MACHINE-LEARNING TECHNIQUE FOR DRAFTING REGIONAL POPULATION POLICY: CLUSTER-BASED DEMOGRAPHIC TYPOLOGY OF RURAL MUNICIPALITIES IN THE THREE BALTIC COUNTRIES

Aleksandrs Dahs, Juris Krumins
(University of Latvia, Riga, Latvia)

EXTENDED ABSTRACT

Introduction

Implementation of an evidence-based regional population policy methodology is crucial for achieving regional development goals and monitoring their implementation performance. Previous research (e.g. see Stonawska and Vaishar, 2018; Brezzi *et al*, 2011 or Hasek, 2020) outlines multiple challenges in terms of used methodology for the classification of territorial units according to their demographic characteristics. There is a lack of classification criteria for assessing demographic risks and population development potential for sub-national territorial units. Most used regional demographic classification based on perceived level of urbanization (urban, sub-urban and rural areas) cannot capture the full spectrum of population development aspects.

To ensure implementation of effective and quickly adaptable population policies in the context of diverse demographic situation and limited resources, researchers and policy makers must develop and improve effective tools for differentiation of territorial units, which would allow dividing them into typical groups according to their socio-demographic risks and development potential. This study aims to develop and test a machine-learning technique for categorization of territorial units based on their demographic characteristics that can be applied in drafting regional population policy measures and monitoring their performance over time. Proposed methodology relies on unsupervised non-hierarchical partitioning clustering algorithm. The study focuses on rural municipalities of the three Baltic countries – Estonia, Latvia and Lithuania, which represent diverse sample of regional demographic development modalities.

Data and Methodology

Data are obtained from national statistical databases (Official Statistics Portal of Latvia, 2023; Official Statistics Portal of Lithuania, 2023; and Statistics Estonia, 2023) and attributed to the administrative territorial borders of local municipalities on the 1st July of 2023. Due to small numbers of aggregated data in some administrative units, a five-year period of 2018-2023 is selected for analysis (with the Population Censuses in the middle – 01.01.2021 in Latvia and Lithuania, 31.12.2021 in Estonia). That provides sufficient grounds for capturing and measuring all main demographic parameters for the proof-of-technique for demographic typology. It is also important to note some differences in data collection among the countries. Most noticeably, annual data points for Lithuania are recorded for 1 July, while Latvia and Estonia capture these data for 1 January. However, this presents an insignificant impact on the validity of the study, as the total timeframe of observations is quite wide.

Indicators used in this study include population density (on 01.01.2023. for Latvia and Estonia and 01.07.2023 for Lithuania), share of population (on 01.01.2023. for Latvia and Estonia and 01.07.2023 for Lithuania) in the main aggregated age groups (0-14, 15-64, 65+) and rates of population increase / decrease during 2018-2022 (total and natural change of population, and net migration).

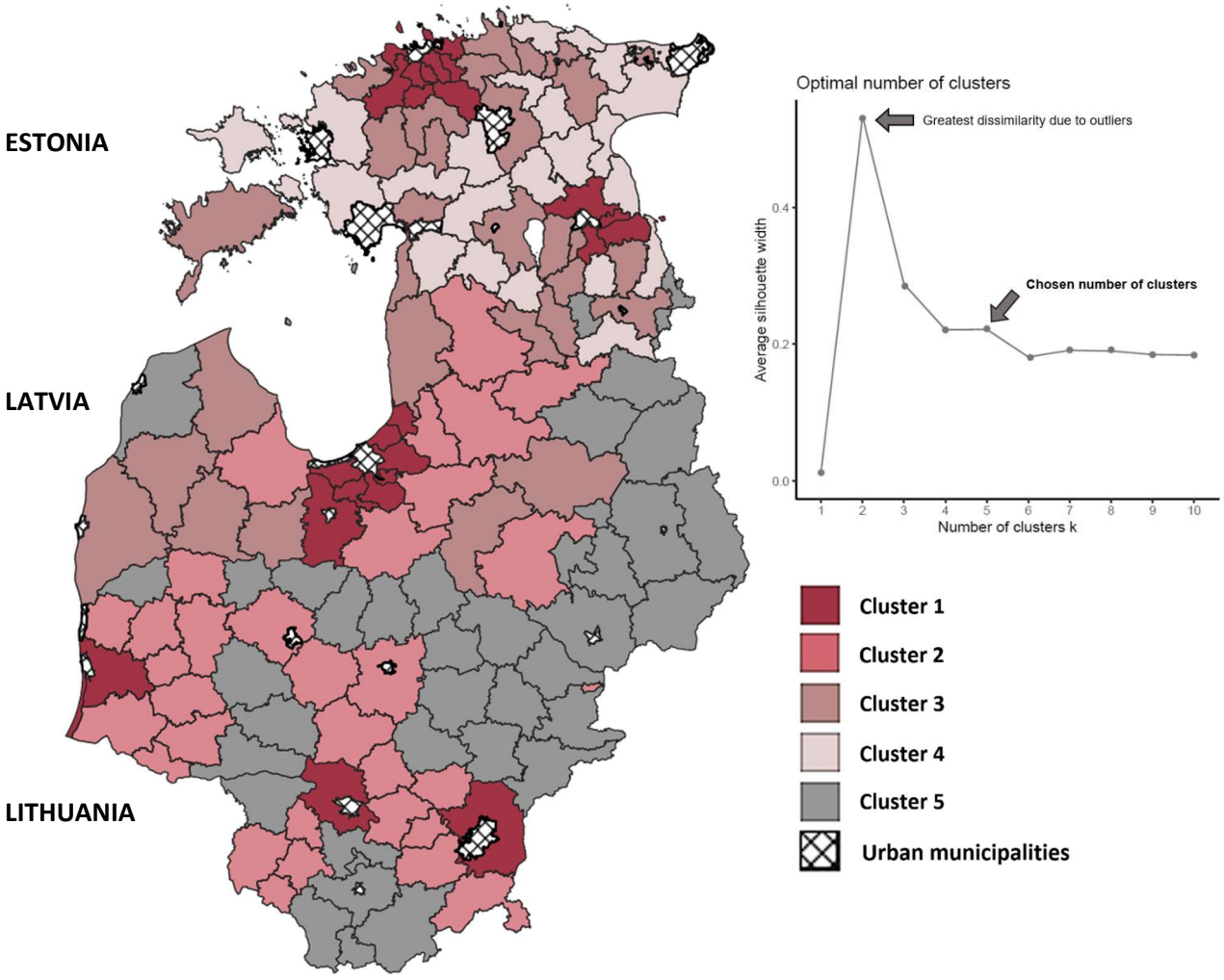
Self-governing municipalities in all three Baltic countries, analysed in this study do not represent the same level of official statistical administrative-territorial division. However, being the smallest self-governing bodies and having approximately similar size, these municipalities serve as comparable test subjects to fulfil the aim of this study. Urban municipalities (municipalities containing only urban territories or considered to be city municipalities) are omitted in this study in order not to disturb the data with too many outliers.

Considering the aim of study and the limited scope of demographic data on municipal level, authors focus on one particular unsupervised machine-learning algorithm - a non-hierarchical partitioning clustering. Clustering approach is a classical first choice as it is not hindered by the data collinearity and is not reliant on any pre-defined measures of similarity between chosen subjects (Peters, 1958). It also allows for great freedom in the number of indicators, types of data used, as well as the desired number of output categories. K-medoids PAM (Partitioning around Medoids) unsupervised clustering algorithm fits this study's aim very well (Kaufman and Rousseeuw, 2009). Although a number of desired clusters can be determined by the policy requirements, there are tools that can suggest the optimal number of groups to be used, for example - average silhouette method, which ranks possible number of groups according to the extent of dissimilarity between them (Rousseeuw, 1987).

Practical use of such technique can allow policy planners to categorise territorial units into a specific number of groups sharing similar demographic indicators and development trends. Targeted policy measures can then be applied to these territorial units based on their assigned group and its main features. By conducting periodical re-assessments using the same indicators, policy planners can then monitor evolution of the territorial units and measure performance of policy interventions.

Results and Discussion

Using data sets and methodology outlined above, one can easily generate a snapshot of demographic typology for rural municipalities in the three Baltic countries (see Figure 1).



Source: authors' construction.

Figure 1. **Results of the unsupervised clustering of rural municipalities of the three Baltic countries based on selected demographic indicators, 2018-2023 (five clusters)**

Analysis of the cluster silhouettes offers multiple options for the number of clusters (top-right diagram in Figure 1). Using two clusters can help singling-out outliers, three clusters would be a good choice for crude classification (good, average, bad situation), while four and five clusters provide greater sensitivity. The dissimilarity of clusters decreases for cluster numbers greater than five. Considering the diverse landscape of the municipalities included in the sample, five-cluster model is selected and used in this study. To describe the produced typology, one should study the representative demographic indicators of typical municipalities from each of the clusters (Table 1).

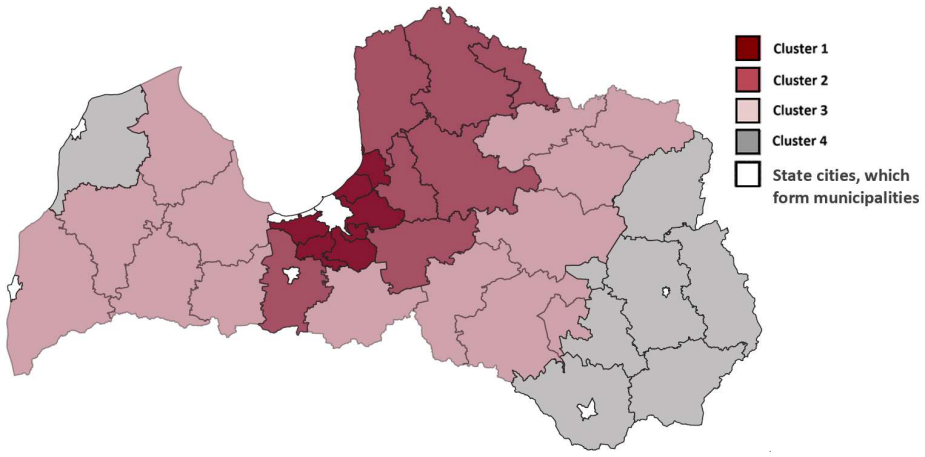
Table 1. **Demographic characteristics of rural municipalities in the three Baltic countries representing five clusters in accordance to the chosen data sets, 2018-2023**

Cluster	Typical municipality representing the cluster	Population denisty (pers/km sq) (2023)	Share of working age population (15-64) (%) (2023)	Share of population below working age (0-14) (%) (2023)	Share of population above working age (65+) (%) (2023)	Total population growth from 2018 to 2022 (%)	Natural population growth from 2018 to 2022 (%)	Change of population due to migration from 2018 to 2022 (%)
1	*LV_Ropažu novads*	63.52	64.97%	20.33%	14.70%	17.37%	0.85%	16.52%
2	*LT_Šilutes rajono savavildybe*	22.48	66.00%	14.20%	19.80%	-5.86%	-3.97%	-1.89%
3	*LV_Dobeles novads*	16.94	62.32%	15.35%	22.33%	-6.39%	-3.92%	-2.46%
4	*EE_Tõrva rural municipality*	9.10	60.06%	14.35%	25.59%	-5.22%	-5.76%	0.53%
5	*LT_Kupiško rajono savivaldybe*	14.93	64.80%	11.40%	23.80%	-9.24%	-6.27%	-2.97%

Source: authors' elaboration

Based on information in Table 1 and considering the available literature (Recano, 2017; Malinen *et al,* 1994; Gabdrakhmanov *et al*, 2017), one can now attempt to describe each cluster and rural municipalities it represents from the demographic policy perspective:

1. Growth areas – municipalities located within the metropolitan areas of big cities and urban centres, showing high population density, stable age structure and positive population growth.
2. Demographically stagnant areas – municipalities with average population density, high share of working age residents and insignificant population decline. Characteristic for economically active regions Latvia and Lithuania.
3. Areas of out-migration – municipalities with average population density, relatively high demographic burden and noticeable negative decline of population due to both negative natural increase and net migration. Characteristic for Latvia.
4. Areas of negative natural increase – municipalities with low population density, high share of senior population and high rate of population decline due to negative natural increase. Characteristic for rural areas of Estonia.
5. Demographic crisis areas – municipalities with low population density, very low share of young residents and extremely high levels of depopulation due to both natural movement and negative net migration. Mostly present in peripheral areas of Latvia and Lithuania, although Lithuanian municipalities in this group show less extremes.



Source: authors' construction.

Figure 2. **Results of the unsupervised clustering of Latvian municipalities based on selected demographic indicators, 2018-2021 (four clusters)**

The same approach applied on a national level can produce clearer results with smaller number of clusters (e.g. see Dahs *et al*, 2021). Figure 2 provides an example of the similar clustering approach for Latvian rural municipalities by using only four clusters and adjusted set of data. In this example, clustering parameters include population density in 2021 (residents per square kilometre), natural population growth (%) and net migration (%) between 2018 and 2021, as well as demographic load in 2021, and number of children (0-14) per 100 senior (64+) residents in 2021.

Provided examples demonstrate that with sufficient data, unsupervised machine-learning tools can be beneficial for drafting and monitoring regional population policy measures. Algorithms like PAM clustering can be used for efficient classification of territorial units according to their demographic characteristics. Using such an approach for smaller datasets (e.g., national, or regional level), provides more sensitive results even with a smaller number of clusters. This study shows that suggested technique provides informative and actionable results useful for policy planners.

Acknowledgements. This study was supported by the National Research Programme "Letonica for the development of Latvian and European society" Project No. VPP Letonika-2021/4-0002 "New solutions in the study of demographic and migration processes for the development of the Latvian and European knowledge society".

Bibliography *(extended abstract only)*

- Brezzi, M., Dijkstra, L. and Ruiz, V., 2011. OECD extended regional typology: the economic performance of remote rural regions. OECD working papers. Available at: https://www.oecd-ilibrary.org/governance/oecd-extended-regional-typology_5kg6z83tw7f4-en
- Dahs, A., Berzins, A. and Krumins, J., 2021. Challenges of Depopulation in Latvia's Rural Areas. In Economic Science for Rural Development Conference Proceedings, Issue 55, pp. 535-545.
- Gabdrakhmanov, N.K., Bagautdinova, N.G., Polyakova, O.V., 2017. Demographic potential of the region: spatial analysis. *The Turkish Online Journal of Design, Art and Communication - Special Edition,* pp. 1700-1705.
- Hasek, O., 2020. Regionální diferenciace plodnosti podle typologie venkova. (The Regional Differentiation of Fertility by Rural Typology in Czechia). *Demografie*, Volume 62, pp. 3-13.
- Kaufman, L. and Rousseeuw, P.J., 2009. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.
- Malinen, P., Keränen, R., Keränen, H., 1994. Rural area typology in Finland - a tool for rural policy. University of Oulu, Research Institute of Northern Finland Research Reports 123. Available at: https://jyu.finna.fi/Record/jykdok.487435?lng=en-gb
- Official Statistics Portal of Latvia, 2023. Population statistics. Available at: https://stat.gov.lv/lv/statistikas-temas/iedzivotaji
- Official Statistics Portal of Lithuania, 2023. Population statistics. Available at: https://osp.stat.gov.lt/statistiniu-rodikliu-analize#/
- Peters, W. S., 1958. Cluster analysis in urban demography. *Social Forces*, Volume 37, pp. 38-48.
- Recano, J., 2017. La sostenibilidad demográfica de la España vacía (The demographic sustainability of empty Spain). *Perspectives demogràfiques*, Issue 7, pp.1-4.
- Rousseeuw, P. J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, Volume 20, pp. 53-65.
- Statistics Estonia, 2023. Population database. Available at: https://andmed.stat.ee/en/stat
- Stonawska, K., Vaishar, A., 2018. Differentiation and Typology of the Moravian Countryside. *European Countryside*, Volume 10, pp. 127-140.