Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 256 (2025) 292–299

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN – International Conference on Project MANagement / HCist – International Conference on Health and Social Care Information Systems and Technologies 2024

Applying Markov Chains in Data Quality Management for GDPR Compliance: A New Perspective

António Gonçalves[1,2]*, Anacleto Correia[1]

[1]*CINAV, Almada, 2810-001, Portugal*
[2]*INESC-ID, Lisboa, 1000-029, Portugal*

Abstract

In today's digital context, data quality management is of critical importance, especially considering the General Data Protection Regulation (GDPR). This regulation imposes the need for accurate and up-to-date data, underlining strict data management to protect the privacy and security off personal data. Markov Chains represent a promising predictive approach, offering a tool for organizations to anticipate and mitigate data security risks while improving GDPR compliance. This study highlights the fundamental connection between data quality and breach prevention, proposing the use of Markov Chains as an innovative method to improve data quality management, regulatory compliance, and personal data security, suggesting a proactive direction for future organizational practices.

Keywords: GDPR; data quality; markov chain; compliance; data breach

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: agoncalveslx@gmail.com

10.1016/j.procs.2025.02.123

1. introduction

At the forefront of the digital era, the transformative shift within the domain of information systems underscores the paramount importance of data quality as a foundational pillar for the successful operation of entities across various sectors. The advent of the General Data Protection Regulation (GDPR) has significantly amplified the complexities inherent in data stewardship, necessitating an unwavering compliance with stringent regulatory frameworks designed to safeguard the privacy and integrity of personal data. This evolving paradigm mandates organizations to maintain rigorous standards of data accuracy, timeliness, and security from inception to ultimate application.

The integration of Markov Chains into this framework represents a significant paradigm shift, offering a structured and predictive methodology by which organizations can proactively identify and address potential vulnerabilities within their data security mechanisms, thereby elevating the value of data throughout its lifecycle.

Amid the heightened value on data quality engendered by GDPR compliance, and the promising potential of Markov Chains in upholding such quality, this article probes the pivotal research question:

"How does the application of Markov Chains in data quality management enhance GDPR compliance and diminish the risk of personal data breaches within organizational infrastructures?"

This investigation endeavours to integrate the theoretical underpinnings of Markov Chains with their practical applications in data quality management, presenting an innovative perspective on the utilization of predictive analytics to bolster data security and ensure compliance with regulatory mandates.

The article progresses with a detailed exposition on the intrinsic value of data quality within the GDPR framework, subsequently transitioning to an exploration of contemporary data management practices. An in-depth analysis of Markov Chains and their theoretical application to enhancing data security sets the stage for the innovative proposition delineated, which posits a predictive model for the anticipation of data breaches. The subsequent sections delve into the ethical and practical considerations of the proposed model, evaluating its implications in the broader context of GDPR compliance. The concluding section aggregates the study's findings, outlining future research avenues and best practices designed to elevate GDPR adherence.

2. Data quality under the GDPR

The concept of data quality is dynamic, constantly evolving to meet the emerging demands and challenges in information management. Echoing the pioneering views of Brodie [1], data quality is heralded as a vital element across the data lifecycle, transcending basic collection and storage to include ongoing maintenance and assessment through the integration of novel tools and cutting-edge concepts and methodologies. This holistic view ensures data remains accurate, trustworthy, and relevant to user needs.

Data quality's role is pivotal, influencing the efficacy of information systems, bolstering user confidence, and ensuring regulatory compliance. As data becomes increasingly fundamental to business operations and decision-making, the imperative for maintaining high-quality standards intensifies. Notably, data quality is essential for adhering to the General Data Protection Regulation (GDPR), which mandates the accuracy and, when necessary, the timeliness of personal data updates. Consequently, data quality management transitions from a best practice to a regulatory mandate across various contexts, highlighting its critical importance for organizational functionality. Hence, GDPR elevates data quality as a key standard for data privacy and security protection, compelling organizations to ensure the processed data's precision, currency, and retention only as required for specified purposes [2].

As outlined by Balau [3], GDPR necessitates strict adherence to data quality standards by data controllers, emphasizing accuracy, relevance, and data minimization. The regulation significantly influences how organizations gather, store, and utilize personal data. Menges [4] elaborates on the significance of pseudonymization in bolstering data security under GDPR, showcasing the interplay between data quality and privacy protection strategies in mitigating security threats.

In conclusion, data quality is a multifaceted discipline that lies at the heart of digital transformation. It is a critical success factor for organizations aiming to thrive in the digital age, requiring a holistic approach that integrates technology, processes, and people. The future of digital innovation and governance will increasingly depend on the ability of organizations to maintain high standards of data quality, underscoring its importance as a strategic asset in the quest for excellence, security, and compliance in the digital domain.

3. Markov chain and data quality

Markov Chains [5], named after the Russian mathematician Andrey Markov, are a class of stochastic processes that have found extensive applications in various fields, from probability theory to computer science and engineering. The essence of these processes lies in the "memoryless" property, or Markov property, where the probability of transitioning to the next state depends only on the current state and not on the sequence of events that preceded it.

Markov Chains are models used to describe systems that undergo transitions between states with defined probabilities, independent of previous history (the "memoryless" property). The "states" represent the different possible conditions of the system, while the "transition probabilities" indicate the chance of moving from one state to another, all summarized in a "transition matrix." Each row of this matrix must sum to 1, representing all possible next conditions. The "time step" refers to the interval between transitions. The "initial state distribution" defines the initial probabilities of each state, and the "steady-state distribution" describes the probabilities that the system reaches after an extensive number of steps, reflecting the long-term behaviour of the system [5].

The application of Markov Chains in data quality management highlights a systematic and predictive approach, essential for the development of robust, accurate, and efficient information systems. By utilizing Markov Chains, organizations gain a detailed understanding of data quality dynamics, essential for making decisions that seek accuracy, reliability, and efficiency in data systems.

4. Data quality and personal data breach under the GDPR

The GDPR represents a significant stride forward in the field of data protection, setting forth comprehensive requirements for the collection, storage, and management of personal data. This regulatory framework ensures that data is processed in a manner that is fair, transparent, and secure. The introduction of GDPR has underscored the importance of data integrity and the rights of individuals, thereby reshaping the landscape of personal data management across the European Union and beyond [6]. This set of precepts are an integral part of the GDPR through a set of articles and recitals presented in Table I.

Table 1. GDPR and data quality.

Article	Recitals	Description
5	39, 58	It defines essential principles for the processing of personal data, emphasizing the accuracy and timeliness of the data.
6	39,58	Establishes the legal conditions for processing personal data, ensuring that they are pertinent and restricted to what is essential for their purposes.
15	66	Defines the right of individuals to verify and correct their personal data processed by the controller.
16	66	It guarantees the right of data subjects to rectify inaccurate or incomplete personal data concerning them.
23	39, 58	It allows member states to establish exceptions to the principle of accuracy of personal data to ensure the protection of the freedom and rights of data subjects.
25	78	Discusses the initial integration of data protection measures, such as pseudonymization and data minimization, into systems and products to improve data quality.
32	39	It stipulates that the controller and the controller must implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, including the accuracy of personal data.

This table summarizes the key components of the GDPR that are related to data quality, reflecting how these principles and legal requirements interconnect to promote accuracy, security, and adequacy in the processing of personal data. Data quality is not just a matter of compliance but also an ethical and practical issue, essential for the trust of data subjects and for the effectiveness of data processing operations.

Under the GDPR, data quality is an integral part of data protection from the outset (data protection by design) and by default (data protection by default), meaning that appropriate measures must be implemented to ensure that personal data is processed to the highest standard of security and only for the necessary period. Furthermore, data subjects have the right to rectify incorrect or incomplete data [7].

A personal data breach occurs when there is a security failure that results, unintentionally or unlawfully, in the destruction, loss, alteration, unauthorized disclosure of, or access to, personal data transmitted, stored, or otherwise processed. This incident can affect the confidentiality, integrity, or availability of personal data and results in various risks to the rights and freedoms of individuals, including financial losses, damage to reputation, loss of confidentiality of data protected by professional secrecy, discrimination, and other significant disadvantages.

The lack of data quality, by contributing to the inaccuracy, incompleteness, or outdatedness of personal data, significantly increases the risk of personal data breaches. Poor quality data can lead to inadequate management and incorrect application of security measures, increasing the chances of a breach and making the detection and response process more difficult. Thus, ensuring data quality is fundamental to reducing the possibility of personal data breaches and to ensuring compliance with data protection regulations such as the GDPR. The relationship between data quality and personal data breaches is fundamentally as follows [8]: **Breach Prevention:** The prevention of personal data breaches starts with maintaining high data quality, a premise based on the triad of accuracy, relevance, and security. Each of these elements plays a crucial role in mitigating risks associated with data management. **Breach Impact**: Data quality also affects the severity of a personal data breach. Inaccurate or outdated data can amplify the negative impact of a breach, affecting decisions based on incorrect information or exposing information of individuals who should not be in the database. **Accountability and Legal Consequences:** The adoption of the GDPR sets strict standards for the management and security of this information. One of the fundamental requirements of the GDPR is the obligation of organizations to ensure data quality, emphasizing the accuracy, relevance, and timeliness of the personal information they process. Failure to maintain data quality is not just an internal management issue; under the GDPR, such negligence is interpreted as a breach of legal obligations. **Breach Response:** Data quality is crucial for effectively managing a data breach, serving as the foundation for response and recovery processes. Accurate and up-to-date data allow organizations to quickly identify which information was affected and which individuals need to be notified. **Transparency and Trust:** Maintaining data quality is essential not only for compliance but also for establishing trust between organizations and individuals. Transparency in data management is crucial to foster this trust. Rigorous practices that ensure data accuracy, timeliness, and relevance demonstrate a commitment to integrity and privacy. **Fulfilment of Data Subject Rights:** Data quality is directly related to an organization's ability to fulfil the rights of data subjects under the GDPR, such as the right to access, rectification, and erasure. Poor data quality poses a significant threat not only to the operational integrity of organizations but also to the protection of fundamental rights of data subjects.

Therefore, data quality is essential for preventing data breaches, mitigating the impact of these breaches when they occur, and ensuring compliance with legal obligations under the GDPR. A proactive approach to maintaining high data quality not only helps avoid data breaches but also strengthens overall compliance with the GDPR, protects individuals' privacy, and upholds the integrity and reputation of organizations.

5. Proposal method

Following the rise in data privacy concerns and GDPR compliance, we propose a method for predicting personal data breaches using Markov Chains. This method aims to enhance organizations' ability to anticipate and mitigate data security risks.

The method is inspired by previous studies demonstrating the effectiveness of Markov Chains in various domains, including cyber-attack detection [9], inventory forecasting [10], air quality prediction 14], and network security forecasting [11]. The proposed method is divided into four stages: **Data Collection and Preparation:** Initially, we

gather and prepare a historical dataset documenting previous security incidents and normal network activities. These data are essential for training the Markov Chain model, **and their quality is crucial for prediction accuracy. Construction of the Markov Chain Model:** We use the prepared data to construct a Markov Chain model, capturing transitions between different data security states. The model is trained to recognize patterns that precede data breaches. **Model Validation and Adjustment:** After constructing the model, we conduct a series of tests to validate its accuracy and reliability. This may involve using test datasets or simulating data breach scenarios. Based on the results, we adjust the model as necessary. **Implementation and Monitoring:** Finally, we implement the model and begin continuous monitoring. The model will predict the likelihood of future data breach occurrences, enabling security teams to proactively respond to mitigate potential threats.

For the Markov Chain model construction, it is crucial to define a set of states and transitions reflecting the various conditions of personal data processing under GDPR principles. These conditions include accuracy, legality of processing, the right to rectification, and data protection measures described in table I. The proposed states are: **Accurate and Updated Data (AUD):** Data complying with Article 5, reflecting the required accuracy and updating. **Legally Processed Data (LPD):** Data that comply with Article 6, processed under an adequate legal basis. **Data Being Corrected (DBC):** Data being corrected by the data subject, according to Articles 15 and 16, to ensure its accuracy. **Data with Accuracy Exception (DAE):** Data that, due to specific circumstances (Article 23), may not fully comply with the accuracy requirement. **Protected and Minimized Data (PMD):** Data that have been subject to initial protection measures, such as pseudonymization and minimization, as per Article 25, to improve their quality. **Data at Risk of Breach (DRB):** Data that, despite implemented security measures (Article 32), present a high risk of breach.

The transitions represent changes in data states resulting from internal actions of data controllers or data subjects, as well as external events that may influence data security. The proposed transitions are (Fig 1): **From AUD to LPD:** Applies when accurate data are subjected to legal processing. **From LPD to DBC:** Begins when data subjects request correction of their data. **From DBC to AUD:** Occurs after data correction, restoring its accuracy. **From AUD/LPD to DAE:** When exceptions to the data accuracy principle are applied. **From AUD/LPD/DBC to PMD:** Reflects the implementation of data protection measures from the start. **From any state to DRB:** Indicates the identification of a high risk of data breach.

This model aims to facilitate the prediction of potential personal data breaches, underscoring the importance of adopting proactive and reactive measures in line with GDPR principles, particularly regarding accuracy, legality, and data security.

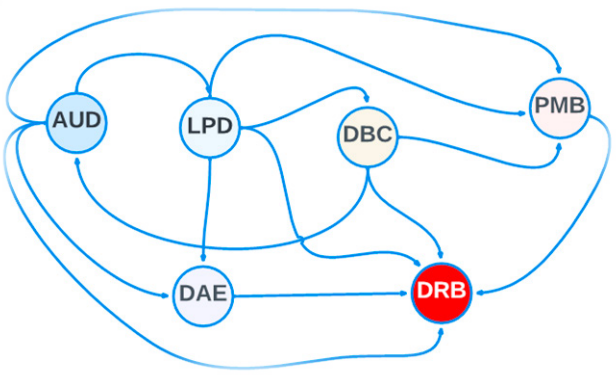

Figure 1 Markov chain model.

To calculate the percentages of transitions in a Markov Chain model's matrix, it is necessary to follow a set of steps that involve collecting and analysing historical data on the transitions between states. With this collection, it will be possible to calculate the transition between states. It's enough to count how many times the transition occurred in the historical data. For example, how many times the data moved from AUD (Accurate and Updated Data) to LPD (Legally Processed Data), from LPD to DBC (Data Being Corrected), and so on. The transition probability from one state to another is calculated by dividing the number of times a specific transition occurred by the total number of exits from the initial state. The formula is:

$$Pi,j = \frac{Ni,j}{\sum_k Ni,k}$$

The transition matrix P is then formed by all the transition probabilities Pij, representing the percentages of change from one state to another within the model. Each row of the matrix will sum to 100%, reflecting all possible transitions departing from a specific state. To validate the model, data must be divided into two sets: a training set and a test set. The training set is used to construct the model, while the test set is used to evaluate its performance. Additionally, we can utilize cross-validation, as it is a robust technique for assessing the model's generalization capability. It consists of dividing the data into several partitions and conducting multiple rounds of training and testing, alternating the partitions used for each purpose. This process helps to identify whether the model is overfitting to the training data or not.

Based on the validation results, adjustments may be necessary to improve the model's performance. This might include revising the transition probabilities, altering the defined states, or incorporating new data. Methods such as sensitivity analysis can be useful for determining which parameters have the greatest impact on the model's performance and should therefore be adjusted.

6. Case study

To conducting a case study, an existing dataset on Kaggle [12] was utilized, which is a compilation of data from various sources detailing data breaches that occurred between 2004 and 2021, encompassing various entities and breach methods. The dataset includes columns such as "Entity", "Year", "Records" affected, "Type of Organization", and "Method" of breach. The analysis focused primarily on the "Method" column, detailing the different ways data were compromised, including, but not limited to cyber-attacks ("hacked"), security failures ("poor security"), and loss or theft of physical data carriers ("lost/stolen media").

In contextualizing the methods of data breach under GDPR, a mapping approach was adopted to align the types of breaches with specific data quality attributes emphasized by the GDPR. This approach aims to understand how different data breach incidents affect compliance and the integrity of personal data. The GDPR attributes considered were Accuracy, Purpose Limitation, Data Minimization, and Integrity and Confidentiality.

To represent this mapping in a Markov model structure, it would be necessary to construct a transition matrix, where each state represents a data quality attribute as defined by the GDPR. However, to build a Markov table based on the dataset data, we need to clearly define the sequences of states (or transitions) from the data breach methods. Nonetheless, the dataset does not provide a direct temporal sequence of transitions between breach methods for specific entities, which is an essential component for constructing a traditional Markov transition matrix that reflects the real probabilities of transition between states.

A possible approach is to create a table that represents the relative frequencies of the different data breach methods present in the dataset. The distribution of simplified data breach methods, based on the dataset data, is presented in the following table:

Table 2. Relative frequencies of data breach.

Attribute	Frequency
Integrity and Confidentiality	100.00%
Accuracy	61.02%
Limitation	26.44%
Minimization	20.34%

.

Here are some observations on this distribution: Integrity and Confidentiality: This attribute is impacted by all categories of data breaches, reflecting the comprehensive nature of security concerns. Therefore, it reaches a total frequency of 100%, indicating that all types of data breaches contribute to the risk to this attribute. Accuracy: Primarily impacted by breaches involving unauthorized access or manipulation of data (such as "Hacked" and "Accidental Exposure"), which represent approximately 61.02% of the breaches. This underscores the importance of protecting data against unauthorized changes to maintain its accuracy. Purpose Limitation and Data Minimization: These attributes are affected by a smaller proportion of breaches, 26.44% and 20.34% respectively, reflecting types of breaches that result in the use or exposure of data beyond what is necessary for the purposes for which they were collected.

To convert the distribution of data breach methods into a Markov transition matrix, we will apply a theoretical assumption. The assumption we will make is that, after each type of data breach, the next breach could be of any type, distributed according to the relative frequencies observed in the dataset. That is, each type of breach "transits" to another type (including itself) with a probability corresponding to its relative frequency in the dataset (table 3):

Table 3. Theoretical Markov transition matrix.

	Integrity and Confidentiality	Accuracy	Limitation	Minimization
Integrity and Confidentiality	48,12%	29,36%	12,72%	9,79%
Accuracy	48,12 %	29,36%	12,72%	9,79%
Limitation	48,12 %	29,36%	12,72%	9,79%
Minimization	48,12 %	29,36%	12,72%	9,79%

Each cell in this matrix indicates the transition probability from one GDPR attribute (indicated by the row) to another (indicated by the column), based on the relative frequencies of data breaches affecting those attributes. For example, the transition probability from "Integrity and Confidentiality" to "Accuracy" is 29.36%, reflecting the proportion of data breaches that impact these two attributes.

The theoretical Markov transition matrix we created is based on the distribution of simplified types of data breaches, mapped to the data quality attributes specified by the General Data Protection Regulation (GDPR): Integrity and Confidentiality, Accuracy, Purpose Limitation, and Data Minimization.

Here are the main interpretations of the matrix: Weighted Distribution of Transitions: Each row in the matrix shows the distribution of transition probabilities from one GDPR attribute to others after a data breach. For example, the high probability (approximately 48.12%) of transition from any GDPR attribute back to "Integrity and Confidentiality" reflects the dominance of this theme in the dataset's data breach incidents. This suggests that, regardless of the specific GDPR attribute initially affected by a breach, there is a significant chance that the next breach will also affect the Integrity and Confidentiality of the data. Widespread Impact on Data Breaches: The model suggests that breaches affecting Integrity and Confidentiality have the highest likelihood of occurring after any type of previous breach, indicating the widespread nature of data security threats across all categories. This aligns with the reality that many data breaches compromise data security and protection, regardless of the initial cause. Accuracy and Data Transitions: Transitions to "Accuracy" represent the second highest probability, indicating that many data breaches, in addition to compromising security, may result in inaccurate or manipulated information. This highlights the importance of measures that ensure data accuracy following a breach, especially in contexts where data are critical for business or personal decisions. Considerations on Purpose Limitation and Data Minimization: While these categories show lower transition probabilities compared to "Integrity and Confidentiality" and "Accuracy", they still represent relevant post-breach concerns. This underscores the importance of limiting data use to the original purposes and ensuring that only necessary data are collected and retained. Strategic Implications for Data Protection: The matrix provides a theoretical framework for understanding how different types of breaches can affect GDPR data quality attributes. Organizations can use this information to prioritize focus areas in data breach prevention and incident response planning, considering the implications in terms of GDPR compliance.

7. Conclusion and future work

Data quality under GDPR stands as a crucial element for the efficiency and effectiveness of contemporary information systems, directly impacting organizations' ability to make informed decisions, enhance operational procedures, and maintain user trust and regulatory compliance. GDPR compliance demands that personal data be accurate, up-to-date, and retained only if necessary, highlighting the importance of stringent data management for protecting privacy and personal data security.

Future work should focus on implementing predictive approaches, such as using Markov Chains, to anticipate and mitigate data security risks, enhancing organizations' ability to predict personal data breaches. This systematic and predictive approach allows for a detailed understanding of data quality dynamics, essential for making decisions aimed at accuracy, reliability, and the efficiency of data systems.

Additionally, analysing the relationship between data quality and personal data breaches underscores the importance of maintaining high data quality to prevent breaches, mitigate their impact when they occur, and ensure compliance with GDPR legal obligations. This includes continuous data verification and updating practices, along with comprehensive security measures and effective data management practices to protect against data misuse and privacy violations.

Moving forward, organizations are recommended to adopt a proactive approach to data quality management, implementing robust data quality policies and utilizing predictive models, like Markov Chains, to anticipate potential data breaches. This approach not only helps prevent data breaches but also strengthens overall GDPR compliance, protects individual privacy, and upholds the integrity and reputation of organizations.

In summary, maintaining data quality and implementing predictive methods for managing data security risks are crucial for GDPR compliance and for protecting privacy and personal data security. Ongoing commitment to these practices is essential for building lasting trust relationships between organizations and data subjects, ensuring the effectiveness of data processing operations, and the sustainability of information systems within organizations.

References

[1] M. L. Brodie, “Data quality in information systems,” *Inf. \& Manag.*, vol. 3, no. 6, pp. 245–258, 1980.

[2] T. Knauer, N. Nikiforow, and S. Wagener, “Determinants of information system quality and data quality in management accounting,” *J. Manag. Control*, vol. 31, pp. 97–121, 2020.

[3] D. P. Ballou and H. L. Pazer, “Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems,” *Manage. Sci.*, vol. 31, pp. 150–162, 1985.

[4] F. Menges *et al.*, “Towards GDPR-compliant data processing in modern SIEM systems,” *Comput. Secur.*, vol. 103, p. 102165, 2021.

[5] J. R. Norris, *Markov chains*, no. 2. Cambridge university press, 1998.

[6] V. Chico, “The impact of the General Data Protection Regulation on health research,” *Br. Med. Bull.*, vol. 128, pp. 109–118, 2018.

[7] K. Hjerppe, J. Ruohonen, and V. Leppänen, “The General Data Protection Regulation: Requirements, Architectures, and Constraints,” *2019 IEEE 27th Int. Requir. Eng. Conf.*, pp. 265–275, 2019.

[8] T. Hoeren, “Big Data and Data Quality,” pp. 1–12, 2018.

[9] N. Ye, Y. Zhang, and C. Borror, “Robustness of the Markov-chain model for cyber-attack detection,” *IEEE Trans. Reliab.*, vol. 53, pp. 116–123, 2004.

[10] Z. He and W. Jiang, “A new belief Markov chain model and its application in inventory prediction,” *Int. J. Prod. Res.*, vol. 56, pp. 2800–2817, 2017.

[11] Y. Wang, W. Li, and Y. Liu, “A Forecast Method for Network Security Situation Based on Fuzzy Markov Chain,” pp. 953–962, 2013.

[12] K. Banachewicz, L. Massaron, and A. Goldbloom, *The Kaggle Book: Data analysis and machine learning for competitive data science*. Packt Publishing Ltd, 2022.