Correlation n-star Graphs

Dudáš, Adam and Modrovičová, Bianka

Abstract: ***When working with correlation analysis and the visualization of its results and processes, the systematic identification of correlation classes is often overlooked. Such classes can be described as sets of attributes of the studied dataset that share similar correlation patterns – like the strength and direction of the relationship measured between a pair of attributes – and can be used as a method of identification of significant and insignificant correlations while introducing an attribute hierarchy for automated predictive analysis. This work proposes correlation n-star graphs, a graph-based visualization model designed to support automatic identification of these classes in multidimensional datasets. The work focuses on the design of the visualization model, a Python implementation of the proposed concept, and an experimental evaluation on three benchmark datasets assessing both qualitative and quantitative aspects of the visualization and correlation classification.***

Index Terms: ***correlation structures, correlation classes, correlation analysis, visualization, predictive analysis, data analysis***

1. Introduction

Correlation analysis represents one of the foundational components of data analysis and machine learning [10] [15] [6]. By quantifying the strength and direction of relationships between pairs of attributes, correlation measures inform tasks such as feature selection, dimensionality reduction, anomaly detection, and the construction of predictive models [13].

Visualization also plays a crucial role in making the correlation analysis more utilizable by translating numerical relationships into visual structures that analysts can examine and gain knowledge from [4] [16] [17]. Techniques such as correlation heatmaps or graphical representations of correlation analysis make patterns of association, heterogeneity, and structure immediately visible and aid in analytical processes and scenarios [2]. Many of these models explicitly target pseudotransitivity – the tendency for pairwise correlations to imply higher-order, chain-like associations across attributes, which can be used as a regression error minimization tool in predictive analysis.

However, there is a task related to correlation analysis, which is often overlooked and neglected – the systematic identification of correlation classes as groups of attributes that share similar correlation patterns across the dataset. Correlation classes are usable as a way of filtering the significant and insignificant correlation values in data and to introduce a type of hierarchy to the relationships present in the studied dataset. Their omission also constrains the automatization of the correlation analysis process, where only significant correlation classes can be utilized for further predictive analysis.

1.1 Objective of the Work

The main focus of the presented work is on the design, implementation, and evaluation of the correlation n-star graphs – a graph-based visualization model usable in the context of automatic identification of correlation classes in multidimensional datasets. Therefore, the main contributions of the work can be described as follows:

Manuscript received 28 August 2025, revised 9 December 2025

The research presented in this work was supported by the University Grant Agency of Matej Bel University in Banská Bystrica project number UGA-14-PDS-2025.

Adam Dudáš (corresponding author) is with the Department of Computer Science, Matej Bel University in Banská Bystrica, Slovakia (e-mail: adam.dudas@umb.sk). Bianka Modrovičová is with the Department of Computer Science, Matej Bel University in Banská Bystrica, Slovakia (e-mail: bianka.modrovicova@umb.sk).

- The design of a novel visualization model for the identification of correlation classes in data based on graph theory and clustering methods.
- Implementation of the proposed model in the $Python$ language as an open-source solution.
- Experimental evaluation of the concept on case studies utilizing three benchmarking datasets, examining qualitative and quantitative aspects of the visualization and correlation classification.

This objective, the basic theoretical background of the area, and the original proposed model are described in the four main sections of the presented work. In Section 2, the basics of correlation analysis – correlation coefficients, matrices and classes – are presented. Section 3 then focuses on the conventional and graphical models of correlation analysis visualization and presents the design of correlation n-star graphs, which are the main focal points of the work. The model proposed in Section 3 is then experimentally evaluated in Section 4 using three benchmarking datasets. Finally, Section 5 concludes the study, summarizes the main findings, and proposes some of the possible future work directions.

2. Correlation Analysis, Coefficients, Matrices, and Classes

Correlation analysis refers to a set of statistical techniques and approaches used for the identification of prediction potential stored in the values of attributes in a studied dataset [11]. Since the presented concept of correlation n-star graphs is used for the identification of correlation classes in data, this section of the work focuses on the introduction to correlation coefficients for numeric measurement of predictive potential in data, correlation matrices as a model of summarization of correlation coefficients in a dataset, and correlation classes for division of values of correlation coefficients.

2.1 Correlation Coefficients

The correlation coefficient measures the strength and direction of functional relationships between the values of two attributes from a common dataset [11]. This coefficient acquires a value from $[-1, 1]$ interval, while [21]:

- The absolute value of the correlation coefficient signifies the strength of the relationship – the closer to the extremes of the interval, the stronger the relationship,
- The sign of the value of the correlation coefficient denotes the direction of the relationship – both values of attributes increase or decrease in unison for positive values of the coefficient (so-called correlation), or the values of one attribute increase while the values of the other decrease for the negative values of the correlation coefficient (so-called anticorrelation).

Commonly, there are three basic correlation coefficients used in data analysis – the Pearson correlation coefficient, used for the measurement of the strength and direction of linear relationships between pairs of attributes, and Spearman and Kendall rank correlation coefficients, used for the measurement of these properties when non-linear monotone relationships are present in data.

Pearson correlation coefficient r measured between the values of attributes A and B is computed as [11]:

$$r(A,B) = \frac{\sum_{i=1}^{m}(A_i - \mu(A))(B_i - \mu(B))}{\sqrt{\sum_{i=1}^{m}(A_i - \mu(A))^2}\sqrt{\sum_{i=1}^{m}(B_i - \mu(B))^2}} \quad (1)$$

where μ denotes the mean value of attribute and m is the number of measurements of attributes A and B in the dataset.

The non-linear monotone coefficients are based on the concept of ranking of values of the studied attributes. This ranking can be described as a simple determination of descending order for values of attributes; hence, the highest value gets the ranking of one, and the lowest value gets the ranking of m. Spearman

rank correlation coefficient ρ for the values of attributes A and B can be computed as [11]:

$$\rho(A,B) = 1 - \frac{6\sum d_i^2}{m(m^2-1)} \tag{2}$$

where d_i denotes the difference between rankings of i-th measurements of the studied attributes.

Lastly, the computation of the Kendall correlation coefficient τ is defined as follows [11]:

$$\tau(A,B) = \frac{n_c - n_d}{\frac{m(m-1)}{2}} \tag{3}$$

where n_c denotes the number of concordant pairs of attribute rankings, and n_d denotes the number of discordant pairs of such rankings.

Regardless of the type of correlation coefficient, one can only measure the strength and direction of the relationship between a pair of attributes. Yet, datasets commonly contain many more than two attributes – often tens or even hundreds – and therefore correlation coefficient analysis of the whole dataset needs to be constructed using some type of summarization of pairwise correlation coefficient measurement. Such a summarization can be done via a correlation matrix, where the rows and columns of the matrix are indexed using attributes of the dataset and individual elements of the matrix contain the values of the correlation coefficient measured between the indexing attributes [22]. Matrix (4) presents a generalized example of this summarization for a dataset of a attributes.

2.2 Correlation Classes

All human-computer interaction requires a certain amount of interpretation of precise numerical outputs resulting from computations into a somewhat vague or fuzzy manner natural for humans. In the context of correlation analysis, the conventional practice is to split the correlation coefficient's $[-1,1]$ interval into subintervals, often called correlation classes [20]. A very simple example of such partitions is the following binary division [13]:

$$corr_{class}(A,B) = \begin{cases} significant, & \text{if } |corr(A,B)| \in [0.8,1], \\ insignificant, & \text{if } |corr(A,B)| \in [0,0.8) \end{cases} \tag{5}$$

Based on this concept, an analyst assigns one of two labels to a pair of attributes A and B – either they are considered to have a significant correlation when the absolute value of the correlation coefficient exceeds 0.8, or they are deemed to have an insignificant correlation in all other cases.

This binary approach, however, is rather coarse and omits finer distinctions; therefore, a more granular classification is proposed in [20]. The authors divide the correlation coefficient range into five categories – neutral, weak, moderate, strong, and very strong – as follows:

$$corr_{class}(A,B) = \begin{cases} neutral, & \text{if } |corr(A,B)| \in [0,0.2), \\ weak, & \text{if } |corr(A,B)| \in [0.2,0.4), \\ moderate, & \text{if } |corr(A,B)| \in [0.4,0.6), \\ strong, & \text{if } |corr(A,B)| \in [0.6,0.8), \\ very\ strong, & \text{if } |corr(A,B)| \in [0.8,1]. \end{cases} \tag{6}$$

Beyond these fixed, rule-based categorizations, several adaptive methods that use standard central tendency measures have been suggested. For example, a two-level classification based on dataset-wide statistics is introduced in [2]:

$$corr_{class}(A,B) = \begin{cases} significant, & \text{if } |corr(A,B)| \geq \\ & \frac{\mu(|corr(\mathbb{C})|)+\max(|corr(\mathbb{C})|)}{2}, \\ insignificant, & \text{otherwise.} \end{cases} \tag{7}$$

Here $\mu(|corr(\mathbb{C})|)$ denotes the mean of the absolute correlation coefficient values summarized in the correlation matrix $\mathbb{C}$, and $\max(|corr(\mathbb{C})|)$ is the largest off-diagonal absolute correlation in that matrix.

Similarly, one can construct dynamic groupings using other central-tendency-based cutoffs – for instance, a three-tier classification based

$$
\begin{array}{cccccc}
 & A_1 & A_2 & A_3 & \dots & A_a \\
A_1 & corr(A_1, A_1) & corr(A_1, A_2) & corr(A_1, A_3) & \dots & corr(A_1, A_a) \\
A_2 & corr(A_2, A_1) & corr(A_2, A_2) & corr(A_2, A_3) & \dots & corr(A_2, A_a) \\
A_3 & corr(A_3, A_1) & corr(A_3, A_2) & corr(A_3, A_3) & \dots & corr(A_3, A_a) \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
A_a & corr(A_a, A_1) & corr(A_a, A_2) & corr(A_a, A_3) & \dots & corr(A_a, A_a)
\end{array} \tag{4}
$$

on tertiles of correlation values in the studied dataset can be defined as:

$$
corr_{class}(A, B) = \begin{cases} low, & \text{if } |corr(A, B)| \in [0, T1), \\ middle, & \text{if } |corr(A, B)| \in [T1, T2), \\ high, & \text{if } |corr(A, B)| \in [T2, 1]. \end{cases} \tag{8}
$$

where $[0, T1)$ denotes the value of the correlation coefficient in the first tertile (lowest third) of all correlation values in the dataset, $[T1, T2)$ denotes the values of the second tertile of the dataset, and $[T2, 1]$ denotes the values of the third tertile of the data.

2.3 Related Works

As stated above, the research presented in this work aims towards the use of graph theory-based visualization in combination with correlation analysis and the use of such a model in the identification of correlation classes in the context of predictive analysis. Therefore, this section of the presented text briefly describes some of the modern results from the related areas.

The main objective of the work presented in [7] is focused on multi-label learning tasks, which typically involve complex correlation between labels, often spanning across multiple levels. The mainstream multi-label methods often only capture statistical label correlations and miss hierarchical, multi-level relationships, which can reduce predictive performance and interpretability of decision-making models. Authors of the work present a model that uses three-way concept-cognitive operators to structurally represent hierarchical label concepts, map their extents into feature concepts, and fuse those feature concepts to form label cognition. Extensive comparisons show the method improves prediction performance and enhances interpretability, demonstrating superiority and versatility over baseline approaches.

In the work [18], authors focus on so-called key performance indicator-related monitoring methods like partial least squares, correlation analysis and orthonormal subspace analysis. The effectiveness of all of these methods declines with high dimensionality of studied data, causing irrelevant information to mix into key performance indicator components and key performance indicator signal to leak into residuals. Hence, the authors propose Independent Variable Analysis, which identifies the true data structure and extracts independent components to avoid this problem. Simulation tests (including a vehicle engine model) show that the proposed method accurately recovers key performance indicator-related components, attains high detection of key performance indicator-related faults, and avoids alarming key performance indicator-unrelated faults.

This work [23] explores correlations among multiple categories of design arts by combining principal component analysis with an intelligent art-image recognition algorithm to build an art-design model. The authors solve a one-dimensional transient image-information transfer equation under diffuse-reflection boundary conditions, analyze albedo effects on hemispherical reflectance and transmittance, and treat transient transfer between double-layer plates, providing a calculation example. The analyses reveal strong correlations across design-art categories and underscore the multi-faceted roles that artistic design can play.

Missing label distributions in semi-supervised label-distribution learning make

mining reliable label correlations difficult and can bias correlation estimates. To address this, the authors of [24] propose two complementary strategies – global matrix completion via Independent Component Analysis and a locally improved $k - NN$ that uses known label-distribution constraints to guide unknown samples – and integrate them into a semi-supervised algorithm. Experiments show the method outperforms existing approaches in 67.27% of cases and yields statistically significant improvements in two-sample t-tests.

3. Visualization in the Context of Correlation Analysis

After the construction of a correlation matrix for the purposes of summarization of correlation coefficient values in the studied dataset, the visualization is conventionally done using correlation heatmaps [1]. This visualization translates the specific values of the correlation coefficient to colors from a previously defined color scale, while the extremes of the scale correspond to the extremes of the correlation coefficient interval using a smooth gradient for all other values. Figure 1 visualizes an example of a correlation coefficient heatmap of a dataset consisting of n attributes using a pink-to-purple color scale for the coding.

Other than the visualization of correlation analysis through correlation heatmap, this work focuses on alternative – less common – visualization approaches based on the concepts from the theory of graphs, which use the correlation matrix as a weighted adjacency matrix for graphical structures called correlation structures [2] [3].

3.1 Graphical Structures and Correlation Analysis

Correlation structures are based on a so-called correlation graph, which is defined as an undirected graph built on a dataset D consisting of #A attributes, while vertices of the graph denote individual attributes of the dataset and edges between these vertices are weighted using the correlation coefficient value [2]. For the purposes of the data analysis, only edges where

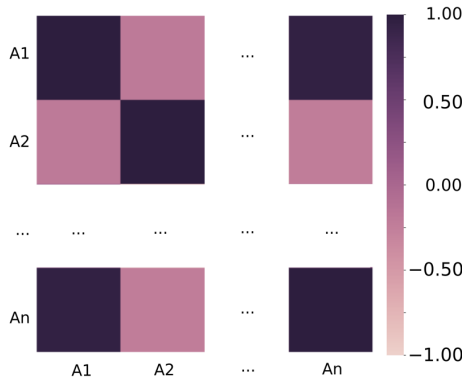

Figure 1: An example of a correlation heatmap of a dataset with $A_1, A_2, ..., A_n$ attributes

the weight is higher than the computed border of significance are visualized, which leads to a pruning of the correlation graph and, therefore, presentation of only interesting relationships in the studied data. Figure 2 presents an example of a correlation graph with the values of the correlation coefficient neglected for better readability.

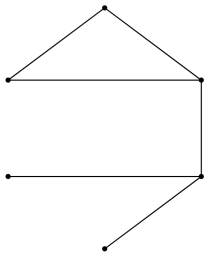

Figure 2: Example of a correlation graph

Several sub-graphs of the correlation graph were identified as structures of interest. First of these sub-graphs are correlation chains defined as a finite sequence of edges of the correlation graph from attribute A_i to A_j without repetition of any of the vertices or edges in the sequence [2]. An example of such a correlation chain is presented in Fig. 3, where the chain is using full lines, the edges, which are not taken into account, are presented using dotted lines, and the starting and ending attributes are denoted by the empty vertices.

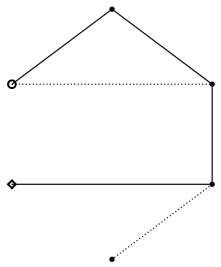

Figure 3: Example of correlation chain

The second of the sub-graphs of the correlation graph are correlation n-ptychs – complete subgraphs of the correlation graph consisting of n vertices – attributes [2]. Example presented in Fig. 4 showcases the correlation n-ptych for $n = 4$ (tetraptych) and $n = 5$ (pentaptych). The original study of n-ptychs considers $n \in \{2, 3, 4, 5, 6\}$ for the construction of readable (yet not always planar) structures.

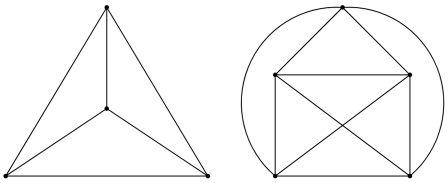

Figure 4: Example of correlation n-ptychs for $n \in \{4, 5\}$

The last of the graphical structures used in the context of correlation analysis can be labelled as correlation cycles. These cycles of size c are typical, as defined in graph theory, identified in the correlation graph, which consist of n vertices (attributes) of the dataset [3]. An example of the correlation cycle of $c = 5$ is presented in Fig. 5.

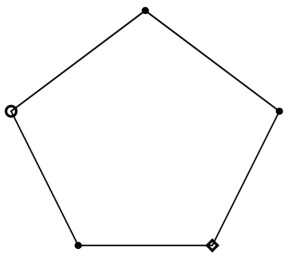

Figure 5: Example of correlation cycle

These graphical structures can be used in a number of analytical scenarios besides the visualization of correlation analysis itself. The most typical task conducted with the use of correlation structures is the study of pseudotransitivity of prediction potential in data, identification of prediction sequences based on correlation analysis, identification of subsets of attributes which influence significant portions of dataset values, or minimization of regression model errors. The concept of representing the attributes of a dataset as vertices of a graph interconnected via weighted edges is key to the visualization model, usable in visualization of correlation classes, presented in this work.

3.2 Correlation n-star Graphs

Based on the previously introduced graphical models for correlation analysis purposes, this work focuses on the design and implementation of the correlation n-star graphs for the identification of correlation classes in a studied dataset. In graph theory, star graphs are composed of one central vertex interconnected via edges to all other vertices of the graph, while these vertices are not interconnected [5] (see Fig. 6).

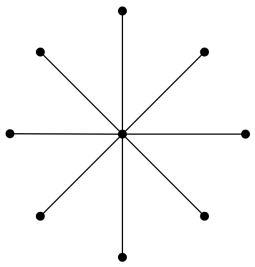

Figure 6: Example of a star graph

For the purposes of correlation analysis, the n-star graphs utilize the concepts from the previous graphical models, where vertices represent attributes of the dataset and edges are weighted by the correlation coefficient value measured between a pair of attributes, while the analysts themselves define the attribute of interest which is studied. This attribute of interest then serves as the central vertex for a set of star graphs, while each star graph represents a single correlation class of the dataset.

Hence, the n-star graph can be defined as a graph consisting of a set of n interconnected star graphs as presented in Fig. 7, while attributes of the dataset with similar values of correlation coefficient in relation to the attribute of interest are grouped to the same star, and therefore class.

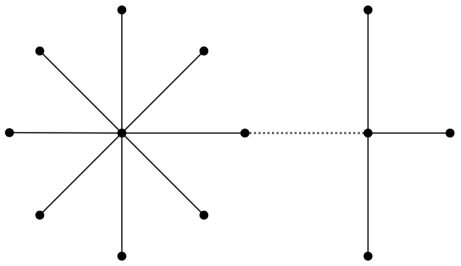

Figure 7: Example of an n-star graph for $n = 2$

This description of n-star graphs directly points to several significant notes related to the implementation of the concept:

- The number of stars or recognised correlation classes (n) in the studied data and the attribute of interest are defined by the user, in this case data analyst.
- The identification of correlation classes is conducted via clustering of correlation coefficient values into n clusters.
- As mentioned above, the n-star graph is constructed from a set of n interconnected star graphs. These components of the graph are interconnected via edges containing the strongest value of the correlation coefficient between a pair of attributes from neighbouring correlation classes (hence not creating any cycles).

In this way, a single graph of n interconnected stars is constructed, in which one can identify two types of edges. The first type can be labelled as $intra - star$ edges constructed between the attribute of interest and other attributes in a selected class, which denote the membership of an attribute to one specific correlation class. The second type of edges are the $inter - star$ edges connecting pairs of correlation class stars to produce one holistic graph and – more critically – to allow for the study of relationships between individual correlation classes of the dataset via pseudo-transitivity of prediction potential [2]. The whole process described above is presented in the schema in Fig. 8.

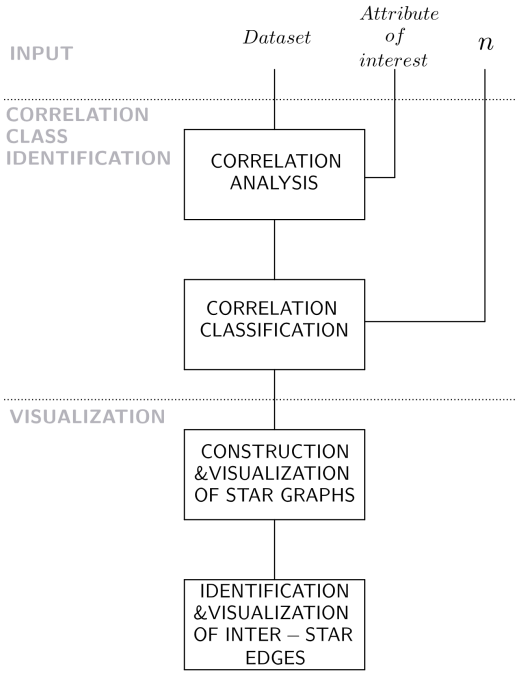

Figure 8: Schema of the proposed approach

In order to construct this visualization, the presented work utilizes $K - Means$ algorithm [9] for the clustering of correlation coefficient values measured between the attribute of interest and all other attributes in the dataset into the correlation classes. Since the correlation classes are commonly labelled with a title of class as presented in section 2.2, for these purposes, the mean absolute value of the correlation coefficient (μ) in the class is used. This μ metric is also used for the sorting of the correlation classes from the weakest correlation class (lowest μ) to the strongest (highest μ).

4. Case Studies of n-star Graphs

Based on the description of the correlation n-star graphs from the previous section of the work, the concept has been implemented in $Python$ language using $pandas$ and $numpy$ packages for the processing of input data,

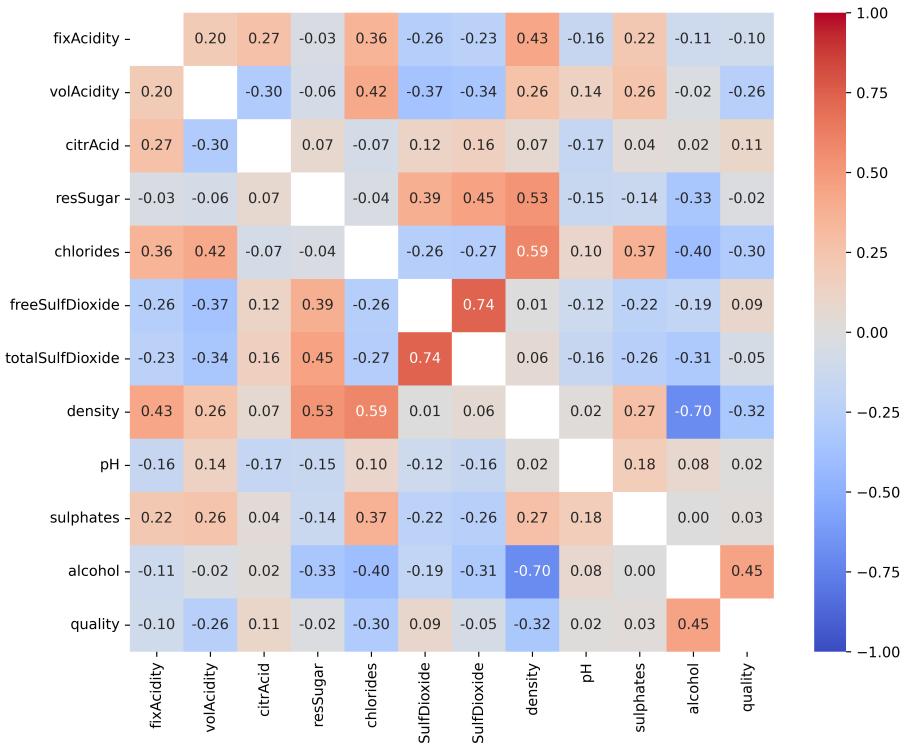

Figure 9: Correlation heatmap of Spearman type for Wine Quality Datasets

$sklearn$ package for clustering of the processed data via $K - Means$ algorithm, and $networkx$ and $matplotlib$ packages for composition and visualization of the graphs themselves.

For the evaluation of the proposed method, three commonly used benchmarking datasets were utilized:

- Wine Quality Dataset [12] – dataset describing chemical properties of red and white vinho verde wine from the northern regions of Portugal. For the presented analysis, this dataset serves as the smallest of the datasets, containing 13 attributes measured on 4 898 entities. The basic visualization of the correlation properties of the dataset is presented in the heatmap in Fig. 9.
- Appliance Energy Dataset [8] – in this dataset, experimental data concerning the use of appliances in a low energy building are stored. The dataset is medium in size, consisting of 29 attributes and 19 735 measurements. Correlation heatmap of the dataset is presented in Fig. 10.
- Superconductivity Dataset [19] – the largest of datasets describes superconductors and their relevant features, such as their critical temperatures or chemical formulas. The dataset describes 21 263 of such superconductors via 82 attributes. To visualize the relationships in this dataset, Fig. 11 presents its correlation heatmap.

These datasets were selected as a way of illustrating data originating in drastically different domains – chemical wine properties, energy performance in green housing, and physical properties of superconductors. The other critical property of the datasets is the variation in their dimensionality, with the number of attributes ranging from 13 to 82. Hence, these datasets provide a broad spectrum of application areas and demonstrate how the proposed method scales across diverse domains, which is crucial from the point of view of the visualization and computation effectiveness.

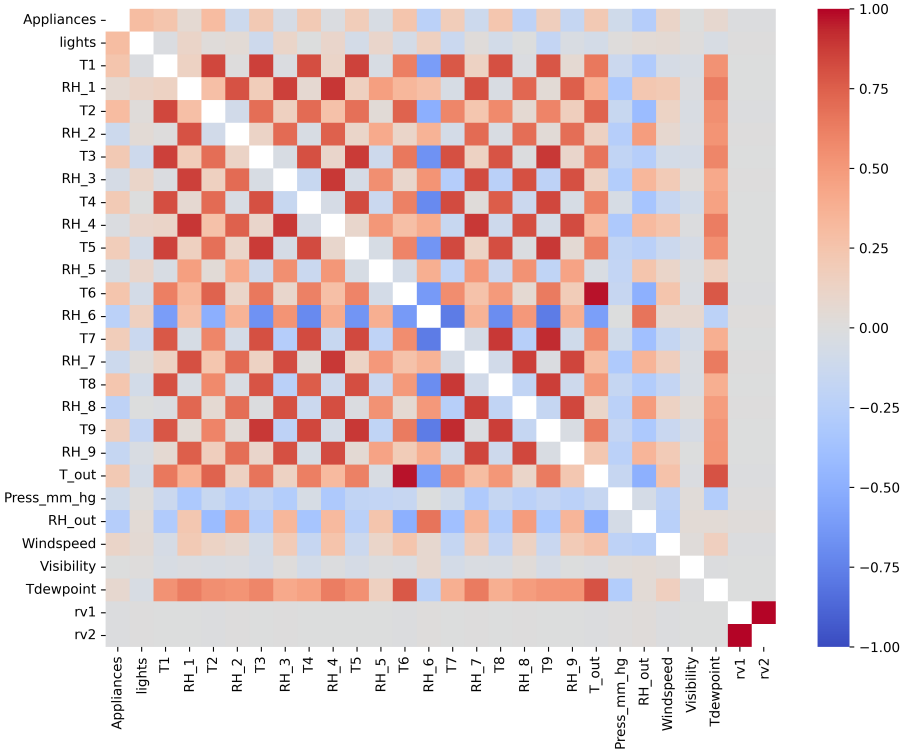

Figure 10: Correlation heatmap of Spearman type for Appliance Energy Dataset

Therefore, using these datasets, two aspects of the proposed visualization model are examined. Firstly, the visualization of these datasets via the model of n-star graphs is conducted for various values of the n parameter (the number of stars) and qualitative aspects of the visualization based on the criteria defined in [14] are evaluated. Secondly, quantitative properties of the proposed model are examined, and the fitting number of classes for each of the studied datasets and attributes of interest is determined.

4.1 Visualization of n-star graphs

Since, for the presented visualization, the $K-Means$ clustering requires input value describing the number of clusters (or stars), the following study presents correlation classification into various number of classes for all three of the considered datasets.

Figures 12 and 13 focus on the visualization of n-star graphs for the Wine Quality Dataset with the attribute of interest set as $density$ of wine. For the study purposes, the n is set to 2 and 4, respectively. As seen in Fig. 12 the two identified correlation classes described by the stars are fairly well distinguishable – the first class labeled as $\mu = 0.145$ contains attributes with low correlation coefficient values measured between the attribute of interest and all other attributes of the dataset, while the class labeled with $\mu = 0.563$ contains medium to high values of the correlation coefficient. The two classes are interconnected using the correlation between $total\ Sulfur\ Dioxide$ contained in the wine sample and the amount of $residual\ Sugar$ in the sample.

Naturally, the classification of data into two correlation classes is often insufficient, since it lacks nuance. Hence, more finer division of attributes based on their correlation coefficient values is presented in Fig. 13, where four correlation classes can be seen – low correlations where $\mu = 0.038$, marginal correlations with $\mu = 0.286$, middle correlations marked by $\mu = 0.517$, and high correlation coefficient values in $\mu = 0.699$.

Classification of correlation data into a

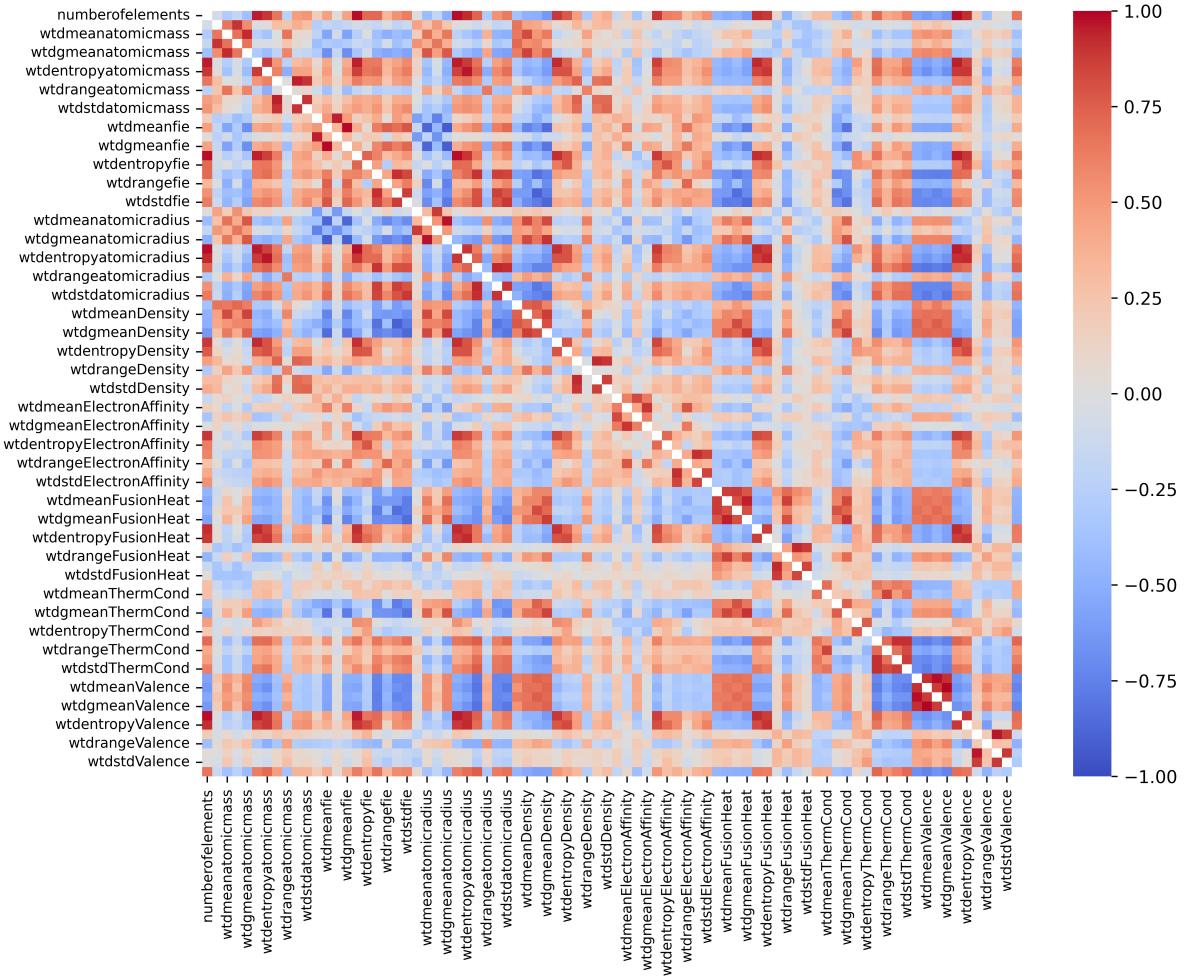

Figure 11: Correlation heatmap of Spearman type for Superconductivity Dataset

higher number of correlation classes is presented in Figures 14 and 15, constructed on Appliance Energy and Superconductivity datasets. Fig. 14 focuses on the visualization of 5 correlation classes for the $Tdewpoint$ attribute selected as the attribute of interest. One can see a fairly clear separation between the identified classes based on the μ value, while the number of attributes in individual classes varies from 2 to 10.

On the other hand, the division of the correlation classes in Fig. 15 can be considered nuanced, but there can also be identified minor redundancy. Specifically, the classes of $\mu = 0.313$ and $\mu = 0.398$ are quite close to each other in a subset of correlation coefficient values (eg, one class contains the correlation coefficient value of 0.33 and the other the value 0.36). A similar, but less severe, situation exists in the $\mu = 0.611$ and $\mu = 0.698$ classes.

4.2 Qualitative Aspects of the Visualization

From the point of view of qualitative analysis of the proposed model, we focus on the evaluation of two distinct aspects relevant to the quality of the method:

- Evaluation of qualitative aspects of visual models via Qualitative Result Inspection and Visual Data Analysis and Reasoning based on [14].
- Comparison of the proposed approach to the conventionally used method of correlation class identification – static 5-tier classification and dynamic tertile-based classification.

The main objective of Qualitative Result Inspection is to evaluate the image quality and visual encoding of information created by the proposed model, while Visual Data Analysis and Reasoning aims to explore how the model supports analysis and reasoning about data and helps to derive relevant knowledge in the domain. In the case of correlation n-star graphs, one can see a clear division of correlation classes into the stars themselves, which are color-coded and interconnected via the dotted edges representing the strongest inter-star correlation value between neighbouring correlation classes. Besides the color, each correlation class is identified by a μ value, which helps in fast understanding of class hierarchy. Since the visualization is inherently partially interactive, meaning there is a possibility of zooming in and out of individual correlation classes, analysts can focus their study on specific parts of the dataset.

When working with datasets containing a high number of attributes – such as the Superconductivity dataset (Fig. 15) –, we can see certain crossings of graph elements. As an example, we can note the class defined by $\mu = 0.398$, where more than 20 attributes were classified, and the labelled vertices of the star cross each other, therefore making the analysis harder. However, this deficiency is easily bypassed by the abovementioned zooming in to the class itself.

The second aspect of the qualitative assessment of the proposed model focuses on the comparison of the results reached with the use of the proposed n-star method and conventional correlation class identification methods. In the presented case, the comparative analysis is conducted using static 5-tier classification and dynamic tertile-based classification presented in Section 2.2.

Tables 1 and 2 contain the division of attributes of all three of the considered datasets into correlation classes based on the correlation coefficient value measured between the attribute of interest and all other attributes based on equations (6) and (8). For comparison purposes, the same attributes of interest were selected across the considered datasets, specifically $density$ for the Wine Quality dataset, $Tdewpoint$ for the Appliance Energy dataset, and $criticalTemp$ for the Superconductivity dataset.

From the tables, we note some of the interesting properties of the basic correlation class identifications. When using the static approach, it is quite common that some of the classes do not contain any attributes, hence are not necessary, eg, the class $strong$ and $very\ strong$ in the Wine Quality dataset or the class $very\ strong$ in the Superconductivity dataset utilizing $criticalTemp$ as the attribute of interest. On the other hand, when the dy-

namic approach based on correlation coefficient value tertiles is utilized, one can note the fact that each class is approximately the same size (similar number of attributes).

Utilizing n-star graphs, the classification yields similar results to other approaches, yet when compared to the static approach, the proposed model does not create empty classes, and from the point of view of comparison to the dynamic method, the classes are of different sizes, which is fitting from the point of view of division of the coefficient values. Additionally, the proposed model has the strong advantage of visualization, which can be done using any of the classifications, but it isn't done natively on any of the models, except the correlation n-star graphs.

4.3 Quantitative Aspects of the Visualization

As a last part of the evaluation of the proposed approach, the quantitative study of the similarity of correlation classes created using n-star graphs is presented. Table 3 contains the results of n-star graph construction for $n \in \{2, 3, 4, 5, 6, 7\}$ for all three of the considered datasets and for the same attributes of interest as in the previous comparative analysis (Tables 1 and 2).

For the identification of correlation classes in data, the objective is to create clearly separable, dissimilar classes. For the purposes of the presented research, the Δ measure is used for the measurement of similarity of correlations in two neighbouring correlation classes. This metric can be described as the lowest difference between the values of the correlation coefficient in neighbouring classes, hence:

$$\Delta(C_i, C_{i-1}) = min(C_i) - max(C_{i-1}), \quad \forall i \in 2, 3, ..., n \tag{9}$$

where C_i is the i-th correlation class of the n-star graph. Then, the natural objective is to maximize Δ value while identifying the highest possible n, which leads to a high number of dissimilar correlation classes. Since the Δ is a simple numerical value used for the overall evaluation of a correlation class, some border of satisfactory dissimilarity ($diss$) of classes needs to be defined. From the perspective of correlation analysis, such a border can be set to 0.05, so that:

$$diss(C_i, C_{i-1}) = \begin{cases} satisfactory, & \text{if } \Delta(C_i, C_{i-1}) \geq 0.05 \\ unsatisfactory, & \text{otherwise.} \end{cases} \tag{10}$$

The results presented in Table 3 point to several interesting findings and properties of the proposed model and the studied data itself. For the Wine Quality dataset and the $density$ as the attribute of interest, the dissimilarity of correlation classes is comparatively high with Δ value close to 0.1 for $n \in \{2, 3, 4, 5\}$. In other cases, such a high value is reached only for $n = 2$, except for the Superconductivity dataset, where the highest Δ reached 0.0224. If adhering to the defined border of 0.05 for the measurement of dissimilarity of two correlation classes, one can observe satisfactory results for $n \in \{2, 3, 4, 5\}$ in the Wine Quality dataset. As mentioned before, in the Superconductivity dataset, there were no Δ values higher than or equal to 0.05 and for the $Tdewpoint$ in the Appliance Energy dataset, only $n = 2$ yielded satisfactory classification of correlation values.

From these results, one can conclude:

- The proposed method creates the clearest classification of correlation coefficient values for $n \in \{2, 3, 4, 5\}$, while the value of Δ dropped significantly in all studied cases, where $n \in \{6, 7\}$.
- The specific value of n used to create the proposed visualization model can be determined via the utilization of the Δ value.
- Correlation values measured between $criticalTemp$ and all other attributes of the Superconductivity dataset are similar, and therefore their proper division into the classes is challenging.

5. Conclusion

The concept of correlation n-star graphs represents a simple and visual method for the iden-

Table 1: Correlation classes of the studied datasets constructed via static 5-tier classification

Dataset	Attribute of interest	Class	Attributes in class
Wine Quality	density	neutral	freeSulfDioxide, pH, totalSulfDioxide, citrAcid
		weak	volAcidity, sulphates, quality, fixAcidity
		moderate	resSugar, chlorides, alcohol
		strong	–
		very strong	–
Appliance Energy	Tdewpoint	neutral	rv2, rv1, Visibility, RH_out, lights, Appliances, RH_5, Windspeed
		weak	RH_6, Press_mm_hg, T8, T7
		moderate	RH_3, T4, RH_8, T9, RH_2, RH_9, T1, T5, T2, T3
		strong	RH_4, RH_1, RH_7, T6
		very strong	T_out
Superconductivity	criticalTemp	neutral	gmeanfie, wtdstdFusionHeat, rangeFusionHeat, meanatomicradius, wtdgmeanElectronAffinity, wtdentropyThermCond, wtdmeanElectronAffinity, entropyThermCond, rangeValence, meanatomicmass, stdFusionHeat, wtdrangeElectronAffinity, stdDensity, meanfie, gmeanatomicradius
		weak	meanElectronAffinity, gmeanatomicmass, stdValence, wtdrangeDensity, wtdentropyElectronAffinity, wtdstdValence, wtdstdDensity, wtdrangefie, rangeDensity, wtdrangeatomicradius, stdatomicmass, stdElectronAffinity, wtdstdatomicmass, wtdmeanatomicradius, wtdrangeatomicmass, gmeanThermCond, wtdmeanatomicmass, rangeElectronAffinity, gmeanElectronAffinity, wtdgmeanfie
		moderate	wtdrangeFusionHeat, wtdstdElectronAffinity, wtdmeanThermCond, wtdmeanfie, wtdgmeanatomicmass, wtdentropyfie, meanThermCond, wtdgmeanThermCond, wtdentropyDensity, wtdgmeanatomicradius, meanFusionHeat, meanDensity, rangeatomicmass, wtdrangeValence, wtdmeanFusionHeat, gmeanFusionHeat, entropyElectronAffinity, wtdgmeanFusionHeat, entropyDensity, wtdrangeThermCond, wtdmeanDensity, stdfie, wtdgmeanDensity, gmeanDensity, wtdstdfie
		strong	entropyatomicradius, stdThermCond, entropyfie, wtdentropyFusionHeat, entropyatomicmass, stdatomicradius, entropyFusionHeat, gmeanValence, rangefie, wtdentropyValence, numberofelements, wtdstdatomicradius, meanValence, wtdentropyatomicradius, entropyValence, wtdentropyatomicmass, wtdstdThermCond, rangeThermCond, wtdgmeanValence, wtdmeanValence, rangeatomicradius
		very strong	–

Table 2: Correlation classes of the studied datasets constructed via dynamic classification based on tertiles

Dataset	Attribute of interest	Class	Attributes
Wine Quality	density	1. tertile	freeSulfDioxide, pH, totalSulfDioxide, citrAcid
		2. tertile	volAcidity, sulphates, quality
		3. tertile	fixAcidity, resSugar, chlorides, alcohol
Appliance Energy	Tdewpoint	1. tertile	rv1, rv2, Visibility, RH_out, lights, Appliances, RH_5, Windspeed, RH_6
		2. tertile	Press_mm_hg, T8, T7, RH_3, T4, RH_8, T9, RH_2, RH_9
		3. tertile	T1, T5, T2, T3, RH_4, RH_1, RH_7, T6, T_out
Superconductivity	criticalTemp	1. tertile	gmeanfie, wtdstdFusionHeat, rangeFusionHeat, meanatomicradius, wtdgmeanElectronAffinity, wtdentropyThermCond, wtdmeanElectronAffinity, entropyThermCond, rangeValence, meanatomicmass, stdFusionHeat, stdrangeElectronAffinity, stdDensity, meanfie, gmeanatomicradius, meanElectronAffinity, gmeanatomicmass, stdValence, wtdrangeDensity, wtdentropyElectronAffinity, wtdstdValence, wtdstdDensity, wtdrangefie, rangeDensity, wtdrangeatomicradius, stdatomicmass, stdElectronAffinity,
		2. tertile	wtdstdatomicmass, wtdmeanatomicradius, wtdrangeatomicmass, gmeanThermCond, wtdmeanatomicmass, rangeElectronAffinity, gmeanElectronAffinity, wtdgmeanfie, wtdrangeFusionHeat, wtdstdElectronAffinity, wtdmeanThermCond, wtdmeanfie, wtdgmeanatomicmass, wtdentropyfie, meanThermCond, wtdgmeanThermCond, wtdentropyDensity, wtdgmeanatomicradius, meanFusionHeat, meanDensity, rangeatomicmass, wtdrangeValence, wtdmeanFusionHeat, meanFusionHeat, entropyElectronAffinity, wtdgmeanFusionHeat, entropyDensity
		3. tertile	wtdrangeThermCond, wtdmeanDensity, stdfie, wtdgmeanDensity, gmeanDensity, wtdstdfie, entropyatomicradius, stdThermCond, entropyfie, wtdentropyFusionHeat, entropyatomicmass, stdatomicradius, entropyFusionHeat, gmeanValence, rangefie, wtdentropyValence, numberofelements, wtdstdatomicradius, meanValence, wtdentropyatomicradius, entropyValence, wtdentropyatomicmass, wtdstdThermCond, rangeThermCond, wtdgmeanValence, wtdmeanValence, rangeatomicradius

Table 3: Study of similarity of correlation classes based on Δ measure

Dataset	Attribute of interest	n	Δ
Wine Quality	density	2	0.1113
		3	0.0925
		4	0.1087
		5	0.0925
		6	0.0416
		7	0.0416
Appliance Energy	Tdewpoint	2	0.1096
		3	0.0378
		4	0.0378
		5	0.0394
		6	0.0219
		7	0.0219
Superconductivity	criticalTemp	2	0.0224
		3	0.01
		4	0.0224
		5	0.0087
		6	0.0091
		7	0.0175

tification of correlation classes in multidimensional datasets. These classes are obtained by clustering of the absolute correlation values (using $K-Means$) and labelled and sorted using the class mean absolute correlation μ, while being interconnected by the strongest inter-class correlations selected so as to avoid cycles, producing an interpretable, hierarchical visualization.

Evaluation of the model on three benchmark datasets demonstrates both qualitative and quantitative strengths of the model. Qualitatively, the visualization clearly separates correlation classes through color coding, μ labels and explicit inter-star connections, and it supports focused analysis through its interactive elements. Compared to static 5-tier and dynamic tertile-based classifications, n-star graphs avoid empty classes common in static classifications and produce class sizes that better reflect the underlying distribution of correlation strengths than simple tertiles, while simultaneously providing a native, easy-to-interpret visual encoding. Quantitative results indicate that the proposed method produces clear and most dissimilar classification for $n \in \{2,3,4,5\}$ across the studied datasets.

Future work should address the limitations of the model and extend the method in several directions:

- develop automated methods to select an appropriate n so analysts are guided rather than required to choose n a priori,
- improve layout and labelling for high-dimensional attribute sets (edge-routing, adaptive label placement, or interactive filtering) to reduce visual crossing of graph elements,
- conduct user studies to quantify the model's effectiveness for real analytical tasks versus conventional correlation-classification methods,
- and study other representations of direct and indirect relationships in correlation structures and substructures, such as embedded or naturally occurring trees.

Code and data availability

The code for the proposed visualization model is available at:
`github.com/AdamDudasUMB/n-starGraphs`

The datasets used in the experiments presented in this study are freely available at:
`archive.ics.uci.edu/dataset/186/wine+quality`
`archive.ics.uci.edu/dataset/374/appliances+energy+prediction`
`archive.ics.uci.edu/dataset/464/superconductivity+data`
In case of any questions, don't hesitate to contact the authors of the work via e-mail `adam.dudas@umb.sk`

References

[1] A. Dudáš. Graphical representation of data prediction potential: correlation graphs and correlation chains. *Visual Computer*, 2024.

[2] A. Dudáš. Non-parametric correlation structures and their respective embeddings in predictive analysis. *Discover Applied Sciences*, 2025.

[3] A. Dudáš et al. Exploration and deconstruction of correlation cycles in multidimensional datasets. *Technologies*, 2025.

[4] J. Oh et al. Tpvis: A temporal path visualization system for intuitive understanding of information diffusion inside temporal networks. *IEEE Access*, 2025.

[5] J. Pach et al. Decomposition of geometric graphs into star-forests. *Computational Geometry - Theory and Applications*, 2025.

[6] J. Rabčan et al. Eeg signal classification based on fuzzy classifiers. *IEEE Transactions on Industrial Informatics*, 2022.

[7] J.M. Wu et al. Multi-level correlation information fusion via three-way concept-cognitive learning for multi-label learning. *Information Fusion*, 2025.
[8] L. Candanedo et al. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, 2017.
[9] L.C. Chen et al. Workload balancing for photolithography machines in semiconductor manufacturing via estimation of distribution algorithm integrating kmeans clustering. *IEEE Transactions on Systems, Man, Cybernetics: Systems*, 2025.
[10] M. Kvet et al. Concept of temporal data retrieval: Undefined value management. *Concurrency Computation - Practice and Experience*, 2020.
[11] M.B. Xu et al. Real-time stitching algorithm of vehicle side view image based on multi-region fast phase correlation. *IEEE Access*, 2025.
[12] P. Cortez et al. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 2009.
[13] S.Y. Jiang et al. Spatial correlation guided cross scale feature fusion for age and gender estimation. *Scientific Reports*, 2025.
[14] T. Isenberg et al. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2013.
[15] V. Beneš et al. Situation model of the transport, transport emissions and meteorological conditions. *Neural Network World*, 2024.
[16] W. Steingartner et al. A visualizing tool for graduate course: Semantics of programming languages. *IPSI BGD Transactions on Internet Research*, 2019.
[17] W. Steingartner et al. Some aspects about visualization of natural semantics for a selected domain-specific language. *IPSI BGD Transactions on Internet Research*, 2023.
[18] Z.J. Lou et al. Independent variable analysis for addressing the dimension dilemma in monitoring key-performance-indicator-related faults. *Expert systems and applications*, 2025.
[19] K. Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational materials science*, 2018.
[20] L.B. Iantovics. Avoiding mistakes in bivariate linear regression and correlation analysis. *Acta Polytechnica Hungarica*, 2024.
[21] D.K. Le. Correlation-based approach for heart rate extraction using uwb impulse radar. *International Journal of Fuzzy Logic and Intelligent Systems*, 2025.
[22] K. Ueda and H. Tanaka. Estimation of the square of distance correlation coefficient. *Communications in Statistics - Theory and Methods*, 2025.
[23] L.L. Wang. Correlation analysis between multi-category design arts based on principal component analysis model. *Systems and Soft Computing*, 2025.
[24] P.Q. Yu and X.Y. Jia. Semi-supervised label distribution learning via global factorization and local constrain. *Neurocomputing*, 2025.

Adam Dudáš is an assistant professor at the Department of Computer Science, Faculty of Natural Sciences of Matej Bel University in Banská Bystrica. He is an author and co-author of more than 45 research works, and his research activities are related to descriptive, explorative and predictive data analysis with emphasis on visualization in the context of statistical analysis of data.

Bianka Modrovičová is a student at the Department of Computer Science, Faculty of Natural Sciences of Matej Bel University in Banská Bystrica, Slovakia and works in IBM Slovakia. She is the author and co-author of 8 research works, and her research activities focus on the use of machine and deep learning models in graphical problems, specifically connected to proper edge coloring of graphs.

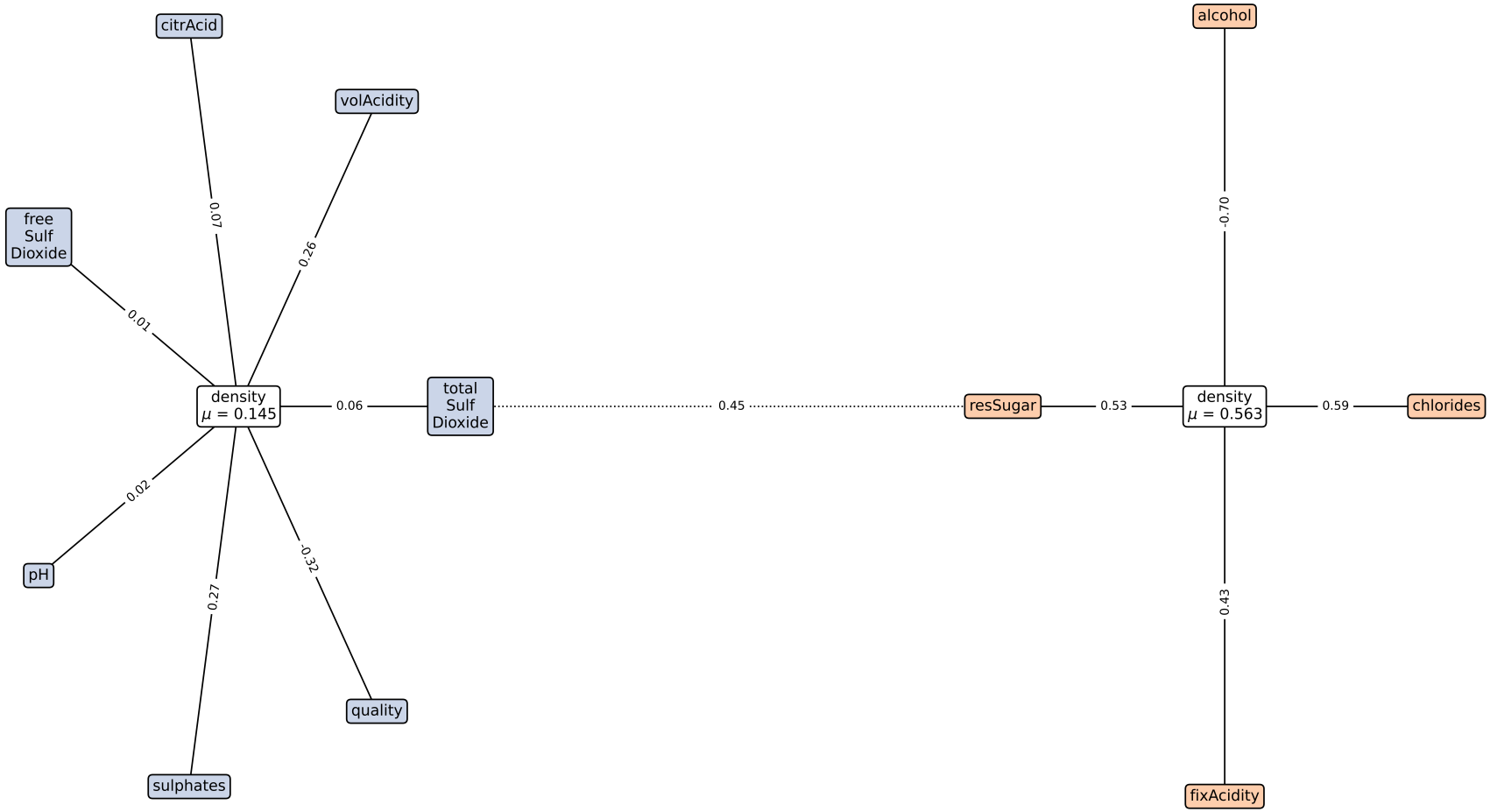

Figure 12: Visualization of Wine Quality Dataset correlation n-star graph for $n = 2$

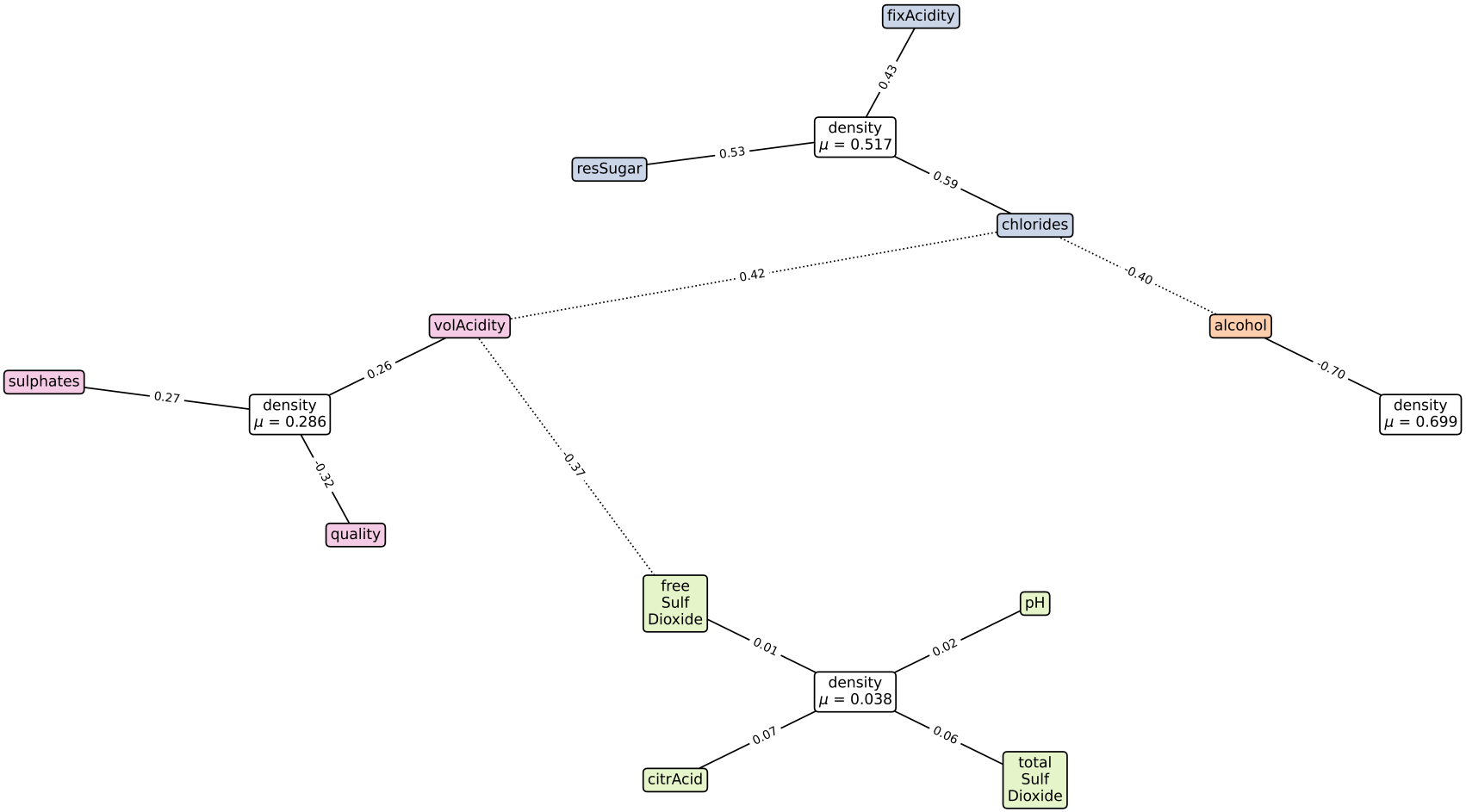

Figure 13: Visualization of Wine Quality Dataset correlation n-star graph for $n = 4$

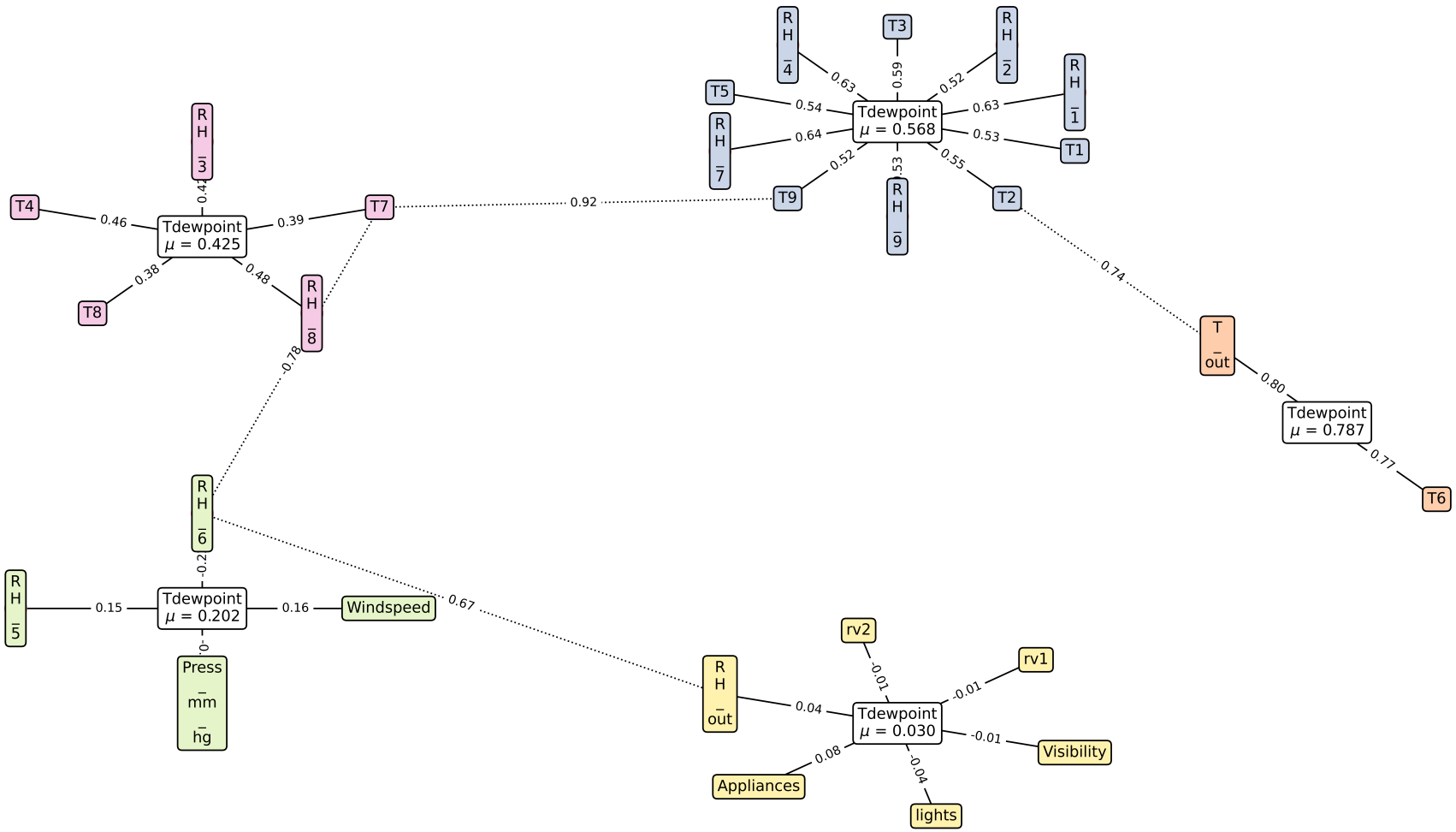

Figure 14: Visualization of Appliance Energy Dataset correlation n-star graph for $n = 5$

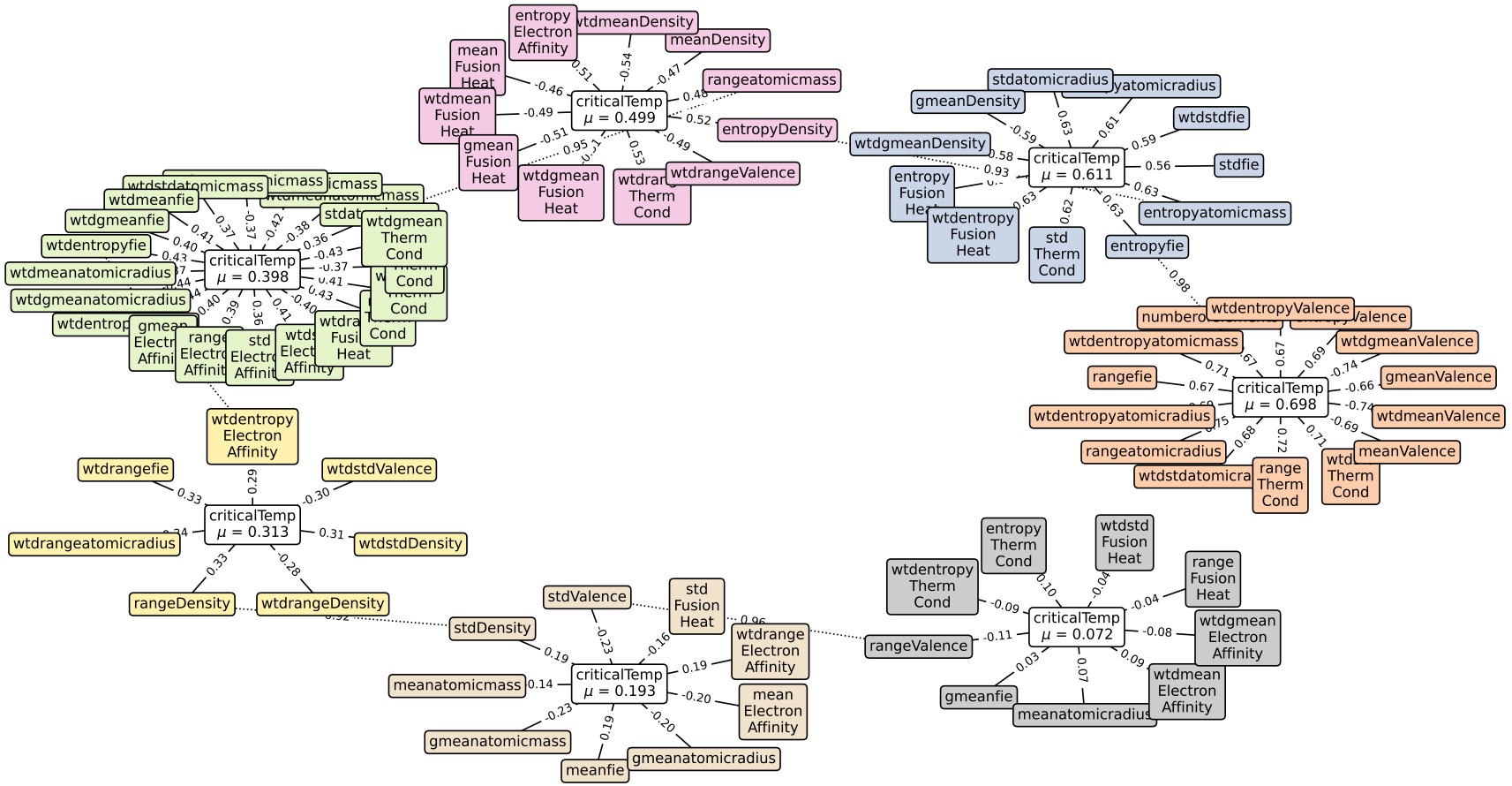

Figure 15: Visualization of Superconductivity Dataset correlation n-star graph for $n = 7$