Adam Dudáš; Tomáš Peregrín

Honeycomb graphs for parametric identification of correlation classes in multidimensional datasets

Kybernetika, Vol. 61 (2025), No. 6, 789–816

Persistent URL: http://dml.cz/dmlcz/153265

KYBERNETIKA — VOLUME *61* (2025), NUMBER 6, PAGES 789–816

HONEYCOMB GRAPHS FOR PARAMETRIC IDENTIFICATION OF CORRELATION CLASSES IN MULTIDIMENSIONAL DATASETS

Adam Dudáš and Tomáš Peregrín

In the process of gaining knowledge from large sets of data, one of the most significant methods from the area of descriptive statistics − correlation analysis − is applied to determine direct functional relationships between pairs of attributes. Even though the results of correlation analysis are measured through a crisp correlation coefficient, whose values belong to the $[-1, 1]$ interval, human interpretation of these values is conventionally vague and uses linguistic classes of correlation to describe the strength of relationships between attribute pairs. However, this interpretative vagueness − and the correlation classes themselves − are not commonly employed in the decision-making processes. Therefore, this work focuses on the design and implementation of so-called Honeycomb Graphs − a visualization method for parametric identification of correlation classes in multidimensional datasets based on graphical models. After implementing the proposed visualization technique, two case studies on benchmark datasets are conducted, and the model is evaluated from both qualitative and quantitative points of view. The results of these studies highlight interactive exploration of correlation analysis while adhering to qualitative and quantitative standards of scientific visualizations and high utilization potential of the method in feature selection tasks, making it a valuable tool for predictive analysis and data exploration.

Keywords: visualization, big data analysis, correlation analysis, correlation classes

Classification: 68P99, 62H20

1. INTRODUCTION

Correlation analysis is a statistical process for identifying prediction potential in multidimensional datasets, which is used in several analytical approaches and scenarios, such as the definition of basic relationships in data in the context of descriptive data analysis, the identification of attribute pairs worthy of visualization in the exploratory analysis of the data, or the selection of features for predictive data analysis [6, 20].

The prediction potential identified in the correlation analysis is determined by the direction and strength of the functional relationship between the values of an attribute pair in a dataset. These properties of the relationship of interest are measured by one of two general types of correlation coefficient − linear measures (such as the Pearson

DOI: 10.14736/kyb-2025-6-0789

correlation coefficient) or nonlinear monotone metrics (such as the rank correlation coefficients) [21].

Even though the correlation coefficient acquires crisp values from $[-1, 1]$ interval, the human interpretation of the values is − somewhat − fuzzy or uncertain, eg. the value of the correlation coefficient equal to 0.79 does not differ significantly from that of value 0.8 even though these values can be perceived as parametrically different when a crisp correlation filter value is set. Therefore, analysts commonly use linguistically-coded classification of the correlation coefficient values, such as *strong correlation* or *marginal correlation* for specified correlation coefficient value intervals [5]. Such correlation classes are then commonly utilized in various data analysis tasks −, mainly feature selection and dimensionality reduction problems, where the large multidimensional space is pruned by the selection of attributes between which the relationships are strongest or most interesting; or anomaly detection tasks, in which correlation classes define the standard behaviour of the data and the measurements deviating from this behaviour are labelled as anomalies [23].

1.1. Objective of the work

In the scope of the presented study, the focus is put on the design, implementation, and experimental evaluation of a novel visualization technique for parametric identification of correlation classes in multidimensional datasets called Honeycomb graphs. Specifically, the proposed visualization method consists of an interconnected graphical model formed from clusters of vertices organized around a user-defined attribute of interest. This attribute represents the main property of the entities modelled in the data for which the prediction potential is measured.

The novelty of the work presented in this study can be summarized into the following points:

- The design of the novel method for visualization of correlation classes and significant relationships in and between these classes.
- The implementation of the proposed visualization model in the *Python* language with the use of freely available packages, such as *matplotlib*, *networkx*, *seaborn*, and so on.
- The case studies of correlation class identification on two openly available benchmarking datasets from the area of chemical analysis of wine samples and electrical engineering related to home appliances.
- The qualitative and quantitative evaluation of the proposed visualization model, based on the standardized criteria used in the area of visual computing, and comparative analysis conducted between the proposed model and conventionally used classification and visualization technniques in the studied area.

The body of this article is structured as follows − the rest of the first section of the work offers a brief description of the previous and related work conducted in the area of visualization of correlation analysis results. In Section 2, the background to correlation analysis, correlation coefficients, correlation classes, and conventional visualization

models, which is necessary for the proposed work, is presented. Section 3 focuses on the design of the proposed visualization model of Honeycomb graphs and the basic properties resulting from this design. In Section 4, the case studies of the proposed visualization technique on the Wine quality and Appliance Energy datasets, and the evaluation of the model are conducted, and finally, Section 5 concludes the work and offers several future work areas.

1.2. Related work

Since the work proposed in this study focuses on the use of visual and graphical models in the correlation analysis processes, mainly concerning correlation classes in multidimensional datasets, in the following section, a brief overview of our own work and the work of other authors in the related areas is presented.

In the studies [9] and [10], we presented the utilization of graphical models in correlation, predictive, and regression analysis based on the pseudo-transitivity of the correlation coefficient. These works concern the design, implementation and experimental verification of three novel graphical models − correlation graphs, correlation chains, and (deconstructed) correlation cycles. All of these visualizations can be used to identify fitting regression sequences, which can, in turn, lower the error rate of regression tasks conducted on multidimensional datasets significantly when compared to conventional regression models.

Other than our own work, in [17] use visualization concepts in the examination of the effect of uncertainty representation on correlation judgment in bivariate visualizations. In this work, the authors propose the so-called Line+Cone model, which is used for simplification of belief elicitation and captures users' uncertainty more effectively when compared with Bayesian cognitive modelling. The article concludes with two main findings of interest − the proposed method reduces complexity and effectively captures users' uncertainty about bivariate relationships, and the utilization of uncertainty visualization makes users more confident in their judgments.

The study [16] presents applied correlation analysis and visualization research in the examination of radon concentrations, environmental factors, and seismic activity in Romania. The authors of the study utilize correlation analysis techniques and smart visualization models to enhance the identification and interpretation of patterns in radon emissions related to seismic activity in the selected area. These models point to two main aspects of interest in the studied problem. Firstly, even though the correlation between radon emissions and environmental factors is strong, their direct relationship with seismic activity is marginal. Secondly, the work highlights the need for visualization and advanced analytics for better interpretation of data analysis results.

Several research projects and studies focus on the specific area of correlation analysis concerned with correlation classes, which are significant from the point of view of the work presented in this article.

In [13], the authors present a methodology containing constraints which need to be satisfied for the bivariate linear regression to be applied correctly. This objective is directly related to the need for the determination of the correct calculus of the bivariate linear correlation coefficient and the correlation class of this coefficient's value for the purposes of the evaluation of the proposed methodology.

The authors of [19] focus on the optimization of Multi-Agent Deep Deterministic Policy Gradient, the effectiveness of which is strongly dependent on local information selection by distance selection method, type selection method and correlation class proximity method. The optimization is then applied in the context of food chain scenarios, where significant improvements are described − specifically, the utilization of the correlation class proximity selection method improves the algorithm's training efficiency and training time consumption.

On the other hand, in [12] the authors propose a new rule for Bayesian updating of classes of precise priors (probability distributions) called the quantile-filtered Bayesian update rule. With the use of this update rule, authors construct a correlation class of precise priors that maintain prescribed precise marginals while allowing for an arbitrary correlation structure, enabling greater flexibility in modelling dependencies between attributes.

2. CORRELATION ANALYSIS AND VISUALIZATION

As noted in the introductory section of this work, correlation analysis is a method of descriptive statistics focused on the identification of predictive potential between the values of two attributes in a common dataset [1]. This predictive potential, measured through the correlation coefficient, defines the strength and direction of the functional relationship between the selected attribute pair [15]. Specifically, the sign of the correlation coefficient value determines the direction of the relationship, and the value itself denotes its strength.

In general, there are two basic types of correlation coefficients categorized on the basis of the type of functional relationship they measure − linear coefficients and nonlinear monotone coefficients. Regardless of its type, the correlation coefficient acquires the values from the $[-1, 1]$ interval, where the closer the values of the coefficient are to the extremes of this interval, the stronger the relationship between the selected pair of attributes [18].

The most commonly used correlation coefficient is the Pearson correlation coefficient (conventionally denoted by r), which measures the strength and direction of the linear functional relationship between attributes A and B as follows [24]:

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

where A_i and B_i denote the ith measurement of the attributes A and B, $\mu(A)$ and $\mu(B)$ are mean values of the attributes A and B, and n denotes the overall number of measurements of the two studied attributes in the dataset.

However, the limitation of the Pearson correlation coefficient to the measurement of strictly linear relationships is − from the point of view of the analysis of data − quite significant. This led to the use of the so-called rank correlation coefficient, which abstracts from the use of values of the attribute pair and, instead, works with the ranking of these values, which ensures the possibility of determining the strength and direction of nonlinear monotone relationships between said attribute pairs. The first of such measures is the Spearman correlation coefficient (ρ) computed as [1]:

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

where d_i is the difference between rankings of the ith values of the studied attributes A and B, and n is the number of measurements of the attributes in the dataset.

Similarly to the Pearson correlation coefficient, the Spearman correlation coefficient has its limitations, mainly the tendency to deviate in the correlation measurement caused by repeated values in an attribute. This issue is solved using the Kendall correlation coefficient (τ), which utilizes combinations of rankings of all values in attributes A and B in the following way [15]:

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

where n_c is the number of so-called concordant pairs of rankings for attributes A and B − pairs of rankings where the relative order of $rank(A)$ and $rank(B)$ is the same, the n_d denotes the number of discordant pairs of rankings for the attribute − meaning the opposite situation to concordance, and n is the number of measurements of the values in the dataset.

2.1. Visualization in the context of correlation analysis

Since any type of correlation coefficient measures the prediction potential stored in a pair of attributes, it is evident that one value of such a coefficient is not able to measure the correlation in the whole dataset consisting of more than two attributes. This leads to the need for a type of summarization technique used in the description of the prediction potential for all possible pairs of attributes in a dataset − correlation matrix (denoted as $\mathbb{C}$). Since the rows and columns of a correlation matrix are indexed by the quantitative attributes of the studied dataset, for a dataset of n attributes, such a correlation matrix is of size $n \times n$. Each element of the $\mathbb{C}$ contains the value of the selected correlation coefficient measured between the attributes which cross-index the matrix element itself.

Therefore, the correlation matrix of a dataset composed of n attributes can be generalized as [10]:

$$\begin{array}{cccccc}
 & attr_1 & attr_2 & attr_3 & \dots & attr_n \\
attr_1 & corr(attr_1, attr_1) & corr(attr_1, attr_2) & corr(attr_1, attr_3) & \dots & corr(attr_1, attr_n) \\
attr_2 & corr(attr_2, attr_1) & corr(attr_2, attr_2) & corr(attr_2, attr_3) & \dots & corr(attr_2, attr_n) \\
attr_3 & corr(attr_3, attr_1) & corr(attr_3, attr_2) & corr(attr_3, attr_3) & \dots & corr(attr_3, attr_n) \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
attr_n & corr(attr_n, attr_1) & corr(attr_n, attr_2) & corr(attr_n, attr_3) & \dots & corr(attr_n, attr_n)
\end{array}$$

Such a correlation matrix has the following natural properties − the main diagonal of the matrix contains the correlation coefficient measured between an attribute and itself, therefore this correlation is always equal to 1; since the order of attributes in a studied pair is irrelevant, the upper and lower triangle of the matrix are equivalent.

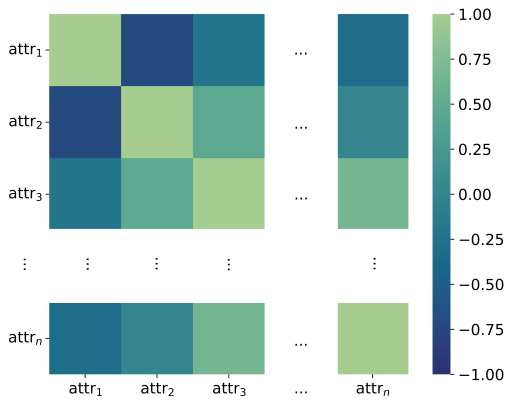

Fig. 1. An example of a correlation heatmap based on the generalized correlation matrix.

One can easily imagine a scenario in which a dataset of high dimensions is studied with the use of the mentioned correlation analysis approaches. Such datasets tend to produce large, unreadable correlation matrices, which are hard to navigate through − even simple task, such as manual identification of the largest value of correlation in the dataset, is not trivial in these matrices. This motivated the visualization of correlation matrices via correlation heatmaps [10] − a visual model, which codes individual elements of the matrix to the color spectrum, while the correlation coefficient value in the matrix element determines the specific color selected for the coloring of the element (see Figure 1).

Even though, when compared to the correlation matrix summarization, the correlation heatmap is a much clearer representation of the prediction potential in large datasets, this model still produces dense visualizations, which can be hard to read in specific cases, eg. in datasets with similar correlation values, the colors of individual map elements merge together.

On the other hand, a set of graphical models used for visualization of correlation analysis results was proposed in the works [9, 10]. In general, these models can be titled correlation structures or correlation graphs − visualization constructs strongly based on graph theory ideas, where the nodes of the correlation graph represent individual attributes of the studied dataset, and the edges connecting these nodes are weighted by the value of the correlation coefficient between a pair of attributes (nodes). Since such a visualization would produce a similarly dense representation as correlation heatmaps, the correlation graphs are commonly pruned using some type of acceptability border for the lowest value of correlation included in the graph. Figure 2 presents an example of such a pruned correlation graph based on the previously presented correlation matrix and heatmap.

Both − the correlation heatmap and correlation graph − visualizations are relevant from the point of view of the presented study, which proposes a novel visualization approach in the context of correlation analysis based on these models. However, instead of visualizing the correlation coefficient values themselves, the proposed model focuses

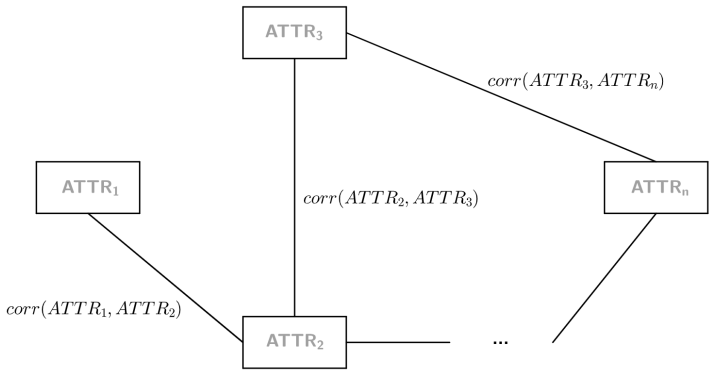

Fig. 2. An example of a correlation graph based on the generalized correlation matrix.

on the visualization of classes of correlation and relationships in and between them.

2.2. Correlation classes

In any human-computer interaction, a certain human fuzzy interpretation of exact values presented by an algorithm is common and natural. When considering correlation analysis, the $[-1, 1]$ interval of correlation coefficient value is conventionally divided into sub-intervals, which can be titled correlation classes [12, 19]. The most basic of such correlation coefficient value divisions is the following [22]:

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

In this correlation classification, the analysts consider only two possible classes of correlation between a pair of attributes A and B − significant correlation between these attributes, in such a case when the absolute value of the correlation coefficient is higher than 0.8, or insignificant correlation in other cases.

However, such a determination of correlation classes is quite limiting and lacks any kind of nuance, which leads to a more detailed classification presented in [13]. The authors of the work consider the division of the correlation coefficient interval into five classes − neutral, weak, moderate, strong, and very strong correlation − as follows:

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

Other than these strict, static definitions of the correlation classes, several dynamic

approaches based on the basic central tendency measures were proposed. The two-tier correlation coefficient value classification is proposed in [11] as:

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

where $\mu(|corr(\mathbb{C})|)$ is the mean value of the correlation coefficient in the datasets described by the correlation matrix $\mathbb{C}$, and $\max(|corr(\mathbb{C})|)$ is the highest value of the correlation matrix (not taking into account the main diagonal).

Similar dynamic classification can be developed on the basis of a combination of any commonly used central tendency measures, eg. three-tier classification of correlation coefficient values based on tertiles ($T1 - T3$):

$$corr_{class}(A,B) = \begin{cases} \text{low}, & \text{if } |corr(A,B)| \in [0, T1), \\ \text{medium}, & \text{if } |corr(A,B)| \in [T1, T2), \\ \text{strong}, & \text{if } |corr(A,B)| \in [T2, T3]. \end{cases}$$

3. HONEYCOMB GRAPHS FOR CORRELATION ANALYSIS

Based on the information stated in the previous section, this part of the work presents the design of the Honeycomb graphs − a novel visualization model for the graphical representation of correlation classes, their internal relationships, and relationships between classes in multidimensional datasets. As already stated, the proposed model is deemed parametric visualization, the mentioned parameter being the so-called attribute of interest ($attr_i$) − an attribute of the studied dataset, which is the main objective of the conducted data analysis (eg. the values of which need to be predicted).

A honeycomb graph is a graphical model of correlation analysis consisting of three main components:

- Cells − The most basic component of a Honeycomb graph is a cell, by the use of which, the individual attributes of the studied dataset are visualized. In the proposed version of the model, there are three considered types of cells distinguished based on their shape − triangular, tetragonal, and hexagonal cells (see Figure 3).

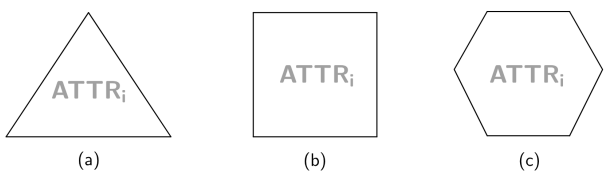

Fig. 3. Cell types considered for the proposed Honeycomb graph visualization − (a) triangular cell, (b) tetragonal cell, (c) hexagonal cell.

The shape of individual cells is determined by the number of attributes, which bear prediction potential to the attribute of interest of the same class (the value of the correlation coefficient), and these cells are constructed into the titular honeycombs.

- Honeycombs – A honeycomb is a star graph, which consists of a set of $n \in [1, 6]$ cells connected to a cell containing an attribute of interest at its centre. Figure 4 presents three basic honeycomb types – triangular honeycomb, where $n = 3$; tetragonal honeycomb, with $n = 4$, and hexagonal honeycomb, where $n = 6$.

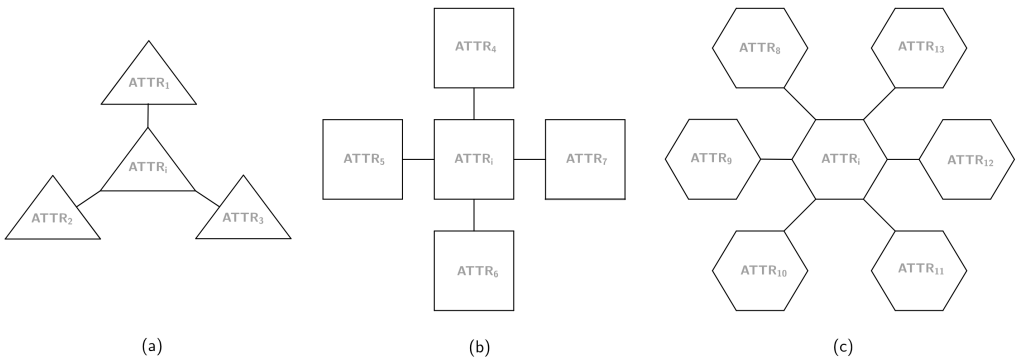

Fig. 4. Honeycomb types considered for the proposed visualization – (*a*) pure triangular honeycomb, (*b*) pure tetragonal honeycomb, (*c*) pure hexagonal honeycomb.

Similar to the shape of cells, the number of cells is determined by the number of attributes in individual correlation classes. Honeycomb dimensions were selected to balance two competing objectives – to visualize as many inter-correlated attributes as possible, and to preserve visual clarity (label legibility, unambiguous edge weights and separation between cells). A six-cell limit of hexagonal honeycombs provides a symmetric, space-efficient packing that maximizes attributes per honeycomb without overlapping or crowding. When a correlation class contains more than six attributes, the class is represented using multiple honeycombs of the same scale and interconnected with inter-honeycomb edges, which keeps individual honeycombs easy to read and facilitates comparison across the class.

- Edges – The last components of the Honeycomb graph are edges – connections between cells of one or more honeycombs, which are weighted by the value of the correlation coefficient measured between attributes in the pair of cells.

Therefore, the process of constructing a Honeycomb graph consists of the following phases:

Phase 1: Correlation class preparation. In the first step of the proposed procedure, the value of the correlation coefficient is measured between the user-defined

attribute of interest and all other attributes in the dataset. The values of the correlation coefficients are, then, classified into correlation classes based on the following pre-defined static borders:

$$corr_{level}(A,B) = \begin{cases} \text{low}, & \text{if } |corr(A,B)| \in [0, 0.5), \\ \text{medium}, & \text{if } |corr(A,B)| \in [0.5, 0.8), \\ \text{strong}, & \text{if } |corr(A,B)| \in [0.8, 1]. \end{cases}$$

for a dataset of 19 or fewer attributes. Or

$$corr_{level}(A,B) = \begin{cases} \text{low}, & \text{if } |corr(A,B)| \in [0, 0.2), \\ \text{marginal}, & \text{if } |corr(A,B)| \in [0.2, 0.4), \\ \text{medium}, & \text{if } |corr(A,B)| \in [0.4, 0.6), \\ \text{significant}, & \text{if } |corr(A,B)| \in [0.6, 0.8), \\ \text{strong}, & \text{if } |corr(A,B)| \in [0.8, 1]. \end{cases}$$

for a dataset of 20 or more attributes.

This division into two correlation class systems is motivated by the convention of identifying an odd number of classes in systems − most typically three or five. Since the largest − hexagonal − honeycombs considered in this work contain an attribute of interest and 6 other attributes, this determines that for visualization of a dataset of 19 attributes, three full hexagonal honeycombs are needed (6 for each honeycomb plus the attribute of interest). Naturally, if the number of attributes is lower than 19, the honeycombs are either incomplete or they change their shape as noted in phase 2.

Phase 2: Honeycomb construction. After the identification of attributes in the dataset and their assignment into defined correlation classes, the honeycombs themselves are constructed. Each honeycomb consists of a cell containing an attribute of interest at its centre, and a set of one to six attributes, which bear the correlation coefficient value of the same class measured between the set of attributes and the defined attribute of interest. Therefore, each honeycomb consists of a set of cells, which are connected to the attribute of interest with edges weighted by the correlation coefficient value measured between the attribute of interest and the other attribute. In this way, each honeycomb defines either the whole or part of a correlation class in the studied dataset and the values of individual correlation coefficients are identified by the edge values.

The type of honeycomb used for the visualization of the (part of) correlation class is defined by the number of attributes in the class (see Figure 4, where the correlation values between cells of the honeycombs are neglected for the purposes of the schema readability):

- triangular honeycombs for classes of up to 3 attributes,
- tetragonal honeycombs for correlation classes containing up to 4 attributes,
- hexagonal honeycombs for classes containing up to 6 attributes.

Naturally, a correlation class of a dataset can contain such a number of attributes that there is no possibility of the construction of full triangular, tetragonal, or hexagonal honeycombs. This leads to the so-called incomplete honeycombs, eg. a hexagonal honeycomb with only five cells connected to the attribute of interest.

On the other hand, the possibility of a correlation class containing more than six attributes (the highest value for the considered honeycombs) is also common. In such a case, one class can be visualized using a set of honeycombs, which can be of mixed types and completeness.

Phase 3: Relationship identification. Other than edges that connect the attribute cells to the cell of the attribute of interest, the honeycomb graphs contain two other types of edges denoting relationships between attributes of the dataset.

The first type of such a relationship is determined by the inter-honeycomb edges, which visualize the strongest relationships between attributes of two adjacent correlation classes, eg. the strongest correlation coefficient values between a set of attributes with marginal and medium correlation measured between the attributes and the defined attribute of interest.

The other type of relationships visualized in the proposed models are the intra-honeycomb edges, which connect pairs of attributes in the same honeycomb between which there is the highest correlation coefficient value. Both of these relationships are visualized using a dashed line in Figure 5, while the correlation values between cells of honeycombs and $attr_i$ cell are neglected for the purposes of the schema readability.

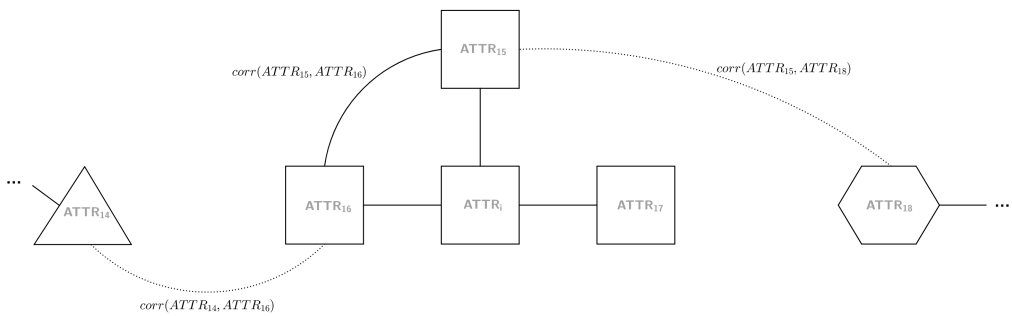

Fig. 5. An example of the use of inter-honeycomb and intra-honeycomb edges.

The flowchart of the proposed process is presented schematically in Figure 6. In this way, the presented method constructs a parametric visual assignment of attribute sets into individual correlation classes, while highlighting the strongest relationships between and in the classes themselves. Such visualization can be utilized in the task of feature (attribute) selection, where the analysts can, on the basis of the visualization, intelligently decide which attributes are used in the decision-making process and which are not, eg. one can only use the attributes of strong correlation classes. Additionally, since the visualization contains information regarding relationships between classes and

attributes in the classes, analysts can use these relationships in order to complete the value sets for the prediction models. For example, in Figure 5 there is a strong relationship between $attr_{15}$ and $attr_{16}$, and both of these attributes influence the values of the attribute of interest to some degree. In such a case, where the values of $attr_{15}$ are unknown or missing, the value of $attr_{16}$ can be used in a regression task to compute this missing value, hence, allowing more precise prediction of the value of the attribute of interest.

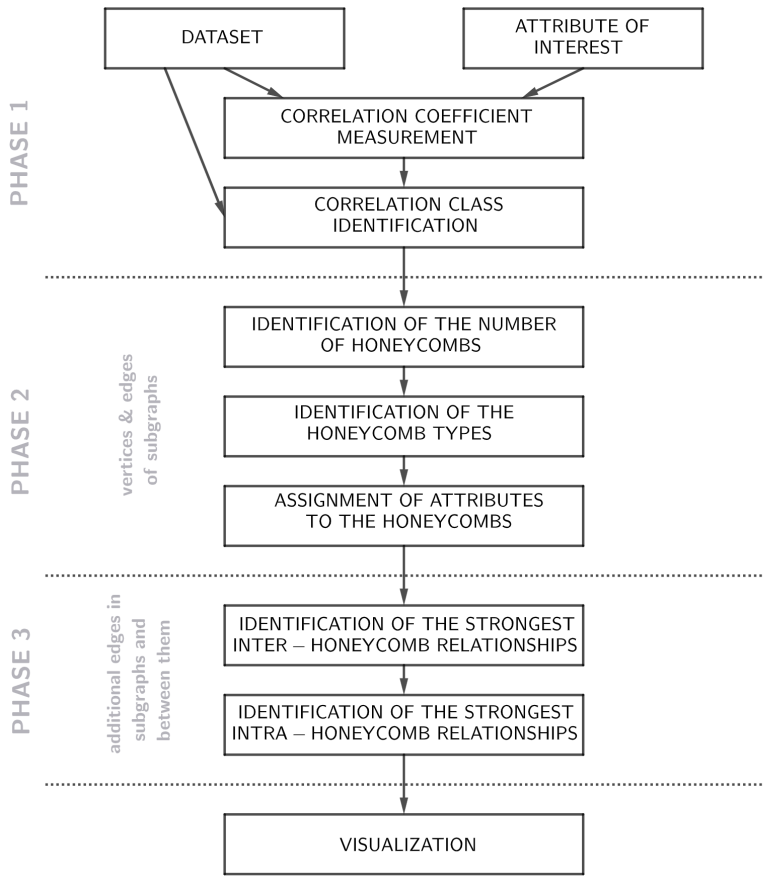

Fig. 6. Flowchart of the proposed process from reading of the input data to the honeycomb visualization.

The following are some of the basic properties of Honeycomb graphs, which are proposed in the design of the method itself:

- Maximization of pureness of individual honeycombs − in the context of a honeycomb graph, no honeycomb consists of more than one type of cell. Therefore, in the proposed method, only pure triangular, tetragonal, or hexagonal honeycombs

are utilized. However, as mentioned above, these honeycombs can be complete or incomplete, meaning a honeycomb can contain the maximum allowed number of cells, but is not required to.

- Selection of proper color mapping − each cell of each honeycomb is color coded based on the strength of the correlation coefficient measured between the attribute, which indexes the cell and the attribute of interest. This color is automatically selected from the colormap presented in Figure 7.

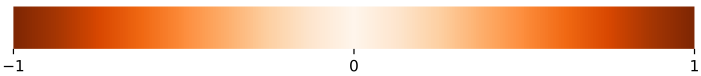

Fig. 7. Color map for coding of correlation coefficient values in the proposed visualization model.

As can be seen, the colormap constructed for the model utilized the same colors in both correlational directions. This is due to the fact that from the point of view of correlation classes, the absolute value of correlation is always examined. Therefore, the color of the cell signifies the level of measured correlation or anticorrelation.

- Automatic construction of legend − the last of the described properties of the proposed visualization model focuses on the automatic construction of the attribute abbreviation legend. Such a legend is necessary for the overall readability of Honeycomb graphs constructed on datasets with attributes labelled with long strings, which can overflow the assigned shapes and dimensions of cells. The abbreviations of attribute titles are created as a set of unique symbols constructed from the beginning symbols of attribute titles, and the legend presents pairs of abbreviations and full attribute titles.

4. CASE STUDIES OF CORRELATION CLASS IDENTIFICATION AND VISUALIZATION

The proposed visualization model is implemented in the *Python* programming language with the use of *pandas* and *numpy* packages for the processing and analysis of input data, *networkx*, *shapely*, and *matplotlib* packages for the construction of graphical elements of the visualization, and *seaborn* package for the stylization of the visualization.

In the following section, the experimental verification of the Honeycomb graph approach is studied on two datasets from the area of chemical analysis and electrical engineering:

- Wine quality dataset [7] − the dataset consists of 4 898 chemical property measurements of red and white variants of the Portuguese wine. Each of these measurements is described by 11 chemical properties of the wine, such as various forms of wine acidity, levels of residual sugars, chlorides, sulphates, pH and so on [8].

- Appliance energy dataset [2] − this dataset consists of measurements of internal temperature and humidity conditions of a specific building done by a wireless sensor network, which are combined with weather information and data on electricity consumption in the building. These values are contained in 19 735 measurements and 28 attributes [3].

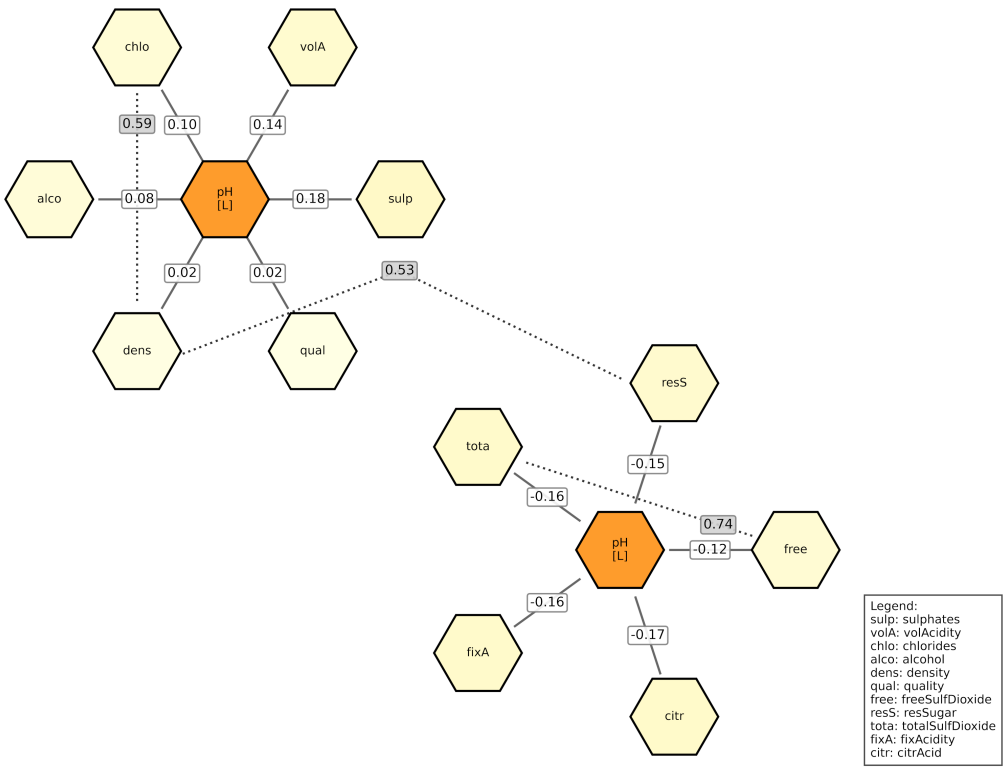

Fig. 8. Honeycomb graph of the Wine quality dataset with pH as the attribute of interest.

After the utilization of the datasets in the presented case studies, the proposed visualization model is evaluated from two points of view − the qualitative aspects of the evaluation are based on the Qualitative Result Inspection (QRI) and Visual Data Analysis and Reasoning (VDAR) approaches [14], while the quantitative elements are focused on generalizability, precision and realism of the visual data representation [4]. Additionally, comparative analysis between the Honeycomb graph method and conventional methods of visualization and correlation class identification is conducted.

4.1. Honeycomb graphs of the Wine dataset

The visualization of the Wine quality dataset presents two specific scenarios of analysis using Honeycomb graphs.

Figure 8 contains a Spearman-type honeycomb graph for the dataset while using pH as the attribute of interest. As can be seen, in this case, the visualization consists of two honeycombs − both of hexagonal type. These honeycombs represent one correlation class of the dataset, specifically low correlation (denoted by the abbreviated title of the correlation class [L]). Naturally, the amount, type and correlation class of honeycombs varies based on the strength of correlation coefficient values between attributes and attributes of interest.

Regarding the relationships between identified classes, one can observe the relationship between *density* and *resSugar* attributes ($\rho(density, resSugar) = 0.53$), which interconnects the two honeycombs of the low correlation class. The inter-honeycomb relationships show significant correlation values between *chlorides* and *density* of value 0.59, and $totalSulfDioxide$ and $freeSulfDioxide$, where

$$\rho(totalSulfDioxide, freeSulfDioxide) = 0.74.$$

On the other hand, Figure 9 presents an alternative visualization of the same dataset, with the attribute of interest set to *density* of the measured wine sample. Unlike the visualization in Figure 8, one can note three honeycombs of two types, two tetragonal and one triangular, representing three correlation classes − low correlation class ([L]), medium correlation class ([ME]), and marginal correlation class ([MA]).

4.2. Honeycomb graphs of the appliance energy dataset

Since the Appliance energy dataset is composed of 28 attributes, its correlational structure − and, therefore, the visualization − is more complex. Figure 10 shows the Honeycomb graph model of the dataset with RH_1 (relative humidity 1) for the attribute of interest.

The number of honeycombs in the presented visualization is of note − the graph is composed of seven honeycombs describing five correlation classes (two low correlation honeycombs ([L]), two marginal honeycombs ([MA]), and one honeycomb of each for the medium ([ME]), significant ([SI]), and strong ([ST]) correlation classes).

When compared to the previous case study, one can see that the Appliance Energy dataset contains both complete and incomplete honeycombs and several interesting inter- and intra-honeycomb relationships.

The visualization model, as proposed and implemented, contains two elements of interactivity:

- scaling − necessitating zooming in and out of the graph and its components,
- and listing − meaning interactive scrolling of the zoomed-in image.

Therefore, the label positioning of correlation coefficient values on edges is dynamically adjusted as needed. In this way, the label crossing presented in Figure 10 is minimized (see Figure 11 for an example of this phenomenon).

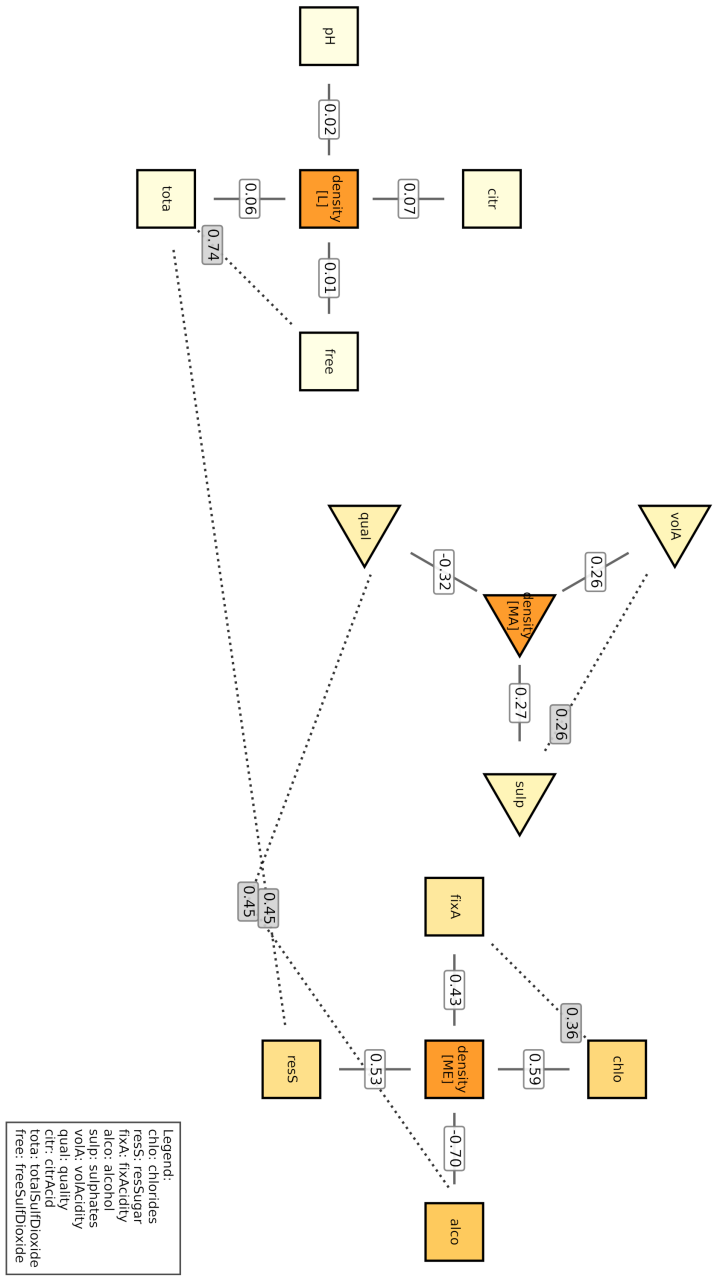

Fig. 9. Honeycomb graph of the Wine quality dataset with *density* concentration as the attribute of interest.

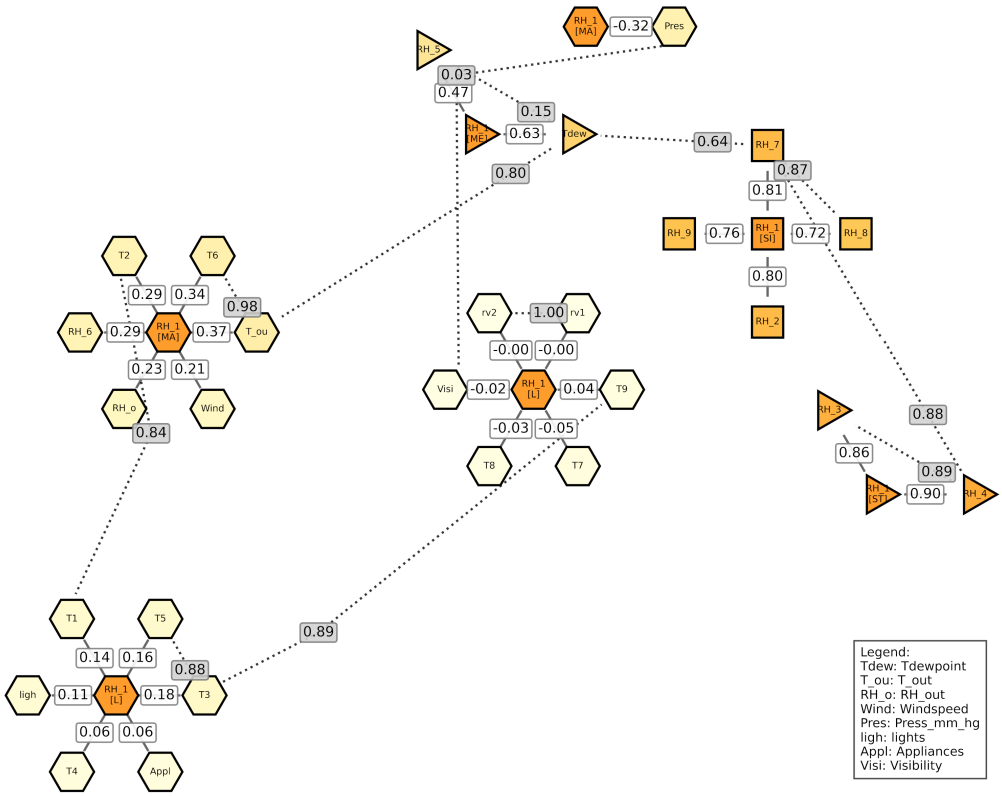

Fig. 10. Honeycomb graph of the Appliance energy dataset with RH_1 as the attribute of interest (raw program output slightly adjusted for better readability).

4.3. Evaluation of the proposed visualization approach

As noted in the introduction of this section, for the evaluation of the presented visualization method, two qualitative and one quantitative criterion are considered. From the qualitative point of view, the following aspects are relevant [14]:

- Qualitative Result Inspection (QRI) – the main objective of this evaluation technique is focused on the quality of the fulfilment of the visualization's purposes. In this case, the purpose of the visualization is the representation of correlation classes and relationships in and between them in the form of a Honeycomb graph. As presented in the case studies, the implemented algorithm parametrically divides attributes of the studied dataset into statically defined classes, visualizes them in the form of honeycombs according to the pre-determined visualization criteria, and interconnects the cells of these honeycombs with inter- or intra-honeycomb relationships (edges). Since the visualization presented the data in such a way that

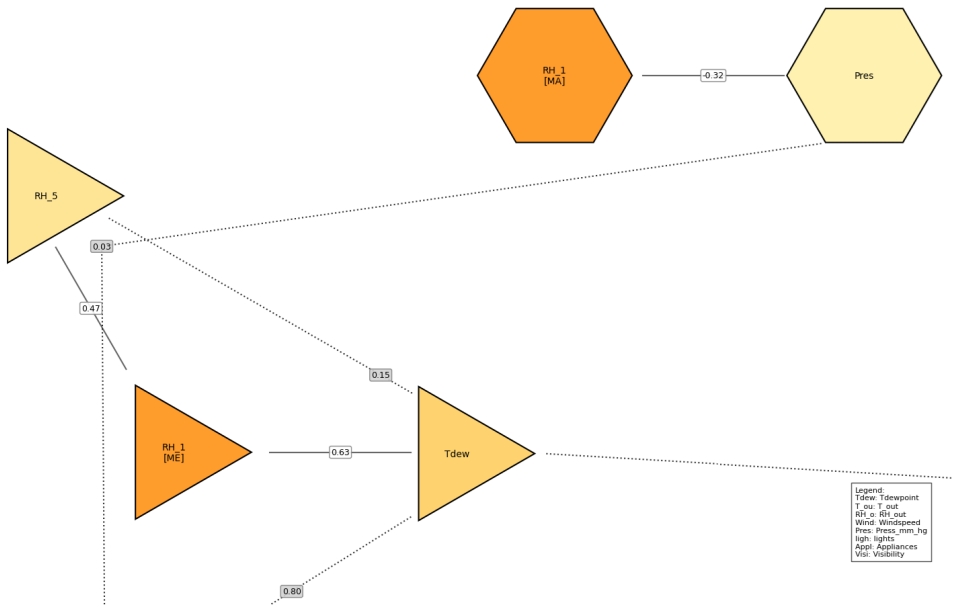

Fig. 11. Zoomed in part of the Honeycomb graph of the Appliance Energy dataset.

the classes of correlation and relationships between them are easily interpretable, the QRI of the model can be considered successful.

- Visual Data Analysis and Reasoning (VDAR) – this method focuses on the evaluation of the proposed visualization based on its quality in the solution of the problem, the visualization model is designed for. In the case of Honeycomb graphs, three interconnected problems are of importance:
  - Correlation analysis – the aspect of highest significance studied in the presented model is a parametric identification of correlation classes in multidimensional datasets, the strongest relationships between these classes and within them. As can be seen in Figures 8 – 10, this objective is reached using the grouping of the attributes into honeycombs, while using edges to highlight the inter- and intra-honeycomb relationships of highest strength (the highest value of correlation coefficient).
  - Feature selection – based on the parametric classification of attributes of a studied dataset into correlation classes, analysts are able to select attribute sets with the strongest direct predictive potential related to the attribute of interest – attributes classified into [ST] or [SI] classes/honeycombs. In this way, the inputs for predictive analysis models can be tuned so the quality of model decision-making is not lowered by irrelevant – not correlated – attributes. Other than direct relationships, Honeycomb graphs visualize in-

teresting indirect relationships in a dataset via intra-honeycomb edges. These edges can be used in order to find predictive sequences of note, which can be used in regression analysis [10].

– Regression analysis – as stated above, the correlation values measured in a dataset have a direct influence on regression models constructed on them. Therefore, using correlation classes identified via Honeycomb graphs, analysts can prune datasets and use only strongly correlated attributes as inputs for such models, thus producing lower error rates when compared to the regression models without the use of feature selection.

The second evaluation focuses on the quantifiable components of the proposed visualization method based on the work presented in [4]. These qualitative components can be the generalizability of the model, precision of the method, or realism of the visual data representation.

From the point of view of the model's generalizability – the behaviour of the method with various datasets and input parameters – Tables 1 and 2 are presented.

In Table 1, the number of correlation classes (*#classes*) and the number of honeycombs in the Honeycomb graph (*#honeycombs*) are quantified for selected attributes of interest of both studied datasets. As can be seen, there are specific cases where the amount of classes is equal to the amount of honeycombs, but also cases where there are fewer classes than honeycombs. One such instance is the utilization of *pH* as the attribute of interest in the Wine dataset, where only a low correlation class is identified, but since the dataset contains more than six attributes, the model, as proposed, uses two hexagonal honeycombs for the visualization (see Figure 8). Similarly, there are only two correlation classes identified in the Appliance Energy dataset when using *lights* as the attribute of interest, yet these two classes are divided into five honeycombs.

Naturally, with the growing number of classes, the number of honeycombs also grows, while the lowest possible ratio of classes to honeycombs is 1 : 1.

Dataset	**Attribute of interest**	**# classes**	**# honeycombs**
Wine	fixedAcidity	3	3
	freeSulfDioxide	3	5
	density	3	3
	pH	1	2
Appliance Energy	appliances	2	5
	lights	2	5
	T1	5	8
	RH_1	5	7

Tab. 1. Quantification of Honeycomb graph components based on selected attributes of interest.

The descriptive summarization of the number of classes and honeycombs in visualizations created over the two studied datasets is presented in Table 2. The table contains minimal (*min*), mode (*mode*), and maximal (*max*) values for the number of correlation classes and honeycombs in both datasets.

Property \ Dataset	Wine	Appliance Energy
number of attributes	11	28
min(#classes)	1	2
mode(#classes)	3	5
max(#classes)	4	5
min(#honeycombs)	2	5
mode(#honeycombs)	3	5
max(#honeycombs)	5	8

Tab. 2. Descriptive summarization of the number of classes and honeycombs in visualizations of Wine and Appliance Energy datasets.

Regarding the other two quantitative elements of the method − the precision of the proposed model is generally high, with only a minor issue related to the slight overlap of certain elements in the visualization (some edges and their labels). However, overall, the data is well-separated into distinct classes. Additionally, the visual representation of the data is realistic, providing an insightful view of the parametric classification of attributes into correlation classes and the relationships between these classes. This enhances interpretability and supports a more intuitive understanding of the dataset's structure.

The feature selection process using Honeycomb graphs offers several advantages. Firstly, in the honeycombs, the identification of the most relevant attributes from the point of view of predictive analysis is trivial and user-friendly. If the values of the attributes with the strongest relationship to the attribute of interest are not explicitly known, it is possible to infer them transitively by leveraging strongly related attributes via inter- or intra-honeycomb relationships. Secondly, attributes that share strong interdependencies can be considered interchangeable in a predictive context. This means that certain attributes can be used in the same way without losing predictive power, allowing for greater flexibility in model construction and interpretation.

4.4. Comparative analysis of the Honeycomb graphs and conventional methods

As a last part of the evaluation of the proposed approach, the comparative analysis between the conventional methods and Honeycomb graphs is conducted. This comparison is done from two points of view:

- the comparison between the visualization of correlation classes in the studied datasets using correlation heatmaps, correlation graphs, and Honeycomb graphs,
- the comparison between the conventional classification of correlation coefficient values presented in Section 2.2 and the proposed method.

Since the proposed model is unique in its application as a correlation class visualization method, the first part of the presented comparative analysis focuses on the visual-

ization in the context of correlation analysis. Figures 12 and 13 present the correlation matrix and graph of the Wine Quality dataset, respectively, while the conventional visualization of correlation analysis in the Appliance Energy dataset is presented in Figures 14 and 15.

As can be seen, both of the used methods focus on general visualization of relationships in the dataset − the heatmap presents the summarization of correlation coefficient measurements of the dataset, while the graph focuses on pseudo-transitivity of these relationships and possible identification of predictive sequences, which can be utilized in regression tasks. On the other hand, the model of Honeycomb graphs aims towards the visualization of correlation classes and relationships in and between them. Naturally, the heatmap-based and graph-based visualization can be used to extract classification of correlation coefficient values (eg. by filtering the heatmap and selecting only the cells of a similar color), but these approaches require additional computations or user activity to do so.

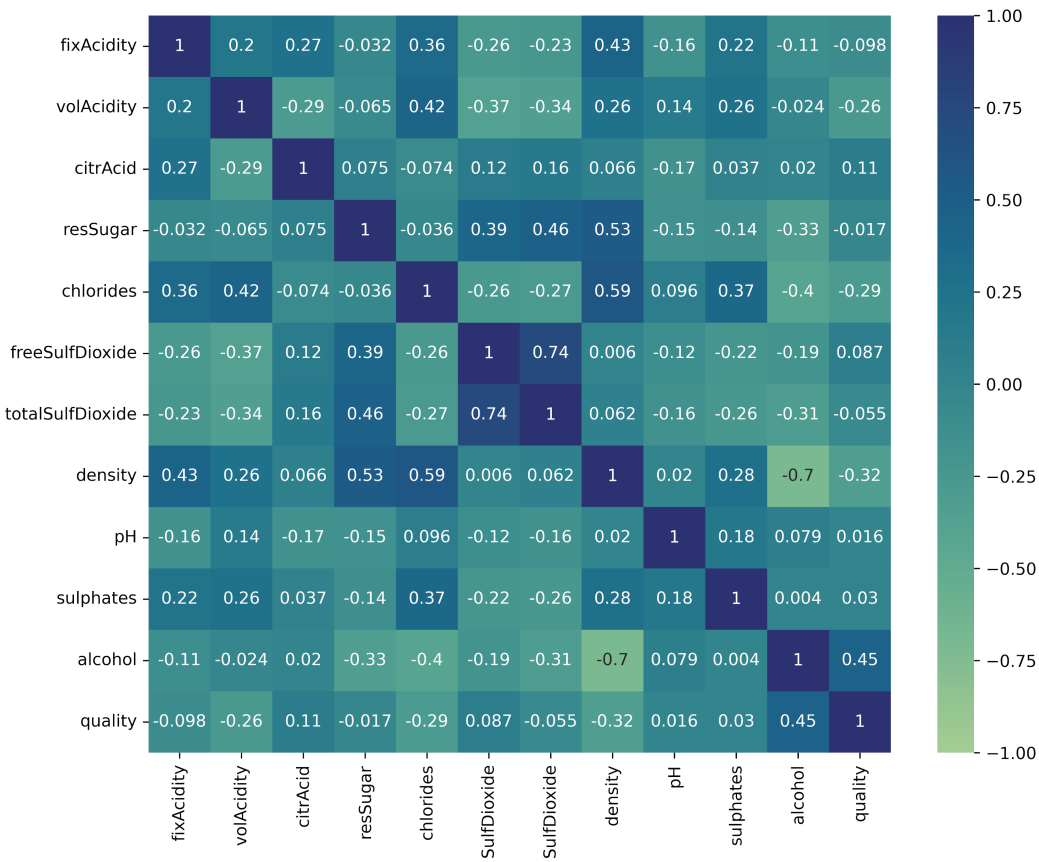

Fig. 12. Correlation matrix of Spearman type for the Wine quality dataset.

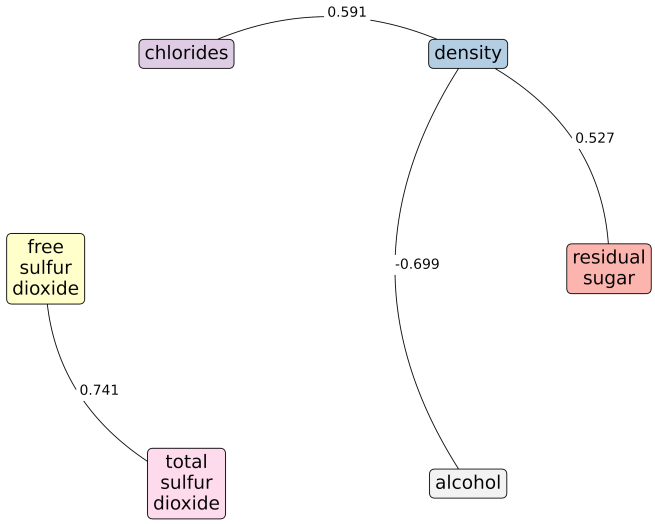

Fig. 13. Correlation graph of Spearman type for the Wine quality dataset.

The second aspect of the proposed visualization model consists of the identification of correlation classes in data. Tables 3 and 4.4 present the assignment of attributes from the considered datasets to correlation classes on the basis of the correlation coefficient computed between the attribute of interest and all other attributes. For comparison, two attributes of interest were selected for each dataset − *density* and *resSugar* for the Wine Quality dataset, and *Tdewpoint* and *T_out* for the Appliance Energy dataset.

The tables highlight several notable features of the basic correlation-class identifications. Under the static approach, some classes are frequently empty and therefore redundant (for example, the *strong* and *very strong* classes in the Wine Quality dataset). By contrast, the dynamic approach − which partitions correlation-coefficient values into tertiles − produces classes that are roughly equal in size, i. e., each contains a similar number of attributes.

Applying the Honeycomb graph model yields classifications broadly consistent with the other methods, but with two important differences. Compared to the static approach, the Honeycomb model does not produce empty classes and thus avoids honeycombs that would contain only the attribute of interest. Compared to the dynamic method, class sizes under the Honeycomb model can vary, which aligns with the natural distribution of correlation coefficient values. A key advantage of the proposed Honeycomb model is its explicit identification of the strongest inter- and intra-class relationships via honeycomb edges, together with a distinctive visualization capability that, while applicable to any classification scheme, is currently unique to Honeycomb graphs.

The main disadvantage of the proposed method when compared to the conventionally utilized correlation classifications is the models computational complexity caused by the

construction and visualization of the honeycombs themselves and by the determination of the strongest connecting relationships. Estimated computational complexity of the approach for the Spearman rank correlation coefficient and dataset consisting of n attributes and m measurements of each attribute is $0(n^4)$, while static five-class model ranks approximately $O(n \times m)$ and $O(n \times \ m \times \ log\ m)$ for the dynamic tertile-based model.

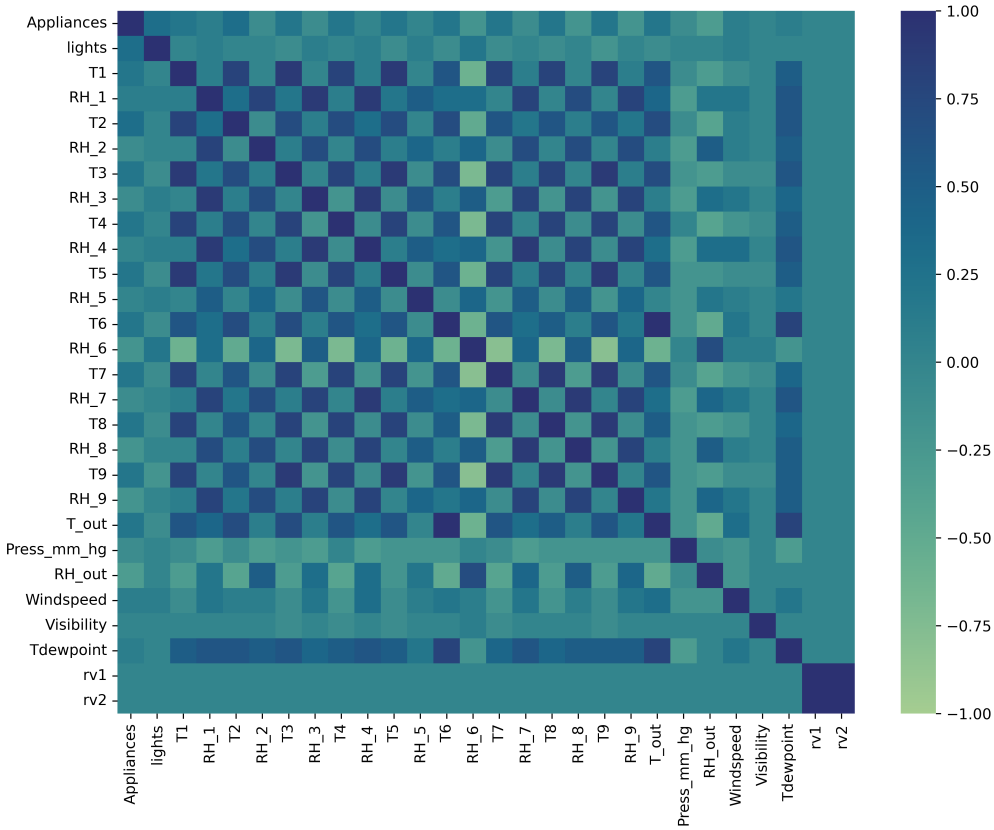

Fig. 14. Correlation matrix of Spearman type for the Appliance Energy dataset.

5. CONCLUSION

In this study, we focus on the design, implementation, and experimental evaluation of a novel visualization technique called Honeycomb graphs for the parametric identification of correlation classes in multidimensional datasets. The proposed method involves an interconnected graphical model consisting of clusters of vertices organized around a user-defined attribute of interest and can be utilized in feature selection problems or as a regression analysis optimization tool.

Dataset	Attribute of interest	Class	Attributes
Wine Quality	density	neutral	freeSulfDioxide, pH, totalSulfDioxide, citrAcid
		weak	volAcidity, sulphates, quality, fixAcidity
		moderate	resSugar, chlorides, alcohol
		strong	–
		very strong	–
	resSugar	neutral	quality, fixAcidity, chlorides, volAcidity, citrAcid,sulphates, pH
		weak	alcohol, freeSulfDioxide
		moderate	totalSulfDioxide, density
		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
	T_out	neutral	rv2, rv1,Visibility, RH_5, lights, RH_8, RH_3, RH_2, Press_mm_hg
		weak	Appliances, RH_9, Windspeed, RH_7, RH_4, RH_1
		moderate	RH_out, T8, T7, RH_6
		strong	T5, T4, T9, T1, T3, T2
		very strong	Tdewpoint, T6

Tab. 3. Static five-level model of correlation class identification in Wine and Appliance Energy datasets.

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
	resSugar	1. tertile	quality, fixAcidity, chlorides, volAcidity
		2. tertile	citrAcid, sulphates, pH
		3. tertile	alcohol, freeSulfDioxide, totalSulfDioxide, density
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
	T_out	1. tertile	rv1, rv2, Visibility, RH_5, lights, RH_8, RH_3, RH_2, Press_mm_hg
		2. tertile	Appliances, RH_9, Windspeed, RH_7, RH_4, RH_1, RH_out, T8, T7
		3. tertile	RH_6, T5, T4, T9, T1, T3, T2, Tdewpoint, T6

Tab. 4. Dynamic three-level model of correlation class identification in Wine and Appliance Energy datasets.

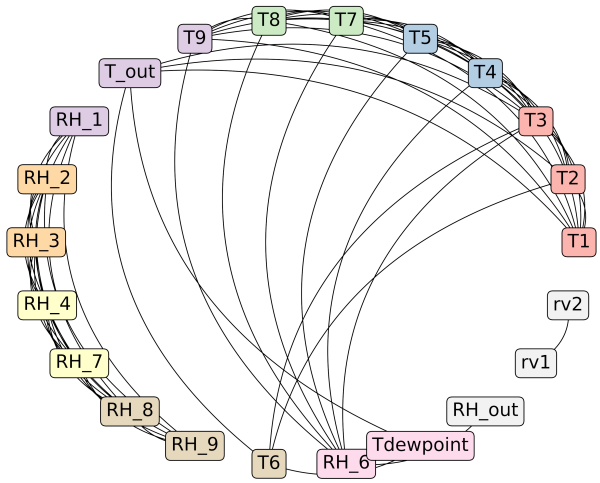

Fig. 15. Correlation graph of Spearman type for the Appliance Energy dataset.

As the results of case studies presented in this work, the model's ability to identify correlation structures and relationships between attributes through Honeycomb graphs was examined. The model supports interactive exploration of correlation analysis and classification while adhering to qualitative and quantitative aspects of visualization design – mainly focused on the effectiveness of the method in representing data relationships with high precision and realism. Additionally, the approach has high potential in feature selection tasks by identifying key attributes and their interdependencies, making it a valuable tool for predictive analysis and data exploration.

From the implementation of the proposed model, several novel areas of potential future work arose:

- Interactivity of the visualization – in the future approaches to the problem, interactive features that allow users to move and adjust the honeycombs and their individual cells should be implemented for better interpretability of the visualization, and for the higher efficiency of the decision-making done on the basis of the model.

- Dynamic definition of correlation classes – enabling the dynamic definition and redefinition of correlation classes based on the input data should allow to construction of dataset-specific visualizations, which would lead to more precise knowledge mining.

- Embedded regression modelling – integrating regression models directly into the visualization to predict, estimate, and analyze data trends in and between the correlation classes would make the visualization more practical and self-contained.

ACKNOWLEDGEMENT

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.

CODE AND DATA AVAILABILITY

Code for the presented visualization methods of Honeycomb graphs is available at:
`https://github.com/AdamDudasUMB/honeycombGraphs`
Data used in case studies presented in this study are freely available at:
`https://archive.ics.uci.edu/dataset/186/wine+quality`
`https://archive.ics.uci.edu/dataset/374/appliances+energy+prediction`

(Received June 10, 2025)

REFERENCES

[1] E. Birihanu and I. Lendák: Explainable correlation-based anomaly detection for industrial control systems. Forntiers Artificial Intelligence *7* (2025). DOI:10.3389/frai.2024.1508821

[2] L. Candanedo: Appliances Energy Prediction [Dataset]. UCI Machine Learning Repository. DOI:10.24432/C5VC8G

[3] L. Candanedo, V. Feldheim, and D. Deramaix: Data driven prediction models of energy use of appliances in a low-energy house. Energy Buildings *140* (2017), 81–97. DOI:10.1016/j.enbuild.2017.01.083

[4] S. Carpendale: Evaluating Information Visualizations. Lecture Notes in Computer Science *4950* (2008). DOI:10.1007/978-3-540-70956-5_2

[5] Y. F. Chen, Y. T. Long, Z. Yang, and J. Long: Correlation embedding semantic-enhanced hashing for multimedia retrieval. Image Vision Comput. *154* (2025). DOI:10.1016/j.imavis.2025.105421

[6] A. Chen, C. D. Wu, and C. J. Leng: Hourglass-GCN for 3D human pose estimation using skeleton structure and view correlation. Computers Materials Continua *82* (2025), Q, 173–191. DOI:10.32604/cmc.2024.059284

[7] P. Cortez, A. Cerdeira, F. Almeida et al.: Wine Quality [Dataset]. UCI Machine Learning Repository. DOI:10.24432/C56S3T

[8] P. Cortez, A. Cerdeira, F. Almeida et al.: Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems *47* (2009), 4, 547–553. DOI:10.1016/j.dss.2009.05.016

[9] A. Dudáš: Graphical representation of data prediction potential: correlation graphs and correlation chains. Visual Computer *40* (2024), 10, 6969–6982. DOI:10.1007/s00371-023-03240-y

[10] A. Dudáš A, E. Kršák, and M. Kvaššay: Exploration and deconstruction of correlation cycles in multidimensional datasets. Technologies *13* (2025), 2. DOI:10.3390/technologies13020085

[11] A. Dudáš and M. Vagač: Diagnostic analysis approach to correlation maps through large language models. In: Proc. 2024 IEEE 17th International Scientific Conference on Informatics 2024. DOI:10.1109/Informatics62280.2024.10900889

[12] H. Held: Quantile-filtered Bayesian learning for the correlation class. In: Proc. 5th International Symposium on Imprecise Probability 2007, pp. 223–232.

[13] L. B. Iantovics: Avoiding mistakes in bivariate linear regression and correlation analysis, in rigorous research. Acta Polytechn. Hungarica *21* (2024), 6, 33–52. DOI:10.12700/APH.21.6.2024.6.2

[14] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Moller: A systematic review on the practice of evaluating visualization. IEEE Trans. Visual. Computer Graphics *19* (2013), 12, 2818–2827. DOI:10.1109/TVCG.2013.126

[15] M. Jamei, N. Bailek, K. Bouchouicha, M. A. Hassan, A. Elbeltagi et al.: Data-driven models for predicting solar radiation in semi-arid regions. Computers Materials Continua *74* (2023), 1, 1625–1640. DOI:10.32604/cmc.2023.031406

[16] 0. Jianu and M. Dragoicea: Enhancing seismic analysis: A fusion of smar visualization and correlation techniques. Univ. Politeh. Bucharest Sci. Bull. Ser. C – Electr. Engrg. Comput. Sci. *86* (2024), 3, 51–66.

[17] A. Karduni, D. Markant, R. Wesslen, and V. W. Dou: A Bayesian cognition approach for belief updating of correlation judgement through uncertainty visualizations. IEEE Trans. Visual. Computer Graphics *27* (2021), 2, 978–988. DOI:10.1109/TVCG.2020.3029412

[18] S. Lee, H. Seong, S. Lee, and E. Kim: Correlation verification for image retrieval and its memory footprint optimization. IEEE Trans. Pattern Anal. Machine Intell. *47* (2025), 3, 1514–1529. DOI:10.1109/TPAMI.2024.3504274

[19] L. C. Li, Z. X. He, B. Z. Wang, Z. Wang, and L. B. Li: Multi-agent reinforcement learning algorithm based on local information. In: Lecture Notes in Electrical Engineering. Proc. 2022 International Conference on Autonomous Unmanned Systems 2022, *1010*, pp. 3080–3091. DOI:10.1007/978-981-99-0479-2_284

[20] B. E. Monroy-Castillo, M. A. Jácome, and R. C. R. Cao: Improved distance correlation estimation. Appl. Intelligence *55* (2025), 4. DOI:10.1007/s10489-024-05940-x

[21] N. Pahuja: Correlations in multispecies PASEP on a ring. Electron. Commun. Probab. *30* (2025). DOI:10.1214/25-ECP666

[22] S. S. Skiena: The Data Science Design Manual. Springer, 2017.DOI:10.1007/978-3-319-55444-0

[23] K. H. Yang, C. W. She, W. Zhang, J. Q. Yao, and S. S. Long: Multi-label learning based on transfer learning and label correlation. Computers Materials Continua *61* (2019), 1, 155–169. DOI:10.32604/cmc.2019.05901

[24] L. Zhang, Q. B. Hou, Y. Liu, J. W. Bian, X. Xu et al.: Deep negative correlation classification. Machine Learning *113* (2024), 10, 7223–7241. DOI:10.1007/s10994-024-06604-0

Adam Dudáš, Department of Computer Science, Faculty of Natural Sciences, Matej Bel University, Tajovského 40, 974 01 Banská Bystrica. Slovak Republic.
e-mail: adam.dudas@umb.sk

Tomáš Peregrín, Department of Computer Science, Faculty of Natural Sciences, Matej Bel University, Tajovského 40, 974 01 Banská Bystrica. Slovak Republic.
e-mail: tomas.peregrin@student.umb.sk