold.

k	Gower + PAM silhouette	k-prototypes silhouette	Hierarchical silhouette	k-means silhouette	Calinski-Harabasz	Davies-Bouldin
2	0.268	0.265	0.458	0.307	282	3.05
3	**0.321**	**0.158**	**0.371**	**0.209**	**838**	**2.08**
4	0.253	0.133	0.297	0.166	730	2.10
5	0.243	0.120	0.263	0.162	589	2.31
6	0.236	0.110	0.270	0.148	484	2.46
7	0.239	0.090	0.282	0.146	433	2.42
8	0.246	0.088	0.297	0.140	415	2.40

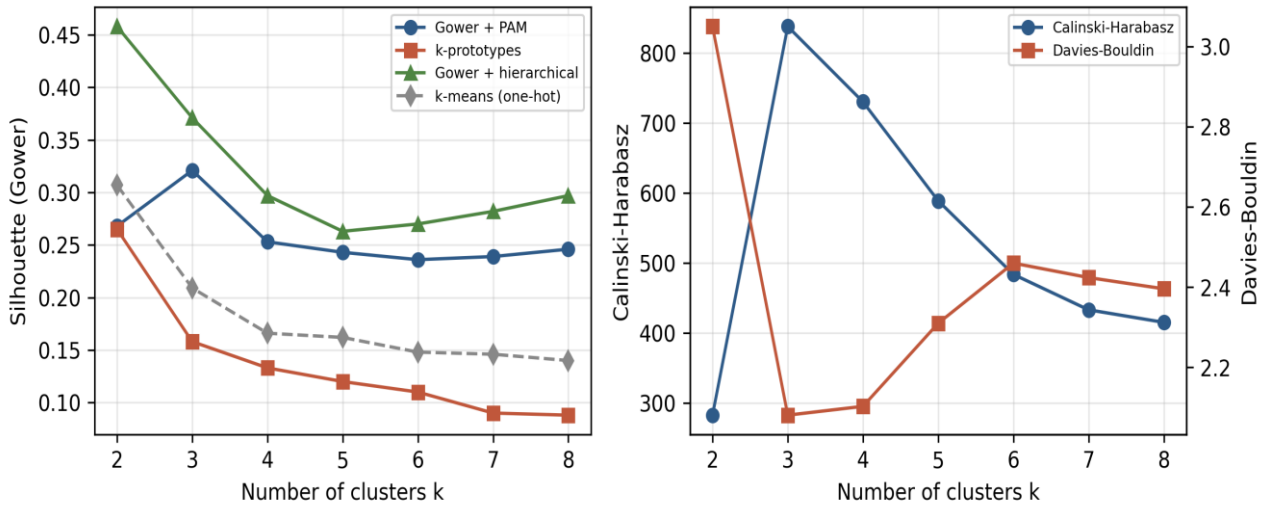

Figure 1. Internal cluster-validity criteria by number of clusters. Left: silhouette width on the Gower distance for the four algorithms. Right: Calinski-Harabasz and Davies-Bouldin indices for partitioning around medoids. The criteria for partitioning around medoids identify three segments.

The three-segment solution is the most stable

The choice of three segments is also the most reproducible. Across bootstrap subsamples, the mean adjusted Rand index between resampled and reference partitions is highest at three segments (about 0.40), whereas two segments are unstable and four or more segments are less reproducible (Table 3, Figure 2). The conclusion is that three segments are a recurring structure rather than an artefact of a single sample, although the absolute adjusted Rand value of about 0.40 indicates only moderate reproducibility.

Table 3. Bootstrap stability of partitioning around medoids by number of clusters k, measured by the mean adjusted Rand index between resampled and reference partitions.

Number of clusters k	Mean adjusted Rand index	Standard deviation
2	0.261	0.294
3	**0.404**	**0.217**
4	0.266	0.095
5	0.326	0.109
6	0.323	0.066

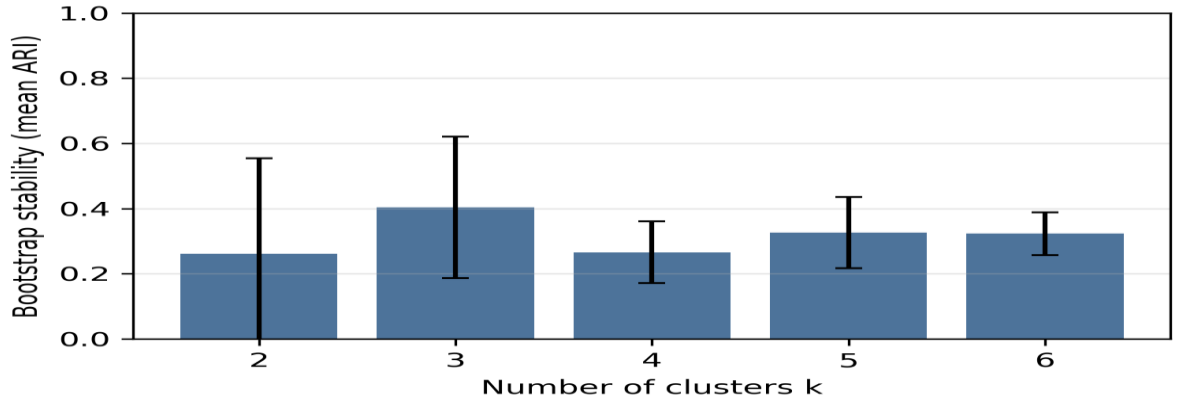

Figure 2. Bootstrap stability (mean adjusted Rand index, with standard-deviation bars) by number of clusters. Stability is highest at three segments.

Representation and algorithm both affect the result

Method choice matters on mixed-type data. At three segments the silhouette differs markedly across methods: average-linkage hierarchical clustering reaches the highest value (0.371), partitioning around medoids follows (0.321), and the one-hot k-means (0.209) and k-prototypes (0.158) baselines are lower (Table 2). The hierarchical silhouette is even higher at two clusters (0.458), but a two-cluster solution splits off only a small residual group and is uninformative as a segmentation. Partitioning around medoids was preferred for reasons beyond the silhouette alone: it produced a balanced and interpretable three-segment solution (Table 4), represented each segment by an actual citizen through its medoid, addressed the mixed-type portrait directly through the Gower distance, and was the most stable under resampling (Section 3.2). Agreement between the algorithms is only moderate (adjusted Rand index between 0.19 and 0.29 at three segments), which shows that the representation and algorithm materially affect the result. The conclusion is that a principled mixed-type distance with medoid-based partitioning is a sound and defensible default here, while the one-hot k-means and k-prototypes baselines separate the segments less well.

Three interpretable citizen archetypes

The three segments admit a clear reading (Table 4, Figure 3). Segment 1, an urban digitally-engaged group, is younger, fully urban, has the highest internet-access and digital-literacy scores and the richest devices, and the highest e-service engagement. Segment 2, a rural moderate-digital group of similar age, is fully rural with moderate digital scores and intermediate engagement. Segment 3, an older low-digital group, is much older, largely composed of pensioners, has the lowest digital scores and devices and the lowest engagement. The held-out engagement rate falls steadily across the three groups, from 67 to 56 to 45 per cent, so the unsupervised segments line up with this held-out outcome and with the well-documented age and connectivity gradients of the digital divide. The two-dimensional projection of the Gower distance shows the older low-digital group separating along the first axis and the urban and rural groups separating along the second (Figure 4). The conclusion is that the portrait supports three demographically coherent, policy-relevant archetypes.

Table 4. Profiles of the three segments (adults, random subsample). E-service engagement is the held-out variable, not used in clustering.

Attribute	Segment 1 (urban engaged)	Segment 2 (rural moderate)	Segment 3 (older low-digital)
Size (n in sample)	1,384	1,170	446
Share of adults (%)	46.1	39.0	14.9
E-service engagement (%)	67.3	55.6	45.1
Mean age (years)	37.2	37.4	70.8

Attribute	Segment 1 (urban engaged)	Segment 2 (rural moderate)	Segment 3 (older low-digital)
Internet-access score (0-100)	64.4	49.5	23.2
Digital-literacy score (0-100)	68.7	61.0	29.9
Daily internet use (hours)	3.7	2.8	1.3
Device richness (0-3)	2.0	1.8	1.1
Urban residence (%)	100	0	56
Pensioner (%)	1	2	87
Social-media use (%)	78	74	39
Dominant occupation	Employed / self-employed	Employed / self-employed	Retired

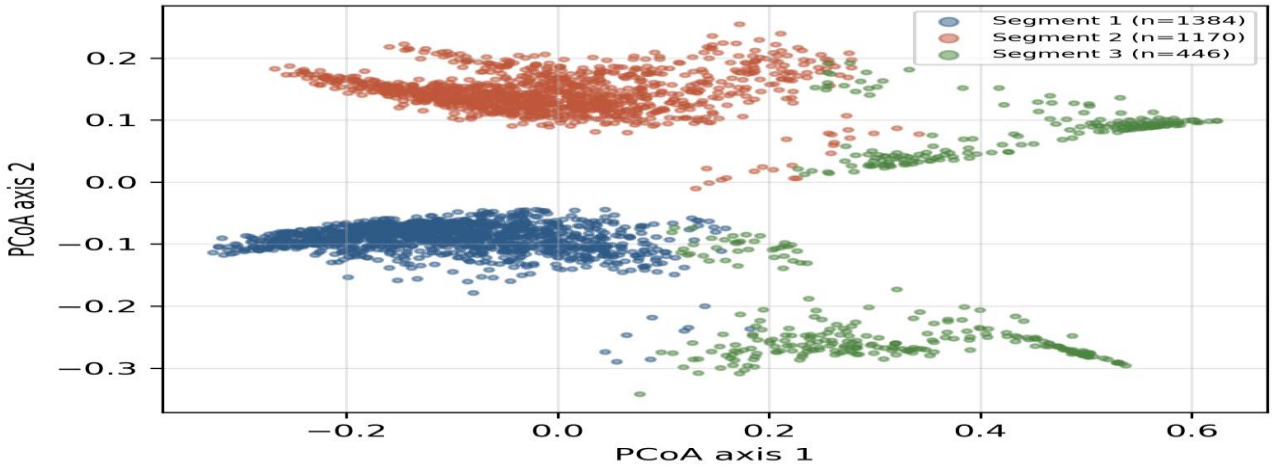

Figure 3. Standardised mean profiles of the three segments across the continuous and ordinal portrait features (z-scores; positive values above the adult average). The older low-digital segment is sharply below average on all digital features and above average on age and work experience.

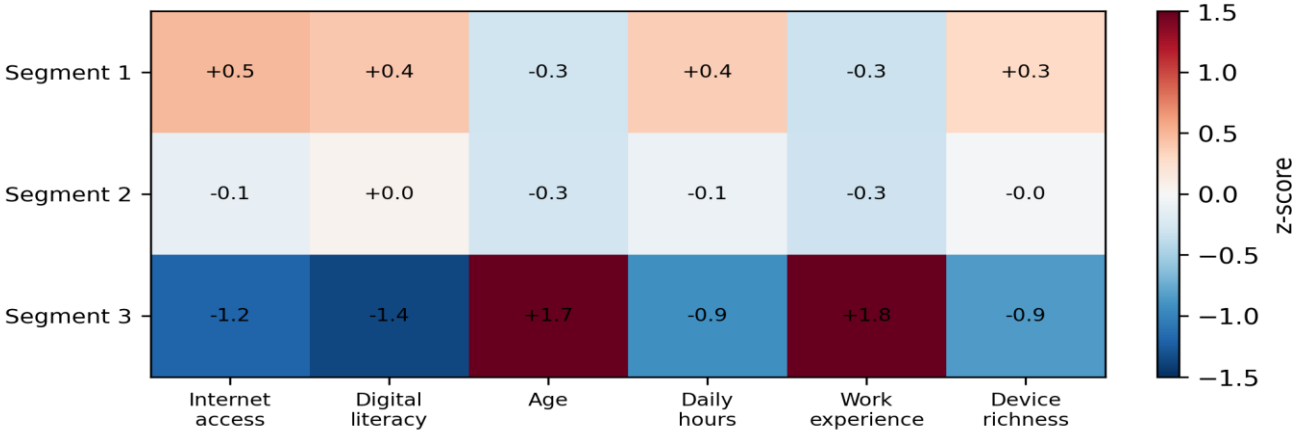

Figure 4. Two-dimensional principal-coordinates projection of the Gower distance, coloured by segment. The first axis separates the older low-digital group; the second axis separates the urban and rural groups.

Discussion

The principal finding is that the mixed-type digital demographic portrait supports three citizen archetypes that are interpretable, coherent and reproducible, but only moderately separated. The validity indices and the stability analysis converge on three segments, and the held-out engagement gradient is consistent with the

segments reflecting real differences in digital participation rather than arbitrary partitions, although, as noted in the limitations, this is a post hoc consistency check rather than independent validation. The three groups, urban digitally-engaged, rural moderate-digital and older low-digital, reproduce the age and connectivity structure that digital-divide research reports repeatedly, including the grey divide between older and younger citizens [9,10], and they resemble the data-driven segments found in population-health studies that cluster linked records [7,8].

A clear methodological message follows from the comparison of algorithms. On data that mix continuous, ordinal, nominal and binary variables, the distance and algorithm are not interchangeable: partitioning around medoids on the Gower distance recovered the most balanced, interpretable and stable structure; average-linkage hierarchical clustering attained a slightly higher silhouette but a less informative partition; and the one-hot k-means and purpose-built k-prototypes baselines separated the segments less well, with k-prototypes giving the lowest silhouette at three segments. The modest agreement between methods reinforces this point. For e-demography portraits, which are inherently mixed-type, a principled dissimilarity such as Gower with medoid-based partitioning is therefore a sounder default than forcing the data into a Euclidean form, which is the main algorithmic contribution of this work [13,16].

The segmentation has direct implications for digital-government practice. The older low-digital group, concentrated among pensioners with low digital-literacy and access scores, is the natural priority for assisted channels and skills support, while the rural moderate-digital group points to connectivity and outreach rather than age as the relevant lever; this is consistent with calls to move digital-inclusion policy from access toward capabilities and to design e-government services around vulnerable groups [11,12]. Because such segments can drive consequential decisions, the workflow is built to be transparent and auditable, with named features and reproducible steps, in line with guidance on explainable and fair machine learning for public policy [22-24]. Fixing the feature set in advance and then clustering with a documented mixed-type method removes a common source of arbitrariness from segmentation pipelines. Several limitations bound these claims. Most importantly, the data are synthetic, so the segment sizes and profiles illustrate the method and must not be read as findings about any real population. The segments are only moderately separated, with silhouette values near 0.32, so they are best understood as overlapping tendencies rather than sharply distinct classes. Because the portrait features were originally selected for their relevance to e-service engagement, the engagement gradient across the segments is a post hoc consistency check rather than an independent external validation. The distance-based analysis uses a random subsample for tractability, with the scalable methods additionally run on the full adult dataset as a sensitivity analysis. Finally, the portrait is tied to a single engagement target inherited from the feature-selection stage; a different target could yield a different portrait and therefore different segments. Applying the method to governed real-register data within a secure environment remains the priority next step.

Within the wider research programme this study is the central algorithmic stage. A first study established the compact mixed-type portrait by feature selection; the present study develops and validates the clustering method that turns that portrait into citizen segments; and a further study can build on these segments to address governance, digital-inclusion targeting and the economic and administrative implications of segment-based service design.

Conclusion

This study developed and validated a clustering method for the mixed-type digital demographic portrait of an electronic demography platform. Using the Gower distance to respect heterogeneous measurement types, it compared partitioning around medoids, k-prototypes, hierarchical clustering and a one-hot k-means baseline, selected the number of segments with the silhouette width and the Calinski-Harabasz and Davies-Bouldin criteria, and assessed reproducibility by bootstrap resampling with the adjusted Rand index. On 8,620 adults from a synthetic dataset, the validity indices and the stability analysis supported a three-segment solution; partitioning around medoids on the Gower distance gave the most balanced, interpretable and stable

segmentation, while the one-hot k-means and k-prototypes baselines separated the segments less well.

The three segments, an urban digitally-engaged group, a rural moderate-digital group and an older low-digital group, are interpretable and consistent with the held-out engagement outcome, with a steady engagement gradient of 67, 56 and 45 per cent, although they overlap rather than separate sharply. Two conclusions carry beyond the synthetic data. Methodologically, clustering an e-demography portrait should use a mixed-type distance and a medoid-based algorithm rather than a Euclidean one, and the number of segments should be fixed by the internal validity criteria together with stability and interpretability rather than by any single index. Substantively, and provisionally, citizen segments are organised mainly by digital capability and age, with residence as a secondary axis. The reproducible pipeline presented here is the core clustering method of a wider programme that proceeds from feature selection through segmentation to governance.

REFERENCES:

1. Yusifov F, Akhundova NE. Analysis of demographic characteristics based on e-demography data. Demography and Social Economy. 2022;47(1):38-54. doi:10.15407/dse2022.01.038.

2. Alguliyev RM, Aliguliyev RM, Imamverdiyev YN, Sukhostat LV. Analysis of the demographic situation in the Republic of Azerbaijan: challenges and opportunities for the country. Electronic Government, an International Journal. 2024;20(2):202-222. doi:10.1504/EG.2024.136982.

3. Kashyap R. Has demography witnessed a data revolution? Promises and pitfalls of a changing data ecosystem. Population Studies. 2021;75(S1):47-75. doi:10.1080/00324728.2021.1969031.

4. Breen CF, Feehan DM. New data sources for demographic research. Population and Development Review. 2025;51(1):539-573. doi:10.1111/padr.12671.

5. Hajirahimova MS, Aliyeva AS. Development of a prediction model on demographic indicators based on machine learning methods: Azerbaijan example. International Journal of Education and Management Engineering. 2023;13(2):1-9. doi:10.5815/ijeme.2023.02.01.

6. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005;27(8):1226-1238. doi:10.1109/TPAMI.2005.159.

7. Pioch C, Henschke C, Lantzsch H, Busse R, Vogt V. Applying a data-driven population segmentation approach in German claims data. BMC Health Services Research. 2023;23:591. doi:10.1186/s12913-023-09620-3.

8. Nnoaham KE, Cann KF. Can cluster analyses of linked healthcare data identify unique population segments in a general practice-registered population? BMC Public Health. 2020;20:798. doi:10.1186/s12889-020-08930-z.

9. Elena-Bucea A, Cruz-Jesus F, Oliveira T, Coelho PS. Assessing the role of age, education, gender and income on the digital divide: evidence for the European Union. Information Systems Frontiers. 2021;23(4):1007-1021. doi:10.1007/s10796-020-10012-9.

10. Mubarak F, Suomi R. Elderly forgotten? Digital exclusion in the information age and the rising grey digital divide. INQUIRY. 2022;59:00469580221096272. doi:10.1177/00469580221096272.

11. Sung WJ, Lee J. A longitudinal study on the diffusion and the divide in the use of e-government services among vulnerable citizens in Korea. Government Information Quarterly. 2024;41(2):101938. doi:10.1016/j.giq.2024.101938.

12. Vassilakopoulou P, Hustad E. Bridging digital divides: a literature review and research agenda for information systems research. Information Systems Frontiers. 2021;25:955-969. doi:10.1007/s10796-020-10096-3.

13. Ahmad A, Khan SS. Survey of state-of-the-art mixed data clustering algorithms. IEEE Access. 2019;7:31883-31902. doi:10.1109/ACCESS.2019.2903568.

14. Gower JC. A general coefficient of similarity and some of its consequences. Biometrics. 1971;27(4):857-874. doi:10.2307/2528823.

15. Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley; 1990. doi:10.1002/9780470316801.

16. Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery. 1998;2(3):283-304. doi:10.1023/A:1009769707641.

17. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20:53-65. doi:10.1016/0377-0427(87)90125-7.

18. Calinski T, Harabasz J. A dendrite method for cluster analysis. Communications in Statistics. 1974;3(1):1-27. doi:10.1080/03610927408827101.

19. Davies DL, Bouldin DW. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1979;PAMI-1(2):224-227. doi:10.1109/TPAMI.1979.4766909.

20. Hennig C. Cluster-wise assessment of cluster stability. Computational Statistics and Data Analysis. 2007;52(1):258-271. doi:10.1016/j.csda.2006.11.025.

21. Hubert L, Arabie P. Comparing partitions. Journal of Classification. 1985;2(1):193-218. doi:10.1007/BF01908075.

22. Amarasinghe K, Rodolfa KT, Lamba H, Ghani R. Explainable machine learning for public policy: use cases, gaps, and research directions. Data and Policy. 2023;5:e5. doi:10.1017/dap.2023.2.

23. Mhasawade V, Zhao Y, Chunara R. Machine learning and algorithmic fairness in public and population health. Nature Machine Intelligence. 2021;3(8):659-666. doi:10.1038/s42256-021-00373-4.

24. Pencheva I, Esteve M, Mikhaylov SJ. Big data and AI, a transformational shift for government. Public Policy and Administration. 2020;35(1):24-44. doi:10.1177/0952076718780537.

VƏTƏNDAŞLARIN ELEKTRON DEMOQRAFİYA PLATFORMASINDA ÇOXPARAMETRLİ KLASTERLƏŞDİRİLMƏSİ: RƏQƏMSAL DEMOQRAFİK PORTRET ÜÇÜN QARIŞIQ TİPLİ MƏSAFƏ YANAŞMASI

Fərhad Firudin YUSİFOV
Elvin Şahin QURBANOV
Azərbaycan Respublikası Elm və Təhsil Nazirliyinin
İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
E-mail: elvingurbanov@list.ru

Xülasə. Elektron demoqrafiya platformalarında hər bir vətəndaş müxtəlif tipli çoxsaylı inzibati göstəricilərlə xarakterizə olunur və vətəndaşların az sayda interpretasiya edilə bilən seqmentlər üzrə qruplaşdırılması rəqəmsal inklüzivlik siyasətinin formalaşdırılması üçün mühüm əsas yaradır. Tədqiqatın məqsədi bu işin əvvəlki mərhələsində xüsusiyyətlərin seçilməsi nəticəsində formalaşdırılmış on göstəricili (beş fasiləsiz, bir sıralı, bir nominal və üç ikili dəyişən) rəqəmsal demoqrafik portret üçün qarışıq tipli verilənlərin klasterləşdirilməsi metodunun işlənib hazırlanması və qiymətləndirilməsidir. Rəqəmsal portret müxtəlif ölçü tiplərini özündə birləşdirdiyindən klassik k-ortalar (k-means) alqoritminin tətbiqi məqsədəuyğun hesab edilmir. Buna görə də analiz Gower məsafəsi əsasında aparılmış, medoidlər ətrafında bölünmə (Partitioning Around Medoids – PAM), k-prototiplər alqoritmi, orta əlaqəli iyerarxik klasterləşdirmə və one-hot kodlaşdırılmış k-ortalar bazis yanaşması müqayisə edilmişdir. Seqmentlərin optimal sayı siluet əmsalı, Kalinski-Harabaş və Devis-Bouldin indeksləri vasitəsilə müəyyən edilmiş, klasterlərin sabitliyi isə düzəldilmiş Rand indeksi əsasında bootstrap təkrar seçmə üsulu ilə qiymətləndirilmişdir. Sintetik verilənlər bazasında yer alan 8620 yetkin fərd üzrə aparılan təhlillər üç seqmentli həllin həm keyfiyyət göstəriciləri, həm də sabitlik baxımından daha məqsədəuyğun olduğunu göstərmişdir. Gower məsafəsinə əsaslanan PAM metodu daha balanslaşdırılmış, interpretasiya edilə bilən və sabit seqmentasiya təmin etmiş, k-prototiplər və one-hot k-ortalar yanaşmaları isə seqmentləri daha zəif fərqləndirmişdir. Müəyyən edilmiş üç seqment — şəhər mənşəli yüksək rəqəmsal fəallığa malik qrup, kənd mənşəli orta rəqəmsal fəallığa malik qrup və yaşlı aşağı rəqəmsal fəallığa malik qrup — elektron xidmətlərdən istifadə səviyyəsində aydın fərqlər nümayiş etdirmişdir (müvafiq olaraq 67%, 56% və 45%). Verilənlər sintetik xarakter daşıdığından əldə olunan nəticələr real əhalinin təsviri deyil, təklif olunan

metodun imkanlarının nümayişi kimi qiymətləndirilməlidir. Təklif olunan yanaşma elektron demoqrafiya sahəsində qarışıq tipli verilənlərin təkrarlana bilən və etibarlı seqmentləşdirilməsi üçün praktik metodoloji çərçivə təqdim edir.
Açar sözlər: elektron demoqrafiya, qarışıq tipli klasterləşdirmə, Gower məsafəsi, medoidlər ətrafında bölünmə, vətəndaş seqmentasiyası, rəqəmsal uçurum.

МНОГОПАРАМЕТРИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ ГРАЖДАН НА ПЛАТФОРМЕ ЭЛЕКТРОННОЙ ДЕМОГРАФИИ: ПОДХОД НА ОСНОВЕ МЕТРИКИ СМЕШАННЫХ ТИПОВ ДАННЫХ ДЛЯ ФОРМИРОВАНИЯ ЦИФРОВОГО ДЕМОГРАФИЧЕСКОГО ПОРТРЕТА

Фархад Фирудин ЮСИФОВ
Эльвин Шахин ГУРБАНОВ
Институт информационных технологий
Министерства науки и образования Азербайджанской Республики, Баку, Азербайджан
E-mail: elvingurbanov@list.ru

Резюме: На платформах электронной демографии каждый гражданин характеризуется множеством административных показателей различной природы, а группировка граждан в ограниченное число интерпретируемых сегментов является важной основой для формирования политики цифровой инклюзии. Целью данного исследования является разработка и валидация метода кластеризации цифрового демографического портрета смешанного типа, включающего десять признаков (пять непрерывных, один порядковый, один номинальный и три бинарных признака), сформированных на предыдущем этапе исследования в результате отбора характеристик. Поскольку цифровой портрет объединяет данные различных типов измерений, применение стандартного алгоритма k-средних является нецелесообразным. В связи с этим анализ основан на метрике расстояния Гауэра, а для сравнения использованы метод Partitioning Around Medoids (PAM), алгоритм k-прототипов, иерархическая кластеризация со средним связыванием и базовый подход k-средних с one-hot кодированием. Оптимальное число сегментов определялось с использованием коэффициента силуэта, индексов Калинского–Харабаша и Дэвиса–Болдина. Устойчивость результатов оценивалась методом бутстреп-перевыборки с использованием скорректированного индекса Рэнда. Исследование, проведённое на синтетическом наборе данных, содержащем сведения о 8620 взрослых гражданах, показало, что наиболее обоснованным является решение с тремя сегментами. Метод PAM на основе расстояния Гауэра обеспечил наиболее сбалансированную, интерпретируемую и устойчивую сегментацию, тогда как алгоритмы k-прототипов и k-средних с one-hot кодированием демонстрировали менее чёткое разделение сегментов. Выделенные три сегмента — городская группа с высокой цифровой активностью, сельская группа со средним уровнем цифровой активности и возрастная группа с низким уровнем цифровой активности — продемонстрировали выраженные различия в использовании электронных услуг (67%, 56% и 45% соответственно). Поскольку использованные данные являются синтетическими, полученные результаты следует рассматривать как иллюстрацию предложенного метода, а не как описание реальной популяции. Разработанный подход представляет собой воспроизводимый метод сегментации смешанных данных для задач электронной демографии.
Ключевые слова: электронная демография, кластеризация смешанных данных, расстояние Гауэра, Partitioning Around Medoids, сегментация граждан, цифровое неравенство.