European Research Studies Journal
Volume XXVIII, Issue 3, 2025

pp. 1566-1586

Bridging the Gap: Performance Evaluation of Classical and Modern Discretization Techniques in Real Estate Data

Submitted 15/08/25, 1st revision 29/08/25, 2nd revision 11/09/25, accepted 30/09/25

Anna Gdakowicz[1], Małgorzata Łatuszyńska[2]

Abstract:

Purpose: *The aim of this study is to evaluate the effectiveness of selected discretization methods for continuous variables using the example of real estate area—one of the key attributes in property analysis. The study addresses the challenge of transforming continuous variables into discrete forms with minimal information loss, a process crucial for data mining, statistical modeling, and classification tasks.*
Design/Methodology/Approach: *Thirteen discretization methods were applied to a dataset of 3,732 residential real estate listings from the Szczecin housing market between 2017 and 2021. The methods include classical approaches with predefined class parameters (equal width and equal frequency), expert-driven methods, quantile-based techniques, clustering (k-means), and supervised learning approaches such as entropy minimization and 1R. The evaluation criteria included the deviation of grouped results from ungrouped data (arithmetic mean difference and loss function), and the number of classes, treated as a nominant. A linear ordering technique (Hellwig's method) was used to rank the methods.*
Findings: *The method based on expert-defined class width (Method 4) showed the highest consistency with the original data, followed by Scott's rule (Method 2) and the entropy-based supervised method (Method 11). Contrary to expectations, quantile-based methods and commonly used rules such as Freedman–Diaconis or square-root yielded unsatisfactory results, either due to oversimplification (too few intervals) or excessive granularity (too many classes).*
Practical Implications: *The results underline the importance of selecting discretization methods tailored to the characteristics of the variable and research context. In particular, they demonstrate the value of domain expertise in guiding discretization decisions in real estate analytics, improving data quality for downstream analysis such as classification, segmentation, or regression.*
Originality/Value: *This study is one of the first to systematically compare a broad spectrum of discretization methods in the context of real estate data. It introduces a comprehensive evaluation framework combining statistical accuracy and interpretability. The findings contribute to both methodological development in data preprocessing and practical decision-making in real estate market research.*

Keywords: *Discretization methods, real estate market, property area, data preprocessing,*

[1]*Corresponding author, ORCID 0000-0002-4360-3755, University of Szczecin, Szczecin, Poland, e-mail: anna.gdakowicz@usz.edu.pl;*
[2]*ORCID 0000-0002-7847-0198, University of Szczecin, Szczecin, Poland, e-mail: malgorzata.latuszynska@usz.edu.pl;*

variable transformation,, grouped frequency distribution

JEL Code: *C38, C89, R31.*

Paper type: *Original research article.*

Acknowledgement: *This research was co-financed by the Minister of Science under the "Regional Excellence Initiative.*

1. Introduction

Real estate is described by a set of different attributes. Among them are qualitative attributes expressed on weak measurement scales, such as: nominal scale – location (favorable, unfavorable), neighborhood (favorable, average, unfavorable), land fertility classes (class I, class II, class III, etc.), and type of land plot (agricultural, residential, commercial, industrial) or ordinal scale – property size (small, medium, large), building height (low, average, high), and distance from the center (small, medium, large).

Additionally, there are quantitative attributes expressed on interval or ratio scale. The latter can be further categorized into: discrete characteristics, which take integer values (e.g., number of rooms), and continuous characteristics, which take real but positive values (e.g., area, price, volume, distance).

Continuous real estate attributes pose specific challenges in real estate market modeling. Many research methods and data mining algorithms require data to be presented in a discrete format. For example, in the application of survival analysis methods (Gdakowicz and Putek-Szeląg, 2022), decision tree algorithms (Fayyad and Irani, 1993), ordinal regression (Ruiz, Angulo and Agell, 2006), or probabilistic models such as Bayesian networks (Wójciak and Łupińska-Dubicka, 2018), variables must be discretized.

The development of information systems across various fields, including real estate, has led to the accumulation of large and complex datasets. Since these databases are compiled by multiple individuals and often lack standardized rules for data entry and interpretation, inconsistencies arise. For example, in some cases, the usable area of a

property is recorded, whereas in others, the total property area is entered, leading to measurement errors and significant variability in real estate attributes.

In both of the aforementioned cases, it is necessary to discretize continuous variables, meaning that continuous values must be divided into a predefined (or dynamically determined) number of disjoint intervals, with each interval assigned a symbolic representation. Discretization is a crucial step in data analysis and data mining (Wójciak and Łupińska-Dubicka, 2018). The advantages of discretization include (Dougherty, Kohavi and Sahami, 1995; Liu, Hussain, Tan, and Dash, 2002; Gatnar and Walesiak, 2011):

- Efficiency – reducing the number of variable values increases computational efficiency.
- Simplicity – decision-making processes become more straightforward, result transparency improves, and interpretation becomes easier.
- Versatility – enables the use of methods that operate on both continuous and symbolic variables and allows for the processing of datasets containing outliers or missing values.

However, a major drawback of discretization is the loss of granular information about variables. Consequently, one of the primary objectives of newly developed discretization methods is to minimize this loss (García, Luengo, Sáez, López, and Herrera, 2013).

Discretization is now a fundamental step in the analysis of continuous variables. However, there is no universally accepted, single best method. On the contrary, as computational power increases, the number of available discretization techniques continues to grow.

The aim of this study is to identify discretization methods applicable to real estate area analysis and to determine which methods should be recommended for practical use in data mining. Given the increasing number of newly developed, advanced discretization algorithms, this study hypothesizes that supervised discretization methods are more effective than unsupervised methods.

This article presents 13 discretization methods, ranging from the simplest and most well-established to the latest machine learning-based approaches. The effectiveness of the selected methods was evaluated using predefined criteria.

2. Literature Review

Numerous discretization methods and algorithms have been described in the literature (Kerber, 1992; Dougherty *et al.,* 1995; Liu *et al.,* 2002; Kotsiantis and Kanellopoulos, 2006; García *et al.,* 2013).

Initially, discretization methods were categorized into equal interval width and equal frequency approaches. However, with advancements in computational techniques, more complex methods have been developed. These can be broadly classified into primary and composite methods (Yang, Webb, and Wu, 2010). Primary methods operate independently and do not reference other methods during the discretization process. Composite methods are built upon one or more base methods.

Primary discretization methods can be further divided into the following categories:

- Supervised vs. unsupervised (Dougherty *et al.*, 1995).
- Splitting vs. merging (Kerber, 1992).
- Global vs. local (Dougherty *et al.*, 1995; Chmielewski and Grzymala-Busse, 1996).
- Static vs. dynamic (Dougherty *et al.*, 1995; Liu *et al.*, 2002).
- Univariate vs. multivariate (Bay, 2000).
- Disjoint vs. non-disjoint (Yang and Webb, 2002).

Discretization methods are developed based on statistical principles, logical inference, machine learning, and information theory. To gain a comprehensive understanding of these methods, it is essential to consider the criteria used for their evaluation and comparison.

In *supervised* discretization, the number of target classes is predefined. In contrast, *unsupervised* methods determine the number of classes dynamically, without prior assumptions.

Splitting discretization (top-down) treats the entire range of values as a single interval and successively divides it into smaller intervals until a predefined stopping criterion is met. In contrast, *merging* discretization (bottom-up) initially assigns each unique value to its own interval and progressively merges adjacent intervals until a stopping threshold is reached.

Local methods perform discretization based on a subset of the variable's variation, often depending on the values of other variables. This allows different ranges to be used in different classification contexts. For example, in decision trees, different discretization schemes may be applied at different nodes (Quinlan, 1992). *Global* methods, on the other hand, discretize a continuous variable once, regardless of other attributes.

Static methods determine a fixed number of intervals for each variable independently, while *dynamic* methods adjust the number of intervals dynamically based on all variables in the dataset.

Univariate discretization processes each variable independently, while *multivariate*

methods consider relationships between attributes when forming discretization intervals.

Disjoint methods assign values to non-overlapping intervals, whereas *non-disjoint* methods allow overlapping intervals.

Both independent (predictor) and dependent (response) variables can undergo discretization (Jiang, Cukic, and Menzies, 2008; Menzies, Milton, Turhan, Cukic, Jiang, and Bener, 2010; Menzies, Bird, Zimmermann, Schulte, and Kocaganeli, 2011; Rajbahadur, Wang, Kamei, and Hassan, 2022).

While discretizing independent variables can improve the performance of machine learning classifiers (Yang *et al.*, 2010; García *et al.*, 2013), the dichotomization of dependent variables may introduce noise and bias (Cohen, 1983; Altman and Royston, 2006; Royston, Altman and Sauerbrei, 2006; Dawson and Weiss, 2012; Krishnan Rajbahadur, Wang, Kamei, and Hassan, 2021).

Researchers emphasize that it is generally safer not to discretize dependent variables (Maccallum, Zhang, Preacher, and Rucker, 2002; DeCoster, Iselin, and Gallucci, 2009), although in some cases, discretization has shown benefits (Farrington and Loeber, 2000; Chmura Kraemer, Noda, and O'hara, 2004; DeCoster, Gallucci, and Iselin, 2011).

García *et al.* (2013) classified more than 80 discretization methods. The most commonly used include: equal width, equal frequency, MDLP (Fayyad and Irani, 1993), ID3 (Quinlan, 1992), ChiMerge (Kerber, 1992), 1R (Holte, 1993), D2 (Catlett, 1991), Chi2 (Liu and Setiono, 1997).

There is no universally accepted measure for evaluating the effectiveness of discretization methods. According to Vannucci *et al.* (2004), an effective discretization process should:

- Accurately represent the original distribution of the attribute.
- Preserve meaningful patterns within the data.
- Avoid excessive generalization (where large intervals obscure significant findings) and over-fragmentation (where small intervals reduce statistical significance).
- Produce semantically meaningful intervals that are interpretable by domain experts.

Liu *et al.* (2002) proposed evaluating discretization processes based on accuracy, computational time, learning time, and interpretability. Vannucci *et al.* (2004) assessed discretization error using a weighted mean of standard deviations within classified intervals.

García *et al.* (2013) applied evaluation criteria such as the number of intervals, inconsistency levels, predictive classification rate, and computational efficiency. Wójciak and Łupińska-Dubicka (2018) assessed discretization quality based on the number of intervals, inconsistency rate, and prediction accuracy.

It is important to note that the effectiveness of a discretization method depends on the variable being discretized and the intended application. No single method consistently outperforms others across all evaluation criteria.

3. Methodology of Research

Thirteen discretization methods were analyzed for partitioning an attribute such as real estate area. These methods include both well-established techniques and relatively recent approaches developed in response to increasing computational power and advances in data processing techniques, particularly machine learning.

The discretization procedure, especially in unsupervised methods, consists of multiple steps and requires determining three key parameters: the number of classes k, the width of the classes h, the lower boundary of the first interval x_{01}.

The resulting intervals must be adjacent but non-overlapping. The method used to determine these parameters varies depending on the partitioning procedure employed. By definition, the upper boundary of each interval serves as the lower boundary of the next. Additionally, it is generally assumed that the intervals are left-closed, meaning they include their lower boundary but exclude their upper boundary.

It is important to emphasize that there is no single optimal number of intervals k. The choice depends on several factors, including the number of observations, the variability of the analyzed attribute, and the selected discretization method.

According to Yule and Kendall (1951), the ideal number of classes ranges between 15 and 25, and in any case, should not be fewer than 10. Conversely, Zając (1994) argues that too few intervals can lead to an excessive concentration of statistical material, thereby masking important patterns in the distribution of the studied variable. On the other hand, too many intervals may result in over-fragmentation, making it difficult to analyze the data and derive meaningful conclusions.

Table 1 presents an overview of the 13 discretization methods examined in this study: methods 1–4: unsupervised methods based on equal frequency intervals, methods 5–9: unsupervised methods based on equal width intervals, method 10: an unsupervised method utilizing the *k-means* clustering algorithm, methods 11–13: supervised discretization methods.

Table 1. *Continuous variable discretization methods*

No.	Name	Number of classes k	Width of classes h	Lower limit of the first interval x_{01}
1	Freedman and Diaconis's rule (Freedman and Diaconis, 1981)	$\left\lceil \frac{x_{max} - x_{min}}{h} \right\rceil$	$2(IQ)n^{-1/3}$	$x_{min} - 0.1h$
2	Scott's rule (Scott, 1979)	$\left\lceil \frac{x_{max} - x_{min}}{h} \right\rceil$	$3.49S_x n^{-1/3}$	$x_{min} - 0.1h$
3	Square-root choice	$\lceil \sqrt{n} \rceil$	$\left\lceil \frac{x_{max} - x_{min}}{k} \right\rceil$	$x_{min} - 0.5h$
4	$h = 20$	$\left\lceil \frac{x_{max} - x_{min}}{h} \right\rceil$	20	10
5	Quartiles	4	–	x_{min}
6	Deciles	10	–	x_{min}
7	Freedman and Diaconis's rule equal frequency	$\left\lceil \frac{x_{max} - x_{min}}{2(IQ)n^{-1/3}} \right\rceil$	–	x_{min}
8	Scott's rule equal frequency	$\left\lceil \frac{x_{max} - x_{min}}{3.49S_x n^{-1/3}} \right\rceil$	–	x_{min}
9	Square-root choice equal frequency	$\lceil \sqrt{n} \rceil$	–	x_{min}
10	*k-means* algorithm	4	–	x_{min}

No.	Name	Description
11	Entropy minimization (Catlett, 1991; Fayyad and Irani, 1993)	$E(X,C,S) = \frac{\lvert S_1 \rvert}{n} Ent(S_1) + \frac{\lvert S_2 \rvert}{n} Ent(S_2)$ where: $Ent(S) = -\sum_{j=1}^{m} p_j \log_2(p_j)$ – entropy function; X – continuous variable under discretization; C – split point; S – training set; $\lvert S_1 \rvert$ – size of the set with values not greater than C, $\lvert S_2 \rvert$ – size of the set with values greater than C; $p_j = l_j / n$ – probability of belonging to jth class; l_j size of jth class. Selection of optimal splitting point using Minimal Description Length Principle – MDLP (stop criterion). The splitting procedure will stop when the inequality is satisfied: $Gain(X,C,S) < \frac{\log_2(n-1) + \Delta(X,C,S)}{n}$ where: $Gain(X,C,S) = Ent(S) - E(X,C,S)$; k, k_1, k_2 – number of classes in the set S, S_1, S_2 respectively.
12	C4.5 algorithm (Quinlan, 1992)	$x_i \leq C$ or $x_i > C$ where: C – the split point. From the possible values of C, the one that maximizes the value of the function is chosen: $J(S,X) = \varphi(\mathbf{p}) - \frac{1}{n}\sum_{j=1}^{m} \varphi(\mathbf{p}_j)$ where: $\varphi(\mathbf{p})$ – the differentiation function, $\mathbf{p}_j$ – the vector of probabilities in the set S_j.

13	One Rule Holte (1R) (Holte, 1993)	Creating disjuncts from adjacent values of a variable that belong to a single class

Note: *x_{max} – the highest value of the variable, x_{min}– the lowest value of the variable, IQ – interquartile range, n – number of observations, S_x – standard deviation.*
Source: *Gdakowicz and Putek-Szeląg, (2022).*

Discretization methods 1–3 differ in how the number of classes *k* and class width *h* are determined. These methods are widely applied in data analysis and are implemented in numerous computational tools, such as Excel, where the default method for determining the number of classes follows method 3. A detailed description of method 3 can be found in Gdakowicz *et al.* (2022).

The lower boundary of the first interval is conventionally determined. In this study, the methods listed in Table 1 were applied, resulting in equal-width intervals. However, a significant drawback of these methods is the large number of classes, which is directly influenced by the number of observations and the dispersion of variable values.

Methods 7–9 are conceptually similar to methods 1–3, but they produce equal-frequency intervals using different formulas.

Method 4 (Table 1) is based on a common heuristic approach used by researchers, where the class width is assigned arbitrarily to simplify the discretization process, particularly when analyzing variable dispersion across different periods. Despite its heuristic nature, studies (Aguilera, Fernández, Fernández, Rumí, and Salmerón, 2011) indicate that expert-based discretization remains one of the most frequently applied approaches.

Methods 5 and 6 enable variable discretization into a predefined number of classes. Breakpoints are determined based on positional measures: method 5 uses quartiles, method 6 uses deciles, ensuring that class frequencies remain relatively balanced.

Method 10 applies the *k-means* clustering algorithm, which partitions the dataset into clusters of similar objects.

Methods 11–13 (Table 1) belong to the category of supervised discretization methods, primarily derived from machine learning techniques. These methods identify breakpoints in continuous variables by minimizing the entropy function.

Method 12 is an algorithm for constructing classification trees. It employs the C4.5 algorithm, which selects the attribute that most effectively partitions the dataset at each decision tree node. The criterion for division is the maximization of the function $J(S, X)$. The C4.5 algorithm then recursively refines the partitioning

process.

Method 13 generates disjoint intervals by merging adjacent values that belong to the same class. This approach follows a one-attribute classification rule, which simplifies classification while maintaining interpretability. A detailed classification of the discretization methods used in this study is presented in Table 2.

Table 2. *Classification of the discretization methods used*

No.	Name	Classification of methods						
		primary vs. composite	supervised vs. unsupervised	splitting vs. merging	global vs. local	static vs. dynamic	univariate vs. multivariate	disjoint vs. non-disjoint
1	Freedman and Diaconis's rule	primary	unsupervised	splitting	global	static	univariate	disjoint
2	Scott's rule	primary	unsupervised	splitting	global	static	univariate	disjoint
3	Square-root choice	primary	unsupervised	splitting	global	static	univariate	disjoint
4	$h = 20$	primary	unsupervised	splitting	global	static	univariate	disjoint
5	Quartiles	primary	unsupervised	splitting	global	static	univariate	disjoint
6	Deciles	primary	unsupervised	splitting	global	static	univariate	disjoint
7	Freedman and Diaconis's rule equal frequency	primary	unsupervised	splitting	global	static	univariate	disjoint
8	Scott's rule equal frequency	primary	unsupervised	splitting	global	static	univariate	disjoint
9	Square-root choice equal frequency	primary	unsupervised	splitting	global	static	univariate	disjoint
10	*k*-means algorithm	primary	unsupervised	merging	global	static	multivar.	disjoint
11	Entropy minimization	primary	supervised	merging	global	dynamic	univariate	disjoint
12	C4.5 algorithm	primary	supervised	merging	local	dynamic	univariate	*
13	One Rule Holte (1R)	primary	supervised	merging	global	static	univariate	*

Note: ** discretization method can take both classification options.*
Source: *Own study.*

The application of the 13 proposed discretization methods results in 13 grouped frequency distributions of the studied variable. In the next stage of the study, an evaluation of the applied discretization methods was conducted. A discretization method was considered effective if the statistical measures computed for the grouped frequency distribution closely matched those calculated for the ungrouped data of the same variable. The following evaluation criteria were adopted:

- Number of classes k – based on Wójciak and Łupińska-Dubicka (2018).
- Arithmetic mean deviation (proposed by the author) – the weighted arithmetic mean computed for the grouped frequency distribution (determined using the proposed discretization methods) should be as close as possible to the arithmetic mean of the ungrouped data:

$$A = |\bar{x}_M - \bar{x}| \rightarrow min \tag{1}$$

  where: $\bar{x}_M$ – arithmetic mean for the grouped frequency distribution obtained using one of the 13 discretization methods M = 1, 2, …, 13; $\bar{x}$ – arithmetic mean for the ungrouped data.
- Loss function LF (proposed by the author) – the loss function value, calculated for the grouped frequency distribution, should be as close as possible to the corresponding value for the ungrouped frequency distribution:
-

$$LF = \left| \sum_{i=1}^{k} (\dot{x}_i - \bar{x}_M)^2 \cdot n_i - \sum_{i=1}^{n} (x_i - \bar{x})^2 \right| \rightarrow min \tag{2}$$

  where: $\dot{x}_i$ – midpoint of the class in the grouped frequency distribution, k – number of classes, n_i – frequency of the interval in the grouped frequency distribution, x_i – value of the variable, n – number of observations, other symbols are the same as in the Equation (1).

The applied discretization methods were ranked using Hellwig's (1968) method according to three evaluation criteria: number of classes k, arithmetic mean deviation A, and loss function LF. According to Bąk (2018), the linear ordering procedure involves the following steps:

- Determine the nature of the variables (stimulant, destimulant, or nominant).
- Assign weights to the variables.
- Normalize the variables.
- Determine the coordinates of the pattern or anti-pattern (depending on the study type).
- Perform pattern or anti-pattern aggregation.

A stimulant is a variable where higher values positively affect classification. A destimulant is a variable where lower values positively affect classification. A nominant (Borys, 1984) is a variable where a specific value or range is desirable (e.g., unimodal or bimodal nominants), while extreme values are undesirable. Since all variables must positively contribute to classification, destimulants were transformed into stimulants using the formula (Gatnar and Walesiak, 2004):

$$x_{ij} = b_j - x_{ij}^D \tag{3}$$

where:
x_{ij} – transformed value of the *j*th variable for the *i*th discretization method,
x_{ij}^D – original value of the destimulant variable,
b_j – maximum value of the *j*th variable.

The number of classes k is treated as a nominant with a preferred range. It was transformed into a stimulant using Kukuła's (2020) normalization formula:

$$x_{ij} = \begin{cases} \frac{1}{c_{1j} - a_j}\left(x_{ij}^N - a_j\right) & for \quad x_{ij}^N < c_{1j} \\ 1 & for \quad c_{1j} \le x_{ij}^N \le c_{2j} \\ \frac{1}{c_{2j} - b_j}\left(x_{ij}^N - b_j\right) & for \quad c_{2j} \le x_{ij}^N \end{cases} \tag{4}$$

where:
x_{ij}^N – nominant value for the *i*th discretization method,
c_{1j}, c_{2j}– lower and upper limits of the preferred range,
a_j, b_j – minimum and maximum values of the variable.

To ensure symmetry, Batóg and Wawrzyniak (2020; 2021; 2022) introduced modified minimum and maximum values:

$$a_j^* = \begin{cases} a_j & for \quad c_{1j} - a_j \ge b_j - c_{2j} \\ c_{1j} - \left(b_j - c_{2j}\right) & for \quad c_{1j} - a_j < b_j - c_{2j} \end{cases} \tag{5a}$$

$$b_j^* = \begin{cases} c_{2j} + \left(c_{1j} - a_j\right)_j & for \quad c_{1j} - a_j \ge b_j - c_{2j} \\ b_j & for \quad c_{1j} - a_j < b_j - c_{2j} \end{cases} \tag{5b}$$

where:
a_j^*, b_j^* – modified minimum and maximum values for the nominant.

All variables were normalized using the Kukuła (1999) formula:

$$x_{ij} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \tag{6}$$

The upper development pattern was established as:

$$z_{0j} = \max_i \{z_{ij}\} \tag{7}$$

The distance of objects from the pattern was computed as:

$$d_{0i} = \sqrt{\sum_{j=1}^{m} (z_{ij} - z_{0j})^2} \tag{8}$$

Finally, the aggregate variable was determined by:

$$q_i = 1 - \frac{d_{0i}}{d_0} \tag{9}$$

where: $d_0 = \bar{d}_0 + 2S_d$; $\bar{d}_0 = \frac{\sum_{i=1}^{n} d_{i0}}{n}$; $S_d = \sqrt{\frac{\sum_{i=1}^{n} (d_{i0} - \bar{d}_0)^2}{n}}$.

The range of q_i values is typically between 0 and 1. The best discretization method corresponds to $\max\{q_i\}$, while the worst corresponds to $\min\{q_i\}$ (Bąk, 2018).

4. Empirical Study

The proposed discretization methods were applied to the variable real estate area. The data pertained to the Szczecin residential real estate market, specifically apartments offered for sale between 2017 and 2021 by real estate agents affiliated with the West Pomeranian Association of Realtors.

Only listings that were either sold during the study period or were still active in the Multiple Listing System (MLS) were included. Listings that were withdrawn from the MLS were excluded from the analysis. The final dataset included 3,732 apartment listings. Table 3 presents the descriptive statistics for the analyzed variable – apartment area.

Table 3. *Descriptive statistics of apartment area for properties offered for sale in Szczecin, 2017–2021*

Statistical measures	Value
Minimum	15.00
1st Quartile	42.40
Median	53.62

Mean	58.09
3rd Quartile	68.15
Maximum	188.00
Standard deviation	23.13
Interquartile range	25.75
Coefficient of variation (%)	39.83
Relative positional coefficient of variation (%)	24.01
Asymmetry	1.326

Source: *Own study.*

The area of residential properties offered for sale in Szczecin during the analyzed period exhibited significant variability (Table 3), with apartment sizes ranging from 15 m² to 188 m². Half of the apartments had an area of no more than 53.62 m². The average apartment area was 58 m², and the distribution of the variable was strongly right-skewed, indicating a predominance of smaller units in the market offer. The real estate area variable was discretized using the methods outlined in Table 1.

The distribution calculations were performed in the R software environment, employing the following packages: grDevices, RWeka (Hornik, Buchta and Zeileis, 2008), and arules (Hahsler, Chelluboina, Hornik and Buchta, 2011). The resulting grouped frequency distributions are illustrated in graphical form as histograms (Figures 1–11).

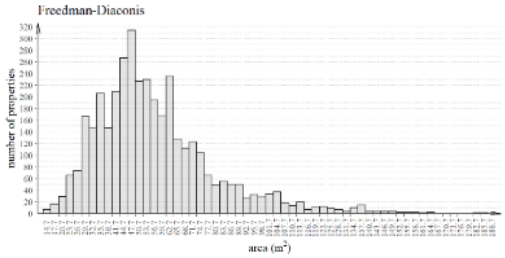

Fig. 1 *Grouped frequency distribution of apartment areas offered for sale in Szczecin from 2017 to 2021, calculated using method 1 – Freedman and Diaconis's rule*
Source: Own study.

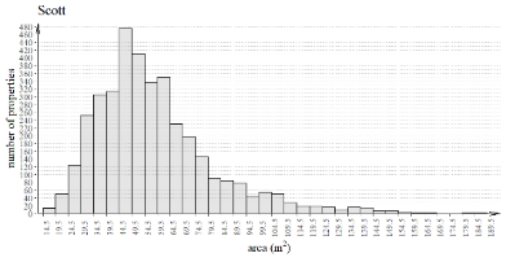

Fig. 2 *Grouped frequency distribution of apartment areas offered for sale in Szczecin from 2017 to 2021, calculated using method 2 – Scott's rule*
Source: Own study.

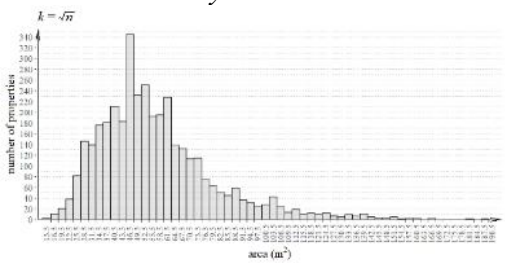

Fig. 3 *Grouped frequency distribution of*

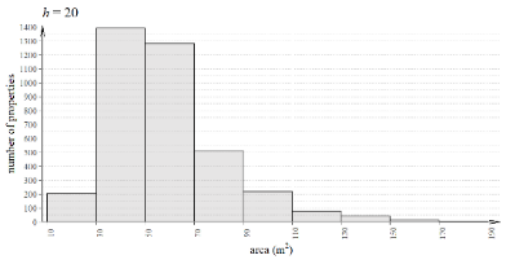

Fig. 4 *Grouped frequency distribution*

apartment areas offered for sale in Szczecin from 2017 to 2021, calculated using method 3 – square-root choice
Source: Own study.

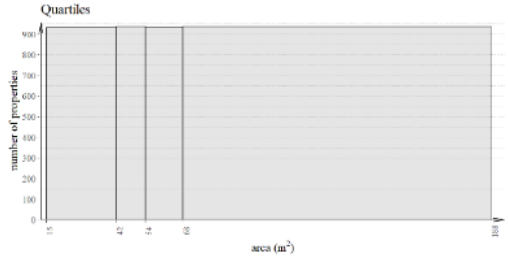

Fig. 5 *Grouped frequency distribution of apartment areas offered for sale in Szczecin from 2017 to 2021, calculated using method 5 – quartiles*
Source: Own study.

of apartment areas offered for sale in Szczecin from 2017 to 2021, calculated using method 4 – h = 20
Source: Own study.

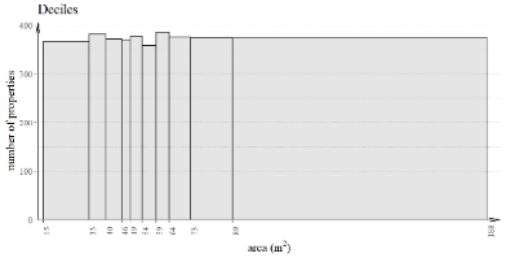

Fig. 6 *Grouped frequency distribution of apartment areas offered for sale in Szczecin from 2017 to 2021, calculated using method 6 – deciles*
Source: Own study.

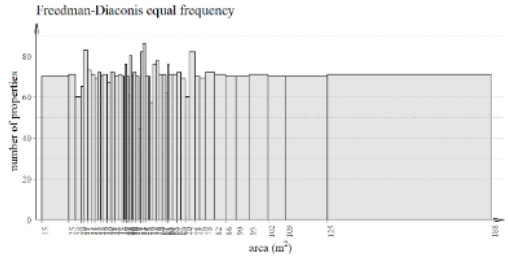

Fig. 7 *Grouped frequency distribution of apartment areas offered for sale in Szczecin from 2017 to 2021, calculated using method 7 – Freedman and Diaconis's rule equal frequency*
Source: Own study.

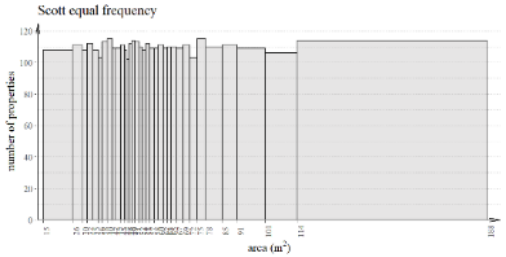

Fig. 8 *Grouped frequency distribution of apartment areas offered for sale in Szczecin from 2017 to 2021, calculated using method 8 – Scott's rule equal frequency*
Source: Own study.

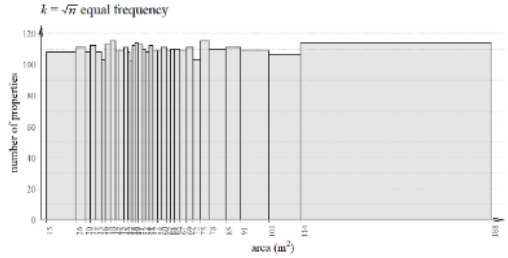

Fig. 9 *Grouped frequency distribution of apartment areas offered for sale in Szczecin from 2017 to 2021, calculated using method 9 – square-root choice equal frequency*
Source: Own study.

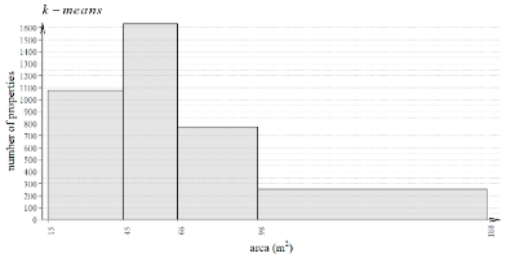

Fig. 10 *Grouped frequency distribution of apartment areas offered for sale in Szczecin from 2017 to 2021, calculated using method 10 – k-means algorithm*
Source: Own study.

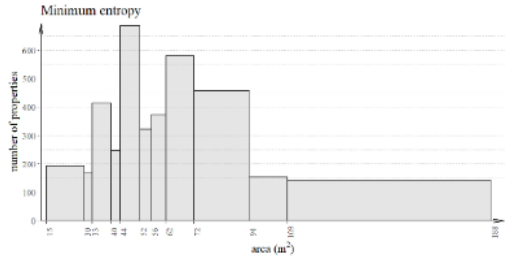

Fig. 11 *Grouped frequency distribution of apartment areas offered for sale in Szczecin from 2017 to 2021, calculated using method 11 – entropy minimization Source: Own study.*

Methods 1–3 (Fig. 1–3) and 7–9 (Fig. 7–9) are characterized by generating a large number of classes for datasets with a high number of observations and substantial variability in the analyzed variable. The resulting number of classes ranged from 28 (Method 2) to as many as 61 (Method 9). Among these, methods based on equal frequencies (Methods 7–9) produced more classes than their counterparts based on equal widths (Methods 1–3).

A notable drawback of Methods 1–3 is the occurrence of empty intervals at the higher end of the area range, specifically between 167 m² and 182 m². This gap suggests a lack of properties within these sizes—an outcome expected given their rarity in the apartment segment—although such units may sporadically appear on the market. In such cases, it is advisable to consider truncating the final interval, for example, at 160 m² (Gdakowicz and Putek-Szeląg, 2022).

Method 4 (Fig. 4) entailed dividing the variable into a predefined number of classes, with the class widths determined based on expert judgment. This approach resulted in seven subgroups. Methods 5 and 6 (Fig. 5–6), which are based on quantiles—recommended for variables exhibiting high dispersion—yielded four and ten groups, respectively, consistent with initial assumptions. Method 10 also utilized a fixed number of classes (four); however, the interval boundaries differed from those generated by other techniques.

For Methods 11–13, it was necessary to designate a secondary variable to guide the optimization of the data partitioning. In this case, the number of rooms served as the reference variable. Using entropy minimization (Method 11, Fig. 11), the dataset was divided into eleven disjoint classes based on property area. Methods 12 and 13 partitioned the property area into seven classes; however, in the case of Method 1R, the final two classes contained zero frequency.

Moreover, for both methods, the resulting class intervals overlapped. For instance, in Method 12, the first class encompassed properties from 15 to 32.89 m², the second

from 29.75 to 55.9 m², the third from 55.94 to 136.5 m², and the fourth from 93.96 to 188 m², among others. A similar issue was observed in Method 13. The lack of disjoint class intervals precludes the construction of graphical representations of the grouped frequency distributions and, consequently, disqualifies these methods from further comparative analysis where non-overlapping classes are required.

The subsequent phase of the research involved evaluating the effectiveness of discretization using Methods 1–11 (Methods 12 and 13 were excluded due to overlapping class boundaries). For this purpose, the linear ordering procedure described by formulas (7) through (9) was applied.

For each grouped frequency distribution, a weighted arithmetic mean was computed, followed by the determination of the evaluation criteria specified by formulas (1) and (2). Both criteria were interpreted as *destimulants*, as lower values (i.e., lower deviation from the arithmetic mean of the ungrouped distribution and a smaller loss function) are preferred. They were subsequently transformed into *stimulants* using formula (3).

The third criterion, the number of classes, is a *nominant*—a variable for which neither excessively few nor excessively many classes are desirable. An optimal number of classes was assumed to lie between 6 and 25, as this range supports the observation of data regularities while avoiding both overgeneralization and excessive detail.

As a nominant, this criterion was also transformed into a stimulant using formula (4), adjusted by formulas (5a) and (5b). After transformation, all variables were normalized according to formula (6). The final evaluation, based on the normalized aggregate variable, is presented in Table 4.

Table 4. *Linear ordering of the applied discretization methods*

No.	Name	Aggregate variable value q_i
4	$h = 20$	0.98
2	Scott's rule	0.90
11	Entropy minimization	0.87
8	Scott's rule equal frequency	0.80
10	*k*-means algorithm	0.73
6	Deciles	0.63
1	Freedman and Diaconis's rule	0.53
3	Square-root choice	0.53
7	Freedman and Diaconis's rule equal frequency	0.45
9	Square-root choice equal frequency	0.29
5	Quartiles	0.00

Source: *Own study.*

The evaluation results were most consistent with those estimated from the raw

(ungrouped) data for the grouped frequency distribution derived using the fourth discretization method, with a class width of $h = 20$. This approach, in which both the class width and the lower bound of the first interval were determined a priori based on expert knowledge, yielded the most accurate representation.

The second-best method in terms of consistency between ungrouped and grouped frequency distributions was Scott's rule (Method 2) — a widely used technique that, as confirmed by this study, provides reliable results. Notably, its variant based on equal frequencies also produced satisfactory outcomes. The third-ranked approach was the entropy minimization method (Method 11), representing one of the supervised discretization techniques.

At the opposite end of the ranking were the quartile-based method (Method 5) and commonly applied techniques relying on formulas proposed by Freedman and Diaconis (Methods 1 and 7), as well as the square-root methods (Methods 3 and 9), regardless of whether equal-width or equal-frequency intervals were used.

The low position of the quartile-based method was primarily due to the small number of generated classes, which limited the ability to observe statistical regularities. Conversely, other low-ranking methods created an excessive number of classes, leading to overly detailed frequency distributions that impeded analytical clarity.

5. Discussion and Conclusions

Transforming continuous variables into discrete forms (i.e., arranging raw data into grouped frequency distributions) inevitably involves some degree of information loss. Nevertheless, discretization remains a crucial step in data analysis, offering practical benefits such as improved efficiency, simplicity, and generalizability.

The key challenge lies in selecting a method that minimizes this information loss. Although no single, universally optimal discretization method exists, over 80 approaches have been documented in the literature, with the choice depending largely on the context and judgment of the researcher.

This study examined the discretization of a continuous variable — real estate area — using 13 selected methods. These included both widely known techniques (Methods 1–3 and 7–9) involving pre-defined parameters such as the number of classes, class width, and lower interval bound (in equal-width and equal-frequency variants); a method based on expert knowledge (Method 4); methods with a fixed number of classes (Methods 5, 6, and 10); and supervised learning methods (Methods 11–13).

For each resulting grouped frequency distribution, measures of discretization quality were computed, evaluating how closely results from grouped data approximated those derived from the raw (ungrouped) data. Additionally, the number of intervals

was assessed to ensure it was neither too small (leading to oversimplification) nor too large (causing excessive granularity). Based on these criteria, a ranking of the discretization methods was developed.

The empirical analysis focused on the area of residential properties — a fundamental attribute in real estate. Data were drawn from property listings in Szczecin from 2017 to 2021. The variable exhibited high dispersion and variability, characteristics typical of real estate attributes due to the unique nature of each property. This high variance posed additional challenges for discretization, as some methods produced an excessive number of intervals.

The resulting classification yielded insightful outcomes. The highest-ranked approach was Method 4 — a heuristic technique based on expert-determined class width — underscoring the value of domain expertise in real estate research. The next best methods were Scott's rule (Method 2), a classical approach modified by a refined class width formula, and the entropy minimization method (Method 11), a supervised technique that also performed well in the study by García et al. (2013). Notably, Methods 4 and 2 were not included in that earlier work.

Given the high variance of the variable, methods using positional measures were expected to perform well. However, contrary to expectations, they yielded the poorest results. This suggests that while statistical theory offers useful guidance, the nuanced understanding of the variable — often held by domain experts — plays a decisive role in effective discretization. The ability to establish class intervals that balance representativeness and statistical accuracy highlights the importance of the researcher's judgment.

Future research will proceed in two directions. First, the evaluation of discretization methods will be extended to other continuous real estate attributes. Second, the study will explore the temporal stability of discretization performance, aiming to identify methods that maintain consistent classification quality across both individual years and the entire analysis period.

References:

Aguilera, P.A., Fernández, A., Fernández, R., Rumí, R., Salmerón, A. 2011. Bayesian networks in environmental modelling. Environmental Modelling, Software. 26(12), 1376-1388. DOI: 10.1016/J.ENVSOFT.2011.06.004.

Altman, D.G., Royston, P. 2006. Statistics Notes: The cost of dichotomising continuous variables. BMJ : British Medical Journal, 332(7549), 1080. DOI: 10.1136/BMJ.332.7549.1080.

Bąk, A. 2018. Analiza porównawcza wybranych metod porządkowania liniowego (Comparative Analysis of Selected Linear Ordering Methods Based on Empirical and Simulation Data). Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu. (nr 508 Taksonomia 21. Klasyfikacja i analiza danych-teoria i zastosowania), 19-28. DOI: 10.15611/PN.2018.508.02.

Batóg, B., Wawrzyniak, K. 2020. Comparison of proposals of transformation of nominants into stimulants on the example of financial ratios of companies listed on the warsaw stock exchange. Studies in Classification, Data Analysis, and Knowledge Organization, 3–17. DOI:10.1007/978-3-030-52348-0_1/COVER.

Batóg, B., Wawrzyniak, K. 2021. Propositions of Transformations of Asymmetrical Nominants into Stimulants on the Example of Chosen Financial Ratios. Studies in Classification, Data Analysis, and Knowledge Organization, 83-100. DOI: 10.1007/978-3-030-75190-6_6/COVER.

Batóg, B., Wawrzyniak, K. 2022. Comparison of Influence of Various Proposals of Transforming Nominants into Stimulants on Linear Ordering and Grouping of Listed Companies, 81-92. DOI: 10.1007/978-3-031-10190-8_7/COVER.

Bay, S.D. 2000. Multivariate discretization of continuous variables for set mining. Proceeding of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 315-319. DOI: 10.1145/347090.347159.

Borys, T. 1984. Kategoria jakości w statystycznej analizie porównawczej. Wrocław.

Catlett, J. 1991. On changing continuous attributes into ordered discrete attribute. In: Y. Kodratoff (ed.). Machine Learning — EWSL-91. EWSL 1991. Lecture Notes in Computer Science, vol. 482, Berlin, Heidelberg: Springer Verlag. 164-178. DOI: 10.1007/BFB0017012.

Chmielewski, M.R., Grzymala-Busse, J.W. 1996. Global discretization of continuous attributes as preprocessing for machine learning. International Journal of Approximate Reasoning, 15(4), 319-331. DOI:10.1016/S0888-613X(96)00074-6.

Chmura Kraemer, H., Noda, A., O'hara, R. 2004. Categorical versus dimensional approaches to diagnosis: methodological challenges. Journal of Psychiatric Research, 38, 17-25. DOI: 10.1016/S0022-3956(03)00097-9.

Cohen, J. 1983. The Cost of Dichotomization. Applied Psychological Measurement, 7(3), 249-253. DOI:10.1177/014662168300700301.

Dawson, N.V., Weiss, R. 2012. Dichotomizing continuous variables in statistical analysis: A practice to avoid. Medical Decision Making, 32(2), 225-226. DOI: 10.1177/0272989X12437605.

DeCoster, J., Iselin, A.M.R., Gallucci, M. 2009. A conceptual and empirical examination of justifications for dichotomization. Psychological Methods, 14(4), 349-366. DOI: 10.1037/a0016956.

DeCoster, J., Gallucci, M., Iselin, A.M.R. 2011. Best Practices for Using Median Splits, Artificial Categorization, and their Continuous Alternatives. Journal of Experimental Psychopathology, 2(2), 197-209. DOI:10.5127/jep.008310.

Dougherty, J., Kohavi, R., Sahami, M. 1995. Supervised and Unsupervised Discretization of Continuous Features, in Machine Learning: Proceedings of the Twelfth International Conference, 194-202. Available: http://130.203.136.95/viewdoc/summary?doi=10.1.1.47.6141.

Farrington, D.P., Loeber, R. 2000. Some benefits of dichotomization in psychiatric and criminological research. Criminal Behaviour and Mental Health, 10, 100-122. DOI:10.1002/cbm.349.

Fayyad, U., Irani, K. 1993. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. IJCAI, 93(2), 1022-1029.

Freedman, D., Diaconis, P. 1981. On the histogram as a density estimator: L2 theory. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 57(4), 453-476. DOI:10.1007/BF01025868.

García, S., Luengo, J., Sáez, J.A., López, V., Herrera, F. 2013. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering, 25(4), 734-750. DOI:10.1109/TKDE.2012.35.

Gatnar, E., Walesiak, M. (eds.). 2004. Metody statystycznej analizy wielowymiarowej w badaniach marketingowych. Wrocław: Wydawnictwo Akademii Ekonomicznej im. Oskara Langego we Wrocławiu.

Gatnar, E., Walesiak, M. (eds.). 2011. Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R. Wydawnictwo C.H. Beck.

Gdakowicz, A., Putek-Szeląg, E. 2022. The influence of discretization of the flat's area on its selling time. Procedia Computer Science, 207, 1881-1890. DOI:10.1016/J.PROCS.2022.09.246.

Gdakowicz, A., Hozer-Koćmiel, M., Markowicz, I. 2022. Zastosowanie metod opisu statystycznego do badania zjawisk społeczno-ekonomicznych. Warszawa: CeDeWu sp. z o.o.

Hahsler, M., Chelluboina, S., Hornik, K., Buchta, C. 2011. The arules R-Package Ecosystem: Analyzing Interesting Patterns from Large Transaction Data Sets. Journal of Machine Learning Research, 12(57), 2021-2025. Available: http://jmlr.org/papers/v12/hahsler11a.html.

Hellwig, Z. 1968. Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju oraz zasoby i strukturę wykwalifikowanych kadr. Przegląd Statystyczny, 4, 307-327.

Holte, R.C. 1993. Very simple classification rules perform well on most commonly used datasets. Machine Learning., 11, 63-91. DOI:10.1.1.67.2711.

Hornik, K., Buchta, C., Zeileis, A. 2008. Open-source machine learning: R meets Weka. Computational Statistics, 224(2), 225-232. DOI:10.1007/S00180-008-0119-7.

Jiang, Y., Cukic, B., Menzies, T. 2008. Can data transformation help in the detection of fault-prone modules? DEFECTS'08: 2008 International Symposium on Software Testing and Analysis - Proceedings of the 2008 Workshop on Defects in Large Software Systems 2008, DEFECTS'08, 16-20. DOI:10.1145/1390817.1390822.

Kerber, R. 1992. ChiMerge: Discretization of Numeric Attributes. In: Proceedings of the 10th National Conference on Artificial Intelligence, 123-128.

Kotsiantis, S., Kanellopoulos, D. 2006. Discretization Techniques: A recent survey. GESTS International Transactions on Computer Science and Engineering, 32(1), 47-58.

Krishnan Rajbahadur, G., Wang, S., Kamei, Y., Hassan, A.E. 2021. Impact of Discretization Noise of the Dependent variable on Machine Learning Classifiers in Software Engineering. IEEE Transactions on Software Engineering, 47(7), 1414-1430. \ DOI:10.1109/TSE.2019.2924371.

Kukuła, K. 1999. Metoda unitaryzacji zerowanej na tle wybranych metod normowania cech diagnostycznych. Acta Scientifica Academiae Ostroviensis, 4, 5-31.

Kukuła, K. 2020. Metoda unitaryzacji zerowanej. Warszawa: Wydawnictwo Naukowe PWN. Available: https://www.empik.com/metoda-unitaryzacji-zerowanej-kukula-karol,319703,ksiazka-p.

Liu, H., Setiono, R. 1997. Feature selection via discretization. IEEE Transactions on Knowledge and Data Engineering, 9(4), 642-645. DOI:10.1109/69.617056.

Liu, H., Hussain, F., Tan, C.L., Dash, M. 2002. Discretization: An Enabling Technique. Data Mining and Knowledge Discovery, 6(4), 393-423. DOI:10.1023/A:1016304305535.

Maccallum, R.C., Zhang, S., Preacher, K.J., Rucker, D.D. 2002. On the Practice of Dichotomization of Quantitative Variables. Psychological Methods, 7(1), 19-40. DOI: 10.1037/1082-989X.7.1.19.

Menzies, T., Milton, Z., Turhan, B., Cukic, B., Jiang, Y., Bener, A. 2010. Defect prediction from static code features: Current results, limitations, new approaches. Automated Software Engineering, 17(4), 375-407. DOI:10.1007/S10515-010-0069-5.

Menzies, T., Bird, C., Zimmermann, T., Schulte, W., Kocaganeli, E. 2011. The inductive software engineering manifesto. Proceedings of the International Workshop on Machine Learning Technologies in Software Engineering. Available: https://www.academia.edu/2699592/The_inductive_software_engineering_manifesto_Principles_for_industrial_data_mining.

Quinlan, J.R. 1992. C4.5 Programs for Machine Learning. San Mateo: Morgan Kaufmann.

Rajbahadur, G.K., Wang, S., Kamei, Y., Hassan, A.E. 2022. The impact of feature importance methods on the interpretation of defect classifiers. (February, 4). DOI:10.1109/TSE.2021.3056941.

Royston, P., Altman, D.G., Sauerbrei, W. 2006. Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in Medicine, 25(1), 127-141. DOI:10.1002/SIM.2331.

Ruiz, F.J., Angulo, C., Agell, N. 2006. A Supervised Discretization Method for Quantitative and Qualitative Ordered Variables. Computación y Sistemas, 9(4), 314-325. Available: http://www.redalyc.org/articulo.oa?id=61590403.

Scott, D.W. 1979. On Optimal and Data-Based Histograms. Biometrika, 66(3), 605-610.

Vannucci, M., Colla Percro -Scuola, V., Sant'anna, S. 2004. Meaningful discretization of continuous features for association rules mining by means of a SOM. In: ESANN'2004 proceedings - European Symposium on Artificial Neural Networks, Bruges (Belgium), 489-499.

Wójciak, M., Łupińska-Dubicka, A. 2018. Empirical Comparison of Methods of Data Discretization in Learning Probabilistic Models. Advances in Computer Science Research, (14), 177-192. DOI:10.24427/ACSR-2018-VOL14-0011.

Yang, Y., Webb, G.I. 2002. Non-Disjoint Discretization for Naive-Bayes Classifiers. In: C.A.G. Sammut (ed.). Proceedings of the Nineteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., 666-673. Available: https://dl.acm.org/doi/abs/10.5555/645531.655997.

Yang, Y., Webb, G.I., Wu, X. 2010. Discretization Methods. In: O. Maimon, L. Rokach, (eds.). Data Mining and Knowledge Discovery Handbook, Second Edition ed., New York Dordrecht Heidelberg. London, Springer.

Yule, G.U., Kendall, M.G. 1951. An Introduction to the Theory of Statistics. Journal of Symbolic Logic, 16(1). DOI:10.2307/2268667.

Zając, K. 1994. Zarys metod statystycznych. Warszawa: Państwowe Wydawnictwo Ekonomiczne.