Quantitative Assessment on the Misuse Risk of Intelligent Connected Vehicle Data

Lu, Yi; Li, Hao; Tu, Huizhao; Liu, Jian; Yuan, Yufei; Van Lint, Hans

DOI
10.1061/JTEPBS.TEENG-9361

Publication date
2025

Document Version
Final published version

Published in
Journal of Transportation Engineering Part A: Systems

Citation (APA)
Lu, Y., Li, H., Tu, H., Liu, J., Yuan, Y., & Van Lint, H. (2025). Quantitative Assessment on the Misuse Risk of Intelligent Connected Vehicle Data. *Journal of Transportation Engineering Part A: Systems*, *152*(2), Article 04025133. https://doi.org/10.1061/JTEPBS.TEENG-9361

Quantitative Assessment on the Misuse Risk of Intelligent Connected Vehicle Data

Yi Lu[1]; Hao Li, Ph.D.[2]; Huizhao Tu, Ph.D.[3]; Jian Liu[4]; Yufei Yuan, Ph.D.[5]; and Hans van Lint, Ph.D.[6]

Abstract: The operation of intelligent connected vehicles (ICVs) is fundamentally data-driven, continuously generating massive amounts of data. Given the significant value of ICV data to enterprises, industries, and nations, promoting data openness and sharing has become essential. However, such data often contain sensitive information, and its misuse can threaten individual privacy, corporate security, and even national interests. To address this dilemma, this paper develops the misuse risk score (MR-score), a novel quantification model and associated evaluation method for assessing the risk of ICV data misuse. The MR-score is constructed based on three core properties of ICV data: sensitivity; scale; and identifiability. The sensitivity score, information quantity, and identifiability factor are designated as the corresponding evaluation indicators, and systematic approaches for their quantification are proposed. The analytic hierarchy process is employed to measure the sensitivity score. Information entropy is adopted to evaluate the information quantity. A combination of *k*-anonymity-based and damage source determination-based methods is utilized to estimate the identifiability factor, considering data incompleteness, imprecision, and invalidity. Two empirical ICV data sets are utilized, and comparative analyses are conducted to demonstrate the effectiveness of the MR-score in capturing misuse risks. Higher MR-scores correspond to greater risk. The model captures the joint influence of all three data properties and reveals the marginal diminishing effect of data scale on misuse risk. This work offers valuable tools for data owners and regulatory agencies to prioritize critical data sets, implement targeted data protection measures, and enable secure data circulation while maximizing the value of ICV data. **DOI: 10.1061/JTEPBS.TEENG-9361.**

Author keywords: Intelligent connected vehicle (ICV) data; Data misuse risk; Data sensitivity; Data scale; Data identifiability.

Introduction

The intelligent connected vehicle (ICV), which deeply integrates automotive systems with information technologies, communications, and artificial intelligence, represents the future trajectory of automotive technology and industry (Fei et al. 2024; Jin et al. 2024; Tu et al. 2024). The development, deployment, and assessment of ICV's functionalities rely on massive data while simultaneously generating substantial data volumes.

ICV data encompass diverse categories across five dimensions: basic vehicle properties; operational status; environmental perception outputs; object recognition results; and occupant-related information (Mlada et al. 2022; Vaniš et al. 2022). It exhibits distinguished characteristics: (1) high risks in use: in general ICV data sets may contain private and sensitive information, particularly information on geolocation, vehicle operation, and infrastructure-related data, of which the misuse could result in serious consequences such as privacy breaches, exposure of ICV's technique functions, or systemic vulnerabilities (Elrose et al. 2022; Zhou et al. 2022; Almaskati et al. 2024); (2) high diversity in data sources: the data attributes contained in ICV data sets across the five aforementioned dimensions are collected from different sources, for instance from vehicle-mounted sensors (LiDAR/cameras), in-vehicle networks (CAN/Ethernet), V2X communication systems (C-V2X/5G), cloud-based services (HD mapping platforms), etc. It leads to diversity in data features like data types, data formats, etc., which results in difficulty in data assessments; (3) high diversity in data contents: besides the data describing the target ICVs, the data set might contain many attributes of infrastructure (e.g., road topology), environment (weather conditions), etc. Its misuse risks may depend on different data regulations and data freshness; (4) massive volume: due to the continuous movement of ICVs, the dynamic motion records, including velocity, acceleration rates, real-time locations, etc., lead to an extensive number of records per vehicle and a massive-volume data set.

These characteristics reveal a practical dilemma in ICV data governance. Insufficient safeguards may result in the unintended exploitation of critical data elements, compromising individual privacy, corporate security, or even national interests. Conversely, excessive protection of high-value ICV data may incur substantial

[1]Ph.D. Candidate, College of Transportation, Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji Univ., 4800 Cao'an Rd., Jiading District, Shanghai 201804, PR China. Email: yi_lu@tongji.edu.cn

[2]Professor, College of Transportation, Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji Univ., 4800 Cao'an Rd., Jiading District, Shanghai 201804, PR China (corresponding author). Email: haolitj@tongji.edu.cn

[3]Professor, College of Transportation, Key Laboratory of Road and Traffic Engineering, Ministry of Education, Tongji Univ., 4800 Cao'an Rd., Jiading District, Shanghai 201804, PR China. Email: huizhaotu@tongji.edu.cn

[4]Manager, Cangzhou Qugang Expressway Construction Co., Ltd., Heibei 062255, PR China. Email: jianliuqg@163.com

[5]Senior Researcher, Dept. of Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft Univ. of Technology, Stevinweg 1, Delft 2628CN, Netherlands. Email: y.yuan@tudelft.nl

[6]Professor, Dept. of Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft Univ. of Technology, Stevinweg 1, Delft 2628CN, Netherlands. Email: j.w.c.vanlint@tudelft.nl

Note. This manuscript was submitted on May 23, 2025; approved on September 11, 2025; published online on November 26, 2025. Discussion period open until April 26, 2026; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Transportation Engineering, Part A: Systems***, © ASCE, ISSN 2473-2907.

storage costs and administrative burdens, while impeding data sharing and contradicting national strategies that promote data circulation, potentially constraining industry growth (DAMA 2017). This dilemma underscores the need for developing robust ICV data misuse risk assessment frameworks that strike a balance between security and openness (CEN 2016; Usmonov 2024).

Existing studies have conducted quantitative risk assessment methodologies across various domains, including personal data (Vavilis et al. 2016; Laurie et al. 2017; Mlada et al. 2022; Soussan and Trovati 2022), corporate information (Soomro and Ahmed 2013; Corallo et al. 2020), government records (Akanfe et al. 2020), etc. Some measurement approaches have been introduced, such as information security risk assessment (ISRA) (Wangen 2017; Gozhyj et al. 2020; Shaikh and Siponen 2023), financial impact modeling (Wang et al. 2019; Corallo et al. 2020), and specialized scoring systems, including M-score (Harel et al. 2012; Vavilis et al. 2014), L-severity (Vavilis et al. 2016), TKL-score (Eng and Stroulia 2021), risk-score (Mlada et al. 2022), etc., which can quantitatively reflect data misuse risk by certain scores. Despite these advancements, the specific quantification of ICV data misuse risk remains unexplored in current literature.

Given the unique characteristics of ICV data that distinguish it from other domains studied in the literature, existing data misuse risk quantification methods cannot be directly applied. Therefore, this study proposes a specific approach for assessing and quantifying ICV data misuse risk, with consideration of its distinct features: (1) massive data volume: ICV data sets may contain an extensive number of records per vehicle due to continuous spatiotemporal movements (e.g., velocity, real-time locations), necessitating specialized methods to quantify the data scale; (2) diverse and underexplored attributes: with over 200 distinct attributes, most of which lack established sensitivity assessments, ICV data demand dedicated research to assess the potential risks and sensitivities; and (3) dynamic versus static data: unlike conventional data sets that primarily capture static characteristics and identity information (Yu and He 2021; Soussan and Trovati 2022), ICV data encompasses static features (e.g., vehicle exterior) and dynamic information (e.g., real-time motion status, environmental perception) (Vemou and Karyda 2018a; Sudhakar and Rao 2020). This requires novel identifiability assessment methods that accommodate dynamic attributes.

This study makes contributions in multiple dimensions: (1) developing a novel quantification model and associated evaluation method, misuse risk score (MR-score), for assessing ICV data misuse risks, which enables comprehensive evaluation across key data properties: sensitivity, identifiability, and scale; and (2) designating suitable indicators and systematic approaches to measure the three key data properties considering the distinguished ICV data characteristics. Sensitivity scores across ICV data attributes are quantified. A novel identifiability evaluation approach that integrates k-anonymity and critical damage source determination is developed. These contributions support data governance by enabling stakeholders to identify critical data sets, formulate protection strategies, establish storage/sharing protocols, and implement access controls. These advancements facilitate secure data ecosystems while promoting efficient circulation and maximizing the utility of ICV data assets.

The structure of the paper is as follows. The literature review summarizes existing studies relevant to ICV data misuse risk. The methodology section introduces the proposed MR-score framework for risk quantification. The section of applications illustrates the implementation of this framework using two real-world data sets collected by an automotive technology enterprise in Shanghai. The discussion section interprets the key findings, and the conclusion section provides a summary of findings and outlines future research directions.

Literature Review

This section presents a systematic review of the concept of data misuse, the quantification methods for data misuse risks, and the properties that affect risk. It further identifies limitations in existing research, which will be explicitly addressed in this study.

Data Misuse

Data misuse is a broad concept encompassing potential causes of cybercrime, data damage, and data leakage, among others (Furnell et al. 2020; Wang et al. 2024). Data misuse can trigger adverse consequences for its owners or the entity described by the data, including individuals, enterprises, and even nations (Jung 2021).

For individuals, the misuse of personal data violates a person's privacy (Alkhalil et al. 2021; Chua et al. 2021; Jain et al. 2021; Hysa et al. 2023). Specifically, vehicle identification information in ICV data is closely related to personally identifiable information, and its misuse will cause privacy breaches (Rannenberg 2016; Chah et al. 2022; Vaniš et al. 2022; Matin and Dia 2024).

The misuse of corporate information has adverse consequences for enterprises. Martin et al. (2017) revealed that such misuse will reduce customers' trust in an enterprise. Park et al. (2016) illustrated that technical information breaches can cause significant financial losses, especially for innovation-driven enterprises, due to the leakage of trade secrets and intellectual property. At the national level, the misuse of critical technical data, e.g., information on automobiles, communications, and artificial intelligence, can undermine industrial development and even harm national economic interests (Dolzhenkova et al. 2020; Vatanparast 2020).

Quantification of Data Misuse Risks

As a vital indicator of data security, data misuse risk describes the severity of potential harm to the data owner or the entity described by the data as well as the probability of such harm occurring (Ethicist 2015; Martin et al. 2017; Solove and Citron 2018; Jung 2021). Ahmad et al. (2014) proposed the threat score to assess data misuse risk by considering the probability of insider employees working errors for enterprises, but the severity of misuse was not taken into account. In contrast, Corallo et al. (2020) and Wang et al. (2019) assessed the impact of data misuse on enterprises based on financial loss measurement, which does not reflect the probability of misuse occurrence.

Additionally, the ISRA method was proposed, which assesses data misuse risk by combining the likelihood of a data misuse event and the resulting loss to the owners (Shameli-Sendi et al. 2016; Wangen et al. 2018; Landoll 2021). ISRA assesses data misuse risk from a macro perspective. Specifically, it multiplies the vulnerability magnitude of the data security system, the frequency of events that can probably cause data misuse, and the value of data assets (Yang 2022; Zhou et al. 2022).

However, identifying all the factors that cause data vulnerability is complicated, and the frequency of historical events may be unavailable. Thus, some quantitative methods for data misuse risk assessment were developed based on three properties of data: data scale; data sensitivity; and data identifiability (Zardari et al. 2014; Laurie et al. 2017; Ormerod 2019; Nayak and Ojha 2020; Reyes et al. 2022). These properties characterize three essential aspects related to data misuse risk: the amount of critical information that could be disclosed through the data set; the likelihood of critical

information being identified and further misused; and the importance and criticality of the misused information to its owner.

Harel et al. (2012) developed the M-score to estimate the misuse risk of a personal data set by multiplying indicators of data scale, data sensitivity, and data identifiability. Vavilis et al. (2016) argued that M-score only accounts for the sensitivity of personally identifiable information without considering its counterparts of other attributes. Therefore, L-severity was further proposed based on M-score by assigning an accurate sensitivity score to each data attribute in the healthcare domain and summing the sensitivity scores of all the contained attributes when assessing the data set misuse risk. Eng and Stroulia (2021) further pointed out that M-score and L-severity can only reflect the identifiability of personal identity information rather than sensitive attributes. Thus, the authors developed the TKL-score, which considers the recognizable likelihood of sensitive attributes in a personal data set. However, it was also pointed out that the misuse risk assessment based on TKL-score is unsuitable for dynamic data sets, such as ICV data.

Moreover, the principle of multiplying indicators across data properties was also adopted in the development of risk-score, which assesses the misuse risk of personal data in autonomous vehicles (Mlada et al. 2022). However, the risk-score remains underdeveloped due to two major limitations: it overlooks other dimensions of ICV data attributes (e.g., basic vehicle properties and operational status); and only considers data scale and sensitivity without incorporating data identifiability. Consequently, data sets with identical data sensitivity and scale but differing identifiability are assigned the same misuse risk. Lower identifiability, which implies that critical information is less likely to be identified, might lead to reduced risks. However, this effect cannot be assessed and quantified by risk-score.

Quantification of Data Properties

As previously introduced, data misuse risk is generally quantified based on three data properties: data sensitivity; data scale; and data identifiability (Vemou and Karyda 2019). However, different quantification methods have been adopted, significantly impacting misuse risk assessment results (Yang et al. 2018; Rios et al. 2020).

Quantification of Data Sensitivity

Data sensitivity refers to the perceived importance of information and its critical degree to the data owner or the entity being described, normally measured by the indicator of sensitivity score (Harel et al. 2012; Vemou and Karyda 2018a). The calculation of sensitivity scores should consider the affected entity (e.g., a person, an enterprise, or a nation), scope (e.g., the number of people), type (e.g., financial loss, reputational damage), and degree (e.g., slight harm, moderate harm, severe harm) when the data are misused (Wagner and Boiten 2018; Gupta and Singh 2019; Gupta et al. 2022). Expert judgment is widely regarded as a practical approach for quantifying sensitivity scores through methods such as using the analytic hierarchy process (AHP) model, sensitivity ranking, multiple linear regression, etc. (Park et al. 2016; Guo et al. 2021; Kuzminykh et al. 2021; Nugraha and Martin 2022). Although some studies have qualitatively classified the ICV data into different sensitivity degrees (CAAM 2021; Ministry of Transport 2023), no research has yet quantified sensitivity scores for ICV data, which contains over 200 attributes with different sensitivity degrees.

Quantification of Data Scale

Data scale refers to the amount of information that can potentially be exposed (Hilbert 2015; Wagner and Boiten 2018). Larger-scale data sets imply broader involvement of described entities and greater potential for sensitive information disclosure (Borek et al. 2013; Miller 2013; Haley 2020). One direct indicator for measuring the data scale is the number of records in the data set (Haycock et al. 2019; Algarni et al. 2021), under the assumption of a linear relationship between record quantity and risk. However, Algarni et al. (2021) pointed out that, in a massive-scale data set, the average misuse risk of each record may marginally decline compared to a small-scale data set.

Information quantity is another applicable indicator of data scale. Data are the carrier of information, and its misuse risk depends on the quantity of information contained in the data set (Shen et al. 2019). Thus, the information quantity can be used to measure the data scale (Li et al. 2017, 2021). Shannon (1948) proposed that information entropy reflects the quantity of information. It represents the mathematical expectation of the information quantity generated by the data source after eliminating the redundancy of a data set (Zhang and Deng 2021).

The widely adopted calculation method of information entropy is marginal entropy, which relies on the number of elements or records (Zurek 2018). It is suitable for data sets where the attribute columns are independent, and repeated values within the same column cannot contribute extra information. However, it is sometimes inapplicable to ICV data. For example, attributes such as velocity and acceleration describe motion status, and their interpretability depends on the consistency of the described entity. Moreover, identical values of these attributes may carry different meanings when recorded at different time points. All the aforementioned are unquantifiable by the marginal entropy. Thus, ICV data sets require attribute-specific approaches for calculating information quantity.

Quantification of Data Identifiability

Data identifiability reflects the degree to which information can be fully recognized and further exposed through the data set (Vemou and Karyda 2018b; Bos 2020; Schlackl et al. 2022). Higher identifiability increases the potential harm in cases of misuse (Notario et al. 2015; Wagner and Boiten 2018). Common approaches to quantify data identifiability include the k-anonymity-based method and the l-diversity-based method.

The k-anonymity-based method quantifies data identifiability by quantifying the extent to which the identity information in a data set can reveal an entity, typically measured by the anonymity level (Vartanian and Shabtai 2014; Esmeel et al. 2020; Chen et al. 2023). Specifically, the data identifiability is quantified as the reciprocal of the number of entities associated with a given identity attribute (Sriramoju et al. 2014; di Vimercati et al. 2023). As the k-anonymity-based method can only measure the identifiability of identity information (Li et al. 2019), the l-diversity-based method is further proposed. It quantifies data identifiability through the probability that the sensitive attributes can be associated with a specific entity, which is approximated by the reciprocal of the l-diverse value (Eng and Stroulia 2021).

However, the sensitive attributes of ICV data are mainly ICVs' technical information. Their identifiability should be evaluated based on the extent to which the data accurately reflect ICVs' static and dynamic technical features rather than the association with specific vehicles (Gao et al. 2024). Besides, these dynamic sensitive attributes are recorded in real-time by ICVs' on-board sensors and change with their spatiotemporal motion. Their identifiability is compromised primarily by data quality problems arising from recording errors, such as incompleteness, imprecision, and invalidity (Keenan Dworak-Fisher et al. 2020; Zhao et al. 2023; Ma et al. 2025). Therefore, the two methods can hardly assess the identifiability of the dynamic attributes in ICV data.

In this context, some studies measure data identifiability by determining the damage sources that impair data integrity, validity, and precision (Jin et al. 2012; Zhang et al. 2018; Gao et al. 2024).

The possibility of accurately identifying the data decreases as the extent of damage from these sources increases (Richardson and Smith 2015; Zhang 2018; Zheng et al. 2021). Thus, the damage source determination-based method is feasible for evaluating data identifiability for ICV data. However, systematic judgment criteria for identifying damage sources in ICV data sets have not yet been established and require further investigation. In addition, specific quantitative methods should be developed to support the identification of damage sources.

Overall, existing studies suggest that the misuse risk of a data set could be assessed based on three key properties: sensitivity; scale; and identifiability. However, current research lacks methods to specifically quantify these properties for ICV data in the following aspects: 1) the sensitivity of over 200 attributes in ICV data sets remains unquantified; 2) appropriate approaches for measuring the data scale not only for attributes with dependent features but also with independent features in ICV data sets are still underdeveloped; and 3) methods for identifying damage sources that affect dynamic attributes such as location, velocity, and acceleration remain underexplored. This study aims to address these gaps and develop a systematic methodology to quantify the three properties and the misuse risks of ICV data sets.

Methodology

In this section, the ICV data misuse risk quantification method, MR-score, is proposed and established based on data sensitivity, data scale, and data identifiability, as mentioned in the previous section.

Description of ICV Data Sets

For an ICV data set $\mathbf{D}$ with n records and m attributes, each record consists of m elements corresponding to different attributes. An element in $\mathbf{D}$ can be represented as $d_{ij}(i \in [1, n], j \in [1, m])$. Moreover, a record $\mathbf{r}_i$ in the data set can be represented as

$$\mathbf{r}_i = (d_{i1}, d_{i2}, d_{i3}, \ldots, d_{im}) \tag{1}$$

Similarly, an attribute column $\mathbf{c}_j$ in the data set can be denoted as

$$\mathbf{c}_j = (d_{1j}, d_{2j}, d_{3j}, \ldots, d_{nj})^{\mathbf{T}} \tag{2}$$

Further, in an ICV data set, the types of data include identity and descriptive data (Mazilu 2020). The identity data are called the "identifiers." It directly describes an ICV's identity that can distinguish it from others (He et al. 2022; Lu and Song 2024), such as VIN codes and vehicle ID (Zhao et al. 2019). The descriptive data are called the "descriptors." It describes the ICV's status, phenomena, characteristics, or other correlated information (Vimercati and Foresti 2011; Nassaji 2015). Subsequently, descriptive data are further classified as static descriptive data and dynamic descriptive data (Stock and Guesgen 2016).

Static descriptive data illustrate the properties of entities (e.g., ICVs' shapes and colors) defined as "static descriptors." It is unchanged in the short term (generally in one day for ICV data) (Gómez Losada 2017). Meanwhile, the dynamic descriptive data in the ICV data set is generated from the ICV's spatial-temporal movement, referred to as "dynamic descriptors" (Scholtes et al. 2021). Dynamic descriptors in an ICV data set can be the vehicle's operating time, operating status, environmental perception results, and entity recognition results, which are meaningless when the entity identifiers are unknown (Gómez Losada 2017).

In this situation, $\mathbf{c}_j^{id}$ denotes the column of identifiers. Moreover, the column of descriptors is denoted by $\mathbf{c}_j^{de}$, and, specifically, the attribute columns for static and dynamic descriptors are further represented by $\mathbf{c}_j^{\text{sde}}$ and $\mathbf{c}_j^{\text{dde}}$, respectively.

Formulation of Misuse Risks of ICV Data

A method is proposed to quantify ICV data misuse risk, namely, MR-score. It measures the misuse risk of ICV data based on its properties, including ICV data sensitivity, scale, and identifiability. The data misuse risk in relation to these properties is formulated in Eq. (3)

$$\text{MR-Score} = S \times I \times IF \tag{3}$$

where S, I, and IF, correspondingly, = data sensitivity, data scale, and data identifiability. Considering ICV data characteristics, this study proposes sensitivity score, information quantity, and identifiability factor as indicators to quantify the data sensitivity, scale, and identifiability, respectively. Specifically, the ICV data sensitivity, indicated by sensitivity scores, is measured and obtained by using the AHP model. Information entropy is adopted to measure information quantity rather than the traditional method of using the number of records, because ICV data are dynamic with a massive amount (Eng and Stroulia 2021). k-anonymity-based and damage source determination-based approaches are utilized to measure the ICV data identifiability of the columns of identifiers and descriptors, respectively.

Further, considering the variations in value types (e.g., characters, floating-point numbers, integers, etc.) and data types (e.g., identifiers and descriptors) as well as differences in their sensitivity and data scale, each attribute column is taken as the basic unit for estimating the MR-score of an ICV data set. The misuse risk for each attribute column is calculated individually and then summed to obtain the misuse risk of the whole data set. The MR-score for a column $\mathbf{c}_j$ based on Eq. (3) is formulated as

$$\text{MR-Score}(\mathbf{c}_j) = S(\mathbf{c}_j) \times I(\mathbf{c}_j) \times IF(\mathbf{c}_j) \tag{4}$$

where $S(\mathbf{c}_j)$, $I(\mathbf{c}_j)$, and $IF(\mathbf{c}_j)$ = the sensitivity score, information quantity, and identifiability factor of column $\mathbf{c}_j$ in an ICV data set, respectively. After the MR-scores for all columns are obtained, the overall MR-score of data set $\mathbf{D}$ can be calculated by Eq. (5)

$$\begin{aligned}\text{MR-Score}(\mathbf{D}) &= \sum_{j=1}^{m} \text{MR-Score}(\mathbf{c}_j) \\ &= \sum_{j=1}^{m} (S(\mathbf{c}_j) \times I(\mathbf{c}_j) \times IF(\mathbf{c}_j))\end{aligned} \tag{5}$$

This column-based summation is theoretically justified and widely adopted. Theoretically, each attribute column represents an independent vector for data misuse, characterized by sensitivity, scale, and identifiability. The cumulative risk model assumes additivity: data sets with more high-risk attributes inherently carry greater total risk, which is a principle widely adopted in privacy impact assessments to reflect risk accumulation across information dimensions (Vavilis et al. 2014; Eng and Stroulia 2021). A larger MR-score ($\mathbf{c}_j$) or MR-score ($\mathbf{D}$) means higher misuse risks for column data $\mathbf{c}_j$ and data set $\mathbf{D}$, respectively.

Quantification of MR-Score

Data Sensitivity

The sensitivity of an attribute column is measured by the sensitivity score, $S(\mathbf{c}_j)$, which reflects the unit loss expectancy of an attribute

Table 1. Definition of each relative sensitivity

Relative sensitivity	Definition
1	Equally sensitive
3	Moderately more sensitive
5	Strongly more sensitive
7	Very strongly more sensitive
9	Extremely more sensitive
2, 4, 6, 8	Intermediate values between adjacent scale values

column in the ICV data set. Its value range is usually within [0, 1] (Wang et al. 2018; Guo et al. 2021).

CAAM (2021) summarizes 223 ICV data attributes and, based on the practical context of the ICV industry, classifies them into five sensitivity degrees (SDs), ranging from 1 to 5. A higher degree means that the data managers and owners believe the attributes of belonging are more critical and important. Attributes belonging to the same SD can be assigned the same sensitivity score and vice versa (Gupta et al. 2022). $S(\mathbf{c}_j)$ equals the sensitivity score of SD_x ($x = 1, 2, 3, 4, 5$) if attribute j belongs to SD_x, as shown in Eq. (6)

$$S(\mathbf{c}_j) = S(\mathrm{SD}_x), \quad \forall\ \mathbf{c}_j \in \mathrm{SD}_x \tag{6}$$

Normally, sensitivity scores for each SD are determined through expert scoring. Due to the complexity of attributes in ICV data for each SD and the variability in expert judgment standards, it is difficult for experts to value the sensitivity of all SDs directly. In this case, AHP is applied to obtain the relative weight of sensitivity of different SDs, denoted as w_x (Lee 2014), and experts' pairwise comparisons with the scale of nine levels are initially conducted (Li et al. 2013; Lee 2014; Attaallah et al. 2022). The definition of each relative sensitivity is given in Table 1.

The relative weights are subsequently transformed into sensitivity scores (Wang et al. 2018). First, the experts invited to participate in this study were to score only for SD_5, the highest sensitivity degree. The value scoring range is [0, 1], and the average scoring results will be assigned to the value of SD_5. After that, the sensitivity scores of other sensitivity degrees can be calculated by Eq. (7)

$$S(\mathrm{SD}_x) = \frac{w_x}{w_5} \times S(\mathrm{SD}_5) \quad (x = 1, 2, 3, 4, 5) \tag{7}$$

Data Scale

As illustrated in the quantification of data scale part in the literature review, the data scale of an ICV data set is quantified in terms of information quantity through the application of information entropy. For a single attribute column $\mathbf{c}_j$ with n elements $\{d_{ij} | i = 1, 2, 3, \ldots, n\}$, the marginal entropy, $H(\mathbf{c}_j)$, is defined by Eq. (8)

$$H(\mathbf{c}_j) = -\sum_{i=1}^{n} p(d_{ij}) \log_b p(d_{ij}) \tag{8}$$

where $p(d_{ij})$ = the probability of obtaining the element d_{ij}; and b = the base of the logarithmic formula, equaling 2 when the unit of measurement for information entropy is bits. Additionally, the information entropy is the average information quantity after eliminating the redundancy of a data set (Shannon 1948).

Further, conditional entropy should be adopted when the information entropy of an attribute column relies on its counterpart in another column. The $\mathbf{c}_j$ and $\mathbf{c}_q$ are two attribute columns in the data set. The conditional entropy of $\mathbf{c}_q$ based on $\mathbf{c}_j$, $H(\mathbf{c}_q|\mathbf{c}_j)$, is defined in Eq. (9) when the information quantity of $\mathbf{c}_q$ is unknown, while its counterpart in $\mathbf{c}_j$ has been measured

$$H(\mathbf{c}_q|\mathbf{c}_j) = -\sum_{d_{iq} \in \mathbf{c}_q} \sum_{d_{ij} \in \mathbf{c}_j} p(d_{ij}, d_{iq}) \log_2 p(d_{iq}|d_{ij}) \tag{9}$$

where $p(d_{ij}, d_{iq})$ = the joint probability of obtaining the element d_{ij} and d_{iq}; and $p(d_{iq}|d_{ij})$ = the conditional probability of obtaining the element d_{iq} based on its counterpart for d_{ij}.

Two calculating scenarios are established to ease the information entropy measurement for attribute columns with different data types.

Scenario 1: When the information quantity is quantified for columns $\mathbf{c}_j^{id}$ and $\mathbf{c}_j^{\mathrm{sde}}$, the marginal entropy is employed, and the column of identifiers is adopted as the computed object for the marginal entropy. This is because the column of identifiers still has practical significance. For example, a column of vehicle ID numbers contributes to the ICVs' identification information. Besides, the significance of the static descriptors depends on the entity. They are still meaningful even if the values are repetitive for different entities. The information it provides is the static properties of different vehicles (Sudhakar and Rao 2020).

In this case, the information quantity is calculated through Eq. (10), derived from Eq. (8). The result is represented as $I(\mathbf{c}_j^{id})$ and $I(\mathbf{c}_j^{\mathrm{sde}})$, respectively, and d_{ij}^{id} denotes an element in the column $\mathbf{c}_j^{id}$

$$I(\mathbf{c}_j^{id}) = I(\mathbf{c}_j^{\mathrm{sde}}) = H(\mathbf{c}_j^{id}) = -\sum_{i=1}^{n} p(d_{ij}^{id}) \log_b p(d_{ij}^{id}) \tag{10}$$

Scenario 2: The dynamic descriptors, i.e., velocity, acceleration, etc., are informative only when they describe the changes in motion status for the same entity. In this case, the conditional entropy measures the information quantity for $\mathbf{c}_j^{\mathrm{dde}}$ (Stock and Guesgen 2016). Although identical values may appear in these columns for the same entity, they correspond to different motion states at distinct time points. They thus cannot be treated as identical elements. Therefore, the column of vehicle operating time, $\mathbf{c}_j^{\mathrm{time}}$, is adopted to distinguish the different dynamic descriptors in $\mathbf{c}_j^{\mathrm{dde}}$ because of its uniqueness, and $\mathbf{c}_j^{id}$ specifies the specific entity. The information quantity of $\mathbf{c}_j^{\mathrm{dde}}$, $I(\mathbf{c}_j^{\mathrm{dde}})$, is computed through the conditional entropy of $\mathbf{c}_j^{\mathrm{time}}$ based on $\mathbf{c}_j^{id}$, according to Eq. (11)

$$\begin{aligned} I(\mathbf{c}_j^{\mathrm{dde}}) &= H(\mathbf{c}_j^{\mathrm{time}}|\mathbf{c}_j^{id}) \\ &= -\sum_{d_{ij}^{\mathrm{time}} \in \mathbf{c}_{\mathrm{time}}} \sum_{d_{ij}^{id} \in \mathbf{c}_{id}} p(d_{ij}^{\mathrm{time}}, d_{ij}^{id}) \log_2 p(d_{ij}^{\mathrm{time}}|d_{ij}^{id}) \end{aligned} \tag{11}$$

Data Identifiability

As discussed in the literature review, the evaluation of identifiability should focus on whether the data accurately reflects ICVs' static and dynamic technical features. Therefore, this study employs two distinct methods to estimate data identifiability, measured by the identifiability factor, separately for identifier and descriptor columns.

1. Identifiability factor of identifier elements

   The k-anonymity-based method is adopted to quantify the identifiability factor for identifier columns because of its effectiveness in assessing identity disclosure (Li et al. 2019). The k-anonymity-based method quantifies the identifiability factor by the reciprocal of the number of possible vehicles that share the same identifier value. Specifically, when the number of vehicles can be revealed is k through the value of an element, d_{ij}^{id}, in the column of identifier, its identifiability factor equals $1/k$, as shown in Eq. (12)

Table 2. Examples of measurement of the identifiability factor for identifiers

ID number	Number of possibly identified vehicles (k)	Identifiability factor
Shanghai A *1234	34	1/34

$$IF(d_{ij}^{id}) = \frac{1}{k} \tag{12}$$

Table 2 illustrates a measuring case involving the identifiability factor of a vehicle license plate ID containing a gable character *. According to Ministry of Public Security (2018), a valid vehicle license plate ID consists of numbers from 0 to 9 or English letters other than O or I. In this case, the number of possibly identified vehicles, i.e., $k = 34$; the corresponding identifiability factor should be 1/34.

2. Identifiability factor of descriptor elements

The quantification of descriptor identifiability is grounded in the determination of the damage source responsible for its degradation within the ICV data set, predominantly including the damage of completeness (Keenan Dworak-Fisher et al. 2020), precision (Schlackl et al. 2022), and validity (Shaikh and Sasikumar 2015). Accordingly, the three damage sources, namely, incompleteness, imprecision, and invalidity, are defined, and corresponding judgment criteria and quantification methods are established.

The identifiability factor of a descriptor element d_{ij}^{de}, its identifiability factor with respect to the damage source z, is expressed as $IF_z(d_{ij}^{de})$, where z could be 1, 2, and 3, indicating incompleteness, imprecision, and invalidity, respectively. The overall $IF_z(d_{ij}^{de})$ is calculated through Eq. (13) by comprehensively integrating the identifiability factors associated with all three damage sources

$$IF(d_{ij}^{de}) = IF_1(d_{ij}^{de}) \times IF_2(d_{ij}^{de}) \times IF_3(d_{ij}^{de}) \tag{13}$$

- The damage source of incompleteness

Completeness refers to whether all the data an element should have is entirely recorded (DAMA 2017). One situation of incompleteness is value missing, and Table 3 presents an example. An element is missing in the velocity attribute column.

When the value is missing, its identifiability factor value equals zero, as formulated in Eq. (14)

$$IF_1(d_{ij}^{de}) = \begin{cases} 0 & \text{the issue of incompleteness exist} \\ 1 & \text{the issue of incompleteness does not exist} \end{cases} \tag{14}$$

- The damage source of imprecision

Imprecision mainly occurs in the numerical ICV data. For example, velocity values should be recorded with a precision of at least three decimal places. If this requirement is not met, the corresponding value is considered imprecise, as shown in Table 4.

Table 3. Example of data value missing

Position time	Subject vehicle ID	Velocity (km/h)
11/12/2022 11:15:32	Shanghai *****	56.725
11/12/2022 11:15:33	Shanghai *****	56.732
11/12/2022 11:15:34	Shanghai *****	57.074
11/12/2022 11:15:35	**Shanghai *******	—
11/12/2022 11:15:36	Shanghai *****	58.036
—	—	—

Note: Bold font indicates that the corresponding row contains missing values in the data set.

Table 4. Examples of imprecise data

Requirement of velocity precision	Precise data example	Examples of imprecise data
Accurate to three decimal places	56.725	56.7
—	—	—

Imprecision does not entirely compromise the identifiability of an element but instead leads to partial information loss. Therefore, the identifiability factor of an element under imprecision can be quantified based on the extent of information loss. Information loss reflects the degree to which the element fails to accurately represent the objective facts of entities (Murakami and Uno 2018; Esmeel et al. 2020). The information loss of the element d_{ij}^{de} due to imprecision, denoted as $IL_2(d_{ij}^{de})$, is defined as follows (Xu et al. 2006)

$$IL_2(d_{ij}^{de}) = \frac{|R_{\max}(d_{ij}^{de}) - R_{\min}(d_{ij}^{de})|}{|R_{\max}(\mathbf{c}_j^{de}) - R_{\min}(\mathbf{c}_j^{de})|} \tag{15}$$

where $R_{\max}(\mathbf{c}_j^{de})$ and $R_{\min}(\mathbf{c}_j^{de})$ indicate the theoretical maximum and minimum values of elements in the attribute column $\mathbf{c}_j^{de}$, which is given by CAAM (2021). $R_{\max}(d_{ij}^{de})$ and $R_{\min}(d_{ij}^{de})$ denote the upper and lower bounds of the element, d_{ij}^{de}, in the case where no imprecision is present, i.e., when the value is assumed to be precise, $|R_{\max}(d_{ij}^{de}) - R_{\min}(d_{ij}^{de})|$ is the corresponding possible value range. All of these ranges represent the interval between the maximum and minimum values when the element is numeric.

The value range of $IL_2(d_{ij}^{de})$ is [0, 1]. If the damage source of imprecision does not exist, the element value is the same as the original value without loss, and $IL_2(d_{ij}^{de})$ equals zero. When imprecision occurs, according to Eq. (15), and the theoretical calculation result of $IL_2(d_{ij}^{de})$ falls within the range of (0, 1]. In this situation, the information loss of imprecise data presented in Table 4 is 0.001. This is because the theoretical range of velocity is 120 km/h, while the upper and lower bounds of the imprecise element are 56.749 and 56.650 km/h, respectively, resulting in a possible value range of 0.099 km/h.

Then $IL_2(d_{ij}^{de})$ is transformed into an identifiability factor $IF_2(d_{ij}^{de})$, which is approximated through Eq. (16) (Marés and Torra 2012)

$$IF_2(d_{ij}^{de}) = \begin{cases} 1 - 2 \times IL_2(d_{ij}^{de}) & IL_2(d_{ij}) \in [0, 0.5] \\ 0 & IL_2(d_{ij}) \in (0.5, 1] \end{cases} \tag{16}$$

Subsequently, the identifiability factor of the example in Table 4 is 0.998, according to Eq. (16).

- The damage source of invalidity

Validity indicates whether the element value is compliantly or numerically acceptable. In the context of ICV data sets, compliance and numerical validity can be divided (DAMA 2017).

Compliance validity refers to whether the data falls within the prescribed range. If the value of an element exceeds the compliance range specified in the regulation or standard, the

Table 5. Examples of an object type value not in the valid range

Position time	Subject vehicle ID	Compliance range	Record value
19/10/2022 12:55:50	Shanghai *****	Round numbers in $[1, 17] \cup \{99\}$	3
19/10/2022 12:55:51	Shanghai *****	Round numbers in $[1, 17] \cup \{99\}$	4
19/10/2022 12:55:52	**Shanghai *******	**Round numbers in $[1, 17] \cup \{99\}$**	**19**
19/10/2022 12:55:53	Shanghai *****	Round numbers in $[1, 17] \cup \{99\}$	5
19/10/2022 12:55:54	Shanghai *****	Round numbers in $[1, 17] \cup \{99\}$	1
—	—	—	—

Note: Bold font indicates that the corresponding row is associated with a compliance validity issue.

record is considered compliantly invalid. The compliance range of ICV data can be referred to Ministry of Industry and Information Technology (2016) and CAAM (2021). An example of a lane type value not in the valid range is presented in Table 5.

Numerical validity is also judged, which verifies whether the data value is numerically valid. The numerical validity is generally judged for continuous ICV data, which continuously changes with the dynamic spatial-temporal moving process. A strong correlation exists among some attribute columns (Joshi 1989; Celko 2010). The correlation could be leveraged to judge whether the value of the elements is valid. For an ICV data set, such attributes include ICVs' acceleration, deceleration, velocity, location, etc.

For instance, the velocity of ICV could be derived from location and time instant data, and the acceleration of ICV could be derived from velocity and time instant data. Calculating the velocity and acceleration using the location and velocity data could be used to judge whether the adjacent location and velocity data are reasonable and valid.

Take the acceleration as an example, which could be calculated using Eq. (17) v_i and v_{i+1} are the velocities in record i and $i+1$ in the unit of km/h, respectively. t_i is the time instant in record i, expressed in the unit of seconds. $\bar{a}_{i,i+1}$ is the average acceleration or deceleration (in a unit of m/s^2) in the time interval between t_i and t_{i+1}. As we know, in China, the numerical valid range of $\bar{a}_{i,i+1}$ is [−4.5 m/s^2, 4 m/s^2] (Liu et al. 2020). The velocity values can be judged as numerically invalid if the calculated $\bar{a}_{i,i+1}$ is not in the valid range

$$\bar{a}_{i,i+1} = \frac{(v_{i+1} - v_i)}{3.6 \times (t_{i+1} - t_i)} \tag{17}$$

Taking velocity as an example, it could be calculated through Eq. (18) (Jiménez-Meza et al. 2013)

$$\bar{v}_{i,i+1} = \frac{2 \times \text{Radius}}{(t_{i+1} - t_i)/3600} \times \frac{\arcsin\sqrt{\sin^2\left(\frac{\text{lat}_{i,i+1}}{2}\right) + \cos(\text{lat}_i)\cos(\text{lat}_{i+1})\sin^2\left(\frac{\text{lon}_{i,i+1}}{2}\right)}}{(t_{i+1} - t_i)/3600} \tag{18}$$

where $\bar{v}_{i,i+1}$ is the average velocity in the interval between t_i and t_{i+1}, in a unit of km/h. The radius is the radius of the Earth, taking 6,371 km. lat_i and lon_i represent the latitude and longitude in record i, respectively. $\text{lat}_{i,i+1}$ and $\text{lon}_{i,i+1}$, represent the changes in latitude and longitude between t_i and t_{i+1} in radians, respectively. The valid range in China of $\bar{v}_{i,i+1}$ is [0, 120 km/h]. If the calculated $\bar{v}_{i,i+1}$ is not in the valid range, the location values are then regarded as numerically invalid.

A discrepancy-based numerical invalidity detection approach is proposed to identify further hidden numerical invalidity issues that cannot be detected by the aforementioned direct verification method. More columns of data are involved jointly. Specifically, considering the fine granularity of the ICV data with a recording frequency usually above 20 Hz (Ministry of Industry and Information Technology 2023), the average velocity derived from adjacent velocity data for a small-time interval and that derived from location data can be assumed approximately equal, and the same for the acceleration. In this situation, discrepancies in velocity (DV) and acceleration (DA) are defined and calculated to judge the validity. $\text{DV}_{i,i+1}$ (km/h) is defined as the absolute value of the difference between the average velocity derived from velocity data $(v_{i+1} + v_i)/2$ and $\bar{v}_{i,i+1}$ from location data by Eq. (18), formulated as Eq. (19)

$$\text{DV}_{i,i+1} = \left|\bar{v}_{i,i+1} - \frac{v_{i+1} + v_i}{2}\right| \tag{19}$$

Similarly, $\text{DA}_{i,i+1}$ (m/s^2) is defined as the absolute value of the difference between the average acceleration derived from acceleration data $(a_{i+1} + a_i)/2$ and $\bar{a}_{i,i+1}$ that derived from velocity data by Eq. (17), as given in Eq. (20)

$$\text{DA}_{i,i+1} = \left|\bar{a}_{i,i+1} - \frac{a_{i+1} + a_i}{2}\right| \tag{20}$$

An unsupervised outlier identification algorithm, isolation forest (*i*Forest), is adopted and trained to identify the numerical invalidity through discrepancy-based analyses. *i*Forest can detect anomaly values by randomly partitioning data into isolation trees, where anomalies are isolated with fewer splits, resulting in shorter path lengths (Xu et al. 2023). The training data set should have good qualities, which satisfy that the average DV is less than 1 km/h, and the average DA is less than 0.2 m/s^2. Then the trained *i*Forest algorithm is used to identify the outliers of DV and DA for the ICV data set.

The identified anomaly of $\text{DV}_{i,i+1}$ could be caused by the errors in velocity v_i and location (lat_i and lon_i). The anomaly of $\text{DA}_{i,i+1}$ could arise from the wrong values of velocity v_i and acceleration a_i. Therefore, it can be inferred that v_i is numerically invalid when the anomaly of $\text{DA}_{i,i+1}$ and $\text{DV}_{i,i+1}$ cooccur. Moreover, the error of a_i or lat_i and lon_i can be identified when only $\text{DA}_{i,i+1}$ or $\text{DV}_{i,i+1}$ is anomalistic, respectively.

Either compliance invalidity or numerical invalidity occurs, and the identifiability factor for that element equals zero, as illustrated in Eq. (21)

$$IF_3(d_{ij}^{de}) = \begin{cases} 0 & \text{the issue of invalidity exist} \\ 1 & \text{the issue of invalidity does not exist} \end{cases} \tag{21}$$

3. The identifiability factor of the column

After the identifiability factors of all elements in an attribute column are obtained, the overall $IF(\mathbf{c}_j)$ can be calculated with Eq. (22) (Harel et al. 2012; Du et al. 2022)

$$IF(\mathbf{c}_j) = \frac{1}{n}\sum_{i=1}^{n} IF(d_{ij}) \tag{22}$$

Applications with Empirical ICV Data Sets

The proposed MR-score model, along with the proposed quantification indicators, is newly developed and utilized to assess the misuse risks of ICV data sets. To our knowledge, no comparable study has been conducted on the misuse risk analysis of ICV data sets. To validate the performance of the proposed method, the MR-score model is applied to two real-world ICV data sets to investigate its feasibility and effectiveness.

The two sample ICV data sets were collected in August and September 2022 by an automotive technology innovation enterprise in Shanghai, China. The data sets cover distinct operating conditions of seven and six ICVs, respectively. Examples of these two data sets are presented in Tables 6 and 7, respectively. The elements of some attributes are anonymized by * to safeguard the enterprise's interests. Data set $\mathbf{D}_1$ comprises eight attribute columns and 42,755 records, while data set $\mathbf{D}_2$ comprises eight attribute columns and 41,873 records. Both data sets have the same seven attributes, including position time, vehicle ID, driving mode, velocity, acceleration, latitude, and longitude, with $\mathbf{D}_1$ having vehicle color and $\mathbf{D}_2$ having VIN code, respectively.

Deriving Sensitivity Scores

All the 223 ICV data attributes in CAAM (2021) are grouped into the five SDs. The number of data attributes in each SD is shown in Table 8, along with example attributes. The majority of attributes belong to SD_2, totaling 108, while SD_5 contains only two attributes: altitude and road curve. This is because the two most sensitive attributes can expose critical information about geography and infrastructure, thereby posing risks to national security (Meteriz-Yildiran et al. 2022).

To derive the sensitivity scores, nine senior experts from the data security management department of an automotive technology research and development company were invited to conduct evaluations. Each expert has more than three years of experience in ICV data management. The experts were instructed to assess the relative sensitivity across all SDs by considering the contextual significance of each attribute in practical ICV application scenarios, ensuring that the evaluation results accurately reflect domain-specific sensitivity. Table 9 illustrates the arithmetic mean weight judgment matrix constructed from the expert feedback for the five SDs.

The validation of the judgment matrix is verified by the consistency ratio (CR). The CR of the judgment matrix is 0.038 and less than 0.1, meaning that the consistency of the judgment matrix is acceptable. Moreover, all the experts believe that SD_5 should equal 1. Subsequently, the sensitivity scores of all the SDs can be calculated through Eq. (7) and presented in Table 10. Consequently, the SDs of the nine attributes in the two sample ICV data sets, their sensitivity scores and corresponding data sets are summarized in Table 11.

Measuring Information Quantity

Table 12 lists each attribute column's calculated information entropy values in $\mathbf{D}_1$ and $\mathbf{D}_2$, respectively. Among these attributes, vehicle ID and VIN code serve as identifiers, and the column of vehicle color is categorized as a static descriptor. The columns of dynamic descriptors include position time, velocity, acceleration, latitude, and longitude.

The marginal entropy of identifier and static descriptor columns (i.e., vehicle ID and vehicle colors) in $\mathbf{D}_1$ is 2.922 bits, whereas that of the identifier columns in $\mathbf{D}_2$ (i.e., vehicle ID and VIN code) is 2.477 bits. This difference is primarily due to less distinct ICVs in $\mathbf{D}_2$.

Table 6. Partial excerpt of sample ICV data sets $\mathbf{D}_2$

Position time	Vehicle ID	Driving mode	Vehicle color	Velocity (km/h)	Acceleration (m/s^2)	Latitude	Longitude
20/08/2022 09:15:32	Shanghai ****%	1	Blue	56.73	−0.072	30.******	121.******
20/08/2022 09:15:33	Shanghai ****%	1	Blue	56.73	−0.15	30.******	121.******
20/08/2022 09:15:34	Shanghai ****%	1	Blue	57.074	0.01	30.******	122.******
—	—	—	—	—	—	—	

Table 7. Partial excerpt of sample ICV data sets $\mathbf{D}_2$

Position time	Vehicle ID	VIN code	Driving mode	Velocity (km/h)	Acceleration (m/s^2)	Latitude	Longitude
02/09/2022 10:09:08	Shanghai *****	L**********244	1	17.218	0.001	30.******	121.******
02/09/2022 10:09:09	Shanghai *****	L**********244	1	17.221	0.081	30.******	121.******
02/09/2022 10:09:10	Shanghai *****	L**********244	1	17.512	0.064	30.******	122.******
—	—	—	—	—	—	—	

Table 8. Results of attributes assigned to different SDs

Sensitivity degree	Example attributes	Number of attributes	Total number of attributes
SD_1	Position time, vehicle color, vehicle length, ...	44	223
SD_2	Driving mode, velocity, acceleration, ...	108	
SD_3	Vehicle ID, VIN code,...	53	
SD_4	Latitude, longitude, charging voltage, charging current, ...	16	
SD_5	Altitude, road curve	2	

J. Transp. Eng., Part A: Systems, 2026, 152(2): 04025133

Table 9. Relative weights of sensitivity of different SDs

w_1	w_2	w_3	w_4	w_5
0.038	0.068	0.132	0.304	0.457

Table 10. Sensitivity scores of all the SDs

SD_1	SD_2	SD_3	SD_4	SD_5
0.084	0.149	0.288	0.665	1.000

Table 11. Sensitivity scores of attribute columns in the example ICV data set

Attribute column	SD	Corresponding data set	Sensitivity score
Position time	SD_1	$\mathbf{D}_1$, $\mathbf{D}_2$	0.084
Vehicle ID	SD_3	$\mathbf{D}_1$, $\mathbf{D}_2$	0.288
VIN code	SD_3	$\mathbf{D}_2$	0.288
Driving mode	SD_2	$\mathbf{D}_1$, $\mathbf{D}_2$	0.149
Vehicle color	SD_1	$\mathbf{D}_1$	0.084
Velocity	SD_2	$\mathbf{D}_1$, $\mathbf{D}_2$	0.149
Acceleration	SD_2	$\mathbf{D}_1$, $\mathbf{D}_2$	0.149
Latitude	SD_4	$\mathbf{D}_1$, $\mathbf{D}_2$	0.665
Longitude	SD_4	$\mathbf{D}_1$, $\mathbf{D}_2$	0.665

Table 12. Information entropy of each attribute column in $\mathbf{D}_1$ and $\mathbf{D}_2$

Data sets	Attribute column	Types of data	Calculating scenario	Information entropy (bit)
$\mathbf{D}_1$	Position time	Dynamic descriptors	Scenario 2	12.462
	Vehicle ID	Identifiers	Scenario 1	2.922
	Driving mode	Dynamic descriptors	Scenario 2	12.462
	Vehicle color	Static descriptors	Scenario 1	2.922
	Velocity	Dynamic descriptors	Scenario 2	12.462
	Acceleration	Dynamic descriptors	Scenario 2	12.462
	Latitude	Dynamic descriptors	Scenario 2	12.462
	Longitude	Dynamic descriptors	Scenario 2	12.462
$\mathbf{D}_2$	Position time	Dynamic descriptors	Scenario 2	11.812
	Vehicle ID	Identifiers	Scenario 1	2.447
	VIN code	Identifiers	Scenario 1	2.447
	Driving mode	Dynamic descriptors	Scenario 2	11.812
	Velocity	Dynamic descriptors	Scenario 2	11.812
	Acceleration	Dynamic descriptors	Scenario 2	11.812
	Latitude	Dynamic descriptors	Scenario 2	11.812
	Longitude	Dynamic descriptors	Scenario 2	11.812

The conditional information entropy of all the dynamic descriptor columns is equal, with a value of 12.462 bits in $\mathbf{D}_1$ and 11.812 bits in $\mathbf{D}_2$. The lower conditional entropy in $\mathbf{D}_2$ results from fewer distinct ICVs and a smaller number of records in the data set. Further, the identical conditional entropy values for all dynamic descriptors within a data set indicate that different attribute columns of dynamic descriptors will not differ in data scale during a certain operating period.

Calculating the Identifiability Factor

Identifiability Factor of Identifiers

As illustrated by Table 6, an undesired character % appears in some elements of the vehicle ID column in $\mathbf{D}_1$. According to Eq. (12), the identifiability factor of these elements is $1/34$. Given that there are 8,469 such elements, the overall identifiability factor for the vehicle ID column of $\mathbf{D}_1$ is 0.804 based on Eq. (22) For $\mathbf{D}_2$, all values in the vehicle ID and VIN code columns are complete and undamaged. In this case, each vehicle can be accurately identified through its corresponding vehicle ID and VIN code. Thus, the parameter k in Eq. (12) equals 1. Consequently, the identifiability factors for vehicle ID and VIN code columns in $\mathbf{D}_2$ equal 1, according to Eq. (22).

Identifiability Factor of Descriptors

1. Identifiability factor associated with incompleteness

   No problem of incompleteness is detected in any descriptor of $\mathbf{D}_1$ or $\mathbf{D}_2$, based on the "isnull()" function of the Pandas tool in Python. Therefore, the $IF_1(d_{ij}^{de})$ for all elements across all attribute columns in both data sets is 0.
2. Identifiability factor associated with imprecision

   In the descriptors of both sample data sets, precision loss mainly occurs in the columns of velocity, acceleration, longitude, and latitude. The theoretical ranges of velocity and acceleration are set based on the reasonable vehicle operating conditions concluded by Liu et al. (2020). The theoretical latitude and longitude ranges are from the geographical coordinates of the allowed operating area for ICVs in Shanghai. Moreover, the possible value range is assigned based on the difference between the actual and required numerical precision.

   As both sample data sets exhibit similar problems of imprecision, Table 13 presents the typical data imprecision problems, along with the corresponding theoretical and possible value ranges, calculated information loss, and the resulting identifiability factors. Imprecision in acceleration causes greater information loss and reduced identifiability due to its smaller theoretical range. In contrast, the imprecision in latitude and longitude has a negligible effect on identifiability, with nearly no information loss and identifiability factors approaching 1.
3. Identifiability factor associated with invalidity

   As introduced previously, invalidity is mainly involved in the columns of numerical data, including velocity, acceleration, and longitude and latitude. Table 14 illustrates the compliance and numerical invalidity results for $\mathbf{D}_1$ and $\mathbf{D}_2$, respectively. $\mathbf{D}_2$ exhibits significantly fewer invalidity problems than $\mathbf{D}_1$. Specifically, the discrepancy-based method identifies 9,378

Table 13. Examples of typical problems of data imprecision in the sample data set

Attribute column	Example element value	Causes of precision loss	Theoretical range	Possible value range	$IL_2(d_{ij}^{de})$	$IF_2(d_{ij}^{de})$
Velocity (km/h)	28.3	Lack of precision (three decimal places required)	[0, 120 km/h]	[27.250, 28.349 km/h]	0.001	0.998
Acceleration (m/s^2)	0.1	Lack of precision (two decimal places required)	[−4.5, 4.0 m/s^2]	[0.05, 0.14 m/s^2]	0.011	0.978
Latitude	30.*****°	Lack of precision (six decimal places required)	[30.******°, 31.******°]	[30.****00°, 30.****99°]	0.000	1.000
Longitude	120.*****°	Lack of precision (six decimal places required)	[120. ******°, 122.******°]	[120.****00°, 120.****99°]	0.000	1.000

Table 14. Number of invalid elements in sample data sets through invalidity identification

Data sets	Attribute column	Compliance invalidity	Direct numerical invalidity	Discrepancy-based numerical invalidity
$\mathbf{D}_1$	Velocity	0	1	454
	Acceleration	0	0	3,706
	Latitude	0	16	2,609
	Longitude	0	16	2,609
	Total	0	33	9,378
$\mathbf{D}_2$	Velocity	0	0	1,138
	Acceleration	0	0	90
	Latitude	0	0	389
	Longitude	0	0	389
	Total	0	0	2,006

elements (5.48%) in $\mathbf{D}_1$ and 2,006 elements (1.20%) in $\mathbf{D}_2$ as spatiotemporal motion state mismatches. Moreover, the discrepancy-based method detects substantially more invalid elements than the direct method in both data sets. It demonstrates its effectiveness in capturing subtle value distortions that are otherwise difficult to detect.

Fig. 1 visualizes the mismatches of the spatiotemporal motion states of ICVs in $\mathbf{D}_1$ and $\mathbf{D}_2$ by using the discrepancy-based method. The average DV and DA of all elements in $\mathbf{D}_1$ are 1.50 km/h and 0.24 m/s^2, respectively. As illustrated in Figs. 1(a and b), a considerable number of data points still exhibit relatively large discrepancies. Although the majority of DV and DA values are concentrated below the average thresholds, as shown in the heatmap in Fig. 1(b), these localized deviations contribute to the frequent mismatches of spatiotemporal motion states observed in Fig. 1(a).

In $\mathbf{D}_2$, the average DV and DA of all elements in the comparison data set are 0.28 km/h and 0.08 m/s^2, respectively. Compared to $\mathbf{D}_1$, the points of matching in Fig. 1(c) are visibly more than those in Fig. 1(a) because of the better match of the spatiotemporal motion states reflected by the lower DV and DA. The heat map of DA and DV in Fig. 1(d) also tends to be closer to 0 compared to Fig. 1(b). These findings appear to be consistent with those found in Table 14.

4. Identifiability factor

Tables 15 and 16 exemplify the form fragments of $IF(d_{ij}^{de})$ for the velocity and acceleration columns in $\mathbf{D}_1$ as examples, calculated through Eq. (13) In $\mathbf{D}_1$, the invalidity problem is more prevalent in the acceleration column, while imprecision

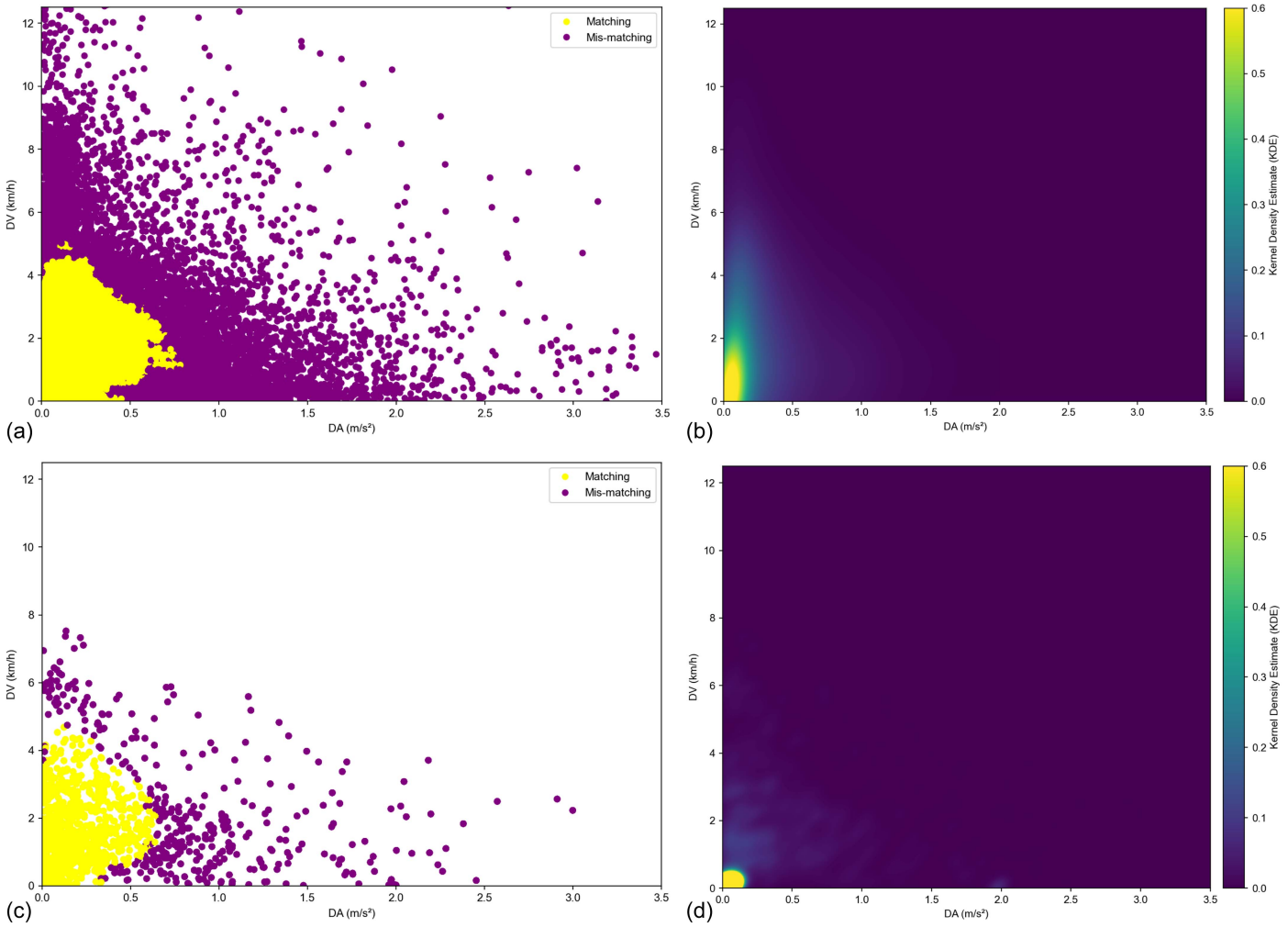

Fig. 1. Spatiotemporal motion state mismatch recognition based on iForest for D1 and D2. (a) Scatter plots of invalid elements identification with $\mathbf{D}_1$; (b) heat map of value distribution of DA and DV with $\mathbf{D}_1$; (c) scatter plots of invalid elements identification with $\mathbf{D}_2$; and (d) heat map of value distribution of DA and DV with $\mathbf{D}_2$.

J. Transp. Eng., Part A: Systems, 2026, 152(2): 04025133

Table 15. $IF(d_{ij}^{de})$ form fragment for the velocity column in $\mathbf{D}_1$

Velocity (km/h)	$IF_1(d_{i5}^{de})$	$IF_2(d_{i5}^{de})$	$IF_3(d_{i5}^{de})$	$IF(d_{i5}^{de})$
—	—	—	—	—
24.9	1.000	0.998	1.000	0.998
27.3	1.000	0.998	1.000	0.998
29.9	1.000	0.998	1.000	0.998
—	—	—	—	—
10.8	1.000	0.998	0.000	0.000
11.7	1.000	0.998	1.000	0.998
15	1.000	0.984	0.000	1.000
17.6	1.000	0.998	1.000	0.998
20.1	1.000	0.998	1.000	0.998
—	—	—	——	—

Table 16. $IF(d_{ij}^{de})$ form fragment for the acceleration column in $\mathbf{D}_1$

Acceleration (m/s^2)	$IF_1(d_{i6}^{de})$	$IF_2(d_{i6}^{de})$	$IF_3(d_{i6}^{de})$	$IF(d_{i6}^{de})$
—	—	—	—	—
−0.059	1.000	1.000	1.000	1.000
0	1.000	0.768	0.000	0.000
−0.125	1.000	0.000	0.000	0.000
—	—	——	—	—
−0.052	1.000	1.000	1.000	1.000
−0.067	1.000	1.000	1.000	1.000
−0.025	1.000	1.000	1.000	1.000
−0.062	1.000	1.000	0.000	0.000
−0.1	1.000	0.978	0.000	0.000
—	—	—	—	—

is more common in the velocity column. The final identifiability factors for all descriptor columns in $\mathbf{D}_1$ and $\mathbf{D}_2$ are summarized in Table 17. The column is more identifiable with a larger identifiability factor. Both data sets exhibit good identifiability, though $\mathbf{D}_2$ outperforms $\mathbf{D}_1$ in most cases.

In more detail, in $\mathbf{D}_1$, the minimum value of the identifiability factor is 0.823 for acceleration, followed by 0.878 (latitude and longitude) and 0.971 (velocity). The corresponding values in $\mathbf{D}_2$ are 0.998, 0.991, and 0.973, respectively. This better performance in $\mathbf{D}_2$ is mainly attributed to the overall higher data precision and validity, as evidenced by Tables 6, 7, and 14.

Calculating MR-Score

According to Eqs. (4) and (5), the column and data set MR-scores for $\mathbf{D}_1$ and $\mathbf{D}_2$ are calculated and provided in Table 18. The MR-scores indicate that the overall estimated misuse risks are 21.874 for $\mathbf{D}_1$ and 23.199 for $\mathbf{D}_2$, suggesting that $\mathbf{D}_2$ has a higher misuse risk. Within each data set, the MR-scores of the identifier and static descriptor columns (vehicle ID, VIN code, and vehicle color) are lower than those of the dynamic descriptor columns, mainly due to their much lower information quantities. The MR-scores of the position time columns (1.047 for $\mathbf{D}_1$ and 0.992 for $\mathbf{D}_2$) are also lower than those of the other dynamic descriptor columns due to their lowest sensitivity scores (0.084). Moreover, the vehicle color column in $\mathbf{D}_1$ shows the lowest MR-score among all columns in the

Table 18. Final assessment results of MR-scores for $\mathbf{D}_1$ and $\mathbf{D}_2$

Data sets	Attribute column	$S(\mathbf{c}_j)$	$\mathrm{I}(\mathbf{c}_j)$	IF $(\mathbf{c}_j)$	MR-score $(\mathbf{c}_j)$	MR-score $(\mathbf{D})$
$\mathbf{D}_1$	Position time	0.084	12.462	1	1.047	21.709
	Vehicle ID	0.288	2.922	0.804	0.677	
	Driving mode	0.149	12.462	1	1.857	
	Vehicle color	0.084	2.922	1	0.245	
	Velocity	0.149	12.462	0.971	1.803	
	Acceleration	0.149	12.462	0.823	1.528	
	Latitude	0.665	12.462	0.878	7.276	
	Longitude	0.665	12.462	0.878	7.276	
$\mathbf{D}_2$	Position time	0.084	11.812	1	0.992	23.199
	Vehicle ID	0.288	2.447	1	0.705	
	VIN code	0.288	2.447	1	0.705	
	Driving mode	0.149	11.812	1	1.760	
	Velocity	0.149	11.812	0.973	1.712	
	Acceleration	0.149	11.812	0.998	1.756	
	Latitude	0.665	11.812	0.991	7.784	
	Longitude	0.665	11.812	0.991	7.784	

two data sets, resulting from its lowest sensitivity score (0.084) and its lowest information quantity (2.922 bits).

Compared to $\mathbf{D}_1$, the higher misuse risk estimated for $\mathbf{D}_2$ is attributed to several aspects. First, the sensitivity score of the VIN code (a different attribute from $\mathbf{D}_1$) in $\mathbf{D}_2$ (0.288) is higher than that of the vehicle color (a different attribute from $\mathbf{D}_2$) in $\mathbf{D}_1$ (0.084). Second, although $\mathbf{D}_2$ shows slightly lower information quantities in most columns, it is offset by the higher identifiability. In particular, the identifiability factors of the latitude and longitude columns increase from 0.878 in $\mathbf{D}_1$ to 0.991 in $\mathbf{D}_2$. Since these two attributes have the highest sensitivity score (0.665), the higher performance in identifiability leads to a substantial increase in the MR-scores from 7.276 to 7.784. A similar situation is observed in the vehicle ID column. Although $\mathbf{D}_1$ contains more distinct ICVs than $\mathbf{D}_2$ (seven versus six), resulting in a higher information quantity for the vehicle ID column (2.922 bits versus 2.447 bits), the higher identifiability in $\mathbf{D}_2$ (1.0 versus 0.804) leads to a slightly higher MR-score for the vehicle ID column in $\mathbf{D}_2$ (0.705 versus 0.677).

In general, the MR-score demonstrates a strong capability for assessing and comparing the misuse risks of ICV data sets. It could capture the contributions of data scale, data sensitivity, and data identifiability to the overall misuse risk, which could facilitate the data owners in managing the data sets.

Discussions

A notable feature of the proposed MR-score model is the use of information quantity rather than the number of records to measure data scale. Fig. 2 compares the MR-scores of a single attribute column with identical identifiability but varying sensitivity scores under two different scale indicators: number of records; and information quantity. As shown in Fig. 2(a), when the number of records is used, the MR-score increases dramatically as the data set expands. This contradicts the characteristic that misuse risk diminishes marginally as the data magnitude increases (Algarni et al. 2021). In contrast, Fig. 2(b)

Table 17. Identifiability factors of all columns of descriptors

Data sets	Position time	Driving mode	Vehicle color	Velocity (km/h)	Acceleration (m/s^2)	Latitude	Longitude
$\mathbf{D}_1$	1.000	1.000	1.000	0.971	0.823	0.878	0.878
$\mathbf{D}_2$	1.000	1.000	No exist	0.973	0.998	0.991	0.991

J. Transp. Eng., Part A: Systems, 2026, 152(2): 04025133

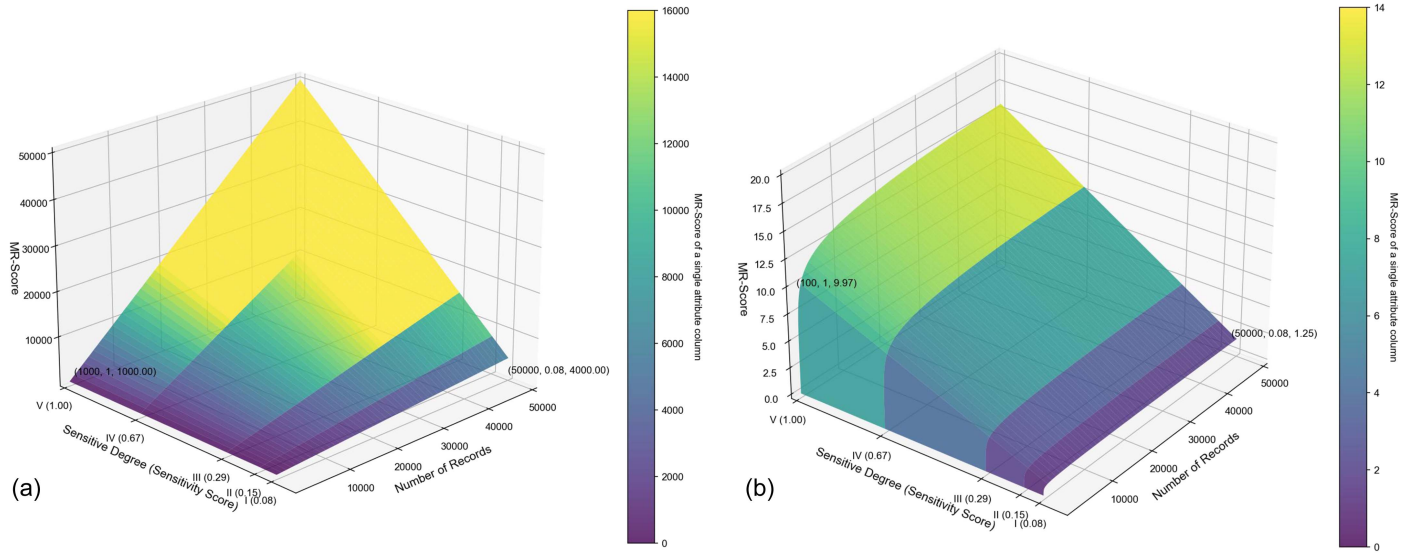

Fig. 2. MR-scores for a single attribute column using different data scale measurements: (a) using the number of records; and (b) using information quantity.

Table 19. MR-scores for a single attribute column based on the number of records and information quantity with the identifiability factors held constant

Sensitivity scores	Number of records	Information quantity	$IF(\mathbf{c}_j)$	MR-score based on the number of records	MR-score based on the information quantity
0.084	200,000	17.610	1	16,000	1.479
0.084	100,000	16.610	1	8,000	1.395
0.149	100,000	16.610	1	15,000	2.475
0.149	50,000	15.610	1	7,500	2.326
0.288	50,000	15.610	1	14,500	4.496
0.288	30,000	14.873	1	8,700	4.283
0.665	30,000	14.873	1	20,100	9.890
0.665	15,000	13.873	1	10,050	9.225
1.000	15,000	13.873	1	15,000	13.873
1.000	8,000	12.966	1	8,000	12.966
1.000	1,000	9.966	1	1,000	9.966
1.000	300	8.229	1	300	8.229
1.000	50	5.644	1	50	5.644

Table 20. MR-Scores for a single attribute column based on the number of records and information quantity with the sensitivity scores held constant

Sensitivity scores	Number of records	Information quantity	$IF(\mathbf{c}_j)$	MR-score based on the number of records	MR-score based on the information quantity
0.149	500,000	18.932	0.15	11,175	0.423
0.149	200,000	17.610	0.15	4,470	0.394
0.149	200,000	17.610	0.35	10,430	0.918
0.149	100,000	16.610	0.35	5,215	0.866
0.149	100,000	16.610	0.65	9,685	1.609
0.149	50,000	15.610	0.65	4,842.5	1.512
0.149	50,000	15.610	0.85	6,332.5	1.977
0.149	30,000	14.873	0.85	3,799.5	1.884
0.149	30,000	14.873	1	4,470	2.216
0.149	15,000	13.873	1	2,235	2.067
0.149	10,000	13.288	1	1,490	1.980
0.149	3,000	11.551	1	447	1.721
0.149	500	8.966	1	74.5	1.336

demonstrates that adopting information quantity captures this marginally diminishing trend in misuse risk more appropriately as the data set enlarges.

Further comparisons with detailed numerical results are shown in Tables 19 and 20, where MR-scores are calculated under varying sensitivity and identifiability conditions using both scale indicators, respectively. In all scenarios, the elements in the attribute column are assumed to be mutually distinct. Initially, all the elements are assumed to be thoroughly identifiable, i.e., $IF(\mathbf{c}_j) = 1$. Under this condition, MR-scores are computed across a range of sensitivity scores using both scale indicators.

Subsequently, the sensitivity score is fixed at 0.149. This value corresponds to SD_2, the most frequently observed degree among ICV data attributes, as shown in Table 8. With fixed sensitivity, MR-scores are then computed across a series of assumed identifiability factor values and different data scales, again using both indicators.

The results in Tables 19 and 20 further demonstrate that, when the number of records measures the data scale, even data with very low sensitivity or identifiability can produce unreasonably high MR-scores if the data scale is large and vice versa. Such outcomes may lead to misleading assessments.

For instance, highly sensitive attributes such as altitude (SD_5, sensitivity score = 1.000) can present substantial misuse risk even at a small scale. In contrast, attributes with sensitivity scores of 0.084, such as vehicle color, should contribute less to the MR-score value regardless of the record number because of their lowest sensitivity degree. Besides, ICV data sets contain many continuous attributes, such as velocity and acceleration. When identifiability is low, such attributes fail to represent actual vehicle dynamics or technical characteristics. Even large volumes of such data pose limited misuse risk because critical information cannot be recognized.

However, as shown in Tables 19 and 20, these expected patterns break down when using the number of records as the scale indicator. The MR-score becomes overly dependent on the data scale, diminishing the contributions of sensitivity and identifiability in misuse risk assessment. This deviates from the theoretical rationale of balanced misuse risk assessment.

In summary, the proposed MR-score model presents a theoretically robust and interpretable framework for quantifying ICV data misuse risk. By leveraging information quantity to represent data

scale, the model ensures that sensitivity, identifiability, and scale contribute proportionally to the overall assessment.

The model has potential applications in identifying high-risk data sets and guiding data protection strategies. A data set with a higher MR-score requires more stringent protection measures to prevent its misuse risk. Risk-based curation of ICV data, where sensitive or high-risk information is selectively withheld, poses profound ethical dilemmas. While such measures mitigate privacy infringements (e.g., geotracking prevention) and cybersecurity threats (e.g., autonomous control hacks), they unintentionally exacerbate systemic inequities. For instance, withholding real-time sensor data may compromise collision-avoidance systems, affecting vulnerable road users, such as pedestrians in low-income regions with non-well-developed infrastructures. Moreover, overly restrictive policies impede independent AI safety audits, eroding public trust in autonomous systems. Balancing risk mitigation with societal welfare necessitates participatory governance, integrating civil society in data-sharing frameworks to uphold equity and public safety imperatives.

Conclusions

This study proposes and develops an MR-score model to quantify ICV data misuse risks, employing three key metrics: sensitivity score; information quantity; and identifiability factor to assess data sensitivity, scale, and identifiability, respectively. It also explores the sensitivity scoring mechanism and proposes a discrepancy-based method for detecting data invalidity specific to ICV data sets. The key findings include:

- The MR-score model effectively quantifies the ICV data set misuse risks and enables the disentanglement of contributions from sensitivity, scale, and identifiability.
- Information quantity outperforms the number of records as an indicator of data scale, better capturing the diminishing marginal impact of data scale on misuse risks.
- The damage source determination-based method provides a valid and robust approach for assessing the identifiability of dynamic descriptor columns within ICV data sets.
- The discrepancy-based method excels at identifying numerical invalidities overlooked by traditional direct detection and compliance checking methods.

These results offer practical guidance for ICV data owners and regulators to safeguard data assets, including identifying critical data sets, formulating protection strategies, establishing storage/sharing protocols, and implementing access controls.

However, due to limitations in actual data acquisition, only nine column attributes are presented in the case study. Follow-up research can further verify the scientific validity of this work once real data sets with more attributes become available. Moreover, only the properties significantly emphasized in previous misuse-related research (i.e., scale, sensitivity, and identifiability) are adopted in this paper. Additional relevant properties, such as data freshness, might also be incorporated in the future research. Last, it is eager to develop misuse risk classification methods based on MR-score values to facilitate the data management and to secure ICV data-sharing.

Data Availability Statement

Some or all data, models, or codes that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (NSFC, Grant No. 52372339), QuGang ({ZF} 2024-013). The contents of this paper reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein.

Author Contributions

Yi Lu: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Validation; Writing – original draft; Writing – review and editing. Hao Li: Formal analysis; Methodology; Resources; Supervision; Validation; Writing – review and editing. Huizhao Tu: Methodology; Resources; Supervision; Writing – review and editing. Jian Liu: Funding acquisition; Project administration. Yufei Yuan: Writing – review and editing. Hans van Lint: Writing – review and editing.

References

Ahmad, M. B., A. Akram, and M. Asif. 2014. "Towards a realistic risk assessment methodology for insider threats of information misuse." In *Proc., 12th Int. Conf. on Frontiers of Information Technology*, 176–181. New York: IEEE.

Akanfe, O., R. Valecha, and H. R. Rao. 2020. "Assessing country-level privacy risk for digital payment systems." *Comput. Secur.* 99 (Dec): 102065. https://doi.org/10.1016/j.cose.2020.102065.

Algarni, A. M., V. Thayananthan, and Y. K. Malaiya. 2021. "Quantitative assessment of cybersecurity risks for mitigating data breaches in business systems." *Appl. Sci.* 11 (8): 3678. https://doi.org/10.3390/app11083678.

Alkhalil, Z., C. Hewage, L. Nawaf, and I. Khan. 2021. "Phishing attacks: A recent comprehensive study and a new anatomy." *Front. Comput. Sci.* 3 (Mar): 563060. https://doi.org/10.3389/fcomp.2021.563060.

Almaskati, D., S. Kermanshachi, and A. Pamidimukkala. 2024. "Convergence of emerging transportation trends: A comprehensive review of shared autonomous vehicles." *J. Intell. Connected Veh.* 7 (3): 177–189. https://doi.org/10.26599/JICV.2023.9210043.

Attaallah, A., H. Alsuhabi, S. Shukla, R. Kumar, B. K. Gupta, and R. A. Khan. 2022. "Analyzing the big data security through a unified decision-making approach." *Intell. Autom. Soft Comput.* 32 (2): 1071–1088. https://doi.org/10.32604/iasc.2022.022569.

Borek, A., A. K. Parlikad, J. Webb, and P. Woodall. 2013. *Total information risk management: Maximizing the value of data and information assets*. Waltham, MA: Newnes.

Bos, J. 2020. *Research ethics for students in the social sciences*. Cham, Switzerland: Springer.

CAAM (China Association of Automobile Manufacturers). 2021. *Data format and definition of intelligent connected vehicles*, 34. Beijing: CAAM.

Celko, J. 2010. "Temporal data." Chap. 41 in *Joe Celko's data, measurements and standards in SQL*, edited by J. Celko. Boston: Morgan Kaufmann.

CEN (European Committee for Standardization). 2016. *The general data protection regulation*. EU 2016/679. Brussels, Belgium: CEN.

Chah, B., A. Lombard, A. Bkakria, R. Yaich, A. Abbas-Turki, and S. Galland. 2022. "Privacy threat analysis for connected and autonomous vehicles." *Procedia Comput. Sci.* 210 (Jan): 36–44. https://doi.org/10.1016/j.procs.2022.10.117.

Chen, H., M. Cai, and X. Chen. 2023. "Privacy protection method for cellular signaling data based on genetic algorithm." *J. Transp. Eng. Part A. Syst.* 149 (4): 04023016. https://doi.org/10.1061/JTEPBS.TEENG-7129.

Chua, H. N., J. S. Ooi, and A. Herbland. 2021. "The effects of different personal data categories on information privacy concern and disclosure." *Comput. Secur.* 110 (Nov): 102453. https://doi.org/10.1016/j.cose.2021.102453.

Corallo, A., M. Lazoi, and M. Lezzi. 2020. "Cybersecurity in the context of Industry 4.0: A structured classification of critical assets and business

impacts." *Comput. Ind.* 114 (Jan): 103165. https://doi.org/10.1016/j.compind.2019.103165.
DAMA (Data Administration Management Association). 2017. *Dama-dmbok: Data management body of knowledge*. 2nd ed. Basking Ridge, NJ: Technics Publications.
Di Vimercati, S. D. C., S. Foresti, G. Livraga, and P. Samarati. 2023. "*k*-anonymity: From theory to applications." *Trans. Data Privacy* 16 (1): 25–49.
Dolzhenkova, E., D. Mokhorov, and T. Baranova. 2020. "National and international issues of cyber security." *IOP Conf. Ser.: Mater. Sci. Eng.* 940 (1): 012015. https://doi.org/10.1088/1757-899X/940/1/012015.
Du, Y., Y. Shi, C. Zhao, Z. Du, and Y. Ji. 2022. "A lifelong framework for data quality monitoring of roadside sensors in cooperative vehicle-infrastructure systems." *Comput. Electr. Eng.* 100 (May): 108030. https://doi.org/10.1016/j.compeleceng.2022.108030.
Elrose, F., I. Lewis, H. Hassan, and C. Murray. 2022. "Insights into the effectiveness of messaging promoting intentions to use connected vehicle technology." *Transp. Res. Part F Psychol. Behav.* 88 (Jul): 155–167. https://doi.org/10.1016/j.trf.2022.05.018.
Eng, K., and E. Stroulia. 2021. "The *tkl*-score for data-sharing misuseability." In *Proc., IFIP Annual Conf. on Data and Applications Security and Privacy*, 312–324. Cham, Switzerland: Springer.
Esmeel, T. K., M. M. Hasan, M. N. Kabir, and A. Firdaus. 2020. "Balancing data utility versus information loss in data-privacy protection using k-anonymity." In *Proc., IEEE 8th Conf. on Systems, Process and Control (ICSPC)*, 158–161. New York: IEEE.
Ethicist, P. 2015. "Simplifying the complexity of confidentiality in research." *J. Empirical Res. Hum. Res. Ethics* 10 (1): 100–102. https://doi.org/10.1177/1556264614568783.
Fei, Y., P. Shi, Y. Liu, and L. Wang. 2024. "Critical roles of control engineering in the development of intelligent and connected vehicles." *J. Intell. Connected Veh.* 7 (2): 79–85. https://doi.org/10.26599/JICV.2023.9210040.
Furnell, S., H. Heyburn, A. Whitehead, and J. N. Shah. 2020. "Understanding the full cost of cyber security breaches." *Comput. Fraud Secur.* 2020 (12): 6–12. https://doi.org/10.1016/S1361-3723(20)30127-5.
Gao, J., C. Hu, L. Wang, and N. Ding. 2024. "Data validity analysis based on reinforcement learning for mixed types of anomalies coexistence in intelligent connected vehicle (ICV)." *Electronics* 13 (2): 444. https://doi.org/10.3390/electronics13020444.
Gómez Losada, A. 2017. *Data science applications to connected vehicles*. Seville, Spain: European Commission. https://doi.org/10.2760/822136.
Gozhyj, A., I. Kalinina, V. Vysotska, S. Sachenko, and R. Kovalchuk. 2020. "Qualitative and quantitative characteristics analysis for information security risk assessment in E-commerce systems." In *Proc., Information-Communication Technologies and Embedded System*, 177–190. Mykolaiv, Ukraine: CEUR Workshop Proceedings.
Guo, J., L. Qi, and J. Suo. 2021. "Research on data classification of intelligent connected vehicles based on scenarios." In *Proc., Int. Conf. on E-Commerce and E-Management (ICECEM)*, 153–158. New York: IEEE.
Gupta, I., and A. K. Singh. 2019. "Dynamic threshold based information leaker identification scheme." *Inf. Process. Lett.* 147 (Jul): 69–73. https://doi.org/10.1016/j.ipl.2019.03.005.
Gupta, I., A. K. Singh, C.-N. Lee, and R. Buyya. 2022. "Secure data storage and sharing techniques for data protection in cloud environments: A systematic review, analysis, and future directions." *IEEE Access* 10 (Jul): 71247–71277. https://doi.org/10.1109/ACCESS.2022.3188110.
Haley, T. D. 2020. "Data protection in disarray." *Wash. Law Rev.* 95 (3): 1193.
Harel, A., A. Shabtai, L. Rokach, and Y. Elovici. 2012. "M-score: A misuseability weight measure." *IEEE Trans. Dependable Secure Comput.* 9 (3): 414–428. https://doi.org/10.1109/TDSC.2012.17.
Haycock, B., J. Campos, N. Koenraad, M. Potter, and S. Advani. 2019. "Creating headlight glare in a driving simulator." *Transp. Res. Part F Psychol. Behav.* 61 (Feb): 93–106. https://doi.org/10.1016/j.trf.2017.10.006.
He, Y.-L., G.-L. Ou, P. Fournier-viger, J. Z. Huang, and P. N. Suganthan. 2022. "A novel dependency-oriented mixed-attribute data classification method." *Expert Syst. Appl.* 199 (Aug): 116782. https://doi.org/10.1016/j.eswa.2022.116782.
Hilbert, M. 2015. "A review of large-scale 'how much information' inventories: Variations, achievements and challenges." *Inf. Res.* 20 (4): 20–24.
Hysa, X., M. D'arco, and J. Kostaqi. 2023. "Misuse of personal data: Exploring the privacy paradox in the age of big data analytics." In *Big data and decision-making: Applications and uses in the public and private sector.* Bingley, UK: Emerald Publishing.
Jain, A. K., S. R. Sahoo, and J. Kaubiyal. 2021. "Online social networks security and privacy: Comprehensive review and analysis." *Complex Intell. Syst.* 7 (5): 2157–2177. https://doi.org/10.1007/s40747-021-00409-7.
Jiménez-Meza, A., J. Arámburo-Lizárraga, and E. De la Fuente. 2013. "Framework for estimating travel time, distance, speed, and street segment level of service (LOS), based on GPS data." *Procedia Technol.* 7 (Jan): 61–70. https://doi.org/10.1016/j.protcy.2013.04.008.
Jin, P., S. Parker, J. Fang, B. Ran, and C. M. Walton. 2012. "Freeway recurrent bottleneck identification algorithms considering detector data quality issues." *J. Transp. Eng.* 138 (10): 1205–1214. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000424.
Jin, Z., H. Mao, D. Chen, H. Li, H. Tu, Y. Yang, and M. Attard. 2024. "Multi-objective optimization model of autonomous minibus considering passenger arrival reliability and travel risk." *Commun. Transp. Res.* 4 (Dec): 100152. https://doi.org/10.1016/j.commtr.2024.100152.
Joshi, K. D. 1989. *Foundations of discrete mathematics*. New Delhi, India: New Age International.
Jung, K. 2021. "Extreme data breach losses: An alternative approach to estimating probable maximum loss for data breach risk." *North Am. Actuarial J.* 25 (4): 580–603. https://doi.org/10.1080/10920277.2021.1919145.
Keenan Dworak-Fisher, L. M., D. P. Jennifer, P. John, P. Mark, R. S. Rolf, M. Marilyn, and L. J. Y. Seastrom. 2020. *A framework for data quality.* Washington, DC: Federal Committee on Statistical Methodology.
Kuzminykh, I., B. Ghita, V. Sokolov, and T. Bakhshi. 2021. "Information security risk assessment." *Encyclopedia* 1 (3): 602–617. https://doi.org/10.3390/encyclopedia1030050.
Landoll, D. 2021. *The security risk assessment handbook: A complete guide for performing security risk assessments*. Boca Raton, FL: CRC Press.
Laurie, G., L. Stevens, C. Dobbs, and K. H. Jones. 2017. "Risks of harm from data misuse." Accessed June 30, 2014. https://wellcome.figshare.com/articles/journal_contribution/Risks_of_harm_from_data_misuse/5613187/1?file=9770317.
Lee, M.-C. 2014. "Information security risk analysis methods and research trends: AHP and fuzzy comprehensive method." *Int. J. Comput. Sci. Inf. Technol.* 6 (1): 29. https://doi.org/10.5121/ijcsit.2014.6103.
Li, F., K. K. Phoon, X. Du, and M. Zhang. 2013. "Improved AHP method and its application in risk identification." *J. Constr. Eng. Manage.* 139 (3): 312–320. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000605.
Li, Q., A. G. D'souza, C. Schmit, and H.-C. Kum. 2019. "Increasing transparent and accountable use of data by quantifying the actual privacy risk in interactive record linkage." Preprint, submitted June 7, 2019. https://arxiv.org/abs/1906.03345.
Li, X., J. Yao, X. Liu, and H. Guan. 2017. "A first look at information entropy-based data pricing." In *Proc., IEEE 37th Int. Conf. on Distributed Computing Systems (ICDCS)*, 2053–2060. New York: IEEE.
Li, Y., J. Yang, and J. Wen. 2021. "Entropy-based redundancy analysis and information screening." *Digital Commun. Networks* 9 (5): 2352–8648. https://doi.org/10.1016/j.dcan.2021.12.001.
Liu, Y., Z. X. Wu, H. Zhou, H. Zheng, N. Yu, X. P. An, J. Y. Li, and M. L. Li. 2020. "Development of China light-duty vehicle test cycle." *Int. J. Automot. Technol.* 21 (5): 1233–1246. https://doi.org/10.1007/s12239-020-0117-5.
Lu, X., and W. Song. 2024. "Efficient multi-source anonymity for aggregated internet of vehicles datasets." *Appl. Sci.* 14 (8): 3230. https://doi.org/10.3390/app14083230.
Ma, J., C. Roncoli, G. Ren, Y. Yang, Q. Cao, Y. Deng, and J. Li. 2025. "Vehicle trajectory reconstruction from sparse data using a hybrid approach." *J. Transp. Eng. Part A. Syst.* 151 (2): 04024108. https://doi.org/10.1061/JTEPBS.TEENG-8569.

Marés, J., and V. Torra. 2012. "An evolutionary optimization approach for categorical data protection." In *Proc., 15th Int. Conf. on Database Theory*, 148–157. New York: Association for Computing Machinery.
Martin, K. D., A. Borah, and R. W. Palmatier. 2017. "Data privacy: Effects on customer and firm performance." *J. Mark.* 81 (1): 36–58. https://doi.org/10.1509/jm.15.0497.
Matin, A., and H. Dia. 2024. "Public perception of connected and automated vehicles: Benefits, concerns, and barriers from an Australian perspective." *J. Intell. Connected Veh.* 7 (2): 108–128. https://doi.org/10.26599/JICV.2023.9210028.
Mazilu, M. 2020. *Schema mapping generation for autonomous data sources*. Manchester, UK: Univ. of Manchester.
Meteriz-Yildiran, U., N. F. Yildiran, J. Kim, and D. Mohaisen. 2022. "Learning location from shared elevation profiles in fitness apps: A privacy perspective." *IEEE Trans. Mob. Comput.* 23 (1): 581–596. https://doi.org/10.1109/TMC.2022.3218148.
Miller, H. E. 2013. "Big-data in cloud computing: A taxonomy of risks." *Inf. Res.* 18 (1): 1.
Ministry of Industry and Information Technology. 2016. *Technical specifications of remote service and management system for electric vehicles—Part 3: Communication protocol and data format*. GB/T 32960.3-2016. Beijing: Standardization Administration of China.
Ministry of Industry and Information Technology. 2023. *Intelligent and connected vehicle-data storage system for automated driving*. GB 20214420-Q-339. Beijing: Standardization Administration of China.
Ministry of Public Security. 2018. *License plates of motor vehicles of the People's Republic of China*. GA 36-2018. Beijing: Ministry of Public Security of China.
Ministry of Transport. 2023. *Security classification and protection requirements for transportation data*. JT/T 1522-2024. Beijing: Ministry of Transport of China.
Mlada, M., R. Holý, J. Jirovský, and T. Kasalický. 2022. "Protection of personal data in autonomous vehicles and its data categorization." In *Proc., Smart City Symp. Prague (SCSP)*, 1–5. New York: IEEE.
Murakami, K., and T. Uno. 2018. "Optimization algorithm for *k*-anonymization of datasets with low information loss." *Int. J. Inf. Secur.* 17 (6): 631–644. https://doi.org/10.1007/s10207-017-0392-y.
Nassaji, H. 2015. "Qualitative and descriptive research: Data type versus data analysis." *Lang. Teach. Res.* 19 (2): 129–132. https://doi.org/10.1177/1362168815572747.
Nayak, S. K., and A. C. Ojha. 2020. "Data leakage detection and prevention: Review and research directions." In *Proc., Machine Learning and Information Processing: Proc. of ICMLIP 2019*, 203–212. Singapore: Springer.
Notario, N., A. Crespo, Y.-S. Martín, J. M. Del Alamo, D. Le Métayer, T. Antignac, A. Kung, I. Kroener, and D. Wright. 2015. "PRIPARE: Integrating privacy best practices into a privacy engineering methodology." In *Proc., IEEE Security and Privacy Workshops*, 151–158. New York: IEEE.
Nugraha, Y., and A. Martin. 2022. "Cybersecurity service level agreements: Understanding government data confidentiality requirements." *J. Cybersecur.* 8 (1): tyac004. https://doi.org/10.1093/cybsec/tyac004.
Ormerod, P. C. 2019. "A private enforcement remedy for information misuse." *Boston Coll. Law Rev.* 60 (7): 1893.
Park, Y., W. Teiken, J. R. Rao, and S. N. Chari. 2016. "Data classification and sensitivity estimation for critical asset discovery." *IBM J. Res. Dev.* 60 (4): 1–2. https://doi.org/10.1147/JRD.2016.2557638.
Rannenberg, K. 2016. "Opportunities and risks associated with collecting and making usable additional data." In *Autonomous driving: Technical, legal and social aspects*. Berlin: Springer.
Reyes, J., W. Fuertes, P. Arévalo, and M. Macas. 2022. "An environment-specific prioritization model for information-security vulnerabilities based on risk factor analysis." *Electronics* 11 (9): 1334. https://doi.org/10.3390/electronics11091334.
Richardson, J. K., and B. L. Smith. 2015. "Application of maximum entropy sampling design to traveler information system data-quality evaluations." *J. Transp. Eng.* 141 (7): 04015006. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000765.
Rios, E., A. Rego, E. Iturbe, M. Higuero, and X. Larrucea. 2020. "Continuous quantitative risk management in smart grids using attack defense trees." *Sensors* 20 (16): 4404. https://doi.org/10.3390/s20164404.
Schlackl, F., N. Link, and H. Hoehle. 2022. "Antecedents and consequences of data breaches: A systematic review." *Inf. Manage.* 59 (4): 103638. https://doi.org/10.1016/j.im.2022.103638.
Scholtes, M., L. Westhofen, L. R. Turner, K. Lotto, M. Schuldes, H. Weber, N. Wagener, C. Neurohr, M. H. Bollmann, and F. Körtke. 2021. "6-layer model for a structured description and categorization of urban traffic and environment." *IEEE Access* 9 (Apr): 59131–59147. https://doi.org/10.1109/ACCESS.2021.3072739.
Shaikh, F. A., and M. Siponen. 2023. "Information security risk assessments following cybersecurity breaches: The mediating role of top management attention to cybersecurity." *Comput. Secur.* 124 (Jan): 102974. https://doi.org/10.1016/j.cose.2022.102974.
Shaikh, R., and M. Sasikumar. 2015. "Data classification for achieving security in cloud computing." *Procedia Comput. Sci.* 45 (Jan): 493–498. https://doi.org/10.1016/j.procs.2015.03.087.
Shameli-Sendi, A., R. Aghababaei-Barzegar, and M. Cheriet. 2016. "Taxonomy of information security risk assessment (ISRA)." *Comput. Secur.* 57 (Mar): 14–30. https://doi.org/10.1016/j.cose.2015.11.001.
Shannon, C. E. 1948. "A mathematical theory of communication." *Bell Syst. Tech. J.* 27 (3): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.
Shen, Y., B. Guo, Y. Shen, X. Duan, X. Dong, and H. Zhang. 2019. "Pricing personal data based on information entropy." In *Proc., 2nd Int. Conf. on Software Engineering and Information Management*, 143–146. New York: Association for Computing Machinery.
Solove, D. J., and D. K. Citron. 2018. "Risk and anxiety: A theory of data-breach harms." *Tex. Law Rev.* 96 (4): 737–786.
Soomro, I., and N. Ahmed. 2013. "Towards security risk-oriented misuse cases." In *Proc., Int. Conf. on Business Process Management*, 689–700. Berlin: Springer.
Soussan, T., and M. Trovati. 2022. "Social media data misuse." In *Proc., Int. Conf. on Intelligent Networking and Collaborative Systems*, 183–189. Cham, Switzerland: Springer.
Sriramoju, S. B., A. C. Naik, and N. S. S. Rao. 2014. "Predicting the misusability of data from malicious insiders." *Int. J. Comput. Eng. Appl.* 5 (1): 46–51.
Stock, K., and H. Guesgen. 2016. "Geospatial reasoning with open data." In *Automating open source intelligence*. Amsterdam, Netherlands: Elsevier.
Sudhakar, R. V., and T. C. M. Rao. 2020. "Security aware index based quasi–identifier approach for privacy preservation of data sets for cloud applications." *Cluster Comput.* 23 (4): 2579–2589. https://doi.org/10.1007/s10586-019-03028-7.
Tu, H., M. Attard, Y. Yang, K. Scerri, A. Muscat, and H. Li. 2024. "On-demand automated bus services: Opportunities and challenges." *Commun. Transp. Res.* 4 (Dec): 100134. https://doi.org/10.1016/j.commtr.2024.100134.
Usmonov, M. 2024. "Basic concepts of information security." *Indexing* 1 (1): 81–85. https://doi.org/10.1007/978-3-662-70639-8_7.
Vaniš, M., T. Zelinka, T. Ščerba, and A. Stárková. 2022. "Classification of non-personal data in autonomous vehicles." In *Proc., Smart City Symp. Prague (SCSP)*, 1–6. New York: IEEE.
Vartanian, A., and A. Shabtai. 2014. "TM-score: A misuseability weight measure for textual content." *IEEE Trans. Inf. Forensics Secur.* 9 (12): 2205–2219. https://doi.org/10.1109/TIFS.2014.2359370.
Vatanparast, R. 2020. "Data governance and the elasticity of sovereignty." *Brooklyn J. Int. Law* 46 (1): 1. https://doi.org/10.2139/ssrn.3609579.
Vavilis, S., M. Petković, and N. Zannone. 2014. "Data leakage quantification." In *Proc., IFIP Annual Conf. on Data and Applications Security and Privacy*, 98–113. Berlin: Springer.
Vavilis, S., M. Petković, and N. Zannone. 2016. "A severity-based quantification of data leakages in database systems." *J. Comput. Secur.* 24 (3): 321–345. https://doi.org/10.3233/JCS-160543.
Vemou, K., and M. Karyda. 2018a. "An evaluation framework for privacy impact assessment methods." In *Proc., Mediterranean Conf. on Information Systems (MCIS 2018)*. Atlanta: Association for Information Systems Electronic Library.

Vemou, K., and M. Karyda. 2018b. "An organizational scheme for privacy impact assessments." In *Proc., European, Mediterranean, and Middle Eastern Conf. on Information Systems*, 258–271. Cham, Switzerland: Springer.

Vemou, K., and M. Karyda. 2019. "Evaluating privacy impact assessment methods: Guidelines and best practice." *Inf. Comput. Secur.* 28 (1): 35–53. https://doi.org/10.1108/ICS-04-2019-0047.

Vimercati, S. D. C. D., and S. Foresti. 2011. "Quasi-identifier." In *Encyclopedia of cryptography and security*, edited by H. C. A. Van Tilborg and S. Jajodia. Boston: Springer.

Wagner, I., and E. Boiten. 2018. "Privacy risk assessment: From art to science, by metrics." In *Proc., Int. Workshop on Data Privacy Management*, 225–241. Cham, Switzerland: Springer.

Wang, D., B. Guo, and Y. Shen. 2018. "Method for measuring the privacy level of pre-published dataset." *IET Inf. Secur.* 12 (5): 425–430. https://doi.org/10.1049/iet-ifs.2017.0341.

Wang, L., L. Jin, H. Ji, Y. Chen, P. He, J. Wang, and J. Fang. 2024. "Research on data security risk analysis method of intelligent and connected vehicles based on data asset." In *Proc., CICTP 2024: Resilient, Intelligent, Connected, and Low-Carbon Multimodal Transportation*. Reston, VA: ASCE.

Wang, P., H. D'cruze, and D. Wood. 2019. "Economic costs and impacts of business data breaches." *Issues Inf. Syst.* 20 (2): 162–171. https://doi.org/10.48009/2_iis_2019_162-171.

Wangen, G. 2017. "Information security risk assessment: A method comparison." *Computer* 50 (4): 52–61. https://doi.org/10.1109/MC.2017.107.

Wangen, G., C. Hallstensen, and E. Snekkenes. 2018. "A framework for estimating information security risk assessment method completeness: Core unified risk framework, CURF." *Int. J. Inf. Secur.* 17 (6): 681–699. https://doi.org/10.1007/s10207-017-0382-0.

Xu, H., G. Pang, Y. Wang, and Y. Wang. 2023. "Deep isolation forest for anomaly detection." *IEEE Trans. Knowl. Data Eng.* 35 (12): 12591–12604. https://doi.org/10.1109/TKDE.2023.3270293.

Xu, J., W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu. 2006. "Utility-based anonymization using local recoding." In *Proc., 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 785–790. New York: Association for Computing Machinery.

Yang, M. 2022. "Information security risk management model for big data." *Adv. Multimedia* 2022 (1): 3383251. https://doi.org/10.1155/2022/3383251.

Yang, S., M. Ishtiaq, and M. Anwar. 2018. "Enterprise risk management practices and firm performance, the mediating role of competitive advantage and the moderating role of financial literacy." *J. Risk Financ. Manage.* 11 (3): 35. https://doi.org/10.3390/jrfm11030035.

Yu, H., and X. He. 2021. "Corporate data sharing, leakage, and supervision mechanism research." *Sustainability* 13 (2): 931. https://doi.org/10.3390/su13020931.

Zardari, M. A., L. T. Jung, and M. N. B. Zakaria. 2014. "Research article data classification based on confidentiality in virtual cloud environment." *Res. J. Appl. Sci. Eng. Technol.* 8 (13): 1498–1509. https://doi.org/10.19026/rjaset.8.1128.

Zhang, D. 2018. "Big data security and privacy protection." In *Proc., 8th Int. Conf. on Management and Computer Science (ICMCS 2018)*, 275–278. Dordrecht, Netherlands: Atlantis Press.

Zhang, H., and Y. Deng. 2021. "Entropy measure for orderable sets." *Inf. Sci.* 561 (Jun): 141–151. https://doi.org/10.1016/j.ins.2021.01.073.

Zhang, M., T. Wo, and T. Xie. 2018. "A platform solution of data-quality improvement for internet-of-vehicle services." In *Proc., IEEE Int. Conf. on Pervasive Computing and Communications (PerCom)*, 1–7. New York: IEEE.

Zhao, X., Y. Fang, H. Min, X. Wu, W. Wang, and R. Teixeira. 2023. "Potential sources of sensor data anomalies for autonomous vehicles: An overview from road vehicle safety perspective." *Expert Syst. Appl.* 236 (Feb): 121358. https://doi.org/10.1016/j.eswa.2023.121358.

Zhao, Y., C. Shen, H. Wang, and S. Chen. 2019. "Structural analysis of attributes for vehicle re-identification and retrieval." *IEEE Trans. Intell. Transp. Syst.* 21 (2): 723–734. https://doi.org/10.1109/TITS.2019.2896273.

Zheng, K., W. Liu, L. He, T. Mei, J. Luo, and Z.-J. Zha. 2021. "Group-aware label transfer for domain adaptive person re-identification." In *Proc., IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 5310–5319. New York: IEEE.

Zhou, S., X. Yang, M. Li, H. Yang, and H. Ji. 2022. "Data security risk assessment method for connected and automated vehicles." In *Proc., IEEE 7th Int. Conf. on Intelligent Transportation Engineering (ICITE)*, 387–379. New York: IEEE.

Zurek, W. H. 2018. *Complexity, entropy and the physics of information*. Boca Raton, FL: CRC Press.