Ph.D. hab. Joanna Dębicka prof. EU

Department of Statistics

University of Economics and Business in Wrocław

February 7, 2024

Review of the habilitation thesis

HIERARCHICAL CLUSTER ANALYSIS OF CATEGORICAL DATA

by Ing. Zdeňk Šulc Ph.D.

I. LEGAL BASIS FOR PREPARING THE REVIEW

1. § 72, paragraph 7 of the Act on Higher Education 111/98 Coll.
2. The request of prof. Ing. Jakub Fischer, Ph.D. (Dean of the Faculty of Informatics and Statistics of the Prague University of Economics and Business), based on the recommendation of the habilitation commission to appoint Ing. Zdeňka Šulce, PhD, assistant professor of the Department of Statistics and Probability, associate professor of Statistics.

II. EVALUATION OF SCIENTIFIC ACHIEVEMENTS

The habilitation thesis "Hierarchical Cluster Analysis of Categorical Data" aims to comprehensively address the topic of Hierarchical cluster analysis (HCA) of qualitative data, encompassing steps such as dissimilarity matrix calculation, application of a given HCA algorithm, and cluster quality evaluation. While covering all three steps, the thesis primarily focused on similarity measures for dissimilarity matrix calculation and assessing internal evaluation criteria for categorical data.

The research topic undertaken by Zdeněk Šulc is highly significant and still requires systematization and methodological verification of the selection of certain methods. Furthermore, from an applied perspective, Hierarchical Cluster Analysis for categorical data finds application in various research areas, including psychology (behavior classification, consumer preference studies, market segmentation), social sciences (segmentation of social groups, public opinion research) and economics (customer classification. In each of these areas, this analysis serves to identify similarities and patterns in data, leading to a better understanding of the phenomena under investigation and enabling more precise decision-making.

In the Introduction, three main research objectives were outlined. The first objective concerned the comparison of similarity measures for categorical data, the second focused on evaluating internal evaluation criteria for categorical data, and the third aimed at developing a second generation of the nomclust package.

In Chapter 1, the habilitation candidate refers directly to the goals set in the Introduction. In three sections, he embeds each goal in the literature on the subject, rightly not omitting his contributions to date in this area of research. Section 1.1 provides a comprehensive overview of clustering techniques for categorical data, considering both model- and distance-based approaches. The Author points out that the most common method currently involves transforming variables into binary form and using HCA with similarity measures for binary variables, such as the Jaccard coefficient. The passage also reviews historical and contemporary similarity measures for binary-coded data and distance-based algorithms like k-modes and k-prototypes clustering for flat clustering in categorical data, highlighting their advantages and drawbacks. Model-based clustering, specifically LCA, is discussed as an alternative approach, as well as the TwoStep method or the ROCK and the COOLCAT algorithms. The Autor also notes ongoing research, particularly in mixed-type data clustering. Section 1.2 discusses cluster assessment in categorical data, focusing on external and internal evaluation criteria. External criteria compare cluster assignments to a known class variable, but they are often impractical for real-world applications. Internal criteria, utilizing intrinsic properties of datasets, are more suitable for practical use. The author mentions various internal evaluation criteria designed to suggest the optimal number of clusters or assess the quality of created clusters. He then stresses that there is a lack of comprehensive analysis and comparison of these criteria for categorical data, unlike in quantitative data studies. Even modifications to criteria dedicated to quantitative data have not been sufficiently examined, leading to limited usage in evaluating categorical clustering results, what emphasizing the need for more research in this area. Section 1.3

provides an overview of software options for categorical data clustering, emphasizing the importance of accessible implementations for broader adoption. The passage highlights the diverse options available in R and a few commercial software for performing categorical data clustering.

The habilitation candidate embedded his research well in the literature on the subject. Chapter 1 demonstrates the meticulous execution of the literature review and comprehensive exploration of available software for the practical implementation of the described methods. This process ultimately confirmed that the habilitation candidate precisely identified research gaps.

Chapter 2 focuses on expanding the research conducted by Šulc and Rezanková in 2019. It introduces 16 similarity measures (other than in the mentioned article) designed for categorical data, specifically for nominal variables with more than two categories that don't require a dummy transformation. The author describes the properties of the introduces the similarity measures and a method for adjusting them to variable weighting and also outlines linkage methods compatible with the similarity measures for categorical data. The chapter lays the groundwork for employing these measures in subsequent research experiments in Chapter 5.

The presentation of introduced similarity measures between categorical data and objects, as well as the concept of agglomerative hierarchical clustering, is clear, although at varying levels of detail for individual stages. The similarity measures are expressed using formal mathematical notation, and the clustering idea (from the moment of the creation of the dissimilarity matrix) is explained based on an empirical example. For the coherence of the entire chapter, it would have been beneficial to create a small example comprising, for instance, 10 objects (n=10) and 2 variables (m=2), and to illustrate selected measures and individual stages based on it. Additionally, I would move Appendix A to this chapter. In Subsection 2.2.2-2.2.4, the ranges of values that the individual similarity measures can take should be checked. For example, the upper limit for the ES measure is $1-\frac{2}{n^2+2}$ (not $1-\frac{n^2}{n^2+2}$ like the Author wrote), and for the OF measure is equal $\frac{1}{1+(\ln n)^2-\ln(n-1)\cdot\ln n}$ (instead of $\frac{1}{1+(\ln 2)^2}$). In summary, if the work will to be published, for example, as a monograph, it would be advisable to refine this chapter.

Chapter 3 concerns the criteria for assessing the selection of the most appropriate division of clusters. It provides a comprehensive overview of the diverse approaches within the field, and it effectively highlights the commonality among these methods by emphasizing the potential divergence of clusters resulting from different algorithms or methods. The chapter underscores the importance of evaluating cluster assignments using appropriate criteria. One of the strengths of this chapter lies in its grounding in the research conducted by Šulc et al. (2018), which is a natural continuation of the research conducted by the Habilitation Candidate. The division of the chapter into two sections is logical and facilitates a structured exploration of external and internal evaluation criteria.

In particular, in Section 3.1 external evaluation criteria commonly used by many researchers (the Rand index and the adjusted Rand index) are presented. Both indexes are based on an approach that can be interpreted as a series of decisions regarding the cluster memberships of pairs of objects. The author adeptly discussed the Rand index and its adjusted version, explaining their application in cluster partition evaluation. The reader can better understand these indices and their practical application in research scenarios through presented examples and detailed computations. In Section 3.2, the Author describes internal evaluation criteria in cluster analysis as crucial for assessing the quality of clustering solutions, primarily focusing on the concepts of compactness and separation of clusters. The Habilitation Candidate categorizes the evaluation criteria into three main types: variability-based (quantifying the variability within clusters such as mutability and entropy), likelihood-based (which maximize the likelihood function while penalizing complex models, aiding in determining the optimal number of clusters such as the Bayesian information criterion (BIC) and the Akaike information criterion (AIC)), and distance-based (evaluate the relative difference between within-cluster and between-cluster distances, providing insights into cluster separation and compactness, such as the silhouette index (SI) and the Dunn index (DI)). It is worth emphasizing that Subsection 3.2.4 introduces novel variability-based evaluation criteria, namely Hartigan mutability (HM) and Hartigan entropy (HE), tailored for categorical data. These criteria, inspired by Hartigan's rule for quantitative data, assess the marginal gain in cluster compactness with increasing cluster numbers, offering valuable insights into optimal cluster selection and cluster quality assessment.

Chapter 4 provides an overview of the second generation of the nomclust R package (introduced by Šulc et al. (2022)). The package encompasses the entire hierarchical clustering process, from dissimilarity matrix computation to evaluating resultant clusters, utilizing specialized similarity measures, clustering methods, and evaluation criteria tailored specifically for categorical data. More precisely, in Section 4.1 details the methods employed in nomclust 2.0, including calculating dissimilarity matrices using various similarity measures tailored for categorical data. Furthermore, the package utilizes agglomerative clustering from the cluster package, employing three linkage methods suitable for categorical data: average, complete, and single linkage methods. The resultant clusters can be evaluated using up to 13 criteria, including new variability-based coefficients introduced in Subsection 3.2.4 (HM and HE). The section provides insights into the properties of these evaluation criteria, highlighting their significance in cluster assessment and decision-making regarding the optimal number of clusters. Moreover, the optimization of dissimilarity matrix calculation, a computationally intensive task in hierarchical clustering, was addressed by implementing critical code segments in C++. An experiment comparing the computational performance of the new package version with its predecessor demonstrated significant speed improvements, with the new version performing on average 154 times faster. Section 4.2 provides a detailed overview and practical demonstration of the nomclust package.

The section is well-organized, with clear subsections delineating different aspects of the nomclust package. Each subsection focuses on a specific aspect of the package's functionality, making it easy for readers to follow along. Moreover, The Autor does an excellent job of explaining the various functions available in the nomclust package, such as nomclust(), nomprox(), and evalclust(). It provides clear syntax and explanations of each function's parameters, making it easy for users to understand how to use them. From an application point of view, using examples throughout the section increases understanding by providing practical demonstrations of how to use the package. Code snippets are provided along with explanations, allowing readers to replicate the analyses on their datasets. This section also discusses graphical functions like eval.plot() and dend.plot() for visualizing evaluation criteria and dendrograms. These visualizations help interpret clustering results and make informed decisions about the number of clusters. In addition comparisons between different clustering approaches, such as weighted and non-weighted clustering, as well as comparisons between different similarity measures, help understand the implications of their choices. The addition of support for standard generic functions such as summary() and print() is very helpful in analysis, which increases the usability of the package by allowing users to quickly evaluate the grouping results and obtain the necessary information.

To sum up, after reading Chapter 4, I admit that the third goal of the habilitation thesis set by the Habilitation Candidate was achieved. The mentioned chapter provides a comprehensive overview of the nomclust package, covering its functionality, usage, and practical applications. The explanations are clear, and the examples are helpful for users looking to apply hierarchical clustering to categorical data in R. Compared to its predecessor (Šulc and ˇ Rezanková, 2015), the second generation of the nomclust package represents a significant advancement by enhancing its capabilities for hierarchical clustering analysis of categorical data and significantly improved computational efficiency. Key enhancements include a comprehensive redesign of evaluation criteria based on variability, likelihood (adjusted for categorical data), and distance. Additionally, performance issues regarding the computational speed of hierarchical clustering were addressed by rewriting critical code segments in C++, resulting in substantial speed improvements. Furthermore, support for generic functions and the capability to visualize dendrograms and evaluation criterion values were incorporated into the package. The nomclust package is accessible on the Comprehensive R Archive Network (CRAN) website.

Chapter 5 comprehensively explores the comparison of similarity measures for categorical data. Based on the updated gen_object() function introduced by the author in the paper (Šulc, 2016), a detailed description of how the datasets for the experiment were generated, including the methodological framework and the rationale behind the chosen parameters, is provided in Section 5.1. This clarity helps readers understand the properties of the datasets and how they contribute to the experimental design. Section 5.2 is devoted to research methodology and outlines the approach to evaluating the similarity measures. Using HCA and internal evaluation criteria provides a robust framework for assessing clustering performance. Incorporating relative (mean ranked scores methodology) and absolute (boxplot assessment) comparison offers a comprehensive analysis from different perspectives. In particular, the Mean ranked scores methodology allows for a relative comparison of similarity measures based on their clustering performance. The methodology is well-explained, including the rationale for its use and the steps involved in its implementation. It is possible to compare and interpret the results by averaging over replications and other properties. The inclusion of boxplots to visually represent the absolute differences among similarity measures adds depth to your analysis. By breaking down criterion values by specific dataset properties, you can identify potential dependencies and variations across different scenarios. This approach enhances the understanding of how similarity measures perform under various conditions. Moreover, selecting PSFE and CU criteria for evaluating clustering quality is justified based on their relevance and widespread usage in previous studies. Section 5.3 presents a comprehensive experiment aimed at evaluating the performance of different similarity measures in creating clusters under various conditions, particularly considering different linkage methods. The experiment is divided into five subsections. The first one investigates the influence of the linkage method on cluster quality. The subsequent three subsections analyze similarity measures regarding the quality of the created clusters separately for each linkage method. The final subsection recommends which combinations of similarity measures and linkage methods are suitable for a specific dataset with specified properties. This section clearly outlines the experiment's aim to determine the conditions under which certain similarity measures generate high-quality clusters. It also highlights the importance of considering dataset properties and researcher decisions in clustering analysis. Novel aspects compared to previous research, such as using boxplot assessment for evaluating similarity measures, investigating mutual interactions between linkage methods and similarity measures, exploring the influence of minimal between-cluster distances, and utilizing a larger number of datasets, are introduced. The experiment is well-structured because each part focuses on another aspect of the evaluation process. This organization helps systematically analyse the influence of linkage methods and similarity measures on cluster quality. Tables and figures effectively present the experimental results. Mean ranked scores (MRS) are provided for each similarity measure and linkage method combination, facilitating comparisons. Additionally, boxplots visually represent the distribution of evaluation criteria values across different linkage methods. The author points to various aspects of the analysis, for example, pointing out differences in the performance of similarity measures between linkage methods and highlighting exceptions or noteworthy observations, such as the poorness of individual measures. The final part of the experiment is dedicated to summarizing the results obtained in the preceding subsections. To assess which combinations of similarity measures and linkage methods yield the best clusters, mean PSFE and CU scores of 48 combinations of 16 similarity measures and three linkage methods were ordered and ranked in descending order. Moreover, recommendations are provided regarding selecting the most suitable similarity measures and linkage methods based on specific dataset properties. These recommendations provide practical guidance for conducting clustering analysis effectively.

In the context of achieving the first goal of the habilitation thesis regarding similarity measures for categorical data, I conclude that it has been achieved. The content of Chapter 2 and Chapter 5 made this possible. In particular, Chapter 5 represents a significant contribution to the field of clustering analysis, providing valuable insights into the performance of similarity measures under different conditions. The methodological rigour and clear presentation of results strengthen the credibility of the experiment's findings.

Chapter 6 focused on the assessment of evaluation criteria for categorical data, which is a key objective of the thesis. It discusses the comparison of selected internal evaluation criteria for categorical data, excluding certain criteria not applicable to the task being perfermed. The chapter aims to analyze the relationships between these criteria from various perspectives and their effectiveness in recommending the optimal number of clusters. The experiment evaluates 11 internal evaluation criteria to aid researchers in selecting suitable criteria for specific situations or identifying criteria that assess cluster quality similarly. Additionally, the Author explores the relationship between choosen internal criteria and the adjusted Rand index, a representative external criterion. The chapter is structured into three sections covering data generation and chois of similarity measures (Section 6.1), comparing and evaluation methods of internal criteria (Section 6.2), and the conducted experiment (Section 6.3).

Section 6.1 provides a detailed description of data generation and the selection of similarity measures within the context of the second experiment of the habilitation thesis. To obtain datasets for the experiment, the Habilitation Candidate utilizes the updated gen_object() function, introduced by himself in paper Šulc (2016) . The author conducted the experiment using various dataset settings, repeating each setting combination 100 times to ensure the reliability of the obtained results. Additionally, he presents the selection of six different similarity measures for categorical data to be compared in the analysis. These described similarity measures are carefully chosen based on their effectiveness, contributing to safeguarding the experiment against the influence of poorly performing measures. The section is clearly written and contains essential information regarding the data preparation process and the selection of similarity measures what providing a solid foundation for the experiment.

Section 6.2 presents methods for assessing evaluation criteria (focusing on internal criteria) their application and interpretation within the context of the thesis's objectives. The Author arbitrary chose: adjusted Rand index (ARI), Pearson correlation coefficient (PCC) and analysis of variance (ANOVA). ARI was described in Section 3.1 and the other two methods are discussed in this section. In my opinion, the presentation of selected measures is oversimplified. It is not about the lack of a formal mathematical notation of methods or interpretation of results. First of all, there is a lack of discussion about the assumptions underlying the selected methods, which are important to ensure the correctness, reliability and effectiveness of the analysis and research results. It also lacks depth in explaining their applicability and limitations within the context of evaluating criteria for categorical data. In the context of PCC, the author noted that "*Nonlinear relationships cannot be expressed by this coefficient.*" but did not emphasize that in order to use this measure, the data should be distributed in a way that suggests a linear relationship between them. Moreover, the data should not contain outliers that may falsely increase or decrease the value of the correlation coefficient and their distribution should be close to normal. Similarly, the discussion on ANOVA lacks critical analysis. While ANOVA is mentioned as a method for analyzing relationships between quantitative and qualitative variables, the explanation provided is basic. There is a lack of discussion on the assumptions underlying ANOVA, such as the homogeneity of variances and normality of residuals, which are crucial for its validity. The described concerns do not imply that the dependency measures chosen by the Author cannot be applied but definitely require verification of criteria and assumptions (for proper interpretation and application in the experiment).

Section 6.3 consists of four subsections, each addressing different aspects of the experiment. The first subsection examines dependencies and differences among the evaluation criteria under investigation. The second subsection evaluates the relationships between internal and external criteria. The third subsection investigates the dependencies of internal evaluation criteria on dataset properties and similarity measures used. Finally, the fourth subsection summarizes the results obtained from the analysis. The experiment was conducted on 81 types of datasets and for each of them HCAs were performed (ranging from two to seven clusters, utilizing six selected similarity measures and the average linkage method, chosen for its ability to provide high-quality clusters). This resulted in a total of 48,600 HCA outputs. Each output was evaluated using 11 internal criteria (excluding WCM and WCE). Evaluation criteria were primarily analyzed using PCC and ANOVA, as well as ARI. In particular, Subsection 6.3.1 presents an analysis of the similarity of evaluation criteria, employing correlation analysis and multidimensional scaling. This subsection provides valuable insights into whether the choosen criteria assess cluster quality similarly. Due to methodological doubts that I mentioned assessing Section 6.2, its seems that analysis could benefit from a more nuanced approach. Considering potential limitations and uncertainties inherent in the methodology, would enhance the rigor and validity of the obtained results. Subsection 6.3.2 presents an analysis of the relationships between internal evaluation criteria and external criteria, focusing on the assessment of cluster quality. The very goal of determining the ability of internal criteria to recognize the original number of clusters in datasets is commendable. However, the obtained results may be somewhat disturbing, as some of they indicate a significant dependence of the effectiveness of the criteria on the original number of clusters, with some criteria working well only in a two-cluster solution. It must be admitted that the Author approaches the obtained results with the honesty of a scientist. He notices this fact and suggests that one should not strictly rely on the results obtained in practical tasks and should always examine at least one solution with fewer and one with a larger number of clusters than the recommended one. The second part of this subsection examines the relationship between internal criteria and ARI. This part provides a more deeper analysis especially in case of relationship between the BIC criterion and ARI. The observed dependence of the BIC value on the number of variables and categories directly introduces the topic of the next subsection. Subsection 6.3.3 provides valuable discussion on the dependencies of evaluation criteria on dataset properties and similarity measures. The subsection appropriately addresses the importance of understanding these dependencies for accurate interpretation of clustering results. One strength of this part of habilitation thesis lies in its thorough exploration of various dataset properties and their impact on evaluation criteria values. The analysis of dependencies on the number of clusters, variables, and categories provides a comprehensive understanding of how these factors influence the assessment of clustering quality. The final subsection summarizes an experiment that compares 11 internal evaluation criteria for categorical data, with a focus on assessing cluster quality and determining the optimal cluster number. It identifies correlations between specific criteria and recommends task-specific metrics. While acknowledging the absence of a universal criterion, the author discusses the strengths and limitations of individual metrics, including newly proposed ones, which can assist researchers in selecting suitable criteria.

Overall, Chapter 6 provides valuable insights while it is clear that easily interpretable, and unambiguous results cannot be expected solely from empirical (simulation) research. Therefore there remains potential for enhancing the depth of analysis, interpreting results, and considering potential non-linear dependencies. Future research and clarification of these aspects could enhance the robustness and utility of the obtained results.

The second goal was realized in Chapter 3, where the criteria were introduced and categorized, and Chapter 6 presented an experiment on 8,100 generated datasets. Despite the methodological doubts regarding the choice of dependency measures, the experiment allowed for the assessment of the effectiveness of the criteria in identifying the optimal number of classes and the dependence of the criteria on the properties of the dataset. This also highlights the complexity of the issues discussed in the habilitation thesis.

III QUESTIONS FOR DISCUSION

1. The distribution of variable value frequencies affects the choice of distance measure in cluster analysis, particularly with categorical data. If the frequency distribution of categorical variable values is uneven (for example, when some categories occur much more frequently than others), selecting a distance measure that considers this unevenness is crucial. In the literature, a simple structure where the structure indices of all categories are equal is called an egalitarian distribution (a distribution completely devoid of inequality). Structural diversity can be measured using a selected inequality structure index, determining the distance of a given structure from a structure with zero diversity (egalitarian distribution).
   Could similarity measures sensitive to frequency distribution be classified based on the value taken by the inequality structure index of the frequency distribution of variable values?
   It seems that such classification would facilitate the selection of an appropriate similarity measure and reduce the need to test various distance measures to choose the one that best reflects the data structure and the purpose of cluster analysis.

2. *What other measures of dependence could be proposed for evaluation criteria assessment in Section 6.2 ?*

3. At the end of Chapter 6, the Author stated that the newly proposed criteria based on the modification of the Hartigan's rule (i.e., HE and HM criteria) did not perform very well in determining the optimal number of clusters, and these criteria can only be recommended for specific tasks. *In what cases would the Habilitation Candidate prefer their use instead of the internal evaluation criteria based on variability described in Subsection 3.2.1?*

IV FINAL CONCLUSIONS

After reading the habilitation thesis prepared by Ing. Zdeněk Šulc, Ph.D., entitled "Hierarchical Cluster Analysis of Categorical Data", I state that, despite the comments reported in the review, the habilitation thesis contains original elements that enrich the range of tools used for cluster analysis of categorical data, not only from the methodological but also from the application point of view. Moreover, is a valuable resource for a deeper understanding of the evaluation of cluster analysis for qualitative data. In particular, introducing two new internal criteria and expanding the numclust package demonstrates an Autor's commitment to advancing the field and addressing its needs. Its clear organization subordinated to implementing the goals set by the Habilitation Candidate, insightful discussions, and incorporation of recent research make it a commendable contribution to the literature in this area. In addition, the writing style is clear and concise, making it easy to follow the methodological part as well as experimental procedures and interpret the results. Additionally, the logical flow of information enhances the coherence of the habilitation thesis.

Based on the above-mentioned facts, I conclude that the work of Ing. Zdeněk Šulc, Ph.D. "Hierarchical Cluster Analysis of Categorical Data" meets the requirements of the habilitation thesis and I recommend it for defense at the meeting of the Scientific Board of the Faculty of Informatics and Statistics of the Prague University of Economics and Business.

Joanna Dębicka