FACULTY OF SCIENCES
DEPARTMENT OF STATISTICS
COLLECTING AND ANALYZING BIG DATA FOR SOCIAL SCIENCES

Deconstructing the Study Design: Review of a Research Article on Social Determinants as Predictors of Homelessness

Valentin Duprez, Gunho Lee, Lennert Vanhaeren

May 2023

INTRODUCTION

The phenomenon of homelessness is a complex, multidimensional issue with vast societal implications. The homelessness numbers may have reduced over the last couple of years, but the gap between the rich and the poor is increasing by the day. A recent research article tried to investigate the role of social determinants of health as early predictors of homelessness, offering valuable insights. While the authors have shown a lot of competence in their research, some key areas of concern need critical attention. This critical analysis aims to shed light on these issues, primarily focusing on the study's design, reproducibility, and validity.

DATA SOURCE

The first point of concern is the study's design. The authors have used a cross-sectional study design, which is appropriate for the research question. However, the choice of data source and sampling design raises concerns. The researchers rely solely on 2-1-1 San Diego's database, which only covers individuals who have reached out for assistance. By focusing on a sample that actively sought help, the research potentially overlooks those at risk of homelessness who did not or could not contact 2-1-1 San Diego. The latter may be due to a lack of awareness of the service, negative past experiences, or even people who are too ashamed to disclose their situation. This already raises questions about the representativeness of the sample and, consequently, the generalizability of the findings.

Another issue with the Data source concerns the study's reproducibility.In this case, the choice of data source might limit the possibility of replicating the study in other regions or countries that lack similar databases. Reproducibility is an important aspect of robust research, and this limitation has the potential to compromise the research's credibility [2].

Furthermore, the potential bias in the data sample could also influence the reproducibility of the findings. This bias would come from the fact that 71% of the total sample was represented by females against 26% for the males. The authors then mention the following "In San Diego's actual 2019 Homeless Point in Time Count Report, 69% of all recorded homeless were men" and suggest that the latter could point to evidence that men are overcounted in in traditional studies of the homeless that depend on interviews or surveys of those homeless that are observed on the street [3]. While this is a possibility, it has not been demonstrated yet and should thus not be used as a justification for the bias in the sample, ultimately leading to a potential bias in the findings.

The authors do try to acknowledge some of the study's limitations. However, they don't seem to fully address the issue of reproducibility or the potential biases in their sample. This is why a more direct recognition of these limitations and concrete strategies to mitigate them would significantly strengthen the research's validity.

METHODOLOGY

In what concerns the methodology, the research utilized the K-means cluster analysis and decision trees motivated by previous studies [1] as their primary concern aligned with the main research question of the study. Hence, the adoption of the same analyses seems valid, however, the study does not explicitly clarify why the considered methods are more appropriate than the other alternatives, leaving it ambiguous for readers. For instance, the scikit-learn Python package offers various clustering algorithms, but the research does not sufficiently explain why other methods, such as DBSCAN and Hierarchical Clustering, were excluded. It is widely acknowledged that the K-means cluster analysis may be the safest option for cluster analysis, however, testing with multiple methods may provide more interesting insights and give clear evidence why the chosen ones were valid. Especially in unsupervised learning, researchers should be cautious about potential uncertainties.

Moreover, despite the advantages of the elbow method, its drawback lies in the ambiguity of the elbow point selection. Providing a plot of the result would have been helpful to justify the selection. More precisely, it is often not crystal clear to find an elbow point, and hence, measuring the silhouette index alongside the elbow method could extend the depth of the analysis and enable researchers to point out potential differences.

Nonetheless, the integration of unsupervised and supervised learning appears to be a captivating approach. As the study is interested in applying machine learning techniques to the introduced social challenges, the interest would be limited to exhibiting the potential of machine learning techniques. However, it is essential to validate the resulting clusters before incorporating them into a decision tree model, which the article fails to provide clear explanations behind the process. Importantly, if the clusters are erroneous, the interpretability of the decision tree model will be challenging. In short, multiple clustering algorithms could be adopted, followed by a proper validation tool, such as cross-validation, to determine which clusters yield the best outcomes in following the decision tree analysis.

CONCLUSION

In summary, while the study offers valuable insights into the role of social determinants of health in predicting homelessness, the noted methodological issues limit its overall impact. The study's design, lack of reproducibility, and the potential biases within the sample raise significant concerns about the validity of the findings. The choice of analytical methods also demands further justification and potential exploration of alternative approaches. Therefore, the study underscores the importance of rigorous research design, sampling strategies, and methodological transparency for future research in this area. Addressing these limitations can increase its utility for informing policies and practices aimed at preventing homelessness.

References

[1] Randall Kuhn and Dennis Culhane. Applying cluster analysis to test a typology of homelessness by pattern of shelter utilization: Results from the analysis of administrative data. *American journal of community psychology*, 26:207–32, 05 1998.

[2] J. W. Moody, L. A. Keister, and M. C. Ramos. Reproducibility in the social sciences. *Annual Review of Sociology*, 48:65–85, 2022.

[3] Andrea Yoder Clark, Nicole Blumenfeld, Eric Lal, Shikar Darbari, Shiyang Northwood, and Ashkan Wadpey. Using k-means cluster analysis and decision trees to highlight significant factors leading to homelessness. *Mathematics*, 9(17), 2021.