

**JOURNAL OF INFORMATION
SYSTEM AND TECHNOLOGY
MANAGEMENT
(JISTM)**www.gaexcellence.com/jistm**DISCRETIZING BINARY VARIABLES FROM
CONTINUOUS VARIABLES FOR SIMULATION DATA
GENERATION: A CRITICAL LITERATURE REVIEW**Kartini Kasim^{1,2*}, Hashibah Hamid³, Ayu Abdul-Rahman⁴¹School of Quantitative Sciences Universiti Utara Malaysia, Malaysia ms.kartini@gmail.com <https://orcid.org/0009-0001-1712-1736>²Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Kedah, Malaysia kartini2929@uitm.edu.my <https://orcid.org/0009-0001-1712-1736>³Institute of Strategic Industrial Decision Modelling (ISIDM), School of Quantitative Sciences, Universiti Utara Malaysia, Kedah, Malaysia hashibah@gmail.com <https://orcid.org/0000-0002-7908-9495>⁴School of Quantitative Sciences Universiti Utara Malaysia, Malaysia ayurahman@gmail.com <https://orcid.org/0000-0002-9836-9919>

*Corresponding Author

Article Info:**Article history:**

Received date: 03.08.2026

Revised date: 24.08.2026

Accepted date: 08.09.2026

Published date: 28.09.2026

To cite this document:

Kasim, K., Hamid, H., & Abdul-Rahman, A. (2026). Discretizing Binary Variables from Continuous Variables for Simulation Data Generation: A Critical Literature Review. *Journal of Information System and Technology Management*, 11 (44), 546-562.

Abstract:

Generating simulation datasets with both continuous and categorical variables is essential for evaluation of statistical methods involving mixed variables. Continuous variables can be generated directly from specified distribution, however, generating categorical variables requires procedures, as they cannot be generated directly in the same manner as continuous variables. Several reviews have addressed general data discretization or categorical encoding as procedures for generating categorical variables, but there are limited studies synthesised continuous-to-binary discretization methods for simulation data generation. Therefore, this review aims to synthesise the literature on continuous-to-binary discretization methods for simulation data generation, comparing their strengths, limitations, and suitability across simulation contexts. This literature review was searching using Scopus AI with no restriction on publication year, in which the searching completed in August 2026 yielded 18 peer-reviewed journal articles retained as representative of the major methodological approaches identified. Four methodological themes were identified which are threshold-based discretization, latent-variable and probabilistic modelling, dependence-preserving simulation frameworks, and considerations relating to threshold selection, information loss, and model compatibility. Threshold-based methods are simple and efficient but prone to information loss and biased estimation when thresholds are

poorly justified, whereas latent-variable, Gaussian copula, and power-polynomial approaches better preserve dependence structures at the cost of stronger assumptions and greater computational complexity. This review offers a useful methodological reference for selecting discretization methods suited to specific simulation objectives, while identifying limitations and priorities for future research.

DOI: 10.35631/JISTM.1144033

Keyword:

Continuous-To-Binary, Discretization; Mixed-Variable Simulation; Review



© The authors (2026). This is an Open Access article distributed under the terms of the Creative Commons Attribution (CC BY NC) (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact jistm@gaexcellence.com.

Introduction

Simulation studies have become a cornerstone of evaluation methodological performance for statistical research. Through simulation studies, data conditions can be controlled for the purpose of investigating estimator properties, classification accuracy, robustness, statistical power, and model sensitivity across diverse scenarios that may be difficult or impossible to observe empirically. Besides that, simulation frameworks capable to reflect the complexity of real-world observations. To reflect real dataset containing mixed variables, generating mixed-variable simulation data has become an essential component of contemporary statistical methodology, motivating continuous improvements in data generation techniques and simulation frameworks (Amatya & Demirtas, 2016; Demirtas, 2017; Fialkowski & Tiwari, 2019). These developments have placed increasing emphasis not only on generating synthetic data but also on ensuring that simulated variables adequately represent realistic data characteristics.

Despite these advances, generating binary variables within mixed-variable simulation studies remains a methodological challenge since binary variables cannot simply be generated independently as continuous variables. In mixed-variable simulation, binary variables can be generated from the discretization of latent continuous variables with the aim is to retain specified marginal and dependence characteristics. This discretization procedure, however, involves more than simply converting continuous values into two categories. The appropriate discretization strategy therefore depends on the purpose and design of simulation study. If these strategies are not carefully specified, the generated data may differ from the intended statistical structure, which can affect parameter estimates, covariance patterns, and classification results. These effects are particularly important when simulated data are used to evaluate statistical methods, as conclusions about their performance depend on how well the data reflect the conditions being studied (Jiryaie et al., 2016; Grobler & Lee, 2020).

Methods for discretize binary variables from continuous data have developed considerably over the past two decades. Earlier studies often used a simple procedure in which an underlying continuous variable was converted into a binary outcome using a predetermined threshold (Morlini, 2011, 2012). Thresholding continues to be used because it provides a straightforward way to generate binary outcomes. As simulation requirements became more complex, other approaches were introduced to account for relationships among mixed variables. These include Gaussian copula models for representing dependence across different variable types, Bayesian nonparametric approaches based on Dirichlet process mixtures for modelling latent responses, and simulation algorithms for generating correlated mixed data with specified marginal characteristics (DeYoreo & Kottas, 2015; Jiryaie et al., 2016; Fialkowski & Tiwari, 2019). The availability of these different methods allows continuous-to-binary discretization to be applied across a wider range of statistical settings and simulation designs.

Although numerous simulation studies employ binary variables, the procedures used to obtain them are frequently described only briefly, which leaving insufficient methodological detail for replication or adaptation for future studies (Demirtas, 2017; Fialkowski & Tiwari, 2019). Therefore, it makes difficult to compare existing approaches systematically, identify advantages and limitations, and determine the suitability for different simulation contexts. This gap leaves researchers without a unified basis for selecting an appropriate method for a given simulation objective.

Against this background, this review examines the literature on discretizing binary variables from continuous variables for simulation data generation. This review also synthesizes the discretization approaches reported in the literature—and further evaluates their underlying assumptions, methodological characteristics, strengths, limitations, and applicability to mixed-variable simulation studies. By integrating evidence from diverse methodological, this review seeks to provide a useful methodological reference that supports researchers in selecting appropriate discretization strategies.

Methodology

Data collection was conducted using Scopus, an established multidisciplinary abstract and citation database, together with Scopus AI, a subscription-based artificial intelligence tool embedded within Scopus that assists researchers in exploring literature more efficiently (Madsen et al., 2026). Scopus AI combines natural language processing, keyword-based search, and visual thematic (concept) mapping, allowing conceptual trends and interconnections across the relevant literature to be identified alongside a conventional database search.

The first step involved formulating a natural language query addressing the central aim of the review followed by an advanced keyword search of the Scopus database using structured Boolean strings developed with Boolean operators, phrase searching, and synonym expansion. There are two search strings were employed in this study:

Search String 1

("discretization" OR "quantization" OR "binarization" OR "thresholding") AND ("binary variable" OR "discrete variable" OR "categorical variable" OR "0-1 variable") AND ("continuous variable" OR "real number" OR "floating point" OR "numeric variable") AND

("simulation" OR "modeling" OR "data analysis" OR "statistical analysis") AND ("data transformation" OR "data conversion" OR "data processing" OR "data representation")

Search String 2

("continuous variable" OR "continuous data" OR "real-valued variable" OR "numeric variable") AND ("binary variable" OR "binary data" OR "dichotomous variable" OR "categorical variable") AND ("conversion" OR "transformation" OR "discretization" OR "thresholding") AND ("method" OR "technique" OR "approach" OR "procedure")

The search was limited to peer-reviewed journal articles published in English, while conference papers, editorials, book chapters, notes, and other reviews were excluded. However, no restriction was imposed on the publication year. The final search was completed in August 2026, ensuring that the review reflected the most current literature available at the time of the study.

The outputs retrieved from Scopus AI included a narrative summary of the query results, pivotal citations linked to the corresponding article entries in Scopus and (3) thematic linkages between continuous-to-binary discretization. Articles surfaced through Scopus AI, together with those retrieved through the conventional Boolean search, were considered for inclusion where they proposed, evaluated, compared, or applied a method for converting continuous variables into binary variables within a simulation, statistical modelling, or machine learning context, and provided sufficient methodological detail to describe the discretization process. Reference lists of retrieved articles were also examined through backward citation searching to identify additional relevant studies. Following this process, 18 articles were retained as representative as methodological developments in continuous-to-binary discretization for simulation data generation.

The collected data were analysed using a thematic qualitative approach. Prominent themes surfaced through Scopus AI's concept mapping and narrative summaries were categorised, examined for recurrence and consistency across publications, and cross-checked against the full text of each article. Four major themes were identified, and each theme was then examined with respect to its theoretical assumptions, computational complexity, preservation of statistical properties, advantages, limitations, and application contexts.

Consistent with the intended use of Scopus AI as an exploratory and synthesis-supporting tool, its limitations should be noted including it assesses only article abstracts, does not support filtering by date range or journal, and does not retain search history across sessions (Madsen et al., 2026). To mitigate these constraints, the natural language and keyword search outputs were supplemented with backward citation searching and full-text reading of each retained article before its methodological characteristics were extracted and synthesised.

The review examines the main approaches used to discretize continuous variables into binary variables, with attention to how the work and the settings in which they are applied. It also considers the theoretical basis of these methods, together with their strengths and limitations.

Results

Study Selection and Overview of the Included Literature

The final review included 18 articles addressing the discretization of continuous variables into binary outcomes for simulation data generation. The studies came from different research areas which shared common methodological interest in developing or applying procedures for continuous-to-binary discretization. Taken together, these articles provide a useful basis for examining the main developments in this area with their strengths and limitations.

The included review showed the considerable methodological diversity in the objectives, applications, and statistical methods used across the studies. Earlier work included threshold-based discretization and latent-variable modelling (Morlini, 2011, 2012), while subsequent studies considered a broader range of methods. These included Gaussian copula models (Jiryaie et al., 2016), Bayesian nonparametric approaches (DeYoreo & Kottas, 2015), power-polynomial simulation methods (Demirtas et al., 2012; Demirtas, 2017), algorithms and software for mixed-data simulation (Amatya & Demirtas, 2016; Fialkowski & Tiwari, 2019), and model-compatible imputation procedures (Grobler & Lee, 2020). Other studies examined issues such as threshold optimization, discretization in machine learning, and approximation error (Nelson et al., 2017; Nojavan et al., 2017; Ma et al., 2022; Alazaidah, 2023; Zhao et al., 2025). The variety of methods and issues considered across these studies reflects the different statistical requirements involved in generating binary variables from continuous data.

Based on recurring methodological patterns, the findings were organized into four themes: (1) threshold-based discretization methods, (2) latent-variable and probabilistic modelling approaches, (3) dependence-preserving simulation frameworks, and (4) methodological considerations and implementation issues. This grouping provides a basis for comparing how the methods generate binary variables, the statistical principles on which they rely, and the practical issues involved in their use.

Theme 1: Threshold-Based Discretization Methods

From the selected articles, the threshold-based discretization was the most frequently approach for generating binary variables from continuous data. In this approach, observations are divided into two categories according to whether their values fall above or below a specified threshold. The choice of threshold varied across studies and included fixed cut-off values, median or quantile splits, clinically meaningful values, and statistically optimized cut-off points selected to improve classification performance (Hadjicostas, 2006; Nelson et al., 2017; Liquet & Riou, 2019). Thresholding was generally favoured for its simplicity and ease of implementation in different simulation settings (MacCallum et al., 2002; Hadjicostas, 2006; Nelson et al., 2017; Ma et al., 2022). Despite these practical advantages, the choice of threshold can affect information retention, statistical power, prevalence rates, and classification performance (Rucker et al., 2015; Ma et al., 2022). Therefore, careful threshold selection is important part of the simulation design.

Theme 2: Latent-Variable and Probabilistic Modelling Approaches

Another theme identified in the literature involves the methods that generates binary variables from underlying continuous variables. The continuous variables are first generated and then

divided into binary outcomes using a specified threshold. The idea has been used in latent Gaussian mixture models (Morlini, 2011, 2012), Bayesian nonparametric binary regression models (DeYoreo & Kottas, 2015), and simulation procedures for mixed-data generation (Fialkowski & Tiwari, 2019). In these methods, the observed binary outcome remains connected to an underlying continuous variable rather than being generated separately. This connection provides a statistical basis for representing the relationship between continuous and binary variables within the simulated data. The use of this idea across different modelling approaches shows that latent continuous variables have an important role in continuous-to-binary discretization.

Theme 3: Latent-Variable and Probabilistic Modelling Approaches

Preserving dependence structures between continuous and binary variables constituted another prominent methodological theme. In generating correlated mixed variables datasets, Gaussian copula has been addressed as models that separated marginal distributions from dependence modellings (Jiryaie et al., 2016). Other studies have used power-polynomial transformations and specialized simulation algorithms to generate continuous and binary variables with specified marginal distributions and correlation structures (Demirtas et al., 2012; Amatya & Demirtas, 2016; Demirtas, 2017; Fialkowski & Tiwari, 2019). These methods extend beyond simple dichotomization by allowing greater control over the statistical characteristics of the generated data. Particularly, the methods retain specified relationships among variables which makes the preservation of dependence structure. Maintaining this dependence structure is important characteristics when generating mixed-variable data for simulation studies.

Theme 4: Methodological Considerations and Practical Implementation

The final theme focused on broader methodological issues related to discretization. Several concerns appeared across the reviewed studies, including information loss (Rucker et al., 2015; Nojavan et al., 2017), approximation error (Zhao et al., 2025), compatibility with multiple-imputation procedures (Grobler & Lee, 2020), the use of domain knowledge when defining meaningful thresholds (Ahlemeyer-Stubbe & Müller, 2021), and changes in predictive modelling performance following discretization (Alazaidah, 2023). Information loss was a particular concern when continuous variables were converted into binary outcomes without careful consideration of the threshold used. This suggests that discretization should not be treated simply as a step for changing the form of a variable. The decisions made during this process can influence the resulting data structure, model performance, and ultimately how the simulation results are interpreted.

Across the four themes, the literature demonstrates progression from simple threshold-based discretization toward approaches that integrate latent-variable theory, dependence modelling, and computational simulation techniques. Threshold methods provide practical solutions for binary generation, whereas latent-variable and copula-based frameworks enhance the preservation of statistical relationships among mixed variables. Simulation algorithms further operationalise these principles by enabling realistic mixed-data generation, while methodological studies emphasise the importance of threshold selection, information preservation, and model compatibility. These findings address the aims of the review by identifying the methods used for continuous-to-binary transformation, summarising their theoretical foundations, synthesising their methodological strengths and limitations for further methodological development in mixed-variable simulation research.

Table 1 summarizes the characteristics of the 18 selected studies included in this review. The studies are presented according to their bibliographic information, research context, methodological, dimension, discretization approach, principal contribution, and relevance to the present review. The table serves as the foundation for the review by consolidating a diverse body of methodological literature into a consistent framework for comparison.

Table 1: Characteristics of Selected Studies

Authors (Year)	Title	Source	Sector / Context	Study Focus	Key Dimension	Method / Approach Used to Discretize	Main Contribution
Fleishman (1978)	A Method for Simulating Non- Normal Distributions	Psychometrika	Statistical simulation	Simulation of continuous variables	Data generation	Polynomial transformation	Foundation for simulation of non-normal data
Emrich & Piedmonte (1991)	A Method for Generating High- Dimensional Multivariate Binary Variates	The American Statistician	Statistical simulation	Correlated binary generation	Binary simulation	Multivariate binary generation algorithm	Generated correlated binary variables
Higham (2002)	Computing the Nearest Correlation Matrix – A Problem from Finance	IMA Journal of Numerical Analysis	Numerical analysis	Correlation estimation	Correlation preservation	Nearest correlation matrix algorithm	Valid correlation matrices
MacCallum et al. (2002)	On the Practice of Dichotomization of Quantitative Variables	Psychological Methods	Research methodology	Effects of dichotomization	Information loss	Critical evaluation	Explained limitations of dichotomization
Wang & Liu (2005)	Method of the Discretization of Continuous Probability Distribution for Risk Analysis	Journal of Xi'an Shiyu University	Risk analysis	Distribution discretization	Probability representation	Continuous distribution discretization	Applied discretization in risk analysis
Agresti (2006)	An Introduction to Categorical Data Analysis (2nd ed.)	Wiley	Statistical methodology	Categorical data analysis	Binary modelling	Statistical framework	Reference text
Royston et al. (2006)	Dichotomizing Continuous Predictors in Multiple	Statistics in Medicine	Medical statistics	Predictor dichotomization	Methodological limitations	Evaluation of cut-points	Showed drawbacks

Authors (Year)	Title	Source	Sector / Context	Study Focus	Key Dimension	Method / Approach Used to Discretize	Main Contribution
Hadjicostas (2006)	Regression: A Bad Idea Maximizing Proportions of Correct Classifications in Binary Logistic Regression	Journal of Applied Statistics	Classification	Threshold selection	Threshold optimisation	Classification threshold	Improved classification thresholds
Lustgarten et al. (2008)	Improving Classification Performance with Discretization on Biomedical Datasets	AMIA Symposium Proceedings	Biomedical informatics	Classification	Preprocessing	Discretization techniques	Evaluated performance
Demirtas & Hedeker (2011)	A Practical Way for Computing Approximate Lower and Upper Correlation Bounds	The American Statistician	Statistical simulation	Correlation constraints	Correlation modelling	Correlation- bound estimation	Feasible correlation bounds
Morlini (2011)	Mixed Mode Data Clustering: An Approach Based on Tetrachoric Correlations	Studies in Classification, Data Analysis and Knowledge Organization	Clustering	Mixed-variable clustering	Latent-variable modelling	Threshold- based latent variables	Applied latent- variable concepts
Demirtas et al. (2012)	Simulation of Massive Public Health Data by Power Polynomials	Statistics in Medicine	Public health simulation	Mixed-data generation	Simulation	Power- polynomial transformation	Generated correlated mixed datasets
Morlini (2012)	A Latent Variables Approach for Clustering Mixed Binary and Continuous Variables	Advances in Data Analysis and Classification	Model-based clustering	Latent modelling	Continuous-to- binary transformation	Gaussian latent-variable model	Formalized latent-variable discretization

Authors (Year)	Title	Source	Sector / Context	Study Focus	Key Dimension	Method / Approach Used to Discretize	Main Contribution
DeYoreo & Kottas (2015)	Within a Gaussian Mixture Model A Fully Nonparametric Modeling Approach to Binary Regression	Bayesian Analysis	Bayesian statistics	Binary regression	Bayesian latent modelling	Dirichlet process mixture	Flexible latent- response framework
Amatya & Demirtas (2016)	Concurrent Generation of Multivariate Mixed Data with Variables of Dissimilar Types	Journal of Statistical Computation and Simulation	Simulation	Mixed-variable generation	Correlated simulation	Simulation algorithm	Generated mixed variables
Jiryaie et al. (2016)	Gaussian Copula Distributions for Mixed Data, with Application in Discrimination	Journal of Statistical Computation and Simulation	Classification	Mixed-data modelling	Dependence preservation	Gaussian copula	Joint distributions
Demirtas (2017)	Concurrent Generation of Binary and Nonnormal Continuous Data Through Fifth-Order Power Polynomials	Communications in Statistics: Simulation and Computation	Simulation	Binary and continuous generation	Distribution preservation	Fifth-order power polynomial	Extended simulation methods
Grobler & Lee (2020)	Multiple Imputation in the Presence of an Incomplete Binary Variable Created from an Underlying Continuous Variable	Biometrical Journal	Missing-data methodology	Binary imputation	Model compatibility	SMC-FCS	Compared imputation strategies

Table 2 presents a comparative synthesis of the principal methods identified for transforming continuous variables into binary variables, The methods are compared against their advantages, limitations and suitability for different simulation contexts.

Table 2: Comparison of Continuous to Binary Discretization Methods

Method	Description	Advantages	Limitations	Typical Applications	Representative Studies
Thresholding (Dichotomization)	Assigns a value of 1 when the continuous variable exceeds a predefined threshold and 0 otherwise. Thresholds may be fixed, percentile-based, or statistically optimized.	Simple, intuitive, computationally efficient	Information loss; reduced statistical power; threshold-dependent	Logistic regression, simulation, epidemiology	MacCallum et al. (2002); Hadjicostas (2006); Nelson et al. (2017); Ma et al. (2022)
Latent Variable Approach	Generates binary variables from underlying unobserved continuous variables using threshold functions.	Preserves theoretical relationships and dependence structures	More computationally demanding	Gaussian mixture models, clustering, simulation	Morlini (2011, 2012); DeYoreo & Kottas (2015)
Gaussian Copula Models	Constructs joint distributions for continuous and binary variables while preserving dependence.	Maintains realistic multivariate dependence	Requires correlation estimation and model specification	Mixed-data simulation, discriminant analysis	Jiryaie et al. (2016)

Method	Description	Advantages	Limitations	Typical Applications	Representative Studies
Power Polynomial Simulation	Simulates continuous variables using polynomial transformations before generating binary variables.	Preserves marginal distributions and target correlations	Higher computational complexity	Monte Carlo simulation, public health	Demirtas et al. (2012); Demirtas (2017); Fialkowski & Tiwari (2019)
Simulation Algorithms (SimCorrMix)	Implements algorithms for generating correlated mixed datasets containing continuous, binary, ordinal, and count variables.	Flexible and reproducible implementation	Depends on accurate parameter specification	Statistical software, simulation studies	Amatya & Demirtas (2016); Fialkowski & Tiwari (2019)
Imputation-Based Methods	Performs multiple imputation while maintaining consistency between underlying continuous and derived binary variables.	Reduces bias due to missing data	Applicable mainly to missing-data settings	Missing-data analysis	Grobler & Lee (2020)
Threshold Optimization	Determines optimal cut-off values by maximizing classification performance or balancing sensitivity and specificity.	Improves classification accuracy	Thresholds may be data-specific	Binary classification, predictive modelling	Hadjicostas (2006); Lique & Riou (2019)

Method	Description	Advantages	Limitations	Typical Applications	Representative Studies
General Quantization / Discretization	Divides continuous values into discrete intervals that may subsequently be collapsed into binary categories.	Reduces computational complexity	Approximation error and information loss	Machine learning, risk analysis	Wang & Liu (2005); Alazaidah (2023); Zhao et al. (2025)

Discussion

The review shows that a suitable discretization method depends on the purpose of the simulation, the statistical assumptions being made, and the distributional or dependence characteristics that need to be preserved. Threshold-based discretization remains the most commonly used approach for converting continuous variables into binary outcomes. The commonly reason for the preference of the method because of its simplicity and relatively easy to implement (MacCallum et al., 2002; Hadjicostas, 2006; Nelson et al., 2017; Ma et al., 2022). However, thresholding is not applied in a single form and the way the threshold is defined varies across studies. Fixed cut-off values, quantile-based thresholds, and statistically optimized thresholds have all been used depending on the study design and intended purpose (Hadjicostas, 2006; Nelson et al., 2017). Other studies have moved beyond conventional dichotomization by using latent-variable formulations, copula-based models, Bayesian approaches, simulation algorithms, or procedures designed to work within imputation frameworks. This varied method reflects the different requirements of simulation studies, particularly when more complex data structures and relationships among variables need to be considered.

Most of the methods reviewed is the generation of binary variables from the discretization of latent continuous variables. The binary value is obtained by applying a threshold to these continuous variables (Morlini, 2011, 2012; DeYoreo & Kottas, 2015). Such a formulation offers a practical statistical basis for generating binary outcomes while maintaining their relationship with continuous variables. Gaussian copula models approach, for example, extend the idea by separating the modelling of marginal distributions from the dependence structure (Jiryaie et al., 2016). This is useful when mixed continuous and binary variables need to be generated according to a specified pattern of dependence. The use of these frameworks illustrate how continuous-to-binary discretization has moved beyond simple cut-off rules toward approaches that explicitly account for the statistical structure of the generated data.

The reviewed methods also differ in their strengths and limitations, suggesting that the choice of discretization method should be guided by the requirements of the simulation study. Threshold-based methods are relatively simple and computationally efficient, but their performance can depend on how the threshold is specified. An inappropriate cut-off may result in information loss, reduced statistical power, or biased parameter estimates (MacCallum et al., 2002; Rucker et al., 2015). More flexible approaches, such as latent-variable models, Gaussian

copula frameworks, and power-polynomial simulation methods, offer greater control over the marginal characteristics and dependence structure of the generated data. This flexibility, however, comes with additional modelling assumptions and computational demands (Demirtas, 2017; Fialkowski & Tiwari, 2019). These differences suggest that there is no single approach that is suitable for every simulation setting; rather, the method should be selected according to the statistical properties that need to be preserved.

Overall, this review brings together the main methodological developments in continuous-to-binary discretization for simulation data generation. The findings, however, need to be considered alongside several limitations. The search was limited to peer-reviewed articles identified through the selected database strategy and written in English, so relevant work published elsewhere or in other languages may not have been captured. Direct comparison between methods was also difficult because the reviewed studies differed considerably in their objectives, simulation settings, and evaluation criteria. In addition, the scope was deliberately restricted to the discretization of continuous variables into binary outcomes within mixed-variable simulation. Methods developed for variables with more than two categories and highly application-specific forms of discretization were therefore outside the scope of this review. There are also limitations associated with the use of Scopus AI. The platform searches article abstracts rather than full texts, provides limited options for filtering by publication date or journal, and does not retain search histories across sessions (Madsen et al., 2026). For this reason, Scopus AI was used to support the initial search and screening process, while decisions about study inclusion and the interpretation of methodological themes were based on full-text examination of the 18 articles retained for the review.

Conclusion

The review show that continuous variables can be discretize into binary outcomes in several ways, ranging from simple thresholding to more structured approaches based on latent variables, copulas, and simulation algorithms. Each method serves a different purpose and involves different assumptions about the data being generated. For some simulation settings, a simple threshold may be sufficient, while others require greater attention to marginal distributions and the dependence between variables. This means that the choice of discretization method should be guided by the structure of the data and the characteristics simulation design.

-
- Acknowledgements:** The authors would like to express their sincere gratitude to Universiti Teknologi MARA (UiTM) for providing access to academic databases and research resources that supported this review. pecial appreciation is extended to colleagues and peers who contributed valuable insights and constructive feedback, which greatly enhanced the quality of this paper.
- Funding Statement:** No Funding
- Conflict of Interest Statement:** The authors declare that there is no conflict of interest regarding the publication of this paper. All authors have contributed to this work and approved the final version of the manuscript for submission to the Journal of Information System and Technology Management (JISTM).
- Ethics Statement:** This study did not involve any human participants, animals, or sensitive data requiring ethical approval. The authors confirm that the research was conducted in accordance with accepted academic integrity and ethical publishing standards.
- Author Contribution Statement:** Kartini Kasim was responsible for the conceptualization, literature search, methodology, synthesis of the reviewed studies, and preparation of the overall manuscript. Hashibah Hamid and Ayu Abdul Rahman provided supervision and methodological guidance, particularly on the discretization approach.
-

References

- Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). John Wiley & Sons. <https://doi.org/10.1002/0470114754>
- Ahlemeyer-Stubbe, A., & Müller, A. (2021). The importance of domain knowledge for successful and robust predictive modelling. *Applied Marketing Analytics*, 6(4), 344–352.
- Alazaidah, R. (2023). A comparative analysis of discretization techniques in machine learning. In *2023 24th International Arab Conference on Information Technology (ACIT)*. IEEE. <https://doi.org/10.1109/ACIT58888.2023.10453749>
- Amatya, A., & Demirtas, H. (2016). Concurrent generation of multivariate mixed data with variables of dissimilar types. *Journal of Statistical Computation and Simulation*, 86(18), 3595–3607. <https://doi.org/10.1080/00949655.2016.1177530>
- Demirtas, H. (2017). Concurrent generation of binary and nonnormal continuous data through fifth-order power polynomials. *Communications in Statistics: Simulation and Computation*, 46(1), 344–357. <https://doi.org/10.1080/03610918.2014.963613>
- Demirtas, H., & Hedeker, D. (2011). A practical way for computing approximate lower and upper correlation bounds. *The American Statistician*, 65(2), 104–109. <https://doi.org/10.1198/tast.2011.10090>
- Demirtas, H., Hedeker, D., & Mermelstein, R. J. (2012). Simulation of massive public health data by power polynomials. *Statistics in Medicine*, 31(27), 3337–3346. <https://doi.org/10.1002/sim.5362>
- DeYoreo, M., & Kottas, A. (2015). A fully nonparametric modeling approach to binary regression. *Bayesian Analysis*, 10(4), 821–847. <https://doi.org/10.1214/15-BA963SI>
- Emrich, L. J., & Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician*, 45(4), 302–304. <https://doi.org/10.1080/00031305.1991.10475828>
- Fialkowski, A., & Tiwari, H. (2019). SimCorrMix: Simulation of correlated data with multiple variable types including continuous and count mixture distributions. *The R Journal*, 11(1), 248–264. <https://doi.org/10.32614/RJ-2019-022>
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43(4), 521–532. <https://doi.org/10.1007/BF02293811>
- Grobler, A. C., & Lee, K. (2020). Multiple imputation in the presence of an incomplete binary variable created from an underlying continuous variable. *Biometrical Journal*, 62(2), 467–478. <https://doi.org/10.1002/bimj.201900011>
- Hadjicostas, P. (2006). Maximizing proportions of correct classifications in binary logistic regression. *Journal of Applied Statistics*, 33(6), 629–640. <https://doi.org/10.1080/02664760600723367>
- Higham, N. J. (2002). Computing the nearest correlation matrix—A problem from finance. *IMA Journal of Numerical Analysis*, 22(3), 329–343. <https://doi.org/10.1093/imanum/22.3.329>
- Jiryaie, F., Withanage, N., Wu, B., & de Leon, A. R. (2016). Gaussian copula distributions for mixed data, with application in discrimination. *Journal of Statistical Computation and Simulation*, 86(9), 1643–1659. <https://doi.org/10.1080/00949655.2015.1077386>
- Liquet, B., & Riou, J. (2019). CPMGLM: An R package for p-value adjustment when looking for an optimal transformation of a single explanatory variable in generalized linear models. *BMC Medical Research Methodology*, 19, Article 79. <https://doi.org/10.1186/s12874-019-0711-2>

- Lustgarten, J. L., Gopalakrishnan, V., Grover, H., & Visweswaran, S. (2008). Improving classification performance with discretization on biomedical datasets. *AMIA Annual Symposium Proceedings*, 445–449.
- 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. <https://doi.org/10.1037/1082-989X.7.1.19>
- Morlini, I. (2011). Mixed mode data clustering: An approach based on tetrachoric correlations. In F. Palumbo et al. (Eds.), *Studies in Classification, Data Analysis, and Knowledge Organization* (pp. 95–103). Springer. https://doi.org/10.1007/978-3-642-13312-1_9
- Morlini, I. (2012). A latent variables approach for clustering mixed binary and continuous variables within a Gaussian mixture model. *Advances in Data Analysis and Classification*, 6(1), 5–28. <https://doi.org/10.1007/s11634-011-0101-z>
- Nelson, S. L. P., Ramakrishnan, V., Nietert, P. J., Wolf, B. J., & Taylor, M. H. (2017). An evaluation of common methods for dichotomization of continuous variables to discriminate disease status. *Communications in Statistics—Theory and Methods*, 46(21), 10823–10834. <https://doi.org/10.1080/03610926.2016.1202274>
- Nojavan, F. A., Qian, S. S., & Stow, C. A. (2017). Comparative analysis of discretization methods in Bayesian networks. *Environmental Modelling & Software*, 87, 64–71. <https://doi.org/10.1016/j.envsoft.2016.10.007>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, 25(1), 127–141. <https://doi.org/10.1002/sim.2331>
- Rucker, D. D., McShane, B. B., & Preacher, K. J. (2015). A researcher's guide to regression, discretization, and median splits of continuous variables. *Journal of Consumer Psychology*, 25(4), 666–678. <https://doi.org/10.1016/j.jcps.2015.04.004>
- Wang, J.-H., & Liu, B. (2005). Method of the discretization of continuous probability distribution for risk analysis. *Journal of Xi'an Shiyong University (Natural Sciences Edition)*, 20(2), 83–85.
- Zhao, J., Liu, X., Du, B., & Liu, Y. (2025). Approximation error from discretizations and its applications. *The Annals of Statistics*, 53(2), 589–614. <https://doi.org/10.1214/24-AOS2470>