ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - XIII^th SISMEC Congress - Vol. 20 Suppl. 1

Evaluating the Use of Cluster Analysis to Detect and Correct Selection Bias in Single-Center Observational Research on Voluntary Terminations of Pregnancy

Letizia Lorusso[1], Nicola Bartolomeo[2], Paolo Trerotoli[2]

(1) School of Medical Statistics and Biometry, Interdisciplinary Department of Medicine, University of Bari "Aldo Moro", Bari, Italy
(2) Interdisciplinary Department of Medicine, University of Bari "Aldo Moro", Bari, Italy

CORRESPONDING AUTHOR: *Lorusso L., letizia.lorusso08@gmail.com*

INTRODUCTION

Voluntary termination of pregnancy (VTP) has always been debated within the general population. The circumstances leading women to choose VTP are multifaceted, and it remains challenging to identify patterns leading women to opt for this procedure [1, 2]. Up to now, numerous studies have focused on characterizing demographic or psychological profiles of women within specific geographic regions [3, 4]. Due to the sensitive nature of the decision and privacy requirements, studying population clusters in VTP presents inherent methodological limitations. While these challenges can be partially mitigated, they risk a potential selection bias during cohort formation. In these cases, a monocentric study may introduce biases that could arise primarily from the selection of the study centre and secondarily from the difficulty of properly recruiting a sample representative of the entire population.

In this context, cluster analysis conducted first on the general population and subsequently applied to the selected sample could clarify the nature of the selection process that occurred, thereby enabling better estimation of the final conclusions [1, 2]. In general, clustering demonstrated efficiency and feasibility to handle large datasets [5, 6]. In particular, the use of hierarchical clustering procedure is able to choose the optimal number of clusters in the population with a readily available stopping rules, without a priori selection of number of clusters [7]. This approach aligns with contemporary efforts to leverage data-driven methodologies in reproductive health research while addressing ethical considerations inherent to studies of private medical decisions.

OBJECTIVE

The aim is to evaluate the effectiveness of a clustering method in detecting selection bias within a sample from a monocentric observational study, to generalize the findings from the small sample to the broader population and draw valid inferences from the results.

METHODS

We used a retrospective observational database of 6559 women that carried out VTP in Apulia, from January 2023 to January 2024. Data was collected according to the regulations established by the National Surveillance System and were taken from the Health Information System of the Apulia region. The second database was collected, as a cross-sectional, observational study, conducted on 122 women > 18 years who underwent VTP in the Family Planning Unit of the "Di Venere-Fallacara" hospital, between November 2023 and January 2024. The assumption made were that the women of the selected sample are drawn from the same population.

Listwise deletion was applied, to handle missing value. A Multiple Correspondence Analysis (MCA) was first conducted on the population sample to identify the structure of relationships among variables, reducing dimensionality to two principal dimensions. The socio-demographic characteristics and the characteristic of the VTP event were chosen as variables.

DOI: 10.54103/2282-0930/29231

We did a hierarchical cluster analysis using Hierarchical Clustering on Principle Components (HCPC) procedure [8, 9]. We use the Euclidean metric for calculating distances between observations and the Ward's method. The procedure was applied only on the population dataset. When the cluster has been defined, we predicted the MCA output on the selected sample [11, 12].

We selected 10 casual sample of the 111 women from the entire population to compare differences in percentage among clusters. Chi-square tests were employed to compare the proportion of women in each cluster in both samples. We use the z-tests to account the difference in proportions among clusters between the selected sample of women and the population sample [13].

Results with a two-sided p-value < 0.05 were considered statistically significant. The statistical comparisons were done using SAS/STAT® Statistics version 9.4. The cluster analyses were developed using R software version 4.5 [8, 9, 14, 15].

RESULTS

Following listwise deletion, MCA was conducted on a total of 6353 women. The total women in the selected sample with no missing value were 111. The optimal number of clusters was determined using hierarchical clustering procedures implemented through built-in package functions. The p-values across all examined variables are highly significant ($p << 0.001$), providing robust evidence that the HCPC analysis has identified clinically meaningful clusters, representing authentic subgroups within the population. Table 1 shows the percentage distribution of women across cluster.

Cluster 1 is overrepresented in both samples and subsamples (38% vs. 24%, median 41%). Cluster 2 shows minimal deviation between population and sample (8% vs. 9%), but the median percentage in 10 random subsamples deviates substantially (26%). Cluster 3 is significantly overrepresented in the sample selected (53% vs. 28%, median 27%). Cluster 4 is severely underrepresented in both the samples (0.9% vs. 39%, median 8%).

CONCLUSIONS

The selected sample overrepresents Cluster 3 (53% vs. 28%) and nearly excludes Cluster 4 (0.9% vs. 39%). Random subsamples did not align with population proportions, suggesting possible random deviations in the selected sample or from sample size. The observed selection bias might originate from the a priori selection in the monocentric study design. This methodological limitation introduces systematic differences between the sampled cohort and the target population, particularly affecting the generalizability of findings to underrepresented clusters. Beyond ensuring transparency regarding the sample's generalizability, a solution involves implementing covariate balancing techniques [16, 17], balancing alongside synthetic oversampling methods [19] to address overrepresentation of specific variables. This dual approach aligns with causal inference frameworks while mitigating selection bias inherent in monocentric observational designs [20, 21].

REFERENCES

1. Lapresa-Alcalde M. V., Cubo A. M., Alonso-Sardón M. et al., «Reproductive Health Practices in Spanish Women Who Underwent Voluntary Termination of Pregnancy», Diseases, 2023 Mar; vol. 11, no. 1: 1, doi: 10.3390/diseases11010037.
2. Tesema G. A., Mekonnen T. H. and Teshale A. B., «Spatial distribution and determinants of abortion among reproductive age women in Ethiopia, evidence from Ethiopian Demographic and Health Survey 2016 data: Spatial and mixed-effect analysis», PloS One, 2020; vol. 15, no. 6, p. e0235382, doi: 10.1371/journal.pone.0235382.
3. Sánchez-Páez D. A. and Ortega J. A., «Reported patterns of pregnancy termination from Demographic and Health Surveys», PloS One, 2019; vol. 14, no. 8, doi: 10.1371/journal.pone.0221178.
4. Ohashi Y., Takegata M., Takeda S. et al., «Is Your Pregnancy Unwanted or Unhappy? Psychological Correlates of a Cluster of Pregnant Women Who Need Professional Care», Healthcare Basel Switzerland, Aug. 2023; vol. 11, no. 15: 15, doi: 10.3390/healthcare11152196.
5. Nnoaham K. E., Cann K. F., «Can cluster analyses of linked healthcare data identify unique population segments in a general practice-registered population?», BMC Public Health, May 2020; vol. 20, no. 1: 1, doi: 10.1186/s12889-020-08930-z.
6. Croezen, S., Haveman-Nies A., Alvarado V. J., et al., «Characterization of different groups of elderly according to social engagement activity patterns», J. Nutr. Health Aging, Nov. 2009, vol. 13, no. 9, pp. 776–781, doi: 10.1007/s12603-009-0213-8.
7. Caliński T., Harabasz J., «A dendrite method for cluster analysis», Communications in Statistics - Theory and Methods, Jan. 1974; vol. 3, no. 1: pp. 1–27, doi: 10.1080/03610927408827101.
8. «FactoMineR: Exploratory Multivariate Data Analysis with R». Consulted: 15 giugno 2025. [Online]. http://factominer.free.fr/
9. Lê S., Josse J. and Husson F., «FactoMineR: An R package for multivariate analysis», Journal of Statistical Software, mar. 2008; vol. 25, no. 1: pp. 1-18.
10. Husson F., Josse J. and Pagès J., «Principal component methods - hierarchical clustering - partitional clustering: why would we need to choose for visualizing data? », Applied Mathematics Department, set. 2010; vol. 17.
11. Patra B. Kr., Nandi S. and Viswanath P., «A distance based clustering method for arbitrary shaped clusters in large datasets», Pattern Recognit., 2011 Dec, vol. 44, no. 12, doi: 10.1016/j.patcog.2011.04.027
12. Wu W. et al., «Multiview Cluster Analysis Identifies Variable Corticosteroid Response Phenotypes in Severe Asthma», Am. J. Respir. Crit. Care Med., Jun 2109, vol. 199, no. 11, doi: 10.1164/rccm.201808-1543OC

DOI: 10.54103/2282-0930/29231

13. Lakens D., Scheel A. M. and Isager P. M., «Equivalence Testing for Psychological Research: A Tutorial», Adv. Methods Pract. Psychol. Sci., Jun 2108 vol. 1, no. 2, doi: 10.1177/2515245918770963.

14. «R: The R Project for Statistical Computing». Consulted: 15 June 2025. [Online]. https://www.r-project.org/

15. «Extract and Visualize the Results of Multivariate Data Analyses». Consulted: 15 June 2025. [Online]. https://rpkgs.datanovia.com/factoextra/

16. Hainmueller J., "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies," Political Analysis, Jan. 2012, vol. 20, no. 1, pp. 25–46, doi: 10.1093/pan/mpr025.

17. Austin P. C., "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies," Multivariate Behav Res, May 2011, vol. 46, no. 3, pp. 399–424, , doi: 10.1080/00273171.2011.568786.

18. J. M. Robins, M. A. Hernán, and B. Brumback, "Marginal structural models and causal inference in epidemiology," Epidemiology, Sep. 2000, vol. 11, no. 5, pp. 550–560, , doi: 10.1097/00001648-200009000-00011.

19. SMOTE: Synthetic Minority Over-sampling Technique," ResearchGate, doi: 10.1613/jair.953.

20. E. A. Stuart, "Matching methods for causal inference: A review and a look forward," Stat Sci, Feb. 2010, vol. 25, no. 1, pp. 1–21, doi: 10.1214/09-STS313.

21. "Learning from Imbalanced Data Sets | SpringerLink." Accessed: Jun. 15, 2025. [Online]. Available: https://link.springer.com/book/10.1007/978-3-319-98074-4

Table 1. Cluster Distribution Comparison: Population, Selected Sample, and Random Subsamples

	GENERAL POPULATION (N=6353)	**SAMPLE SELECTED (N=111)**	**MEDIAN PERCENTAGE OF 10 RANDOMLY GENERATED SUBSAMPLES (N=111)**
Cluster			
1	*1527 (24.0%)*	*42 (37.8%)*	*41.0%*
2	*540 (8.5%)*	*9 (8.1%)*	*26.0%*
3	*1786 (28.1%)*	*59 (53.2%)*	*28.0%*
4	*2500 (39.4%)*	*1 (0.9%)*	*8.1%*

DOI: 10.54103/2282-0930/29231